Marginal Quantile Treatment Effect*

Ping Yu
University of Auckland

Started: April 2013
First Version: October 2013
This Version: April 2014

Abstract

This paper studies estimation and inference based on the marginal quantile treatment effect. First, we illustrate the importance of the rank preservation assumption in the quantile treatment effects evaluation, show the identifiability of the marginal quantile treatment effect, and clarify the relationship between the marginal quantile treatment effect and other quantile treatment parameters. Second, we develop sharp bounds for the quantile treatment effect with and without the monotonicity assumption, and also sufficient and necessary conditions for point identification. Third, we estimate the marginal quantile treatment effect and associated quantile treatment effect and integrated quantile treatment effect based on the distribution regression, derive the corresponding weak limits and show the validity of the bootstrap inferences. The inference procedure can be used to construct uniform confidence bands for quantile treatment parameters and test unconfoundedness and stochastic dominance. We also develop goodness of fit tests to choose regressors in the distribution regression. Fourth, we conduct two counterfactual analyses: deriving the transition matrix and developing the relative marginal policy relevant quantile treatment effect parameter under the policy invariance. Fifth, we compare the identification schemes in some important literature with that by the marginal quantile treatment effect, and point out advantages and also weaknesses of each scheme, e.g., Chernozhukov and Hansen (2005) concentrate mainly on the quantile treatment effect with the selection select but without the essential heterogeneity; Abadie, Angrist and Imbens (2002), Aakvik, Heckman and Vytlacil (2005) and Chernozhukov and Hansen (2006) suffer from some obvious misspecification problems. Meanwhile, an alternative estimator of the local quantile treatment effect is developed and its weak limit is derived. Finally, we apply the estimation methods to the famous return to schooling dataset of Angrist and Krueger (1991) to illustrate the usefulness of the techniques developed in this paper to practitioners.

KEYWORDS: marginal quantile treatment effect, local quantile treatment effect, rank preservation, selection effect, essential heterogeneity, sharp bound, point identification, distribution regression, two-step estimator, Hadamard differentiability, weak limit, uniform confidence band, unconfoundedness, completeness, stochastic dominance, goodness of fit test, transition matrix, relative marginal policy relevant quantile treatment effect, counterfactual analysis, policy invariance, bootstrap validity, return to schooling
JEL-CLASSIFICATION: C12, C13, C14, C21, C26

*Email: p.yu@auckland.ac.nz.

1 Introduction

Treatment effect evaluation is one main task of econometric analysis. Most literature concentrates on the average treatment effect evaluation; see Heckman and Vytlacil (2007a,b) for a comprehensive summary. Meanwhile, as illustrated in Heckman (1992), Heckman et al. (1997) and Heckman and Smith (1993, 1998), questions of political economy or "social justice" requires knowledge of the distribution of the treatment effect. As a result, distributional treatment effects (especially when unconfoundedness does not hold) become natural parameters of interest among econometricians. Actually, distributional treatment effects have been studied extensively in the empirical literature. For example, Card (1996) uses a panel data set to study the effects of unions on the structure of wages; DiNardo et al. (1996) presents a semiparametric procedure to analyze the effects of institutional and labor market factors on changes in the U.S. distribution of wages; Bitler et al. (2006) estimate quantile treatment effects using random-assignment data from Connecticut's Job First waiver.

Distributional treatment effects are usually estimated based on quantile regression initiated by Koenker and Bassett (1978) (see Koenker (2005) for an introduction to quantile regression). One related field that recently attracts much attention is the "general" semiparametric and nonparametric quantile regression with endogeneity. For the semiparametric setups, see, e.g, Hong and Tamer (2003), Honoré and Hu (2004), Ma and Koenker (2006), Lee (2007), Sakata (2007) and Jun (2008) among others. For nonparametric setups, see, e.g., Chesher (2003), Chernozhukov et al. (2007), Horowitz and Lee (2007), Imbens and Newey (2009), Chen and Pouzo (2012), and Gagliardini and Scaillet (2012) among others. However, the main interest of this paper concentrates on the special structure of the treatment model, namely, the endogenous variable is binary. A key parameter we will develop is the *marginal quantile treatment effect* (MQTE), which is the counterpart of the *marginal treatment effect* (MTE) in the average treatment effect estimation.

The idea of the MTE was first introduced in the context of a parametric normal generalized Roy model by Björklund and Moffitt (1987), and was analyzed more generally by Heckman (1997). In a choice (or selection, or participation) model with the latent variable structure, Heckman and Vytlacil (1999, 2001a) express the conventional average treatment effect parameters as different weighted averages of the MTE, and also identify the MTE by the *local instrumental variable* (LIV) estimator. Actually, Heckman and Vytlacil (2007b) use the MTE to unify the econometric literature on the evaluation of social programs, so it is well recognized that the MTE is a convenient tool to organize the nonparametric literature on the average treatment effect evaluation. An embarrassing situation is that the counterpart of the MTE in the quantile treatment effect literature, the MQTE, is yet to be well understood. The purpose of this paper is to integrate the relevant literature on the quantile treatment effect evaluation without unconfoundedness into one framework and provide some useful estimation and inference methods to practitioners based on the MQTE.

There are two strands of literature concerning about the distributional treatment effects, and they are interwined. Before reviewing the relevant literature, we must emphasize that the distributional treatment effects are functionals of the distribution of $Y_1 - Y_0$, which requires the joint distribution of Y_1 and Y_0, where Y_1 and Y_0 are the outcome under the treatment status and the control status, respectively. As mentioned in Section II.B of Manski (1996) or footnote 5 of Manski (1997), "knowledge of $F(Y_1 - Y_0)$ neither implying nor being implied by knowledge of $F(Y_1)$ and $F(Y_0)$", where $F(X)$ is the cumulative distribution function (CDF) of X for a random variable X. Due to the fundamental problem of causal inference (page 947 of Holland (1986)), Y_0 and Y_1 cannot be observed simultaneously. As a result, even in a random experiment, the joint distribution $F(Y_1, Y_0)$ or $F(Y_1 - Y_0)$ cannot be identified if without further restrictions although $F(Y_1)$ and $F(Y_0)$ can be identified. On the other hand, marginal distributions $F(Y_1)$ and $F(Y_0)$ are also of interest in econometric analysis. For example, in Atkinson (1970), Sen (1997, 2000), Manski (1996, p714), Imbens and Rubin (1997, p558), Imbens and Wooldridge (2009, p17), and Imbens (2010, p409), the marginal distributions of outcomes are more relevant for a social planner choosing between two programs; see also the Introduction of Abadie (2002) for an example where only the marginal distributions are relevant.

The first strand of literature tackles the joint distribution $F(Y_1, Y_0)$ directly. This strand of literature is mainly interested in the *quantile of differences* of Y_1 and Y_0. First, using the classical probability results due to Hoeffding (1940) and Fréchet (1951), Heckman et al. (1997) and Heckman and Smith (1993, 1998) bound $F(Y_1, Y_0)$ using $F(Y_1)$ and $F(Y_0)$ in a random experiment. It turns out that this kind of bounds are too wide to be useful. Later, Aakvik et al. (2005) impose more structures on the problem to get more stringent identification results. Basically, two key structures are imposed: (i) the error terms in the outcome equations and the choice equation are independent of all the covariates; (ii) the error terms follow a one-factor structure, i.e., the correlation among the error terms is only through the factor. Under these two assumptions, $F(Y_1, Y_0)$ can be identified, so all interesting functionals of $F(Y_1, Y_0)$ can be identified. For example, the proportion of people who benefit from participation in the program ($P(Y_1 > Y_0)$), gains to participants at selected levels of the no-treatment distribution ($F(Y_1 - Y_0|Y_0 = y_0)$) or treatment distribution ($F(Y_1 - Y_0|Y_1 = y_1)$), and a variety of other questions including the *quantile treatment effect* (QTE) and the *quantile treatment effect on the treated* (QTT) can be answered. Aakvik et al. (2005) consider only two binary potential outcomes, and parametric one-factor models with cross-sectional data. Extensions to multiple (possibly continuous or mixed discrete and continuous) outcomes, and semiparametric (or nonparametric) multiple-factor models with possibly panel data can be found in Aakvik et al. (1999) and Carneiro et al. (2001, 2003). See Section 2 of Abbring and Heckman (2007) for a summary of this strand of literature.

The second strand of literature concentrates on the marginal distribution of Y_1 and Y_0 (maybe also conditional on some covariates or some specified population). However, as noted above, the distributional treatment effects require the joint distribution of Y_1 and Y_0. To circumvent this problem, this strand of literature explicitly or implicitly assume some type of *rank preservation* (RP) condition. Such type of condition was initiated by Lehmann (1974) and Doksum (1974). Under this assumption, the distributional treatment effects can be described by the *difference of quantiles* of Y_1 and Y_0. The first part of this strand of literature bounds $F(Y_1)$ and $F(Y_0)$ without imposing any restrictions on the choice process. Under the RP assumption, these bounds imply bounds on the QTE. Such literature are summarized in Manski (1994, 1995, 2003). When further restrictions on the choice process are imposed, point identifying some type of quantile treatment effects is possible. The second part of this strand of literature estimates and conducts inferences on some type of quantile treatment effects under point identification. Firgo (2007) estimates the QTE and the QTT in observational studies under the unconfoundedness assumption. When the unconfoundedness assumption fails while only the selection effect exists, Chernozhukov and Hansen (2005) show that the QTE can be identified under some completeness assumption, and Chernozhukov and Hansen (2006) provide a specific estimation scheme; see also Chernozhukov and Hansen (2013) for most updated developments along this line. When there is also the essential heterogeneity, the monotonicity assumption of Imbens and Angrist (1994) or the uniformity assumption of Heckman and Vytlacil (2005) is usually imposed. Under this assumption, Abadie et al. (2002) estimate the *local quantile treatment effect* (LQTE), which is the counterpart of the *local average treatment effect* (LATE),[1] using the identification results in Abadie (2003); see also Imbens and Rubin (1997) for identification of the marginal potential distributions of compliers when no covariates are present, and Abadie (2002) for bootstrap tests of distributional treatment effects in a similar framework. Carneiro and Lee (2009) deal with the essential heterogeneity in an alternative way. They borrow a key assumption, the independence assumption (i) in the last paragraph, from the first strand

[1] The LATE parameter is first introduced by Imbens and Angrist (1994). The MTE is a limit form of the LATE, see, e.g., Björklund and Moffitt (1987), Heckman (1997) and Angrist et al. (2000).

of literature to identify the (conditional) marginal distributions of potential outcomes. These distributions imply the MQTE, which is also the main objective of this paper but we do not need the independence assumption. The above-mentioned literature concentrates on the cross-sectional data; Athey and Imbens (2006) also use the panel data to identify the QTT through what they called change-in-change approach under the RP condition on the treated.

Although these two strands of literature use different identification assumptions, their targets are the same, namely, identifying the joint distribution of Y_1 and Y_0. This paper can be put in the second strand of literature, i.e., we impose some RP assumptions to identify $F(Y_1, Y_0)$. Consequently, the quantile treatment effect in this paper refers to the difference of quantiles rather than the quantile of differences. Meanwhile, we employ the framework in the first strand of literature to study the difference of quantiles.

The rest of this paper is structured as follows. Section 2 sets up our treatment model, illustrates the importance of the RP assumption in the quantile treatment effect evaluation, shows the identifiability of the MQTE, and clarifies the relationship between the MQTE and other quantile treatment parameters. Section 3 develops sharp bounds and sufficient and necessary conditions for point identification of the QTE with and without the monotonicity assumption. In Section 4, we estimate the MQTE based on the *distribution regression* introduced by Foresi and Peracchi (1995), derive its weak limit and show the validity of the bootstrap inferences, and we also develop goodness of fit tests to choose regressors. In Section 5, we conduct two counterfactual analyses: deriving the transition matrix and developing the relative marginal policy relevant quantile treatment effect parameter under the policy invariance. In Section 6, we comment some key literature in the two strands above, pointing out their weaknesses, underlying assumptions, and interactions with this paper. Section 7 presents an empirical application to the return to schooling and Section 8 concludes. All proofs are contained in an appendix.

Some notations are collected here for future reference. d is always used for indicating the two treatment statuses, so is not written out explicitly as "$d = 0, 1$" throughout the paper. $\mathrm{supp}(X)$ for a random variable X denotes the support of the distribution of X. Both $Q_X(\tau)$ and $Q_\tau(X)$ denote the τth quantile of a random variable X. The capital letters such as X denote random variables and the corresponding lower case letter such as x denote the potential values they may take. For any parameter θ, d_θ is the dimension of θ. The space $\ell^\infty(\mathcal{F})$ represents the space of real-valued bounded functions defined on the index set equipped with the supremum norm $\|\cdot\|_{\ell^\infty(\mathcal{F})}$. $C(\mathcal{Y})$ is the space of continuous functions on $\mathcal{Y}$.

2 The Setup and Parameters of Interest

We use the nonlinear and nonseparable outcome model as in Heckman and Vytlacil (2005),

$$\begin{aligned} Y_1 &= \mu_1(X, U_1), \\ Y_0 &= \mu_1(X, U_0). \end{aligned} \tag{1}$$

Actually, the additively separable setup, $Y_d = \mu_d(X) + U_d$, does not lose generality since we can define the new U_d as $Y_d - Q_{Y_d|X}(\tau|X)$ and all our analysis in this paper is conditional on X. The distribution of Y_d may be discrete (e.g., employment status), continuous (e.g., wage), or mixed discrete and continuous (e.g., in the national JTPA study 18 month impact sample used in Heckman et al. (1997), a substantial proportion of persons has zero earnings in both distributions of Y_0 and Y_1). The participation decision

$$D = 1(\mu_D(X, Z) - V \geq 0), \tag{2}$$

where Z includes the instruments for the choice process. Both X and Z appearing as the arguments of μ_D does not lose generality since $\mu_D(X,Z)$ may not depend on all elements of X. By transforming $\mu_D(X,Z)$ and V by $F_{V|X,Z}$, we can rewrite

$$D = 1(p(X,Z) - U_D \geq 0), \tag{3}$$

where $U_D|X,Z \sim U(0,1)$ and $p(X,Z)$ is the propensity score. We use these two formulations of D interchangeably throughout the paper. As shown in Vytlacil (2006), there is a larger class of latent index models that will have a representation of this form. Also, this setup of D implies the monotonicity assumption of Imbens and Angrist (1994) as shown in Vytlacil (2002).

We impose the following assumptions on the outcome equation and the choice equation.

(A1) $\mu_D(X,Z)$ is a nondegenerate random variable conditional on X.
(A2) The random vectors (U_1,V) and (U_0,V) are independent of Z conditional on X.
(A3) The distribution of V is absolutely continuous with respect to Lebesgue measure.
(A4) $X_1 = X_0$ almost everywhere, where X_d denote a value of X if D is set to d.
(A5) $1 > P(D=1|X) > 0$.
(A6) Conditional on $X=x$, $V=v$, Y_0 and Y_1 have the same rank.

(A1)-(A5) corresponds to (A-1)-(A-3), (A-6) and (A-5) in Heckman and Vytlacil (2005), respectively. These assumptions are prevalent in the literature with heterogeneous treatment effects. A necessary condition for (A1) is that Z contains a continuous variable. (A2) allows for both the selection effect ($U_0 \not\perp D|X$) and the essential heterogeneity ($(U_1-U_0) \not\perp D|X$). Also, (A2) implies the usual assumption in the control function approach, say, $Z \perp (U_1,U_0)|(X,V)$. (A1)-(A5), combined with (1) and (2), impose testable restrictions on the distribution of (Y,D,Z,X); see Heckman and Vytlacil (2005) (page 678) for the index sufficiency restriction and the monotonicity restriction. We refer to Heckman and Vytlacil (2005) for more detailed discussions on (A1)-(A5). The assumption (A6) deserves further examination.

2.1 The Rank Preservation Condition

The key extra assumption beyond those in Heckman and Vytlacil (2005) is the RP condition (A6). Chernozhukov and Hansen (2005) state the RP assumption via the Skorohod representation. We try to do the same thing here although unlike them, this representation is not essential for the development of our identification scheme. Suppose Y_d is continuous, and the τth conditional quantile of Y_d given X and V is $q(d,X,V,\tau)$; then we can represent

$$Y_d = q(d,X,V,R_d)$$

by the Skorohod representation, where $R_d|(X,V) \sim U(0,1)$ is the rank variable which represents some unobserved characteristic of Y_d, e.g., ability or proneness, among the slice of people with a specific value of X and V. The RP assumption (A6) can be restated as $R_1|(X,V) = R_0|(X,V)$. We now clarify two key points of the Skorohod representation. First, the Skorohod representation decomposes the information in U_d of (1) into two components: the value information and the rank information. The former is incorporated in the quantile function $q(\cdot)$ and the later is included in R_d. Second, because $R_d|(X,V) \sim U(0,1)$ does not depend on (X,V), it may be suspected that R_d is independent of (X,V). This is incorrect. This mistake is immediately clear if we rewrite $Y_d = q(d,X,V,R_d(X,V))$; in other words, R_d must be understood as a conditional random variable. Suppose there are N distinct points on the support of (X,V), and then there are N rank variables $R_d(X,V)$. Although $R_d(X,V)|(X=x,V=v) \sim U(0,1)$ does not depend on (x,v), the unconditional random variable R_d may depend on (X,V). The RP condition does not restrict the dependence between R_d and (X,V); rather, it restricts the total number of conditional rank variables

$R_d(X, V)$ from $2N$ to N. To be consistent with the notation in the literature, R_d is replaced by U_d in the rest of this paper. The meaning of U_d should be clear from the context. For example, when V appears as an argument of the representation of Y_d, or Y_d is represented as $Y_d = q(\cdot)$, U_d means the rank variable. In this paper, we do not consider the quality of the evidence supporting the assumption (A6). Instead, we consider the evaluation of specific programs under this assumption.

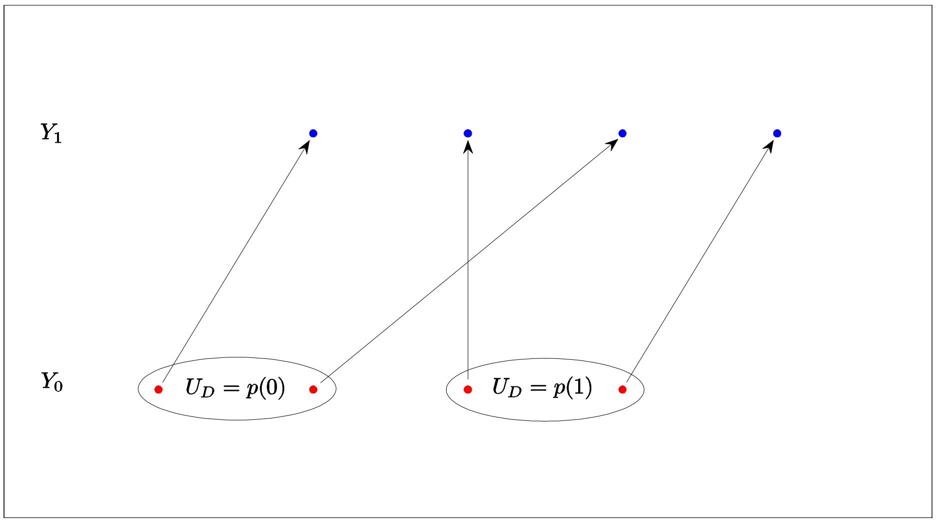

Figure 1: Rank Preserved Conditional on U_D BUT Unconditionally Unpreserved

Our RP assumption is weaker than the usual assumption that Y_0 and Y_1 have the same rank conditional on $X = x$. Think about the following example. Suppose Z is the only covariate in the determination of D, and Z can take only 0 and 1. So the only nontrivial values for U_D are $p(0)$ and $p(1)$. Suppose for each value of U_D, there are only two persons. Figure 1 shows that although the rank is preserved among the people with a specific U_D value, the rank is unpreserved if all people are taken into account. In other words, (A6) only requires the RP condition to hold locally $(X = x, U_D = u_D)$ instead of globally $(X = x)$. Conditional on $X = x$, the rank may not be maintained under the treatment. However, for a finer slice of individuals, the rank is maintained. Local rank preservation is much weaker than global rank preservation. The larger the conditional set on which the RP condition is imposed, the harder for the RP condition to hold. Actually, the analysis in Heckman et al. (1997) show that the unconditional RP condition cannot hold although substantial departures from the perfect positive dependence across Y_1 and Y_0 are not credible in their context; see also Carneiro et al. (2003) for further evidences against the unconditional RP condition.

The RP condition also imposes a restriction on the joint distribution of Y_1 and Y_0 given $X = x$ and $U_D = u_D$, namely, the joint distribution is fully determined by the marginal distribution. It is not hard to see that when the RP condition holds,

$$\begin{aligned}
&P\left(Y_1 \le y_1, Y_0 \le y_0 | X = x, U_D = u_D\right) \\
&= \min\left\{P\left(Y_1 \le y_1 | X = x, U_D = u_D\right), P\left(Y_0 \le y_0 | X = x, U_D = u_D\right)\right\},
\end{aligned}$$

which implies that the joint distribution of Y_1 and Y_0 given $X = x, U_D = u_D$ is degenerate. To see how this joint distribution looks like, suppose $Y_d|(X = x, U_D = u_D)$ is continuously distributed and $\text{supp}(Y_d|X = x, U_D = u_D) = [0,1]$ to simplify the discussion. It turns out that only on the line $\left(y_0, F^{-1}_{Y_1|X,U_D}\left(F_{Y_0|X,U_D}(y_0|x,u_D)|x,u_D\right)\right)$ with $y_0 \in [0,1]$ there is probability. In other words, only on the Q-Q plot, (Y_0, Y_1) can occur simultaneously. An implication of this result is that if $F_{Y_0|X,U_D}(\cdot|x,u_D)$ is the same as $F_{Y_1|X,U_D}(\cdot|x,u_D)$, then the correlation between Y_0 and Y_1 conditional on $X = x, U_D = u_D$ must be 1. Figure 2 shows a typical Q-Q plot of (Y_0, Y_1) conditional on $X = x, U_D = u_D$. In Figure 2, $P(Y_1 \geq Y_0|Y_0 = y_0, X = x, U_D = u_D) = 1$ when $y_0 \leq 0.6$ and $P(Y_1 \geq Y_0|Y_0 = y_0, X = x, U_D = u_D) = 0$ when $y_0 > 0.6$. In other words, for the slice of people with $Y_0 = y_0, X = x, U_D = u_D$, the participant always benefits as long as $y_0 \leq 0.6$, and vice versa. Nevertheless, it is more likely that $P(Y_1 \geq Y_0|Y_0 = y_0, X = x) \in (0,1)$, $P(Y_1 \geq Y_0|X = x, U_D = u_D) = F_{Y_0|X,U_D}(0.6|x,u_D) \in (0,1)$ and $P(Y_1 \geq Y_0|X = x) = \int P(Y_1 \geq Y_0|X = x, U_D = u_D)du_D \in (0,1)$.

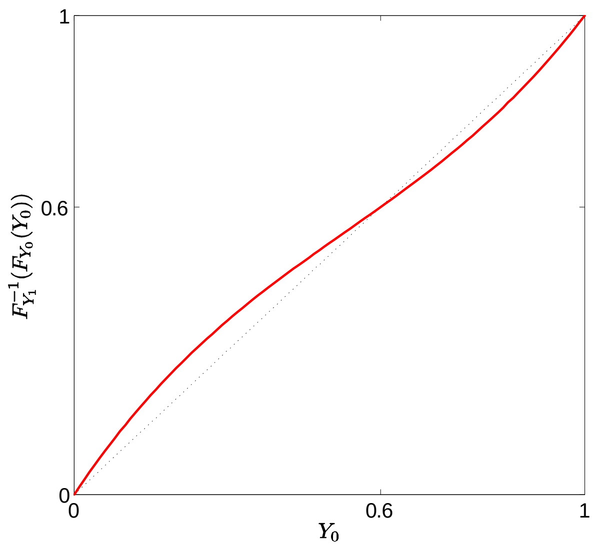

Figure 2: Q-Q Plot of (Y_0, Y_1) Conditional on $X = x, U_D = u_D$

It should be emphasized that the RP condition is only for defining various quantile treatment effects. Even without this condition, we can still identify various marginal distributions which, as argued in the introduction, are useful for many other purposes. Under the RP assumption, we define the MQTE in Carneiro and Lee (2009) as

$$\Delta^{MQTE}_{\tau}(x, u_D) = Q_{Y_1|X,U_D}(\tau|x,u_D) - Q_{Y_0|X,U_D}(\tau|x,u_D).$$

If we strengthen the RP assumption to be conditional on $X = x$ or on $X = x, D = 1$, then we can define the QTE in Chernozhukov and Hansen (2005, 2006) and the QTT as

$$\Delta^{QTE}_{\tau}(x) = Q_{Y_1|X}(\tau|x) - Q_{Y_0|X}(\tau|x)$$

and

$$\Delta^{QTT}_{\tau}(x) = Q_{Y_1|X,D}(\tau|x,1) - Q_{Y_0|X,D}(\tau|x,1),$$

respectively. If the RP assumption is conditional on $X = x, u_D < U_D \leq u'_D$, then the LQTE of Abadie et al. (2002)[2] is defined as

$$\Delta_\tau^{LQTE}(x, u_D, u'_D) = Q_{Y_1|X,U_D}(\tau|x,(u_D,u'_D]) - Q_{Y_0|X,U_D}(\tau|x,(u_D,u'_D]).$$

Finally, if the RP assumption holds unconditionally (with respect to X),[3] then we define the integrated QTE (IQTE)

$$\Delta_\tau^{IQTE} = Q_{Y_1}(\tau) - Q_{Y_0}(\tau),$$

the integrated QTT (IQTT)

$$\Delta_\tau^{IQTT} = Q_{Y_1|D}(\tau|1) - Q_{Y_0|D}(\tau|1)$$

as in Firpo (2007),[4] and the integrated LQTE (ILQTE)

$$\Delta_\tau^{ILQTE}(u_D, u'_D) = Q_{Y_1|U_D}(\tau|(u_D,u'_D]) - Q_{Y_0|U_D}(\tau|(u_D,u'_D]).$$

2.2 Identification of the MQTE

The following theorem states that the MQTE can be identified for a range of u_D.

Theorem 1 *Suppose assumptions (A1)-(A6) hold. If u_D is not an isolated point of $\mathcal{P}_x^1 \cap \mathcal{P}_x^0$, then $\Delta_\tau^{MQTE}(x, u_D)$ can be identified for any $\tau \in (0,1)$, where $\mathcal{P}_x^d$ =supp$(p(X,Z)|X = x, D = d)$.*

Proof. To simplify notations, we depress the conditioning on $X = x$. Given the RP assumption (A6), we need only identify $Q_{Y_d|U_D}(\tau|u_D)$ whose identification is equivalent to the identification of $F_{Y_d|U_D}(\cdot|u_D)$. We provide two methods to identify $F_{Y_d|U_D}(\cdot|u_D)$.
Method 1: Note that

$$\begin{aligned} P\left(Y \leq y|p(Z) = p, D = 1\right)p &= P\left(Y_1 \leq y|p(Z) = p, D = 1\right)P(D = 1|p(Z) = p) \\ &= P\left(Y_1 \leq y|U_D \leq p\right)p = \int_0^p F_{Y_1|U_D}(y|u_D)du_D, \end{aligned}$$

and similarly, $P\left(Y \leq y|p(Z) = p, D = 0\right)(1-p) = \int_p^1 F_{Y_0|U_D}(y|u_D)du_D$, so

$$\begin{aligned} \frac{d\left[P\left(Y \leq y|p(Z) = p, D = 1\right)p\right]}{dp} &= F_{Y_1|U_D}(y|p), \\ -\frac{d\left[P\left(Y \leq y|p(Z) = p, D = 0\right)(1-p)\right]}{dp} &= F_{Y_0|U_D}(y|p). \end{aligned}$$

[2] Abadie et al. (2002) conflate issues of definition of parameters with issues of identification; see Section 6.2 below for their definition. Actually, $\Delta_\tau^{LQTE}(x, u_D, u'_D)$ can be defined for any $u_D, u'_D \in (0,1)$ although it can only be identified for u_D, u_D on the support of $p(x, Z)$.

[3] Note that if the RP assumption holds on $X = x$, $Y_d|X$ can be expressed as $Y_d = q(d, X, U)$ by the Skorohod representation, where $U|X = U_1|X = U_0|X$. If the RP assumption holds unconditionally, then Y_d can be expressed as $Y_d = q(d, U)$ by the Skorohod representation, where $U = U_1 = U_2$. This by no means implies that information in X and Z is useless to the identification or efficiency improvement in the quantile treatment effect evaluation.

[4] Be careful about the terminology in the literature. Our IQTE and IQTT are the QTE and QTT of Firpo (2007). Also, the MQTE of Cattaneo (2010) means $Q_\tau(Y_0)$ and $Q_\tau(Y_1)$ rather than $\Delta_\tau^{MQTE}(x, u_D)$, and the MQTE, QTE and QTT in the first strand of literature mentioned in the introduction means $Q_{Y_1-Y_0|X,U_D}(\tau|x, u_D)$, $Q_{Y_1-Y_0|X}(\tau|x)$ and $Q_{Y_1-Y_0|X,D}(\tau|x, 1)$ rather than $\Delta_\tau^{MQTE}(x, u_D)$, $\Delta_\tau^{QTE}(x)$ and $\Delta_\tau^{QTT}(x)$.

Method 2: As in Hahn (1998), we can use DY and $(1-D)Y$ to identify $F_{Y_d|U_D}(\cdot|u_D)$. Note that for any $y \geq 0$

$$
\begin{aligned}
& P\left(DY \leq y|p(Z)=p\right)-(1-p) \\
& = P\left(DY \leq y|p(Z)=p, D=1\right) P(D=1|p(Z)=p)+P\left(DY \leq y|p(Z)=p, D=0\right) P(D=0|p(Z)=p)-(1-p) \\
& = P\left(Y_1 \leq y|U_D \leq p\right) p+P\left(0 \leq y|U_D > p\right) P(D=0|p(Z)=p)-(1-p) \\
& = \int_0^p F_{Y_1|U_D}(y|u_D) du_D+(1-p)-(1-p)=\int_0^p F_{Y_1|U_D}(y|u_D) du_D,
\end{aligned}
$$

so

$$\frac{d\left[P\left(DY \leq y|p(Z)=p\right)-(1-p)\right]}{dp}=F_{Y_1|U_D}(y|p). \tag{4}$$

For $y<0$, repeat the analysis above, we have

$$\frac{dP\left(DY \leq y|p(Z)=p\right)}{dp}=F_{Y_1|U_D}(y|p). \tag{5}$$

Similarly,

$$-\frac{d\left[P\left((1-D)\, Y \leq y|p(Z)=p\right)-p\right]}{dp}=F_{Y_0|U_D}(y|p) \text{ when } y \geq 0, \tag{6}$$

and

$$-\frac{dP\left((1-D)\, Y \leq y|p(Z)=p\right)}{dp}=F_{Y_0|U_D}(y|p) \text{ when } y<0. \tag{7}$$

Inverting $F_{Y_d|U_D}(y|p)$ as a function of y, we can get $Q_{Y_d|U_D}(\tau|p)$. Since $p(Z)$, $P\left(Y \leq y|p(Z)=p, D=d\right)$, $P\left(DY \leq y|p(Z)=p\right)$ and $P\left((1-D)\, Y \leq y|p(Z)=p\right)$ for $y \in \mathbb{R}$ and $p \in \mathcal{P}_x^d$ can be identified, $\Delta_\tau^{MQTE}(u_D)=Q_{Y_1|U_D}(\tau|u_D)-Q_{Y_0|U_D}(\tau|u_D)$ for all $\tau \in (0,1)$ and u_D not being an isolated point of $\mathcal{P}_x^1 \cap \mathcal{P}_x^0$ can be identified. ■

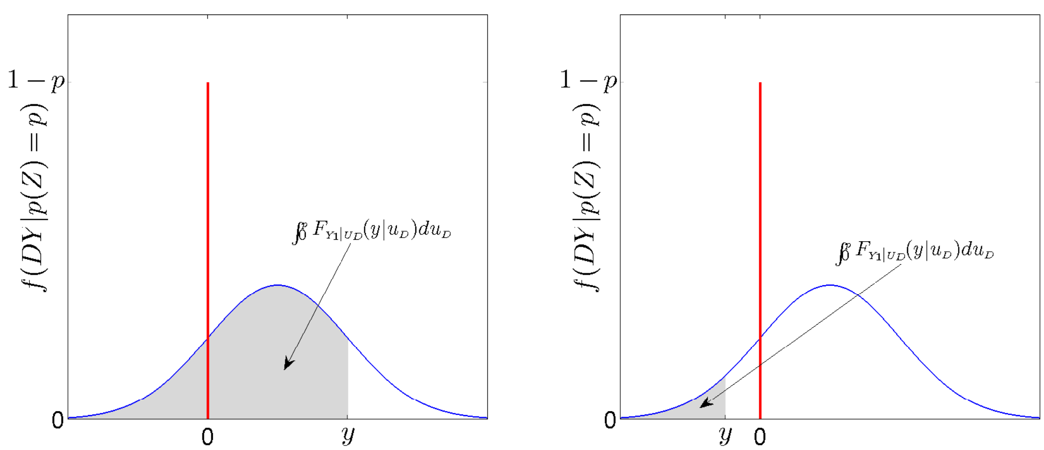

Figure 3: Intuition for Identification of $F_{Y_1|U_D}(y|p)$ with $y \geq 0$ and $y<0$

Figure 3 provides some intuition for the arguments in the second method. For $y \geq 0$, $P\left(DY \leq y|p(Z)=p\right)$ includes a point mass $1-p$ at 0, and the remaining probability is $\int_0^p F_{Y_1|U_D}(y|u_D) du_D$, while for $y<0$,

$P(DY \leq y|p(Z) = p)$ does not include the point mass. This intuition is similar in spirit to that of the censored quantile regression models discussed in Powell (1984, 1986).

The arguments in Theorem 1 can be applied to the discrete Y_d case. Suppose Y_1 and Y_0 have the same support $\{\mathsf{y}_1, \cdots, \mathsf{y}_S\}$, and then the counterpart of the MQTE is $P_{Y_1|U_D}(\mathsf{y}_s|u_D) - P_{Y_0|U_D}(\mathsf{y}_s|u_D)$, $s = 1, \cdots, S$, where $P_{Y_d|U_D}(\mathsf{y}_s|u_D)$ is the point mass of $Y_d|(U_D = u_D)$ at y_s. We can still identify $F_{Y_d|U_D}(\mathsf{y}_s|p)$ by (4), (5), (6) and (7), and then $P_{Y_d|U_D}(\mathsf{y}_1|p) = F_{Y_d|U_D}(\mathsf{y}_1|p)$ and $P_{Y_d|U_D}(\mathsf{y}_s|p) = F_{Y_d|U_D}(\mathsf{y}_s|p) - F_{Y_d|U_D}(\mathsf{y}_{s-1}|p)$ for $s = 2, \cdots, S$ can be sequentially identified. If Y_d can take only 0 and 1, then the parameter of interest is $P_{Y_1|U_D}(1|u_D) - P_{Y_0|U_D}(1|u_D)$ which coincides with the MTE. Of course, we can also consider the case with mixed discrete and continuous outcomes. Both the discrete case and the mixed case are easier to handle than the continuous case, so we will concentrate on the continuous case in the rest of this paper unless stated otherwise.

If we use the idea of LIV as in Heckman and Vytlacil (2001a), we have

$$\begin{aligned} P(Y \leq y|p(Z) = p) &= P(Y \leq y|p(Z) = p, D = 1)\, p + P(Y \leq y|p(Z) = p, D = 0)(1 - p) \\ &= \int_0^p F_{Y_1|U_D}(y|u_D) du_D + \int_p^1 F_{Y_0|U_D}(y|u_D) du_D, \end{aligned}$$

and

$$\frac{\partial P(Y \leq y|p(Z) = p)}{\partial p} = F_{Y_1|U_D}(y|p) - F_{Y_0|U_D}(y|p),$$

which is the difference of CDFs in the two treatment statuses. So it is hard to identify the MQTE from $\partial P(Y \leq y|p(Z) = p)/\partial p$. From Theorem 1, we can identify $E[Y_1|U_D = p]$ and $E[Y_0|U_D = p]$ separately, not just their difference $E[Y_1 - Y_0|U_D = p]$ as in the LIV method of Heckman and Vytlacil (2001a).

Method 1 of the proof is a special case of Theorem 1 in Carneiro and Lee (2009). We also discuss Method 2 to distinguish the difference between the identification scheme of the MTE and the MQTE. For the MTE, $E[DY|p(Z) = p] = E[Y|p(Z) = p, D = 1]\, p = \int_0^p E[Y_1|U_D = u_D]\, du_D$, and $E[(1 - D)Y|p(Z) = p] = E[Y|p(Z) = p, D = 0](1 - p) = \int_p^1 E[Y_0|U_D = u_D]\, du_D$, so the two methods in the proof are the same in the MTE identification.

We close this subsection by a concrete example. Suppose $Y_1 = V + 2U$, $Y_0 = 2V + U$, and $D = 1(Z - V > 0)$, where

$$\begin{pmatrix} U \\ V \\ Z \end{pmatrix} \sim N(0, \Sigma) \text{ with } \Sigma = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It can be shown that $\Delta_\tau^{MQTE}(u_D) = -0.5\Phi^{-1}(u_D) + \sqrt{0.75}\Phi^{-1}(\tau)$. Figure 4 shows $\Delta_\tau^{MQTE}(u_D)$ for $\tau = 0.1, 0.25, 0.5, 0.75$ and 0.9. In this simple model, the spreading measure of the MQTE, e.g., $\Delta_{1-\tau}^{MQTE}(u_D) - \Delta_\tau^{MQTE}(u_D)$ for $\tau \in (0, 0.5)$, is the same for any u_D, which may not be standard in practice. Also, $\Delta_\tau^{MQTE}(u_D)$ is a decreasing function of p, which indicates that the more likely will an individual participate in the program, the higher benefit will she receive.[5] In the figure, we also show $\Delta^{MTE}(u_D)$, Δ_τ^{QTE} and $\Delta^{ATE}(\equiv E[Y_1] - E[Y_0])$ for comparison. Note that in this example, $\Delta^{MTE}(u_D) = \Delta_{.5}^{MQTE}(u_D)$, and $\Delta_\tau^{QTE} = 0 = \Delta^{ATE}$ does not depend on τ.[6] Obviously, $\Delta_\tau^{MQTE}(u_D)$ provides more information than $\Delta^{MTE}(u_D)$, Δ_τ^{QTE}, and Δ^{ATE}.

[5] Aakvik et al. (2005) provide a converse example.

[6] It should be emphasized that Δ_τ^{QTE} is not well defined in this example since the RP condition does not hold unconditionally given that Y_1 and Y_0 have the same marginal distribution but $Corr(Y_1, Y_0) = 6.5/7 < 1$.

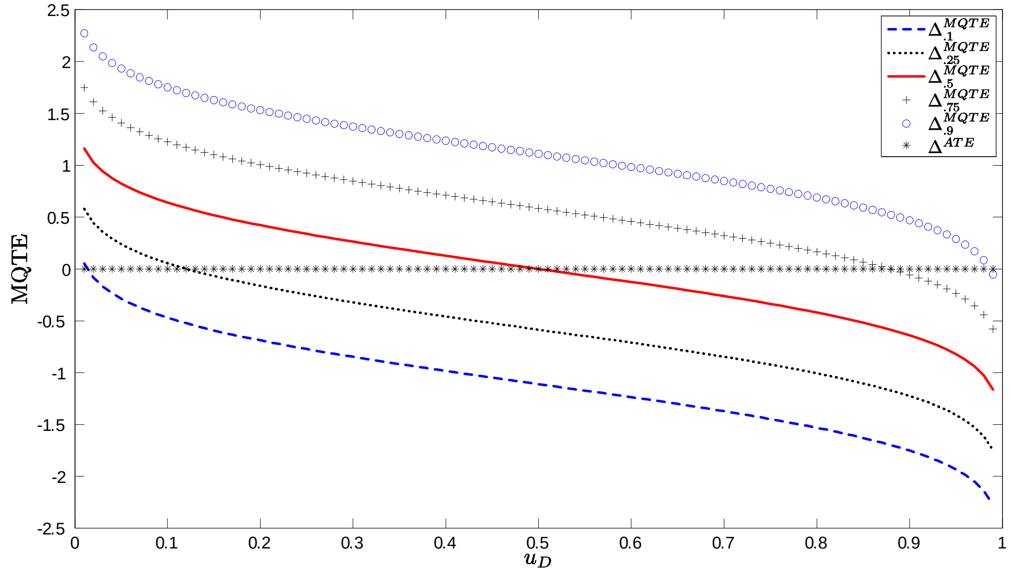

Figure 4: $\Delta_\tau^{MQTE}(u_D)$ for $\tau = 0.1, 0.25, 0.5, 0.75$ and 0.9 in a Simple Example

2.3 Relationship with Other Parameters of Treatment Effects

In this subsection, we first discuss the relationship between $\Delta_\tau^{MQTE}(x, u_D)$ and $\Delta_\tau^{QTT}(x)$, $\Delta_\tau^{QTE}(x)$, $\Delta_\tau^{LQTE}(x, u_D, u'_D)$, Δ_τ^{IQTE}, Δ_τ^{IQTT}, Δ_τ^{ILQTE}. It turns out that the building block is $F_{Y_d|X,U_D}(y_d|x, u_D)$ rather than $\Delta_\tau^{MQTE}(x, u_D)$. Actually, $\Delta_\tau^{MQTE}(x, u_D)$ is more relevant to the (conditional) quantile of $Y_1 - Y_0$.

From the supplementary materials, we can show that

$$\Delta_\tau^{QTT}(x) = F^{-1}_{Y_1|X,D}(\tau|x, 1) - F^{-1}_{Y_0|X,D}(\tau|x, 1)$$

and the quantile treatment effect on the untreated (QTUT)

$$\Delta_\tau^{QTUT}(x) = F^{-1}_{Y_1|X,D}(\tau|x, 0) - F^{-1}_{Y_0|X,D}(\tau|x, 0),$$

where

$$F_{Y_d|X,D}(y_d|x, 1) = \int_0^1 F_{Y_d|X,U_D}(y_d|x, u_D) h_{TT}(x, u_D) du_D,$$
$$F_{Y_d|X,D}(y_d|x, 0) = \int_0^1 F_{Y_d|X,U_D}(y_d|x, u_D) h_{TUT}(x, u_D) du_D,$$

with $h_{TT}(x, u_D) = \left(1 - F_{p(X,Z)|X}(u_D|x)\right)/E[p(X, Z)|X = x]$ and $h_{TUT}(x, u_D) = F_{p(X,Z)|X}(u_D|x)/E[1 - p(X, Z)|X = x]$. Also,

$$\Delta_\tau^{QTE}(x) = F^{-1}_{Y_1|X}(\tau|x) - F^{-1}_{Y_0|X}(\tau|x), \Delta_\tau^{IQTE} = F^{-1}_{Y_1}(\tau) - F^{-1}_{Y_0}(\tau),$$
$$\Delta_\tau^{IQTT} = F^{-1}_{Y_1|D}(\tau|1) - F^{-1}_{Y_0|D}(\tau|1), \Delta_\tau^{IQTUT} = F^{-1}_{Y_1|D}(\tau|0) - F^{-1}_{Y_0|D}(\tau|0)$$

where

$$F_{Y_d|X}(y_d|x) = \int_0^1 F_{Y_d|X,U_D}(y_d|x,u_D)du_D, F_{Y_d}(y_d) = \int F_{Y_d|X}(y_d|x)dF_X(x),$$

$$F_{Y_d|D}(y|1) = \int F_{Y_d|X,D}(y|x,1)dF_{X|D}(x|1) = \int\int_0^1 F_{Y_d|X,U_D}(y|x,u_D)\frac{1-F_{p(X,Z)|X}(u_D|x)}{P(D=1)}du_D dF_X(x),$$

$$F_{Y_d|D}(y|0) = \int F_{Y_d|X,D}(y|x,1)dF_{X|D}(x|0) = \int\int_0^1 F_{Y_d|X,U_D}(y|x,u_D)\frac{F_{p(X,Z)|X}(u_D|x)}{P(D=0)}du_D dF_X(x).$$

Finally,

$$\Delta_\tau^{LQTE}(x,u_D,u'_D) = F^{-1}_{Y_1|X,U_D}(\tau|x,(u_D,u'_D]) - F^{-1}_{Y_0|X,U_D}(\tau|x,(u_D,u'_D]),$$

$$\Delta_\tau^{ILQTE}(u_D,u'_D) = F^{-1}_{Y_1|U_D}(\tau|(u_D,u'_D]) - F^{-1}_{Y_0|U_D}(\tau|(u_D,u'_D]),$$

where

$$F_{Y_d|X,U_D}(y_d|x,(u_D,u'_D]) = \frac{1}{u'_D-u_D}\int_{u_D}^{u'_D} F_{Y_d|X,U_D}(y_d|x,u_D)du_D,$$

$$F_{Y_d|U_D}(y_d|(u'_D,u_D]) = \int F_{Y_d|X,U_D}(y_d|x,(u_D,u'_D])dF_{X|p(X,Z)}(x|(u_D,u'_D])$$

$$= \int F_{Y_d|X,U_D}(y_d|x,(u_D,u'_D])\frac{P\left(p(X,Z)\in(u_D,u'_D]|X=x\right)}{P\left(p(X,Z)\in(u_D,u'_D]\right)}dF_X(x).$$

See Appendix B.1 of Carneiro and Lee (2009) for implementation of some of these parameters in practice. Note that $\Delta_\tau^{QTE}(x) \neq \int_0^1 \Delta_\tau^{QTE}(x,u_D)du_D$, and $\Delta_\tau^{IQTE} \neq \int \Delta_\tau^{QTE}(x)dF_X(x)$, so it is hard to find a relationship between these quantile treatment parameters and $\Delta_\tau^{MQTE}(x,u_D)$.

We can also identify the MTE

$$\Delta^{MTE}(x,u_D) = \int y_1 dF_{Y_1|X,U_D}(y_1|x,u_D) - \int y_0 dF_{Y_0|X,U_D}(y_1|x,u_D) = \int \Delta_\tau^{MQTE}(x,u_D)d\tau,$$

so all the parameters that can be identified by $\Delta^{MTE}(x,u_D)$ as listed in Table IA of Heckman and Vytlacil (2005) can also be identified by $\Delta_\tau^{MQTE}(x,u_D)$. In other words, $\Delta_\tau^{MQTE}(x,u_D)$ is a more basic building block of the average treatment parameters. Note that to identify Δ^{MTE}, we do not need the RP assumption, but we need to assume $E\left[|Y_d|\right] < \infty$.

Heckman et al. (1997) consider also the following parameters of treatment effects: (a) the proportion of people taking the program who benefit from it, $P(Y_1 > Y_0|D=1)$; (b) the proportion of the total population that benefits from the program, $P(Y_1 > Y_0|D=1)P(D=1)$; (c) selected quantiles of the impact distribution, $Q_{Y_1-Y_0|D}(\tau|1)$; (d) the distribution of gains at selected base state values, $F_{Y_1-Y_0|Y_0,D}(\cdot|y_0,1)$;[7] These parameters can be identified from $\Delta_\tau^{MQTE}(x,u_D)$. It is not hard to show that under the RP assumption (A6),

$$P\left(Y_1 - Y_0 \le y|D=1\right) = \int\int_0^1 \left[\int_0^1 1(\Delta_\tau^{MQTE}(x,u_D)\le y)d\tau\right]\frac{1-F_{p(X,Z)|X}(u_D|x)}{P(D=1)}du_D dF_X(x)$$

[7] They also consider the option value of a social program, $E[\max(Y_0,Z)|D=1]-E[Y_0|D=1]$, where Z is the option provided by the program.

by a similar derivation as in the expression of $F_{Y_d|D}(y|1)$, so $Q_{Y_1-Y_0|D}(\tau|1)$ can be identified and

$$P(Y_1 > Y_0|D=1) = 1 - P\left(Y_1 - Y_0 \leq 0|D=1\right)$$

can also be identified.[8] Actually, we can identify any conditional or unconditional quantile of $Y_1 - Y_0$ of interest, e.g., $Q_{Y_1-Y_0|X,D}(\tau|x,d)$, $Q_{Y_1-Y_0|X}(\tau|x)$, $Q_{Y_1-Y_0|D}(\tau|d)$, $Q_{Y_1-Y_0}(\tau)$ and $Q_{Y_1-Y_0|X,U_D}(\tau|x,(u_D', u_D])$, based on $\Delta_\tau^{MQTE}(x,u_D)$. Since the corresponding weights can be similarly defined as above, we neglect the details.

Note that if only assumption (A6) holds, $P(Y_1 > Y_0|D=1)$ need not equal $\int\int_0^1 1(\Delta_\tau^{QTT}(x) > 0)d\tau dF_{X|D}(x|1)$ or $\int_0^1 1(\Delta_\tau^{IQTT} > 0)d\tau$. They are equal only if the RP assumption holds on $X = x, D = 1$ or $D = 1$. This observation can be used to test whether the RP assumption holds on a larger set than $X = x, U_D = u_D$. Because quantile is not a linear operator of the distribution function, $Q_{Y_1-Y_0}(\cdot)$ and $Q_{Y_1}(\cdot) - Q_{Y_0}(\cdot)$ are generally unequal (and do not have any identifiable relationships), so the quantile treatment effect and the quantile of the impact distribution are two different parameters. On the contrary, since mean is a linear operator of the distribution function, the average treatment effect and the average of the impact distribution are the same parameter. In this paper, we concentrate on three most popular quantile treatment effect parameters in the literature: $\Delta_\tau^{MQTE}(x,u_D)$, $\Delta_\tau^{QTE}(x)$ and Δ_τ^{IQTE}. We concentrate on difference of quantiles rather than quantile of differences because the latter may not be interesting. For example, in the common effect model, the distribution of $Y_1 - Y_0$ is a point mass at a fixed value. Even if the treatment effect is not common, $Y_1 - Y_0$ may still have discrete components in its distribution. See Section 3.2 of Aakvik et al. (2005) for definitions of the distributional counterparts of the MTE, ATE and ATT based on $Y_1 - Y_0$ when the outcomes are binary, and see Section 2 of Abbring and Heckman (2007) for definitions of the distributional treatment effects in more general settings.

Finally, we study $F_{Y_1-Y_0|Y_0,D}(\cdot|y_0,1)$. We have already shown in Section 2.1 that under the RP assumption (A6),

$$P(Y_1 - Y_0 \leq y|Y_0 = y_0, X = x, U_D = u_D) = 1(Q_{Y_1|X,U_D}(F_{Y_0|X,U_D}(y_0|x,u_D)|x,u_D) \leq y + y_0),$$

so

$$\begin{aligned} &P(Y_1 - Y_0 \leq y|Y_0 = y_0, D=1) = P(Y_1 - Y_0 \leq y|Y_0 = y_0, U_D \leq p(X,Z)) \\ &= \int\int \left[\tfrac{1}{F_{U_D|U_0,X}(p(x,z)|F_{Y_0|X}(y_0|x),x)}\int_0^{p(x,z)} P(Y_1 - Y_0 \leq y|Y_0 = y_0, X = x, U_D = u_D)dF_{U_D|U_0,X}(p|F_{Y_0|X}(y_0|x),x)\right] \\ &\qquad dF_{Z|X}(z|x)dF_{X|Y_0}(x|y_0), \end{aligned}$$

where the equality is from the fact that $F_{(U_D,X,Z)|Y_0=y_0} = F_{U_D|Y_0=y_0,X,Z}\cdot F_{Z|Y_0=y_0,X}\cdot F_{X|Y_0=y_0} = F_{U_D|U_0=F_{Y_0|X}(y_0|X),X}\cdot F_{Z|X}\cdot F_{X|Y_0=y_0}$, and U_0 is defined in the Skorohod representation of Y_0, $Y_0|X = F_{Y_0|X}^{-1}(U_0|X)$. So this parameter is a complicated functional of $Q_{Y_1|X,U_D}(F_{Y_0|X,U_D}(y_0|x,u_D)|x,u_D)$ and is not easy to estimate. Actually, it is unknown whether it can be point identified since $F_{U_D|U_0,X}$ is hard to be nonparametrically identified without further structures on the model.

3 Sharp Bounds for the QTE

Although $Q_{Y_d|X,U_D}(\tau|x,u_D)$ can be point identified from Theorem 1, we show in this section that $Q_{Y_d|X}(\tau|x)$ generally can only be partially identified, which implies that $\Delta_\tau^{QTE}(x)$ can only be partially identified. Here, we implicitly assume that the RP assumption on $X = x$ holds (i.e., Y_d can be represented as $Y_d = q(d,X,U_d)$

[8] This parameter is useful, e.g., in the median-voter model, we need to check whether $P(Y_1 > Y_0|D=1)P(D=1) > 1/2$.

with $U_0|(X=x) = U_1|(X=x))$, but we do not explicitly explore the information content in this assumption.[9] First, we impose the quantile independence assumption (QIA),

$$Q_{Y_d|X,Z}(\tau|X,Z) = Q_{Y_d|X}(\tau|X) \text{ for all } \tau \in (0,1). \tag{8}$$

This assumption is equivalent to $(Y_1, Y_0) \perp Z|X$. This assumption is parallel to the usual IV assumption $E[Y_d|X,Z] = E[Y_d|X]$ in the average treatment effect evaluation. As in Heckman and Vytlacil (2001b), we assume further that

$$D = 1(p(X,Z) \geq U_D) \text{ and } Z \perp U_D|X \tag{9}$$

to study the improvement on the bounds for $\Delta_\tau^{QTE}(x)$.

3.1 Bounds Under the Quantile Independence Assumption

From Proposition 2 and (36) of Manski (1994), we have sharp bounds for $Q_{Y_d|X}(\tau|x)$ under (8):

$$\sup_{z\in\mathcal{Z}_x} L_\tau^1(x,z) \leq Q_{Y_1|X}(\tau|x) \leq \inf_{z\in\mathcal{Z}_x} R_\tau^1(x,z),$$
$$\sup_{z\in\mathcal{Z}_x} L_\tau^0(x,z) \leq Q_{Y_0|X}(\tau|x) \leq \inf_{z\in\mathcal{Z}_x} R_\tau^0(x,z),$$

where $\mathcal{Z}_x \equiv$supp$(Z|X=x)$,

$$\begin{aligned}
L_\tau^1(x,z) &= \begin{cases} Q_{Y|X,Z,D}\left(1 - \frac{1-\tau}{p(x,z)}\middle| x,z,1\right), & \text{if } p(x,z) > 1-\tau, \\ -\infty, & \text{otherwise,}\end{cases} \\
R_\tau^1(x,z) &= \begin{cases} Q_{Y|X,Z,D}\left(\frac{\tau}{p(x,z)}\middle| |x,z,1\right), & \text{if } p(x,z) \geq \tau, \\ \infty, & \text{otherwise,}\end{cases} \\
L_\tau^0(x,z) &= \begin{cases} Q_{Y|X,Z,D}\left(1 - \frac{1-\tau}{1-p(x,z)}\middle| x,z,0\right), & \text{if } p(x,z) < \tau, \\ -\infty, & \text{otherwise,}\end{cases} \\
R_\tau^0(x,z) &= \begin{cases} Q_{Y|X,Z,D}\left(\frac{\tau}{1-p(x,z)}\middle| x,z,0\right), & \text{if } p(x,z) \leq 1-\tau, \\ \infty, & \text{otherwise.}\end{cases}
\end{aligned} \tag{10}$$

So

$$I_\tau^L(x) \equiv \sup_{z\in\mathcal{Z}_x} L_\tau^1(x,z) - \inf_{z\in\mathcal{Z}_x} R_\tau^0(x,z) \leq \Delta_\tau^{QTE}(x) \leq \inf_{z\in\mathcal{Z}_x} R_\tau^1(x,z) - \sup_{z\in\mathcal{Z}_x} L_\tau^0(x,z) \equiv I_\tau^U(x). \tag{11}$$

This bound is trivial, since $I_\tau^L(x) = -\infty$ and $I_\tau^U(x) = \infty$ if Y_1 and Y_0 are unbounded. Similar phenomena also happen in the average treatment effect evaluation. To avoid such trivial results, we assume that

$$P\left(y_d^l(x) \leq Y_d \leq y_d^u(x)|X=x,Z\right) = 1, \tag{12}$$

where $y_d^l(x), y_d^u(x) \in \mathbb{R}$ does not depend on Z from (8). To simplify notations, we assume that $y_0^l(x) = y_1^l(x)$, denoted as $y^l(x)$, and $y_0^u(x) = y_1^u(x)$, denoted as $y^u(x)$. Then $-\infty$ in (10) is changed to $y^l(x)$ and ∞ is changed to $y^u(x)$.

Let $\mathcal{P}_x \equiv$supp$(p(X,Z)|X=x)$, $p_x^{\sup} = \sup \mathcal{P}_x$ and $p_x^{\inf} = \inf \mathcal{P}_x$. The width of the bounds is $I_\tau^U(x) - I_\tau^L(x)$, a complicated expression to evaluate, especially if $\mathcal{Z}_x$ is uncountable. Note that the above bounds exactly identify $\Delta_\tau^{QTE}(x)$ if $I_\tau^L(x) = I_\tau^U(x)$. Note also that it is neither necessary nor sufficient for $p(x,z)$

[9] In Section 6.1, we will show how Chernozhukov and Hansen (2005) point identify $Q_{Y_d|X}(\tau|x)$ by exploring the information content in this assumption and imposing some completeness conditions.

to be a nontrivial function of z for these bounds to improve upon the bounds when (8) is not imposed (i.e., $\sup_{z\in\mathcal{Z}_x}$ and $\inf_{z\in\mathcal{Z}_x}$ are dropped from (11)); this is because $Q_{Y|X,Z,D}(\tau|x,z,d)$ may depend on z through other channels than $p(x,z)$. Evaluating the bounds for $\Delta_\tau^{QTE}(x)$ requires knowledge of

$$\left\{p(x,z), Q_{Y|X,Z,D}\left(1-\frac{1-\tau}{p(x,z)}\middle| x,z,1\right), Q_{Y|X,Z,D}\left(\frac{\tau}{p(x,z)}\middle| x,z,1\right),\right.$$
$$\left.y^l(x), y^u(x), Q_{Y|X,Z,D}\left(1-\frac{1-\tau}{1-p(x,z)}\middle| x,z,0\right), Q_{Y|X,Z,D}\left(\frac{\tau}{1-p(x,z)}\middle| x,z,0\right)\right\}$$

for each $z \in \mathcal{Z}_x$; estimators of these objects can be constructed in an obvious way, so are omitted here.

The following theorem is parallel to Corollary 1 and 2 of Proposition 6 in Manski (1994). It develops necessary and sufficient conditions on $p_x^{\sup}$ and $p_x^{\inf}$ for point identifying $\Delta_\tau^{QTE}(x)$.

Theorem 2 *Suppose assumptions (8) and (12) hold.*

(i) $p_x^{\sup} \geq \min\{\tau, 1-\tau\}$ *and* $p_x^{\inf} \leq \max\{\tau, 1-\tau\}$ *are necessary for point identification of* $\Delta_\tau^{QTE}(x)$. *Also, when* $p_x^{\inf}$ *and* $p_x^{\sup}$ *are achieved at some values that* Z *can take,* $p_x^{\inf} = 0$ *and* $p_x^{\sup} = 1$ *are sufficient for point identification of* $\Delta_\tau^{QTE}(x)$ *for any fixed* $\tau \in (0,1)$.

(ii) *When* $(Y_1, Y_0) \perp D|\,(X,Z)$, $p_x^{\sup} = 1$ *and* $p_x^{\inf} = 0$ *is sufficient for point identification of* $\Delta_\tau^{QTE}(x)$ *for any fixed* $\tau \in (0,1)$. *If assume further that* $Y_d|\,(X=x)$ *is continuously distributed with a positive density on* $(y^l(x), y^u(x))$, *then* $p_x^{\sup} = 1$ *and* $p_x^{\inf} = 0$ *is also necessary for point identification of* $\Delta_\tau^{QTE}(x)$ *using (11).*

For the average treatment effect evaluation, Corollary 1 of Proposition 6 in Manski (1994) implies that $p_x^{\inf} \leq 1/2$ and $p_x^{\sup} \geq 1/2$ are necessary for point identification of $\Delta^{ATE}(x) \equiv E[Y_1 - Y_0|X=x]$. A key assumption for this result to hold is that the support of $Y_d|\,(X=x,Z)$ does not depend on Z. Our first necessary condition requires only the support independence assumption rather that the full independence assumption (8). So these two sets of necessary conditions are comparable. When $\tau = 1/2$, they are the same.

To understand the sufficient condition for point identification of $\Delta_\tau^{QTE}(x)$ in Theorem 2(i), we need to clarify the meaning of $Q_{Y|X,Z,D}(\tau(x,\overline{z})|x,\overline{z},d)$ when $\overline{z} \in \mathcal{Z}_x$ but cannot be taken by Z, where $\tau(x,z)$ is the quantile index as a function of x and z. Since in this case $Q_{Y|X,Z,D}(\tau(x,\overline{z})|x,\overline{z},d)$ is not defined, it should be understood as the continuous extension of $Q_{Y|X,Z,D}(\tau(x,z)|x,z,d)$ as z converges to $\overline{z}$. The quantile functions in (13) below are similarly understood when $p_x^{\sup}$ and $p_x^{\inf}$ cannot be taken by $p(x,z)$. This extension is not required when Z is discretely distributed. When Z has a continuous component, we can assume that Z can take all values in $\mathcal{Z}_x$, and assume (8) is satisfied for $Z \in \mathcal{Z}_x$. This assumption does not lose generality since redefining a continuous random variable on a set with Lebesgue measure zero will not affect its distribution at all. Under this extension, $p_x^{\inf}$ and $p_x^{\sup}$ must be achieved at some values in $\mathcal{Z}_x$, so $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ are sufficient for point identification of $\Delta_\tau^{QTE}(x)$ for any fixed $\tau \in (0,1)$ regardless of Z is discrete or continuous or a mixture. Finally, note that $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ is essentially the usual large support condition which entails identification-at-infinity.

In the average treatment effect evaluation, there are not sufficient conditions for point identification of $\Delta^{ATE}(x)$ in the literature. From Manski (1994),

$$I^L(x) \leq \Delta^{ATE}(x) \leq I^U(x),$$

where

$$
\begin{aligned}
I^{L}(x) = & \sup_{z\in\mathcal{Z}_x}\left\{p(x,z)E\left[Y_1|X=x,Z=z,D=1\right]+(1-p(x,z))y^{l}(x,z)\right\} \\
& -\inf_{z\in\mathcal{Z}_x}\left\{(1-p(x,z))\,E\left[Y_0|X=x,Z=z,D=0\right]+p(x,z)y^{u}(x,z)\right\}, \\
I^{U}(x) = & \inf_{z\in\mathcal{Z}_x}\left\{p(x,z)E\left[Y_1|X=x,Z=z,D=1\right]+(1-p(x,z))y^{u}(x,z)\right\} \\
& -\sup_{z\in\mathcal{Z}_x}\left\{(1-p(x,z))\,E\left[Y_0|X=x,Z=z,D=0\right]+p(x,z)y^{l}(x,z)\right\},
\end{aligned}
$$

and $y^l(x,z)$, $y^u(x,z)\in\mathbb{R}$ satisfy $P\left(y^l(x,z)\leq Y_d\leq y^u(x,z)|X=x,Z=z\right)=1$. Note here that $y^l(x,z)$ and $y^u(x,z)$ depend on z if only the mean independence assumption, $E[Y_d|X,Z]=E[Y_d|X]$, is imposed. As in Theorem 2(i), when $p_x^{\inf}$ and $p_x^{\sup}$ are achieved at some values that Z can take, $p_x^{\inf}=0$ and $p_x^{\sup}=1$ implies that

$$
\begin{aligned}
E\left[Y_1|X=x,Z=z,D=1\right]-E\left[Y_0|X=x,Z=z,D=0\right]\leq I^L(x)\leq I^U(x) \\
\leq E\left[Y_1|X=x,Z=z,D=1\right]-E\left[Y_0|X=x,Z=z,D=0\right],
\end{aligned}
$$

so $\Delta^{ATE}(x)$ is point identified.

Corollary 2 of Proposition 6 in Manski (1994) implies that when $(Y_1,Y_0)\perp D|\,(X,Z)$, $\Delta^{ATE}(x)$ is point identified using his bound (35) or $[I^L(x),I^U(x)]$ above if and only if $p_x^{\sup}=1$ and $p_x^{\inf}=0$. Our result parallels his result when $Y_d|\,(X=x)$ is continuously distributed with a positive density on $(y^l(x),y^u(x))$. It should be emphasized that when $(Y_1,Y_0)\perp D|\,(X,Z)$, $p_x^{\sup}=1$ and $p_x^{\inf}=0$ is necessary for point identification of $\Delta_\tau^{QTE}(x)$ only when (11) is used. Actually, since $Q_{Y_d|X}(\tau|x)=Q_{Y_d|X,Z}(\tau|x,z)=Q_{Y|X,Z,D}(\tau|x,z,d)$, $Q_{Y_d|X}(\tau|x)$ can be identified directly from $Q_{Y|X,Z,D}(\tau|x,z,d)$.

3.2 Bounds Under the Nonparametric Selection Model

The following theorem states the bounds for $\Delta_\tau^{QTE}(x)$ when assumption (9) is imposed.

Theorem 3 *Suppose assumptions (8), (9) and (12) hold.*

(i) $\Delta_\tau^{QTE}(x)$ *has sharp bounds,*

$$
L_\tau^1(x)-R_\tau^0(x)\leq\Delta_\tau^{QTE}(x)\leq R_\tau^1(x)-L_\tau^0(x), \tag{13}
$$

where

$$
\begin{aligned}
L_\tau^1(x) &= \begin{cases} Q_{Y|X,p(X,Z),D}\left(1-\frac{1-\tau}{p_x^{\sup}}\middle|\,x,p_x^{\sup},1\right), & \text{if } p_x^{\sup}>1-\tau, \\ y^l(x), & \text{otherwise,}\end{cases} \\
R_\tau^1(x) &= \begin{cases} Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x^{\sup}}\middle|\,x,p_x^{\sup},1\right), & \text{if } p_x^{\sup}\geq\tau, \\ y^u(x), & \text{otherwise,}\end{cases} \\
L_\tau^0(x) &= \begin{cases} Q_{Y|X,p(X,Z),D}\left(1-\frac{1-\tau}{1-p_x^{\inf}}\middle|\,x,p_x^{\inf},0\right), & \text{if } p_x^{\inf}<\tau, \\ y^l(x), & \text{otherwise,}\end{cases} \\
R_\tau^0(x) &= \begin{cases} Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{1-p_x^{\inf}}\middle|\,x,p_x^{\inf},0\right), & \text{if } p_x^{\inf}\leq 1-\tau, \\ y^u(x), & \text{otherwise.}\end{cases}
\end{aligned}
$$

(ii) $p_x^{\inf}=0$ *and* $p_x^{\sup}=1$ *are sufficient for point identification of* $\Delta_\tau^{QTE}(x)$ *for any fixed* $\tau\in(0,1)$*. When* $Y|\,(X=x,p(X,Z)=p_x^{\sup},D=1)$ *and* $Y|\left(X=x,p(X,Z)=p_x^{\inf},D=0\right)$ *are continuously distributed with a positive density on* $(y^l(x),y^u(x))$*, they are also necessary.*

(iii) $[I_\tau^L(x), I_\tau^U(x)]$ *in (11) will simplify to the bounds in (13) under assumption (9).*

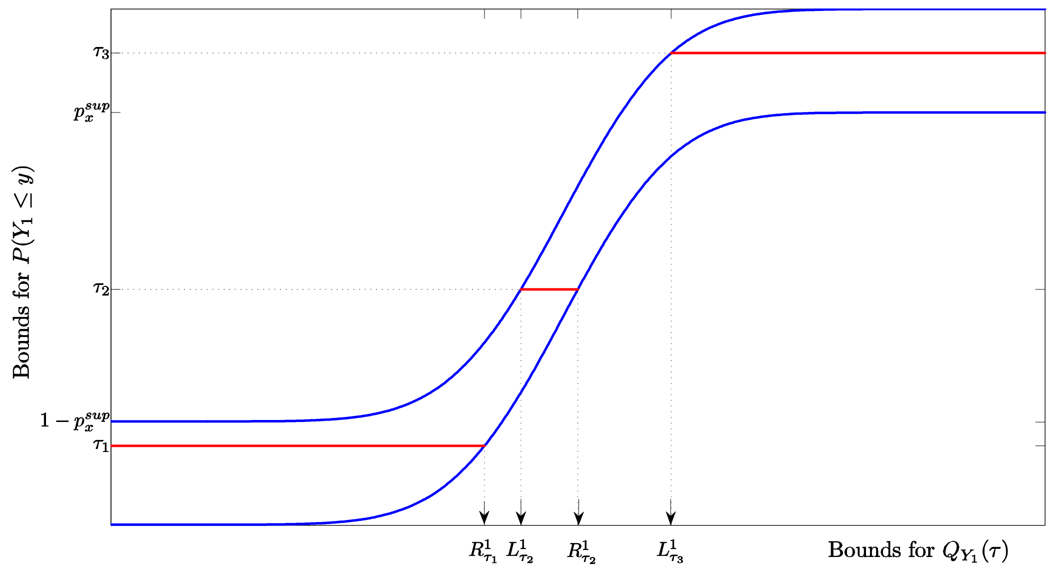

Figure 5: Intuition for $L_\tau^1 \leq Q_{Y_1}(\tau) \leq R_\tau^1$: $p_x^{\sup} = 0.8$, $\rho = 0.5$, $\tau_1 = 0.15$, $\tau_2 = 0.46$ and $\tau_3 = 0.91$

Figure 5 provides some intuition for why $L_\tau^1(x) \leq Q_{Y_1|X}(\tau|x) \leq R_\tau^1(x)$; similar intuition can be applied to the bounds for $Q_{Y_0|X}(\tau|x)$. From the proof of Theorem 3,

$$P\left(Y \leq y|p(Z) = p_x^{\sup}, D = 1\right) p_x^{\sup} \leq P(Y_1 \leq y) \leq P\left(Y \leq y|p(Z) = p_x^{\sup}, D = 1\right) p_x^{\sup} + (1 - p_x^{\sup}),$$

where the conditioning on $X = x$ is depressed. Suppose $(Y_1, V) \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$; then

$$P\left(Y \leq y|p(Z) = p_x^{\sup}, D = 1\right) p_x^{\sup} = \int_0^{p_x^{\sup}} \Phi\left(\frac{y - \rho\Phi^{-1}(u_D)}{\sqrt{1-\rho^2}}\right) du_D.$$

Figure 5 shows the bounds for $P(Y_1 \leq y)$ when $p_x^{\sup} = 0.8$ and $\rho = 0.5$. Inverting the bounds for $P(Y_1 \leq y)$, we can get the bounds for $Q_{Y_1}(\tau)$. When $\tau \leq 1 - p_x^{\sup}$, $L_\tau^1 = y^l$; when $\tau > p_x^{\sup}$, $R_\tau^1 = y^u$. Only if $\tau \in (1 - p_x^{\sup}, p_x^{\sup})$, both bounds are nontrivial. This is not always possible; only if $p_x^{\sup} > \max(\tau, 1-\tau) \geq 1/2$ ($p_x^{\min} < \min(\tau, 1-\tau)$), neither the left nor the right bound for $Q_{Y_1}(\tau)$ ($Q_{Y_0}(\tau)$) is trivial. Pushing $\tau \to 0$ or 1, we can see that there are nontrivial bounds for $Q_{Y_1}(\tau)$ ($Q_{Y_0}(\tau)$) for *all* τ if and only if $p_x^{\sup} = 1$ ($p_x^{\min} = 0$).

Note that $L_\tau^d(x)$ and $R_\tau^d(x)$ are increasing functions of τ; hence the bound for $Q_{Y_d|X}(\tau|x)$ shifts to the right as τ increases. Also observe that

$$1 - \frac{1-\tau}{p_x^{\sup}} \leq \tau \leq \frac{\tau}{p_x^{\sup}} \text{ and } 1 - \frac{1-\tau}{1-p_x^{\inf}} \leq \tau \leq \frac{\tau}{p_x^{\inf}}.$$

Hence $Q_{Y|X,p(X,Z),D}\left(\tau|x, p_x^{\sup}, 1\right)$ and $Q_{Y|X,p(X,Z),D}\left(\tau|x, p_x^{\inf}, 0\right)$ lie within the bound for $Q_{Y_1|X}(\tau|x)$ and $Q_{Y_0|X}(\tau|x)$, respectively. This implies that $F_{Y_1|X}(\cdot|x) = F_{Y|X,p(X,Z),D}(\cdot|x, p_x^{\sup}, 1)$ and $F_{Y_0|X}(\cdot|x) = F_{Y|X,p(X,Z),D}(\cdot|x, p_x^{\inf}, 0)$ are not rejectable in the absence of other information.

Evaluating the bounds for $\Delta_\tau^{QTE}(x)$ requires knowledge only of

$$\left\{p_x^{\inf}, p_x^{\sup}, Q_{Y|X,p(X,Z),D}\left(1-\frac{1-\tau}{p_x^{\sup}}\middle| x, p_x^{\sup}, 1\right), Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x^{\sup}}\middle| x, p_x^{\sup}, 1\right),\right.$$
$$\left. y^l(x), y^u(x), Q_{Y|X,p(X,Z),D}\left(1-\frac{1-\tau}{1-p_x^{\inf}}\middle| x, p_x^{\inf}, 0\right), Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{1-p_x^{\inf}}\middle| x, p_x^{\inf}, 0\right)\right\}.$$

Estimators of these objects can be constructed in an obvious way, so are omitted here. The simpler structure of these bounds results from assumption (9). Also, it is both necessary and sufficient for $p(x,z)$ to be a nontrivial function of z for the bounds in Theorem 3 to improve upon the bounds when (8) and (9) are not imposed.

As shown in Section 4 of Heckman and Vytlacil (2001b), $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ are necessary and sufficient for point identification of $\Delta^{ATE}(x)$ under assumption (9). Our result parallels their result when $Y_d|\,(X = x)$ is continuously distributed with a positive density on $(y^l(x), y^u(x))$. As shown in Section 6 of Heckman and Vytlacil (2001b), the Manski IV bounds of $\Delta^{ATE}(x)$ simplify to their bounds under assumption (9); the last part of Theorem 3 parallels their result.

Finally, note that the bounds for $\Delta_\tau^{QTE}(x)$ can be integrated (with respect to x) to get the bounds for Δ_τ^{IQTE}.

3.3 Some Counterexamples

The bounds for $\Delta_\tau^{QTE}(x)$ in Theorem 3 can be applied to cases with discrete, continuous or mixed response variables. Note that $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ are necessary for point identification of $\Delta_\tau^{QTE}(x)$ only when $Y|\,(X = x, p(X,Z) = p_x^{\sup}, D = 1)$ and $Y|\,\left(X = x, p(X,Z) = p_x^{\inf}, D = 0\right)$ are continuously distributed with a positive density on $(y^l(x), y^u(x))$. The following example illustrates that $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ is not necessary for point identification of $\Delta_\tau^{QTE}(x)$ when Y_d is binary. The supplementary materials include another example in a similar spirit where the distribution of Y_d is a mixture of continuous and discrete.

Example 1 *Suppose $Y_d \in \{0,1\}$. $p_{x1}^{\sup} \equiv P(Y = 0|X = x, p(X,Z) = p_x^{\sup}, D = 1) \in (0,1)$ and $p_{x0}^{\inf} \equiv P(Y = 0|X = x, p(X,Z) = p_x^{\inf}, D = 0) \in (0,1)$. First check the bounds for $Q_{Y_1|X}(\tau|x)$:*

$$L_\tau^1(x) = \begin{cases} 1 & \text{if } p_x^{\sup} > 1-\tau \text{ and } 1-\frac{1-\tau}{p_x^{\sup}} > p_{x1}^{\sup}, \\ 0, & \text{if } p_x^{\sup} \le 1-\tau \text{ or } [p_x^{\sup} > 1-\tau \text{ and } 1-\frac{1-\tau}{p_x^{\sup}} \le p_{x1}^{\sup}], \end{cases}$$
$$R_\tau^1(x) = \begin{cases} 0, & \text{if } p_x^{\sup} \ge \tau \text{ and } \frac{\tau}{p_x^{\sup}} \le p_{x1}^{\sup}, \\ 1, & \text{if } p_x^{\sup} < \tau \text{ or } [p_x^{\sup} \ge \tau \text{ and } \frac{\tau}{p_x^{\sup}} > p_{x1}^{\sup}]. \end{cases}$$

When $\max\left\{1-\tau, \frac{\tau}{p_{x1}^{\sup}}\right\} < p_x^{\sup} \le \frac{1-\tau}{1-p_{x1}^{\sup}}$ or $\frac{\tau}{p_{x1}^{\sup}} \le p_x^{\sup} \le 1-\tau$, $L_\tau^1(x) = R_\tau^1(x) = 0$; when $\frac{1-\tau}{1-p_{x1}^{\sup}} < p_x^{\sup} < \tau$ or $\max\left\{\frac{1-\tau}{1-p_{x1}^{\sup}}, \tau\right\} < p_x^{\sup} < \frac{\tau}{p_{x1}^{\sup}}$, $L_\tau^1(x) = R_\tau^1(x) = 1$. Similarly, when $1-\frac{1-\tau}{1-p_{x0}^{\inf}} \le p_x^{\inf} < \min\left\{\tau, 1-\frac{\tau}{p_{x0}^{\inf}}\right\}$ or $\tau \le p_x^{\inf} \le 1-\frac{\tau}{p_{x0}^{\inf}}$, $L_\tau^0(x) = R_\tau^0(x) = 0$; when $1-\tau < p_x^{\inf} < 1-\frac{1-\tau}{1-p_{x0}^{\inf}}$ or $1-\frac{\tau}{p_{x0}^{\inf}} < p_x^{\inf} < \min\left\{1-\tau, 1-\frac{1-\tau}{1-p_{x0}^{\inf}}\right\}$, $L_\tau^0(x) = R_\tau^0(x) = 1$. Figure 6 shows the point identification combination of $p_x^{\sup}$ ($p_x^{\inf}$) and $p_{x1}^{\sup}$ ($p_{x0}^{\inf}$) for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$. Obviously, $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ are not necessary for point identification of $\Delta_\tau^{QTE}(x)$. Only if $p_{x1}^{\sup} = p_{x0}^{\inf} = \tau$, $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ are necessary. Note also that $p_x^{\sup} \ge \min\{\tau, 1-\tau\}$ and $p_x^{\inf} \le \max\{\tau, 1-\tau\}$ for point identification of $\Delta_\tau^{QTE}(x)$ for any $p_{x1}^{\sup}$, $p_{x0}^{\inf} \in (0,1)$ as predicted by Theorem 2. $\square$

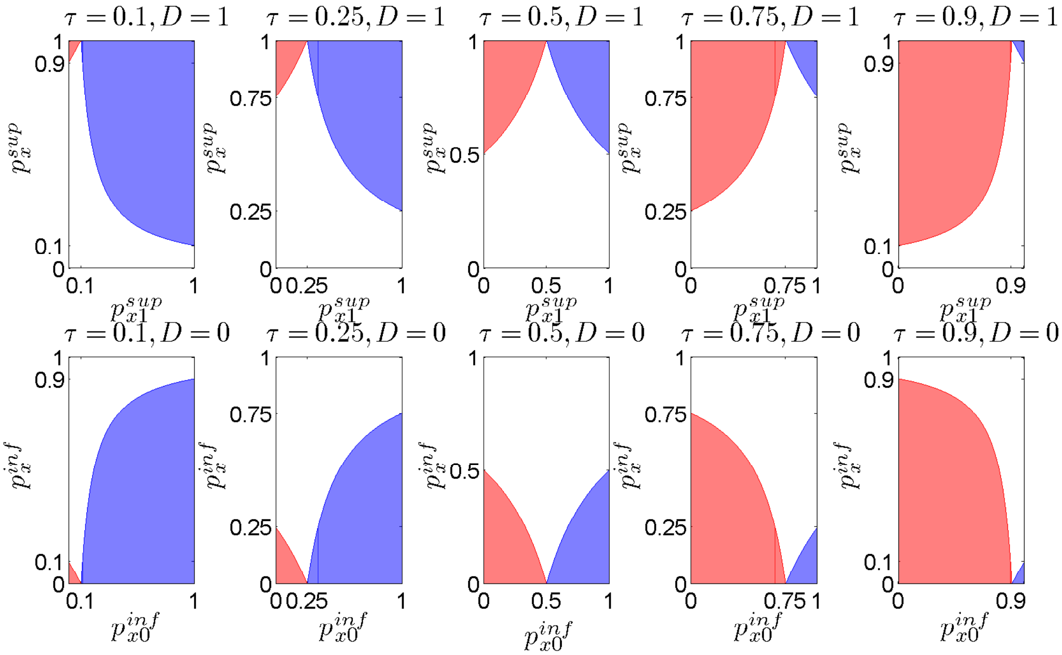

Figure 6: $p_x^{\sup}$ ($p_x^{\inf}$) and $p_{x1}^{\sup}$ ($p_{x0}^{\inf}$) for Point Identification of $Q_{Y_d|X}(\tau|x)$: Red Area for $Q_{Y_d|X}(\tau|x) = 1$ and Blue Area for $Q_{Y_d|X}(\tau|x) = 0$

The next example shows that $[I_\tau^L(x), I_\tau^U(x)]$ in (11) may not simplify to the bounds in Theorem 3 if assumption (9) is not imposed. This example parallels the example in Section 6 of Heckman and Vytlacil (2001b) where they show a similar result for $\Delta^{ATE}(x)$.

Example 2 *Suppose Z is binary and there are no other covariates. Take $\inf_{z\in\mathcal{Z}_x} R_\tau^1(x,z)$ as an example; suppose $y^l(x) = 0$, $y^u(x) = 1$ and $p(1) \equiv p(x,1) > p(x,0) \equiv p(0)$. We want to show that it is possible to have*

$$\begin{aligned}&\min\left\{Q_{Y|Z,D}\left(\frac{\tau}{p(1)}\middle|1,1\right)1(p(1)\geq\tau)+1(p(1)<\tau), Q_{Y|Z,D}\left(\frac{\tau}{p(0)}\middle|0,1\right)1(p(0)\geq\tau)+1(p(0)<\tau)\right\}\\&=Q_{Y|Z,D}\left(\frac{\tau}{p(0)}\middle|0,1\right)1(p(0)\geq\tau)+1(p(0)<\tau)<Q_{Y|Z,D}\left(\frac{\tau}{p(1)}\middle|1,1\right)1(p(1)\geq\tau)+1(p(1)<\tau).\end{aligned}$$

We must assume $\min\{p(0),p(1)\}\geq\tau$ to make this result hold. If $\min\{p(0),p(1)\}\geq\tau$, we need only check

$$q_1 \equiv Q_{Y|Z,D}\left(\frac{\tau}{p(1)}\middle|1,1\right) > Q_{Y|Z,D}\left(\frac{\tau}{p(0)}\middle|0,1\right) \equiv q_0.$$

First, the QIA needs to be satisfied. Without loss of generality, assume $Y_1|Z$ is uniformly distributed. Then the QIA is satisfied if

$$\begin{aligned}F_{Y_1|Z}(y_1|0) &= F_{Y_1|Z,D}(y_1|0,0)(1-p(0)) + F_{Y_1|Z,D}(y_1|0,1)p(0) = y_1,\\ F_{Y_1|Z}(y_1|1) &= F_{Y_1|Z,D}(y_1|1,0)(1-p(1)) + F_{Y_1|Z,D}(y_1|1,1)p(1) = y_1,\end{aligned} \tag{14}$$

for any $y_1 \in [0,1]$*. As long as*

$$F_{Y_1|Z,D}(q_0|0,0) = \frac{q_0 - \tau}{1-p(0)} \in (0,1) \;\; or \; \tau < q_0 < \tau + (1-p(0)),$$
$$F_{Y_1|Z,D}(q_1|1,0) = \frac{q_1 - \tau}{1-p(1)} \in (0,1) \;\; or \; \tau < q_1 < \tau + (1-p(1)),$$

we can find qualified $F_{Y_1|Z,D}(y_1|z,d)$*,* $z = 0,1$*,* $d = 0,1$ *such that (14) is satisfied. For example, let*

$$F_{Y_1|Z,D}(y_1|0,0) = \frac{q_0 - \tau}{(1-p(0))\, q_0} y_1 1(y_1 \le q_0) + \left(\frac{\frac{q_0-\tau}{1-p(0)} - q_0}{1-q_0} + \frac{1 - \frac{q_0-\tau}{1-p(0)}}{1-q_0} y_1 \right) 1(y_1 > q_0),$$
$$F_{Y_1|Z,D}(y_1|0,1) = \frac{\tau}{p(0) q_0} y_1 1(y_1 \le q_0) + \left(\frac{\frac{\tau}{p(0)} - q_0}{1-q_0} + \frac{1 - \frac{\tau}{p(0)}}{1-q_0} y_1 \right) 1(y_1 > q_0),$$
$$F_{Y_1|Z,D}(y_1|1,0) = \frac{q_1 - \tau}{(1-p(1))\, q_1} y_1 1(y_1 \le q_1) + \left(\frac{\frac{q_1-\tau}{1-p(1)} - q_1}{1-q_1} + \frac{1 - \frac{q_1-\tau}{1-p(1)}}{1-q_1} y_1 \right) 1(y_1 > q_1),$$
$$F_{Y_1|Z,D}(y_1|1,1) = \frac{\tau}{p(1) q_1} y_1 1(y_1 \le q_1) + \left(\frac{\frac{\tau}{p(1)} - q_1}{1-q_1} + \frac{1 - \frac{\tau}{p(1)}}{1-q_1} y_1 \right) 1(y_1 > q_1).$$

Figure 7 shows the case with $\tau = 0.5, p(0) = 0.6, p(1) = 0.7, q_0 = 0.65 < 0.75 = q_1$*.* $\square$

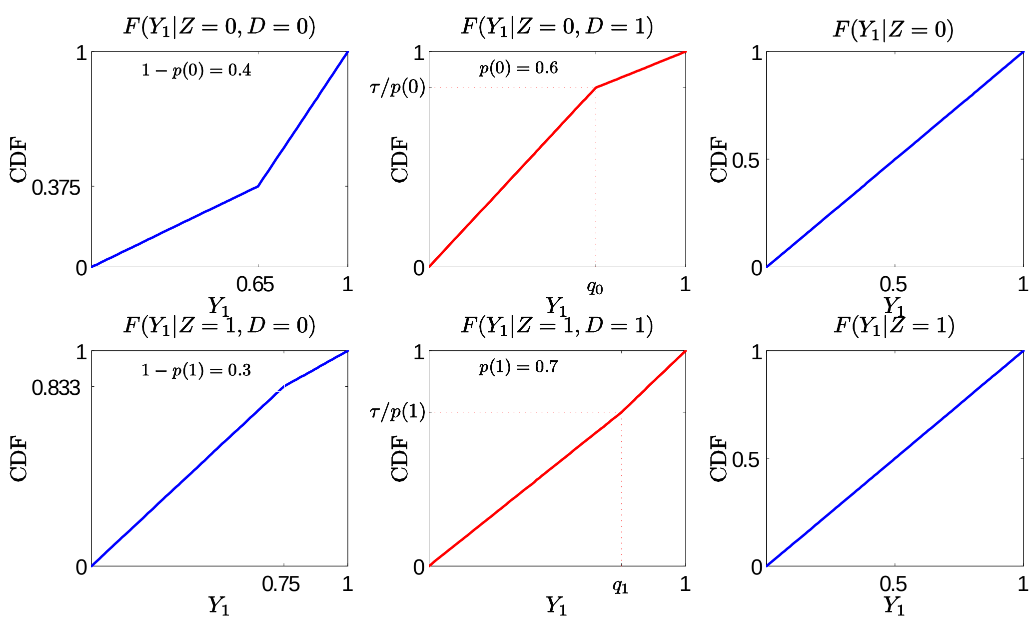

Figure 7: An Illustration of $\inf_{z \in \mathcal{Z}_x} R^1_\tau(x,z) \neq R^1_\tau(x)$ When (9) is NOT Satisfied: $\tau = 0.5$

It is useful to construct a test to check the hypothesis that the bounds $[I^L_\tau(x), I^U_\tau(x)]$ and those in Theorem 3(i) coincide. Since $I^L_\tau(x) \ge L^1_\tau(x) - R^0_\tau(x)$ and $I^U_\tau(x) \le R^1_\tau(x) - L^0_\tau(x)$ always hold, our null hypothesis is

$$L^1_\tau(x) - R^0_\tau(x) - I^L_\tau(x) \ge 0, \text{ and } I^U_\tau(x) - \left[R^1_\tau(x) - L^0_\tau(x)\right] \ge 0,$$

or

$$
\begin{aligned}
L_\tau^1(x) - \sup_{z\in\mathcal{Z}_x} L_\tau^1(x,z) \ge 0,\ \inf_{z\in\mathcal{Z}_x} R_\tau^0(x,z) - R_\tau^0(x) \ge 0,\\
\text{and } L_\tau^0(x) - \sup_{z\in\mathcal{Z}_x} L_\tau^0(x,z) \ge 0,\ \inf_{z\in\mathcal{Z}_x} R_\tau^1(x,z) - R_\tau^1(x) \ge 0.
\end{aligned}
$$

This test can also serve as a test of the monotonicty assumption (9) although (9) is only sufficient (may not be necessary) for the null to hold. The left hand side of the inequalities in the null hypothesis is expressed as intersection bounds, so the estimation and inference procedures in Chernozhukov et al. (2013) are useful for our purpose. The complication here is that the bounds also involve $p(x,z)$ which must be estimated at the first place. This complication also appears in applying the general results in Imbens and Manski (2004) and Stoye (2009) to construct confidence intervals for the bounds in Theorem 3(i).

4 Semiparametric Estimation of the MQTE

Before stating our semiparametric estimation of the MQTE, we first use the familiar parametric specification to motivate our semiparametric setup. Then the MQTE is estimated by the distribution regression and the weak limit of the corresponding estimator is derived. It follows to show that the bootstrap is valid for inferences based on the MQTE. Finally, we put forward some goodness of fit tests to choose regressors in the distribution regression.

4.1 Parametric Motivation

Suppose

$$
\begin{aligned}
&Y_1 = \alpha + \varphi + X'\beta_1 + U_1,\\
&Y_0 = \alpha + X'\beta_0 + U_0,\\
&D = 1(V \le \phi + X'\gamma_1 + Z'\gamma_2).
\end{aligned}
\tag{15}
$$

This model can be interpreted as the Generalized Roy Model (GRM) (Heckman and Vytlacil (2001a)). For example, suppose the cost of receiving treatment is $C = -Z'\gamma_2 + U_C$, and the decision of participation is determined by a benefit-cost analysis: $D = 1(Y_1 - Y_0 - C \ge 0)$; then let $\gamma_1 = \beta_1 - \beta_0$, $\phi = \varphi$, and $V = U_C + U_0 - U_1$, we get the model.

Suppose $(U_0, U_1, V) \sim N(0,\Sigma)$, where Σ represents the variance and covariance matrix with the variance of V being normalized as 1. In what follows, σ_d^2 denotes the variance of U_d, and σ_{Vd} denote the covariance between U_d and V. Due to the RP assumption, U_d can be expressed as $\sigma_{Vd}V + \sqrt{\sigma_d^2 - \sigma_{Vd}^2}U$ for the same U which follows $N(0,1)$ and is independent of V. Now,

$$
P(X,Z) = \Phi(\phi + X'\gamma_1 + Z'\gamma_2) \text{ and } \phi + X'\gamma_1 + Z'\gamma_2 = \Phi^{-1}(P(X,Z)).
$$

Additionally,

$$
\begin{aligned}
P(Y \le y | X = x, Z = z, D = 1) &= P(Y \le y | X = x, p(X,Z) = p, D = 1)\\
&= P(U_1 \le y - \alpha - \varphi - x'\beta_1 | V \le \phi + x'\gamma_1 + z'\gamma_2)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{p}\int_0^p \Phi\left(\frac{y-\alpha-\varphi-x'\beta_1-\sigma_{V1}\Phi^{-1}(u_D)}{\sqrt{\sigma_1^2-\sigma_{V1}^2}}\right)du_D \\
&= \Phi\left(\frac{y-\alpha-\varphi-x'\beta_1-\sigma_{V1}\Phi^{-1}(\overline{u}_D)}{\sqrt{\sigma_1^2-\sigma_{V1}^2}}\right),
\end{aligned}
$$

where $p = p(x, z)$, and $\overline{u}_D$ is a point between 0 and p. Similarly,

$$
P\left(Y \le y | X = x, p(X,Z) = p, D = 0\right) = \Phi\left(\frac{y-\alpha-x'\beta_0-\sigma_{V0}\Phi^{-1}(\widetilde{u}_D)}{\sqrt{\sigma_0^2-\sigma_{V0}^2}}\right),
$$

where $\widetilde{u}_D$ is a point between p and 1. Also, it is easy to see that

$$
\begin{aligned}
\Delta_\tau^{MQTE}(x,p) &= q_\tau\sqrt{\sigma_1^2-\sigma_{V1}^2}+\alpha+\varphi+x'\beta_1+\sigma_{V1}\Phi^{-1}(p)-q_\tau\sqrt{\sigma_0^2-\sigma_{V0}^2}-\alpha-x'\beta_0-\sigma_{V0}\Phi^{-1}(p) \\
&= q_\tau\left(\sqrt{\sigma_1^2-\sigma_{V1}^2}-\sqrt{\sigma_0^2-\sigma_{V0}^2}\right)+\varphi+x'\left(\beta_1-\beta_0\right)+\left(\sigma_{V1}-\sigma_{V0}\right)\Phi^{-1}(p),
\end{aligned}
$$

where q_τ is the τth quantile of $N(0,1)$. Whether $\Delta_\tau^{MQTE}(x,p)$ varies with τ depends on $\sqrt{\sigma_1^2-\sigma_{V1}^2}-\sqrt{\sigma_0^2-\sigma_{V0}^2}$, whether $\Delta_\tau^{MQTE}(x,p)$ varies with p depends on $\sigma_{V1}-\sigma_{V0}$, and whether $\Delta_\tau^{MQTE}(x,p)$ varies with x depends on $\beta_1-\beta_0$.

From the above calculation, we can specify the distribution regression as follows,

$$
P\left(Y \le y | X = x, p(X,Z) = p, D = d\right) = \Lambda\left(T(x,p)'\beta_d(y)\right), y \in \mathcal{Y}_d, \tag{16}
$$

where $\mathcal{Y}_d$ is a compact subset of supp(Y_d), Λ is a known link function, $T(x,p)$ is a vector of transformations of p and X such as polynomials or B-splines, and $\beta_d(y)$ is the unknown function-valued parameters. We divide $T(x,p)$ and $\beta_d(y)$ as $(1, \underline{T}(x,p)')'$ and $\left(\beta_{\alpha d}(y), \underline{\beta}_d(y)'\right)'$ for notational convenience. In the above example, $\Lambda(\cdot) = \Phi(\cdot)$, $\beta_{\alpha 0}(y) = \frac{y-\alpha}{\sqrt{\sigma_0^2-\sigma_{V0}^2}}$, $\beta_{\alpha 1}(y) = \frac{y-\alpha-\varphi}{\sqrt{\sigma_1^2-\sigma_{V1}^2}}$, $\underline{T}(x,p)'\underline{\beta}_0(y) = -\frac{x'\beta_0+\sigma_{V0}\Phi^{-1}(\widetilde{u}_D)}{\sqrt{\sigma_0^2-\sigma_{V0}^2}}$, and $\underline{T}(x,p)'\underline{\beta}_1(y) = -\frac{x'\beta_1+\sigma_{V1}\Phi^{-1}(\overline{u}_D)}{\sqrt{\sigma_1^2-\sigma_{V1}^2}}$. In this example, we can specify $\underline{T}(x,p) = (x', T(p)')'$ without interaction terms of x and p, and $\underline{\beta}_d(y)$ does not depend on y. So the specification of our distribution regression is quite general and covers the existing models as special cases. Another important feature of the distribution regression is that it does not require smoothness of the conditional density, since the approximation is done pointwise in the threshold y, and thus handles continuous, discrete, or mixed Y_d without any special adjustment.

The link function Λ can be the complementary log-log function, $\Lambda(v) = 1 - \exp\left(-\exp(v)\right)$, as in Cox (1972). Other useful link functions include the Logit, Probit, linear, log-log, and Gosset functions (see Koenker and Yoon (2009) for the latter). Note that the distribution regression model is flexible in the sense that, for any given link function Λ, we can approximate the conditional distribution function $F_{Y|X,p(X,Z),D}(y|x,p,d)$ arbitrarily well by using a rich enough $T(x,p)$. Thus, the choice of the link function is not important for sufficiently rich $T(x,p)$. To check whether enough terms are included in $T(x,p)$, we will in the following Section 4.5 develop some goodness of fit tests which extend the usual χ^2 goodness of fit test as suggested in e.g., Carneiro and Lee (2009) (Page 194 and footnote 21), Carneiro et al. (2003) (Section 7) and Hansen et al. (2004) (Section 5).[10] However, as in Abadie et al. (2002) and Chernozhukov and Hansen (2006) (see also Heckman et al. (2006) for the MTE case), we use the semiparametric rather than the fully nonparametric setup, so the dimension of $T(x,p)$ is fixed rather than diverges to infinity slowly. This assumption is valid from a practical point of view: each element of the function space where $F_{Y|X,p(X,Z),D}$

[10] See Section 2.3.6 of Chen (2007) for more suggestions.

stays can be approximated in a suitable norm by a finite-dimensional model, and the approximation error can be made arbitrarily small. Such a setup eases the inference of the marginal quantile treatment process as discussed in Section 4.3 and 4.4. In practice, $T(x,p)$ can include polynomials of p with the highest order 2 to catch the curvature in p since we need to differentiate $P(Y \leq y|X=x, p(X,Z)=p, D=d)$ with respect to p to get the MQTE. This is also a rule-of-thumb strategy in the literature, e.g., in the local polynomial estimation, Fan and Gijbels (1996) suggest that the order of polynomial be equal to one plus the order of the derivative of the function of interest.

In passing, we mention that our setup (16) is more general than that in Carneiro and Lee (2009), where they explore the special structure of (15), i.e., U_d is additively separable ($Y_d = \mu_d(X;\beta_d) + U_d$) and is independent of both X and Z (so there is no heteroskedasticity and $Q_\tau(Y_d|X, U_D)$ is parallel as a function of X for each τ and each U_D). In their setup, $F_{Y_d|X,U_D}(y|x,u_D)$ is completely controlled by the distribution of $U_d|(U_D = u_D)$. They use a two-step control function approach as in Das et al. (2003) to estimate $p(X,Z)$, β_d and U_d; then the density of $U_d|U_D$ is estimated using the estimate of $p(X,Z)$ and U_d, and the MQTE can be derived. Obviously, their estimation procedure is restricted to the continuous Y_d case. It should be emphasized that even if we assume $Y_d = \mu_d(X) + U_d$, and $D = 1(\mu_D(X,Z) - V \geq 0)$ with $(U_1, U_0, V) \perp (X,Z)$, we should add in interaction terms of x and p and let $\underline{\beta}_d(y)$ depend on y. To see why, note from the above derivation that if $F_{U_1|U_D=u_D}$ is continuous in u_D,

$$\begin{aligned} P(Y \leq y|X=x, p(X,Z)=p, D=1) &= \frac{1}{p}\int_0^p F_{U_1|U_D}(y-\mu_1(x)|u_D)du_D = F_{U_1|U_D}(y-\mu_1(x)|\overline{u}_D) \\ &= \Lambda\left(\Lambda^{-1}\left(F_{U_1|U_D}(y-\mu_1(x)|\overline{u}_D)\right)\right), \end{aligned}$$

where $\overline{u}_D \in [0,p]$. So as long as $\Lambda^{-1}\left(F_{U_1|U_D}(y-\mu_1(x)|\overline{u}_D)\right)$ does not degenerate to $a\left[y-\mu_1(x)+f(\overline{u}_D)\right]$, where a is a scalar, and $f(\cdot)$ is a generic function of u_D, the interaction terms should appear and $\underline{\beta}_1(y)$ should depend on y. Similar arguments can be applied to $P(Y \leq y|X=x, p(X,Z)=p, D=0)$.

4.2 Construction of the Quantile Treatment Estimators

As in Chernozhukov et al. (2013), we estimate $\beta_d(y)$, $y \in \mathcal{Y}_d$, by

$$\widehat{\beta}_d(y) = \arg\max_{\beta} \sum_{i=1}^{n} 1(D_i = d)\left[1(Y_i \leq y)\ln\Lambda\left(T(X_i,\widehat{p}_i)'\beta\right) + 1(Y_i > y)\ln\left(1-\Lambda\left(T(X_i,\widehat{p}_i)'\beta\right)\right)\right] \quad (17)$$

where $\widehat{p}_i = p(X_i, Z_i; \widehat{\gamma})$ is a parametric or semiparametric estimator of the propensity score at $(X_i', Z_i')'$. So our estimator is a two-step estimator: the first step estimates the propensity score and the second step estimates the counterfactual distributions. To be specific, we consider $p(X,Z;\gamma) = \Lambda(R(X,Z)'\gamma)$ in what follows, where $R(X,Z)$ is a vector of transformations of X and Z which is similarly defined as $T(x,p)$, and

$$\widehat{\gamma} = \arg\max_{\gamma} \sum_{i=1}^{n}\left[D_i \ln\Lambda\left(R(X_i,Z_i)'\gamma\right) + (1-D_i)\ln\left(1-\Lambda\left(R(X_i,Z_i)'\gamma\right)\right)\right].$$

Then the conditional CDF $P(Y \leq y|X=x, p(X,Z)=p, D=d)$ is estimated as

$$\widehat{F}_{Y|X,p(X,Z),D}(y|x,p,d) = \Lambda\left(T(x,p)'\widehat{\beta}_d(y)\right), \quad (18)$$

so $F_{Y_d|X,U_D}(y_d|x,p)$ is estimated as

$$\widehat{F}_{Y_1|X,U_D}(y_1|x,p) = \widehat{F}_{Y|X,p(X,Z),D}(y_1|x,p,1) + p\frac{\partial T(x,p)'}{\partial p}\widehat{\beta}_1(y_1)\cdot\lambda\left(T(x,p)'\,\widehat{\beta}_1(y_1)\right),$$
$$\widehat{F}_{Y_0|X,U_D}(y_0|x,p) = \widehat{F}_{Y|X,p(X,Z),D}(y_0|x,p,0) - (1-p)\frac{\partial T(x,p)'}{\partial p}\widehat{\beta}_0(y_0)\cdot\lambda\left(T(x,p)'\,\widehat{\beta}_0(y_0)\right),$$

and $\Delta_\tau^{MQTE}(x,u_D)$ is estimated as

$$\widehat{\Delta}_{X,U_D}(\tau|x,u_D) = \widehat{F}^{-1}_{Y_1|X,U_D}(\tau|x,u_D) - \widehat{F}^{-1}_{Y_0|X,U_D}(\tau|x,u_D),$$

where λ is the derivative of Λ, and F^{-1} is the usual left-inverse of F. Although $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$ is defined for all $u_D \in (0,1)$, it is usually only studied on $\text{supp}(p(X,Z)|X=x,D=1)\cap\text{supp}(p(X,Z)|X=x,D=0)$; see page 3 of the Documentation on Estimation Techniques of Heckman et al. (2006) for practical implementations. For notational convenience, we denote the region of interest for p as $\mathcal{P}$ which is compact and does not depend on x. Of course, when $p(X,Z)$ is discrete or has a narrow support, extrapolation is necessary. Note also that $\widehat{F}_{Y_d|X,U_D}(y_d|x,p)$ need not be a monotone function of y_d. Nevertheless, the monotone rearrangement operator developed in Chernozhukov et al. (2010) can be first applied before inverting $\widehat{F}_{Y_d|X,U_D}(y_d|x,p)$. Rearrangement does not affect the weak limit of $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$ under correct specification of the model.

Given $\widehat{F}_{Y_d|X,U_D}(y_d|x,p)$, we can estimate the QTE by

$$\widehat{\Delta}_X(\tau|x) = \widehat{F}^{-1}_{Y_1|X}(\tau|x) - \widehat{F}^{-1}_{Y_0|X}(\tau|x),$$

where $\widehat{F}_{Y_d|X}(y_d|x) = \widehat{F}_{Y|X,p(X,Z),D}(y_d|x,d,d)$. This estimation may involve extrapolation of $p(X,Z)$ out of its support. We can further estimate the IQTE by

$$\widehat{\Delta}(\tau) = \widehat{F}^{-1}_{Y_1}(\tau) - \widehat{F}^{-1}_{Y_0|X}(\tau),$$

where $\widehat{F}_{Y_d}(y_d) = n^{-1}\sum_{i=1}^n \widehat{F}_{Y|X}(y_d|X_i)$.[11][12] We reemphasize here that estimation of the QTE and IQTE requires stronger versions of the RP assumption; otherwise, they only summarize the quantile differences in the two treatment states and do not have a causal interpretation.

4.3 Asymptotics for the Quantile Treatment Estimators

We in this subsection states the weak limit of $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$ indexed by $(\tau,x,u_D)\in\mathcal{TXP}$, where $\mathcal{X}$ =supp(X) and $\mathcal{T}\subset(0,1)$ is compact. Our asymptotic results extend Newey (1984) by allowing the second-step estimator to be a process.

We first specify similar conditions as Condition DR of Chernozhukov et al. (2013).

Condition DR: (a) $p(x,z) = \Lambda(R(x,z)'\gamma)$, $F_{Y|X,p(X,Z),D}(y|x,p,d) = \Lambda\left(T(x,p)'\beta_d(y)\right)$ for all $y\in\mathcal{Y}_d$, $x\in\mathcal{X}$, $z\in\mathcal{Z}_x$ and $p\in\mathcal{P}$, where Λ is either Probit or Logit link function. (b) The region of interest $\mathcal{Y}_d$ is either a compact interval in $\mathbb{R}$ or a finite subset of $\mathbb{R}$. In the former case, the conditional density $f_{Y|X,Z,D}(y|x,z,d)$ exists, is uniformly bounded and uniformly continuous in (y,x,z) in the support of (Y_d,X,Z). $\mathcal{XP}$ is

[11] Note that the average is taken over all X_i's. This is due to Assumption (A4) which states that supp$(X|D=0)$ =supp$(X|D=1)$.

[12] We can also define treatment effects on the usual inequality measures (based on the three kinds of quantile processes above) such as the Lorenz curve and the Gini coefficient, but such inequality measures are not of main interests in this paper; see Bhattacharya (2007) and Barrett and Donald (2009) for related discussions.

compact. (c) $E\left[\|(R,T)\|^2\right] < \infty$ and the minimum eigenvalue of

$$J_p \equiv E\left[\frac{\widetilde{\lambda}^2}{\widetilde{p}\left[1-\widetilde{p}\right]}RR'\right]$$

and

$$J_d(y_d) \equiv E\left[1(D=d)\frac{\lambda_d(y_d)^2}{\Lambda_d\left(y_d\right)\left[1-\Lambda_d\left(y_d\right)\right]}TT'\right]$$

is bounded away from zero uniformly over $y \in \mathcal{Y}_d$, where $R = R(X,Z)$, $\widetilde{p} = \Lambda(R'\gamma)$, $\widetilde{\lambda} = \lambda(R'\gamma)$, $T = T\left(X, p(X,Z)\right)$, $\Lambda_d\left(y_d\right) = \Lambda\left(T'\beta_d(y_d)\right)$, and $\lambda_d(y_d) = \lambda(T'\beta_d(y_d))$.

To ease the statement of our theorem, define

$$J_{0p}(y_0) = E\left[(1-\widetilde{p})\widetilde{\lambda}\lambda_0(y_0)H_0(y_0)\frac{\partial T(X,\widetilde{p})'\beta_0(y_0)}{\partial p}TR'\right],$$
$$J_{1p}(y_1) = E\left[\widetilde{p}\widetilde{\lambda}\lambda_1(y_1)H_1(y_1)\frac{\partial T(X,\widetilde{p})'\beta_1(y_1)}{\partial p}TR'\right],$$

with $H_d(y_d) = H(T'\beta_d(y_d))$ and $H(\cdot) = \lambda(\cdot)/\left\{\Lambda(\cdot)\left[1-\Lambda(\cdot)\right]\right\}$.

$$W_\gamma = \mathbb{G}\left(\kappa_\gamma\right),\ W_0(y_0) = \mathbb{G}(\kappa_0(y_0)), \text{ and } W_1(y_1) = \mathbb{G}(\kappa_1(y_1)),$$

where

$$\kappa_\gamma = \left(\widetilde{p} - D\right)H(R'\gamma)R,$$
$$\kappa_0(y_0) = (1-D)\left[\Lambda(T'\beta_0(y_0)) - 1(Y \le y_0)\right]H(T'\beta_0(y_0))T,$$
$$\kappa_1(y_1) = D\left[\Lambda(T'\beta_1(y_1)) - 1(Y \le y_1)\right]H(T'\beta_1(y_1))T,$$

W_γ is a zero-mean random variable with variance $E[\kappa_\gamma^2]$, $W_1(y_1)$ is a zero-mean Gaussian process with the covariance function $E\left[\mathbb{G}(\kappa_1(y_1))\mathbb{G}(\kappa_1(y_1'))\right] = E\left[\kappa_1(y_1)\kappa_1(y_1')\right] - E\left[\kappa_1(y_1)\right]E\left[\kappa_1(y_1')\right](= E\left[\kappa_1(y_1)\kappa_1(y_1')\right]$ since $E\left[\kappa_1(y_1)\right] = E\left[\kappa_1(y_1')\right] = 0$), and $W_0(y_0)$ is similarly defined. It is easy to check that W_γ, $W_0(\cdot)$ and $W_1(\cdot)$ are independent. Define $\varphi'_{\beta_0(\cdot)}(\alpha_0)(y_0,x,p) : C\left(\mathcal{Y}_0\right)^{d_{\beta_0}} \longrightarrow \ell^\infty\left(\mathcal{Y}_0\mathcal{X}\mathcal{P}\right)$ as

$$\begin{aligned}\varphi'_{\beta_0(\cdot)}(\alpha_0) = &\left[\lambda\left(T\left(x,p\right)'\beta_0(y_0)\right) - (1-p)\frac{\partial T\left(x,p\right)'}{\partial p}\beta_0(y_0)\cdot\lambda'\left(T\left(x,p\right)'\beta_0(y_0)\right)\right]T(x,p)'\alpha_0(y_0)\\ &- (1-p)\frac{\partial T\left(x,p\right)'}{\partial p}\alpha_0(y_0)\cdot\lambda\left(T\left(x,p\right)'\beta_0(y_0)\right),\end{aligned}$$

and $\varphi'_{\beta_1(\cdot)}(\alpha_1)(y_1,x,p) : C\left(\mathcal{Y}_1\right)^{d_{\beta_1}} \longrightarrow \ell^\infty\left(\mathcal{Y}_1\mathcal{X}\mathcal{P}\right)$ as

$$\begin{aligned}\varphi'_{\beta_1(\cdot)}(\alpha_1) = &\left[\lambda\left(T\left(x,p\right)'\beta_1(y_1)\right) + p\frac{\partial T\left(x,p\right)'}{\partial p}\beta_1(y_1)\cdot\lambda'\left(T\left(x,p\right)'\beta_1(y_1)\right)\right]T(x,p)'\alpha_1(y_1)\\ &+ p\frac{\partial T\left(x,p\right)'}{\partial p}\alpha_1(y_1)\cdot\lambda\left(T\left(x,p\right)'\beta_1(y_1)\right).\end{aligned}$$

Theorem 4 *Suppose Condition DR holds, and $F_{Y_d|X,U_D}(y_d|x,u_D)$ admits a positive continuous density $f_{Y_d|X,U_D}(y_d|x,u_D)$ on an interval $[a,b]$ containing an ϵ-enlargement of the set $\left\{Q_{Y_d|X,U_D}(\tau|x,u_D)|\tau \in \mathcal{T}\right\}$*

for all $(x, u_D) \in \mathcal{X}\mathcal{P}$*. Then*

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\Delta}_{X,U_D}(\tau|x,u_D) - \Delta_{\tau}^{MQTE}(x,u_D)\right) \\
&\rightsquigarrow \left.\frac{\varphi'_{\beta_1(\cdot)}\left(J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right)}{f_{Y_1|X,U_D}(y_1|x,u_D)}\right|_{y_1=Q_{Y_1|X,U_D}(\tau|x,u_D)} \\
&- \left.\frac{\varphi'_{\beta_0(\cdot)}\left(J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right)}{f_{Y_0|X,U_D}(y_0|x,u_D)}\right|_{y_0=Q_{Y_0|X,U_D}(\tau|x,u_D)} \quad \text{in } \ell^\infty\left(\mathcal{T}\mathcal{X}\mathcal{P}\right),
\end{aligned}$$

where $\rightsquigarrow$ *means the weak convergence over a metric space.*

Note that $J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)$ and $J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)$ are the weak limits of $\sqrt{n}\left(\widehat{\beta}_0(y_0) - \beta_0(y_0)\right)$ and $\sqrt{n}\left(\widehat{\beta}_1(y_1) - \beta_1(y_1)\right)$, respectively. They are dependent through W_γ which is inherited from the generated regressor $\widehat{p}$. When Y_d is discrete, we'd better state the weak limit indexed by $\mathcal{Y}$ rather than $\mathcal{T}$, where we assume that Y_0 and Y_1 have the same support $\mathcal{Y}$; see the discussion after Theorem 1. From the proof of Theorem 4, the weak limit of $\widehat{F}_{Y_1|X,U_D}(y|x,u_D) - \widehat{F}_{Y_0|X,U_D}(y|x,u_D)$ in $\ell^\infty\left(\mathcal{Y}\mathcal{X}\mathcal{P}\right)$ is

$$-\varphi'_{\beta_1(\cdot)}\left(J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right) + \varphi'_{\beta_0(\cdot)}\left(J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right).$$

This theorem has a trivial corollary. If we fix a point in two of the three index sets, $\mathcal{T}$, $\mathcal{X}$ and $\mathcal{P}$, then the weak limit of the corresponding quantile treatment processes is the same as that stated in the theorem but indexed by only one set. This is often helpful to intuitively illustrate the quantile treatment processes. Suppose $X = (X^{c\prime}, X^{d\prime})'$, where X^c is the continuous component, and X^d is the discrete component with K possible values (say, x_k^d, $k = 1, \cdots, K$). Then we usually fix x at $\left(\overline{X}^c, x_k^d\right)$ to check the quantile treatment effect for an average person in X^c, where $\overline{X}^c$ is the sample mean of X^c.

The following corollary states the weak limit of $\widehat{\Delta}_X(\tau|x)$.

Corollary 1 *Suppose Condition DR holds, and* $F_{Y_d|X}(y_d|x)$ *admits a positive continuous density* $f_{Y_d|X}(y_d|x)$ *on an interval* $[a,b]$ *containing an* ϵ*-enlargement of the set* $\left\{Q_{Y_d|X}(\tau|x)|\tau \in \mathcal{T}\right\}$ *for all* $x \in \mathcal{X}$*. Then*

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\Delta}_X(\tau|x) - \Delta_\tau^{QTE}(x)\right) \\
&\rightsquigarrow \left.\frac{\overline{\varphi}'_{\beta_1(\cdot)}\left(J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right)}{f_{Y_1|X}(y_1|x)}\right|_{y_1=Q_{Y_1|X}(\tau|x)} \\
&- \left.\frac{\overline{\varphi}'_{\beta_0(\cdot)}\left(J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right)}{f_{Y_0|X}(y_0|x)}\right|_{y_0=Q_{Y_0|X}(\tau|x)} \quad \text{in } \ell^\infty\left(\mathcal{T}\mathcal{X}\right),
\end{aligned}$$

where $\overline{\varphi}'_{\beta_0(\cdot)}(\alpha_0)(y_0,x) : C\left(\mathcal{Y}_0\right)^{d_{\beta_0}} \longrightarrow \ell^\infty\left(\mathcal{Y}_0\mathcal{X}\right)$ *is defined as*

$$\overline{\varphi}'_{\beta_0(\cdot)}(\alpha_0) = \lambda\left(T\left(x,0\right)'\beta_0(y_0)\right)T(x,0)'\alpha_0(y_0),$$

and $\overline{\varphi}'_{\beta_1(\cdot)}(\alpha_1)(y_1,x) : C\left(\mathcal{Y}_1\right)^{d_{\beta_1}} \longrightarrow \ell^\infty\left(\mathcal{Y}_1\mathcal{X}\right)$ *is defined as*

$$\overline{\varphi}'_{\beta_1(\cdot)}(\alpha_1) = \lambda\left(T\left(x,1\right)'\beta_1(y_1)\right)T(x,1)'\alpha_1(y_1).$$

By Theorem 4.1 and Theorem 5.2 of Chernozhukov et al. (2013), we also have the following corollary.

Corollary 2 *Suppose Condition DR holds, and $F_{Y_d}(y_d)$ admits a positive continuous density $f_{Y_d}(y_d)$ on an interval $[a,b]$ containing an ϵ-enlargement of the set $\{Q_{Y_d}(\tau)|\tau \in \mathcal{T}\}$. Then*

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\Delta}(\tau) - \Delta_\tau^{IQTE}\right) \\
&\rightsquigarrow \left.\frac{\int \overline{\varphi}'_{\beta_1(\cdot)}\left(J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right)dF_X(x) + \mathbb{G}\left(F_{Y_1|X}(y_1|X)\right)}{f_{Y_1}(y_1)}\right|_{y_1=Q_{Y_1}(\tau)} \\
&- \left.\frac{\int \overline{\varphi}'_{\beta_0(\cdot)}\left(J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right)dF_X(x) + \mathbb{G}(F_{Y_0|X}(y_0|X))}{f_{Y_0}(y_0)}\right|_{y_0=Q_{Y_0}(\tau)} \quad \text{in } \ell^\infty(\mathcal{T}).
\end{aligned}$$

Note that now there are two sources of correlation between the two components of the weak limit of $\sqrt{n}\left(\widehat{\Delta}(\tau) - \Delta_\tau^{IQTE}\right)$: one from W_γ and one from X in $F_{Y_1|X}(y_1|X)$ and $F_{Y_0|X}(y_0|X)$.

4.4 Inferences for Quantile Treatment Processes

The asymptotic theory in the last subsection is not practically useful given that the limit processes are non-pivotal and their covariance functions depend on complicated unknown, though estimable, nuisance parameters. In other words, the Durbin problem (see Durbin (1973)) appears in this context. A popular alternative of the asymptotic methods is the resampling methods, especially, the *exchangeable bootstrap*. This procedure incorporates many popular forms of resampling as special cases, namely the empirical bootstrap, weighted bootstrap, m out of n bootstrap, and subsampling, see Section 3.6.2 of van der Vaart and Wellner (1996) for concrete descriptions. Each bootstrap scheme is useful to a specific application. For example, in small samples, we might want to use the weighted bootstrap to gain good accuracy and robustness to "small cells", whereas in large samples, where computational tractability can be an important consideration, we might prefer subsampling.

Let $(\omega_1, \cdots, \omega_n)$ be a vector of nonnegative random variables that satisfy Condition EB in Chernozhukov et al. (2013) or the conditions (3.6.8) of van der Vaart and Wellner (1996). For example, $(\omega_1, \cdots, \omega_n)$ is a multinomial vector with dimension n and probabilities $(1/n, \cdots, 1/n)$ in the empirical bootstrap. The exchangeable bootstrap uses the components of $(\omega_1, \cdots, \omega_n)$ as random sampling weights in the construction of the bootstrap version of the estimators. Its validity is a trivial application of the Functional Delta method for Bootstrap given the Hadamard-differentiability of various operators in the last subsection. So in what follows we only describe the bootstrap procedures for our estimators and no asymptotic validity results are stated. We will only report the procedure for $\Delta_\tau^{MQTE}(x, u_D)$ since inferences for $\Delta_\tau^{QTE}(x)$ and Δ_τ^{IQTE} are similar.

The bootstrap estimator of $\Delta_\tau^{MQTE}(x, u_D)$ is

$$\widehat{\Delta}^*_{X,U_D}(\tau|x, u_D) = \widehat{F}^{*-1}_{Y_1|X,U_D}(\tau|x, u_D) - \widehat{F}^{*-1}_{Y_0|X,U_D}(\tau|x, u_D),$$

where

$$\begin{aligned}
\widehat{F}^*_{Y_1|X,U_D}(y_1|x,p) &= \widehat{F}^*_{Y|X,p(X,Z),D}(y_1|x,p,1) + p\frac{\partial T(x,p)'}{\partial p}\widehat{\beta}^*_1(y_1)\cdot\lambda\left(T(x,p)'\widehat{\beta}^*_1(y_1)\right), \\
\widehat{F}^*_{Y_0|X,U_D}(y_0|x,p) &= \widehat{F}^*_{Y|X,p(X,Z),D}(y_0|x,p,0) - (1-p)\frac{\partial T(x,p)'}{\partial p}\widehat{\beta}^*_0(y_0)\cdot\lambda\left(T(x,p)'\widehat{\beta}^*_0(y_0)\right)
\end{aligned}$$

with

$$\widehat{F}^*_{Y|X,p(X,Z),D}(y|x,p,d) = \Lambda\left(T(x,p)'\widehat{\beta}^*_d(y)\right),$$

and

$$\widehat{\beta}_d^*(y) = \arg\max_{\beta} \sum_{i=1}^{n} \omega_i 1(D_i = d) \left[1\left(Y_i \leq y\right) \ln \Lambda \left(T\left(X_i, \widehat{p}_i^*\right)' \beta\right) + 1\left(Y_i > y\right) \ln \left(1 - \Lambda\left(T\left(X_i, \widehat{p}_i^*\right)' \beta\right)\right)\right],$$

$$\widehat{p}_i^* = p(X_i, Z_i; \widehat{\gamma}^*),\ \widehat{\gamma}^* = \arg\max_{\gamma} \sum_{i=1}^{n} \omega_i \left[D_i \ln \Lambda\left(R(X_i, Z_i)'\gamma\right) + (1 - D_i) \ln\left(1 - \Lambda\left(R(X_i, Z_i)'\gamma\right)\right)\right].$$

Given $\widehat{\Delta}^*_{X,U_D}(\tau|x, u_D)$, we can conduct uniform inferences for $\Delta_\tau^{MQTE}(x, u_D)$ which cover degenerated cases, e.g., $\Delta_\tau^{MQTE}(x, u_D)$ for a fixed x and τ or a fixed x and u_D, as special cases. An asymptotic simultaneous $(1-\alpha)$ confidence band for $\Delta_\tau^{MQTE}(x, u_D)$ over $\tau \in \mathcal{T}$, $x \in \mathcal{X}$ and $u_D \in \mathcal{P}$ is defined by the end-point functions

$$\widehat{\Delta}^{\pm}_{X,U_D}(\tau|x, u_D) = \widehat{\Delta}_{X,U_D}(\tau|x, u_D) \pm \widehat{t}_{1-\alpha} \widehat{\Sigma}\left(\tau, x, u_D\right)^{1/2} / \sqrt{n},$$

such that

$$\lim_{n\to\infty} P\left(\Delta_\tau^{MQTE}(x, u_D) \in \left[\widehat{\Delta}^{-}_{X,U_D}(\tau|x, u_D), \widehat{\Delta}^{+}_{X,U_D}(\tau|x, u_D)\right] \text{ for all } (\tau, x, u_D) \in \mathcal{TXP}\right) = 1 - \alpha. \quad (19)$$

Here, $\widehat{\Sigma}\left(\tau, x, u_D\right)$ is a uniformly consistent estimator of $\Sigma(\tau, x, u_D)$, the asymptotic variance function of $\sqrt{n}\left(\widehat{\Delta}_{X,U_D}(\tau|x, u_D) - \Delta_\tau^{MQTE}(x, u_D)\right)$. In order to achieve the coverage property (19), we set the critical value $\widehat{t}_{1-\alpha}$ as a consistent estimator of the $(1-\alpha)$th quantile of the maximal t-statistic:

$$t = \sup_{(\tau, x, u_D) \in \mathcal{TXP}} \sqrt{n} \widehat{\Sigma}\left(\tau, x, u_D\right)^{-1/2} \left|\widehat{\Delta}_{X,U_D}(\tau|x, u_D) - \Delta_\tau^{MQTE}(x, u_D)\right|.$$

It remains to obtain $\widehat{\Sigma}\left(\tau, x, u_D\right)$ and $\widehat{t}_{1-\alpha}$. For this purpose, we first get $\widehat{Z}_b^*(\tau, x, u_D)$, $b = 1, \cdots, B$, as iid realization of $\widehat{Z}^*(\tau, x, u_D) = \sqrt{n}\left(\widehat{\Delta}^*_{X,U_D}(\tau|x, u_D) - \widehat{\Delta}_{X,U_D}(\tau|x, u_D)\right)$ for $(\tau, x, u_D) \in \mathcal{TXP}$. Then compute a bootstrap estimate of $\Sigma\left(\tau, x, u_D\right)^{1/2}$ such as the bootstrap interquartile range[13] rescaled with the normal distribution: $\widehat{\Sigma}\left(\tau, x, u_D\right)^{1/2} = \left(q_{0.75}(\tau, x, u_D) - q_{0.25}(\tau, x, u_D)\right)/1.349$ for $(\tau, x, u_D) \in \mathcal{TXP}$, where $q_\alpha(\tau, x, u_D)$ is the αth quantile of $\left\{\widehat{Z}_b^*(\tau, x, u_D), b = 1, \cdots, B\right\}$. Finally, $\widehat{t}_{1-\alpha}$ is set as the $(1-\alpha)$th sample quantile of $\{t_b^*, b = 1, \cdots, B\}$, where $t_b^* = \sup_{(\tau, x, u_D) \in \mathcal{TXP}} \widehat{\Sigma}\left(\tau, x, u_D\right)^{-1/2} \left|\widehat{Z}_b^*(\tau, x, u_D)\right|$. By modifying the procedure above, we can test whether $\Delta_\tau^{MQTE}(x, u_D)$ is constant in τ or in x or in u_D.

Except constructing uniform bands for $\Delta_\tau^{MQTE}(x, u_D)$, the inference procedure above can also be used to test unconfoundedness and stochastic dominance.[14] Under unconfoundedness, $F_{Y|X,p(X,Z),D}(y|x, p, d)$ does not depend on p for any y, x and d,[15] so we can check whether the components of $\beta_d(y)$ associated with all components of $T\left(x, p\right)$ involving p are zero to test unconfoundedness. The test statistic can be the Kolmogorov-Smirnov (KS) statistic

$$K_n = \sqrt{n} \sup_{y_0 \in \mathcal{Y}_0} \left\|\widehat{\beta}_0^p(y_0)\right\| + \sqrt{n} \sup_{y_1 \in \mathcal{Y}_1} \left\|\widehat{\beta}_1^p(y_1)\right\|$$

[13] Here, the interquartile range rather than the standard deviation is used to avoid technical complexities, see Remark 3.2 of Chernozhukov et al. (2013).

[14] See Example 3 and 4 of Section 4.1 in Chernozhukov and Hansen (2006) for alternative tests in their framework.

[15] This is because $F_{Y|X,p(X,Z),D}\left(y|x, p, 1\right) p = \int_0^p F_{Y_1|X,U_D}(y|x, u_D) du_D = p F_{Y_1|X}(y|x)$ under unconfoundedness, which implies $F_{Y|X,p(X,Z),D}\left(y|x, p, 1\right) = F_{Y_1|X}(y|x)$ does not depend on p. This result can be similarly extended to the case with $D = 0$.

or the Cramer-von Mises (CM) statistic

$$C_n = n\int_{\mathcal{Y}_0} \left\|\widehat{\beta}_0^p(y_0)\right\|^2 dy_0 + n\int_{\mathcal{Y}_1} \left\|\widehat{\beta}_1^p(y_1)\right\|^2 dy_1$$

where $\|\cdot\|$ is the Euclidean norm, and $\widehat{\beta}_d^p(y)$ is the component of $\widehat{\beta}_d(y)$ associated with p.[16] Since the bootstrap is valid, we can check whether $K_n > \widehat{c}_{1-\alpha}$ to determine whether unconfoundedness holds, where $\widehat{c}_{1-\alpha}$ is the $(1-\alpha)$-th sample quantile of $\{K_b^*, b = 1, \cdots, B\}$ with

$$K_b^* = \sup_{y_0\in\mathcal{Y}_0} \sqrt{n}\left\|\widehat{\beta}_0^{p*}(y_0) - \widehat{\beta}_0^p(y_0)\right\| + \sup_{y_1\in\mathcal{Y}_1} \sqrt{n}\left\|\widehat{\beta}_1^{p*}(y_1) - \widehat{\beta}_1^p(y_1)\right\|.$$

A similar procedure can also be applied to C_n.

An alternative way to test unconfoundedness is based on $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$. Under unconfoundedness, $\Delta_\tau^{MQTE}(x,u_D)$ does not depend on u_D (in model (15), this is equivalent to $\sigma_{V1} = \sigma_{V0}$), so we can test unconfoundedness by checking whether $\widehat{\Delta}_{X,U_D}(\tau|x,u_D) = \widehat{\Delta}_{X,U_D}(\tau|x,\overline{u}_D)$ for a specified $\overline{u}_D \in \mathcal{P}$. For this purpose, we just change the test statistic to

$$\sqrt{n}\sup_{(\tau,x,u_D)\in\mathcal{T}\mathcal{X}\mathcal{P}} \left|\widehat{\Delta}_{X,U_D}(\tau|x,u_D) - \widehat{\Delta}_{X,U_D}(\tau|x,\overline{u}_D)\right|$$

or

$$n\int_{\mathcal{T}\mathcal{X}\mathcal{P}} \left|\widehat{\Delta}_{X,U_D}(\tau|x,u_D) - \widehat{\Delta}_{X,U_D}(\tau|x,\overline{u}_D)\right|^2 d\tau dx du_D,$$

and bootstrap the critical values. Similar ideas can also be used to test whether D affects only the location of outcome Y conditional on X and U_D, i.e., $\Delta_\tau^{MQTE}(x,u_D)$ does not depends on τ (in model (15), this is equivalent to $\sigma_1^2 - \sigma_{V1}^2 = \sigma_0^2 - \sigma_{V0}^2$), based on, e.g., $\widehat{\Delta}_{X,U_D}(\tau|x,u_D) - \widehat{\Delta}_{X,U_D}(0.5|x,u_D)$. Furthermore, we can also test the setup of Carneiro and Lee (2009): $Y_d = \mu_d(X) + U_d$, $D = 1(\mu_D(X,Z) - V \geq 0)$ and $(U_1, U_0, V) \perp (X, Z)$. In their setup, $\widehat{F}_{Y_d|X,U_D}^{-1}(\tau|x,u_D)$ is parallel as a function of x for each τ and each u_D, so $\widehat{F}_{Y_d|X,U_D}^{-1}(\tau_1|x,u_{D1}) - \widehat{F}_{Y_d|X,U_D}^{-1}(\tau_2|x,u_{D2})$ does not depend on x for any $\tau_1,\tau_2 \in \mathcal{T}$ and $u_{D1}, u_{D2} \in \mathcal{P}$. As a result, the test can be based on

$$\sqrt{n}\sup_{(\tau_1,\tau_2,u_{D1},u_{D2},x)\in\mathcal{T}\mathcal{T}\mathcal{P}\mathcal{P}\mathcal{X}} \left|\left[\widehat{F}_{Y_d|X,U_D}^{-1}(\tau_1|x,u_{D1}) - \widehat{F}_{Y_d|X,U_D}^{-1}(\tau_2|x,u_{D2})\right] - \left[\widehat{F}_{Y_d|X,U_D}^{-1}(\tau_1|\overline{x},u_{D1}) - \widehat{F}_{Y_d|X,U_D}^{-1}(\tau_2|\overline{x},u_{D2})\right]\right|$$

or

$$n\int_{\mathcal{T}\mathcal{T}\mathcal{P}\mathcal{P}\mathcal{X}} \left|\left[\widehat{F}_{Y_d|X,U_D}^{-1}(\tau_1|x,u_{D1}) - \widehat{F}_{Y_d|X,U_D}^{-1}(\tau_2|x,u_{D2})\right] - \left[\widehat{F}_{Y_d|X,U_D}^{-1}(\tau_1|\overline{x},u_{D1}) - \widehat{F}_{Y_d|X,U_D}^{-1}(\tau_2|\overline{x},u_{D2})\right]\right|^2 d\tau_1 d\tau_2 du_{D1} du_{D2} dx$$

for a specified $\overline{x} \in \mathcal{X}$.

As to the test of (first-order) stochastic dominance, the null is $\Delta_\tau^{MQTE}(x,u_D) \geq 0$ for all $(\tau,x,u_D) \in \mathcal{T}\mathcal{X}\mathcal{P}$. In this case, the *least favorable null* is $\Delta_\tau^{MQTE}(x,u_D) = 0$ for all $(\tau,x,u_D) \in \mathcal{T}\mathcal{X}\mathcal{P}$, and one may use the one-sided KS or CM statistics, i.e.,

$$S_n = \sqrt{n}\sup_{(\tau,x,u_D)\in\mathcal{T}\mathcal{X}\mathcal{P}} \max\left(-\widehat{\Delta}_{X,U_D}(\tau|x,u_D), 0\right)$$

[16] The test statistic can be extended to base on $\left\|\widehat{\beta}_d^p(y_d)\right\|_{\widehat{\Lambda}_d^p(y_d)} \equiv \sqrt{\widehat{\beta}_d^p(y_d)'\widehat{\Lambda}_d^p(y_d)\widehat{\beta}_d^p(y_d)}$, where $\widehat{\Lambda}_d^p(y_d)$ is some weight matrix with the probability limit positive definite uniformly on $\mathcal{Y}_d$, e.g., the inverse of an estimator of the asymptotic variance matrix of $\widehat{\beta}_d^p(y_d)$. However, as mentioned in footnote 13, we need to take caution in estimating this weight matrix; see Kato (2011) for more discussions.

or

$$M_n = n \int_{\mathcal{T}\mathcal{X}\mathcal{P}} \left| \max\left(-\widehat{\Delta}_{X,U_D}(\tau|x,u_D), 0 \right) \right|^2 d\tau dx du_D,$$

to test the hypothesis. The bootstrapped critical values can be obtained from $\widehat{Z}^*(\tau, x, u_D)$ in an obvious way. If we only want to test stochastic dominance for x and/or u_D fixed, the procedure can be adjusted correspondingly. Note that the null here is composite, while the bootstrapped critical values are based on the least favorable null, so the test procedure may tend to be conservative (i.e., the type-I error is lower than the nominal level). This testing idea is used in Chernozhukov and Hansen (2006). Other approaches are discussed, for example, by McFadden (1989), Anderson (1996), Davidson and Duclos (2000), Abadie (2002), Barrett and Donald (2003), and Linton et al. (2005).

Similar ideas as in testing stochastic dominance can be applied to test the validity of the our setup (1), (2) and assumptions (A1)-(A5). In our setup, $P(Y \le y|X = x, p(X,Z) = p, D = 1)\, p$ and $-P(Y \le y|X = x, p(X,Z) = p, D = 0)$ as a function of p must have a positive slope of smaller than 1, so the null is

$$\frac{d\left[P\left(Y \le y|p(Z) = p, D = 1\right) p\right]}{dp} \in [0,1] \text{ and } -\frac{d\left[P\left(Y \le y|p(Z) = p, D = 0\right)(1-p)\right]}{dp} \in [0,1].$$[17]

In practice, violation of the range $[0,1]$ can be due to sampling variation or to violations of the assumptions, especially, the exclusion assumption, the ignorability assumption (A2), and the monotonicity assumption.[18] Using the test statistic

$$\begin{aligned}&\sqrt{n} \sup_{(y_1,x,u_D)\in\mathcal{Y}_1\mathcal{X}\mathcal{P}} \max\left(-\widehat{F}_{Y_1|X,U_D}(y_1|x,p), 0\right) + \sqrt{n} \sup_{(y_1,x,u_D)\in\mathcal{Y}_1\mathcal{X}\mathcal{P}} \max\left(\widehat{F}_{Y_1|X,U_D}(y_1|x,p) - 1, 0\right)\\ &+ \sqrt{n} \sup_{(y_0,x,u_D)\in\mathcal{Y}_0\mathcal{X}\mathcal{P}} \max\left(-\widehat{F}_{Y_0|X,U_D}(y_0|x,p), 0\right) + \sqrt{n} \sup_{(y_0,x,u_D)\in\mathcal{Y}_0\mathcal{X}\mathcal{P}} \max\left(\widehat{F}_{Y_0|X,U_D}(y_0|x,p) - 1, 0\right)\end{aligned}$$

or

$$\begin{aligned}&n \int_{\mathcal{Y}_1\mathcal{X}\mathcal{P}} \left|\max\left(-\widehat{F}_{Y_1|X,U_D}(y_1|x,p), 0\right)\right|^2 dy_1 dx du_D + n \int_{\mathcal{Y}_1\mathcal{X}\mathcal{P}} \left|\max\left(\widehat{F}_{Y_1|X,U_D}(y_1|x,p) - 1, 0\right)\right|^2 dy_1 dx du_D\\ &+ n \int_{\mathcal{Y}_0\mathcal{X}\mathcal{P}} \left|\max\left(-\widehat{F}_{Y_0|X,U_D}(y_0|x,p), 0\right)\right|^2 dy_0 dx du_D + n \int_{\mathcal{Y}_0\mathcal{X}\mathcal{P}} \left|\max\left(\widehat{F}_{Y_0|X,U_D}(y_0|x,p) - 1, 0\right)\right|^2 dy_0 dx du_D,\end{aligned}$$

we obtain alternative tests of those in Heckman and Vytlacil (2005).

In practice, the supremum and integration on $\mathcal{Y}_d$, $\mathcal{T}$, $\mathcal{X}$ and $\mathcal{P}$ can be replaced by their corresponding discretized versions by stochastic equicontinuity of the involved processes as long as the distance between adjacent grid points goes to zero as $n \to \infty$.

4.5 Goodness of Fit Tests

In this subsection, we suggest some goodness of fit tests to choose the terms in $T(x,p)$ and $R(x,z)$. The null hypothesis in our case is

$$\begin{gathered}H_0 : F_{Y|X,p(X,Z),D}(y_d|x,p,d) = \Lambda\left(T(x,p)'\beta_d(y_d)\right) \text{ with } p(x,z) = \Lambda\left(R(x,z)'\gamma\right)\\ \text{for some } \beta_d(y_d) \in \mathcal{B}_d(\mathcal{Y}_d),\ \gamma \in \Gamma \text{ and all } (y,x,z,p) \in \mathcal{Y}_d\mathcal{X}\mathcal{Z}\mathcal{P},\end{gathered}$$

[17] See also Section 5 of Imbens and Rubin (1997) for the testable restrictions that the complier's outcome densities when $D = 0$ and $D = 1$ are nonnegative; their restrictions can only be applied for discrete Z.

[18] It may also be due to misspecification of the distribution regression (16); see the next subsection on this issue.

where $\mathcal{B}_d(\mathcal{Y}_d)$ is the class of bounded mappings on $\mathcal{Y}_d$, and Γ is a compact parameter space. The alternative H_1 is the negation of H_0, i.e., $F_{Y|X,p(X,Z),D}(y_d|x,p,d) \neq \Lambda\left(T(x,p)'\beta_d(y_d)\right)$ with $p(x,z) = \Lambda\left(R(x,z)'\gamma\right)$ for all $\beta_d(y_d) \in \mathcal{B}_d(\mathcal{Y}_d)$, $\gamma \in \Gamma$ and some $(y,x,z,p) \in \mathcal{Y}_d\mathcal{X}\mathcal{Z}\mathcal{P}$, where $\mathcal{Z}$ and $\mathcal{P}$ are understood as the support of Z and $p(X,Z)$, respectively. Under H_1 and Condition DR (except (a)), we can still estimate γ and $\beta_d(y)$, but the estimators should be understood as estimating pseudo-true values. Violation of H_0 may be due to either $p(x,z) \neq \Lambda\left(R(x,z)'\gamma\right)$ or $F_{Y|X,p(X,Z),D}(y|x,p,d) \neq \Lambda\left(T(x,p)'\beta_d(y)\right)$ even if $p(x,z) = \Lambda\left(R(x,z)'\gamma\right)$.

As in Heckman (1984), our test is based on the marginal distribution of (Y,D) rather than the conditional distribution $F_{Y|X,p(X,Z),D}$ or the joint distribution $F_{Y,X,p(X,Z),D}$. Our test statistics are constructed by comparing $H^d(y_d) = P(Y \leq y_d, D = d)$ with $F^d(y_d;\beta_d,\gamma) = E\left[1(D=d)\Lambda(T(X,\Lambda\left(R(X,Z)'\gamma\right))'\beta_d)\right]$. Note that under H_0, $H^d(y_d) = F^d(y_d;\beta_d,\gamma)$ for some $\beta_d(y_d) \in \mathcal{B}_d(\mathcal{Y}_d)$, $\gamma \in \Gamma$ and all $(y_d,x,z) \in \mathcal{Y}_d\mathcal{X}\mathcal{Z}$. This is because

$$
\begin{aligned}
H^d(y_d) &= P(Y \leq y_d, D = d) = E\left[P(Y \leq y_d, D = d|X, p(X,Z))\right] \\
&= E\left[P(Y \leq y_d|X, p(X,Z), D = d)P(D = d|X, p(X,Z))\right] \\
&= E\left[F_{Y|X,p(X,Z),D}(y_d|X, p(X,Z), 1)1(D = d)\right] \\
&= F^d(y_d;\beta_d,\gamma) \text{ under } H_0,
\end{aligned}
$$

where the second to last equality is from the law of iterated expectation. We consider $P(Y \leq y_d, D = d)$ rather than $P(Y \leq y_d|D = d)$ to avoid denominators in the CDF estimation. Our test statistics are

$$
T_n^K = \sqrt{n}\sup_{y_1\in\mathcal{Y}_1}\left|\widehat{H}_n^1(y_1) - \widehat{F}_n^1(y_1)\right| + \sqrt{n}\sup_{y_0\in\mathcal{Y}_0}\left|\widehat{H}_n^0(y_0) - \widehat{F}_n^0(y_0)\right|, \tag{20}
$$

or

$$
T_n^C = n\int_{\mathcal{Y}_1}\left(\widehat{H}_n^1(y_1) - \widehat{F}_n^1(y_1)\right)^2 dw_1(y_1) + n\int_{\mathcal{Y}_0}\left(\widehat{H}_n^0(y_0) - \widehat{F}_n^0(y_0)\right)^2 dw_0(y_0), \tag{21}
$$

where

$$
\widehat{H}_n^d(y_d) = \frac{1}{n}\sum_{i=1}^n 1(Y_i \leq y_d)1(D_i = d) \text{ and } \widehat{F}_n^d(y_d) = \frac{1}{n}\sum_{i=1}^n \widehat{F}_{Y|X,p(X,Z),D}(y_d|X_i,\widehat{p}_i,d)1(D_i = d)
$$

with $\widehat{F}_{Y|X,p(X,Z),D}$ and $\widehat{p}_i$ constructed as in Section 4.2, and $w_d(\cdot)$ is a known weighting function on $\mathcal{Y}_d$. Our test extends that of Heckman (1984) in at least three aspects. First, the building elements of T_n^K and T_n^C, $\widehat{H}_n^d$ and $\widehat{F}_n^d$, are indexed by a continuum rather than a few sets, so our tests can detect more possible deviations from H_0. Second, the parameters under the null, $\beta_d(\cdot)$, are function-valued rather than finite-dimensional. Third, the CDFs under the null are estimated by a two-step procedure rather than a one-step procedure.

The following theorem states the consistency of T_n^K and T_n^C and their asymptotic properties under the null and local alternative. To ease our exposition of the theorem, define

$$
W_H^0(y_0) = \mathbb{G}(\kappa_H^0(y_0)), W_H^1(y_1) = \mathbb{G}(\kappa_H^1(y_1)), W_F^0(y_0) = \mathbb{G}(\kappa_F^0(y_0)) \text{ and } W_F^1(y_1) = \mathbb{G}(\kappa_F^1(y_1)),
$$

with

$$
\begin{aligned}
&\kappa_H^0(y_0) = (1-D)\cdot 1(Y \leq y_0) - H^0(y_0), \kappa_H^1(y_1) = D\cdot 1(Y \leq y_1) - H^1(y_1), \\
&\kappa_F^0(y_0) = (1-D)\cdot\Lambda(T'\beta_0) - F^0(y_0), \kappa_F^1(y_1) = D\cdot\Lambda(T'\beta_1) - F^1(y_1),
\end{aligned}
$$

and

$$\begin{aligned}
&\mathbb{G}_H^0(y_0) = W_H^0(y_0), \mathbb{G}_H^1(y_1) = W_H^1(y_1),\\
&\mathbb{G}_F^0(y_0) = A_\gamma^0 J_p^{-1} W_\gamma - A_0^0 J_0(y_0)^{-1}\left(J_{0p}(y_0) J_p^{-1} W_\gamma + W_0(y_0)\right) + W_F^0(y_0),\\
&\mathbb{G}_F^1(y_1) = A_\gamma^1 J_p^{-1} W_\gamma - A_1^1 J_1(y_1)^{-1}\left(J_{1p}(y_1) J_p^{-1} W_\gamma + W_1(y_1)\right) + W_F^1(y_1),
\end{aligned}$$

with

$$\begin{aligned}
A_\gamma^0 &= E\left[(1-D)\frac{\partial T(X,\widetilde{p})'\beta_0}{\partial p}\lambda_0\widetilde{\lambda}R'\right], A_0^0 = E\left[(1-D)\lambda_0 T'\right],\\
A_\gamma^1 &= E\left[D\frac{\partial T(X,\widetilde{p})'\beta_1}{\partial p}\lambda_1\widetilde{\lambda}R'\right], A_1^1 = E\left[D\lambda_1 T'\right].
\end{aligned}$$

Our local alternatives are

$$\begin{aligned}
q_n(x,z) &= (1-\delta_\gamma/\sqrt{n})p_*(x,z) + (\delta_\gamma/\sqrt{n})\, q(x,z),\\
Q_n^d(y_d|x,p) &= (1-\delta_d/\sqrt{n})F_*^d(y_d|x,p) + (\delta_d/\sqrt{n})\, Q^d(y_d|x,p),
\end{aligned} \tag{22}$$

where $p_*(x,z) = \Lambda(R(x,z)'\gamma^*)$ for some $\gamma^* \in \Gamma$ and all $(x,z) \in \mathcal{XZ}$, $q(x,z) \neq \Lambda(R(x,z)'\gamma)$ for all $\gamma \in \Gamma$ and some $(x,z) \in \mathcal{XZ}$, F_*^d is the CDF such that $F_*^d(y_d|x,p) = \Lambda(T(x,p)'\beta_d^*)$ for some $\beta_d^* \in \mathcal{B}_d(\mathcal{Y}_d)$ and all $(y_d,x,p) \in \mathcal{Y}_d\mathcal{XP}$, Q^d is a CDF such that $Q^d(y_d|x,p) \neq \Lambda(T(x,p)'\beta_d)$ for all $\beta_d \in \mathcal{B}_d(\mathcal{Y}_d)$ and some $(y_d,x,p) \in \mathcal{Y}_d\mathcal{XP}$.

Theorem 5 *Under Condition DR (b) and (c), the following statements hold:*

(i) *Under H_0,*

$$T_n^K \xrightarrow{d} \sup_{y_1\in\mathcal{Y}_1}\left|\mathbb{G}_H^1(y_1) - \mathbb{G}_F^1(y_1)\right| + \sup_{y_0\in\mathcal{Y}_0}\left|\mathbb{G}_H^0(y_0) - \mathbb{G}_F^0(y_0)\right|,$$

and

$$T_n^C \xrightarrow{d} \int_{\mathcal{Y}_1}\left(\mathbb{G}_H^1(y_1) - \mathbb{G}_F^1(y_1)\right)^2 dw_1(y_1) + \int_{\mathcal{Y}_0}\left(\mathbb{G}_H^1(y_0) - \mathbb{G}_F^1(y_0)\right)^2 dw_0(y_0).$$

(ii) *Under any fixed alternative such that $P(Y \in S_d(\beta_d,\gamma), D = d) > 0$ for all $\beta_d \in \mathcal{B}_d(\mathcal{Y}_d)$, $\gamma \in \Gamma$,*

$$\lim_{n\to\infty} P\left(T_n^K > c_n\right) = 1 \text{ and } \lim_{n\to\infty} P\left(T_n^C > c_n\right) = 1,$$

for any sequences of random variables $\{c_n : n \geq 1\}$ with $c_n = O_p(1)$, where $S_d(\beta_d,\gamma) = \{y_d \in \mathcal{Y}_d : H^d(y_d) \neq F^d(y_d;\beta_d,\gamma)\}$.

(iii) *If $Q_n^d(y_d) = E\left[Q_n^d(y_d|X, q_n(X,Z))1(D=d)\right]$ is contiguous to the distribution function $Q_*^d(y_d) = E[F_*^d(y_d|X, p_*(X,Z))1(D=d)]$,*[19] *then*

$$T_n^K \xrightarrow{d} \sup_{y_1\in\mathcal{Y}_1}\left|\mathbb{G}_H^1(y_1) - \mathbb{G}_F^1(y_1) + \mu^1(y_1)\right| + \sup_{y_0\in\mathcal{Y}_0}\left|\mathbb{G}_H^0(y_0) - \mathbb{G}_F^0(y_0) + \mu^0(y_0)\right|,$$

and

$$T_n^C \xrightarrow{d} \int_{\mathcal{Y}_1}\left(\mathbb{G}_H^1(y_1) - \mathbb{G}_F^1(y_1) + \mu^1(y_1)\right)^2 dw_1(y_1) + \int_{\mathcal{Y}_0}\left(\mathbb{G}_H^0(y_0) - \mathbb{G}_F^0(y_0) + \mu^0(y_0)\right)^2 dw_0(y_0),$$

[19] Since we concentrate on $\mathcal{Y}_d$, we require only that for any sequence of measurable sets A_n, $Q_*^d(A_n \cap \mathcal{Y}_d) \to 0$ implies $Q_n^d(A_n \cap \mathcal{Y}_d) \to 0$.

where

$$\mu^d(y_d) = \delta_d E\left[\left(Q^d(y_d|X,\widetilde{p}) - \Lambda(T'\beta_d)\right) 1(D=d)\right]$$
$$+ \delta_\gamma \left[A_\gamma^d + A_d^d J_d(y_d)^{-1} J_{dp}(y_d)\right] J_p^{-1} E\left[\left(q(X,Z) - D\right) H(R'\gamma) R\right]$$
$$+ \delta_d A_d^d J_d(y_d)^{-1} E\left[\left[\Lambda(T'\beta_d) - Q^d(y_d|X,\widetilde{p})\right] H(T'\beta_d) T \cdot 1(D=d)\right],$$

and all terms in $\mu^d(y_d)$ are evaluated at γ^* and β_d^*.

We have a few remarks on Theorem 5. First, in (i), $\mathbb{G}_H^d(\cdot)$ and $\mathbb{G}_F^d(\cdot)$ are the weak limits of $\sqrt{n}\left(\widehat{H}_n^d(\cdot) - H^d(\cdot)\right)$ and $\sqrt{n}\left(\widehat{F}_n^d(\cdot) - F^d(\cdot)\right)$ under the null where $H^d(\cdot) = F^d(\cdot)$. Second, (ii) implies that our tests are consistent against some types of fixed alternatives. Third, from (iii), the local powers are from two sources, the deviation of $p(x,z)$ from the form of $\Lambda\left(R(x,z)'\gamma\right)$ (the δ_γ term) and the deviation of $F_{Y|X,p(X,Z),D}(y|x,p,d)$ from the form of $\Lambda(T(x,p)'\beta_d)$ (the two δ_d terms). The first source has a direct effect on the estimation of $F_{Y|X,p(X,Z),D}(y|x,p,d)$ through p (the term associated with A_γ^d) and an indirect effect through β_d (the term associated with A_d^d). The second source may affect the estimation of $H^d(\cdot)$ (the first δ_d term) and also the estimation of $F^d(\cdot)$ through the estimation of $F_{Y|X,p(X,Z),D}(y|x,p,d)$ (the second δ_d term). Fourth, when γ^* and β_d^* take the pseudo-true value under $q(x,z)$ and $Q^d(y_d|x,p)$, $\mu^d(y_d) = \delta_d E\left[\left(Q^d(y_d|X,\widetilde{p}) - \Lambda(T'\beta_d)\right) 1(D=d)\right]$ and the other two terms are equal to zero. In this case, $\mu^d(y_d)$ is proportional to the difference between the marginal distributions implied by $Q^d(y_d|x,p)$ and $F_*^d(y_d|x,p)$. Fifth, from Section 6 of Andrews (1997) and Section 3.2 of Rothe and Wied (2013), when q does not have a larger support than p_*, a sufficient condition for contiguity is

$$\sup_{(y_d,x,p)\in\mathcal{Y}_d\mathcal{X}\mathcal{P}:f_*^d(y_d|x,p)>0} q^d(y_d|x,p)/f_*^d(y_d|x,p) < \infty,$$

where q^d and f_*^d are the density functions corresponding to Q^d and F_*^d. Intuitively, this would be case when $Q^d(\cdot|\cdot)$ has lighter tails than $F_*^d(\cdot|\cdot)$. Sixth, when $\delta_\gamma = \delta_d = \sqrt{n}$, $q_n(x,z) = q(x,z)$ and $Q_n^d(y_d|x,p) = Q^d(y_d|x,p)$, which implies that the powers of our tests against $q(x,z)$ and $Q^d(y_d|x,p)$ when the sample size is n can be approximated by $P\left(T_A^K > c^K(\alpha)\right)$ and $P\left(T_A^C > c^C(\alpha)\right)$ with $\delta_\gamma = \delta_d = \sqrt{n}$, where T_A^K and T_A^C represent the asymptotic distributions of our test statistics under the local alternative, and $c^K(\alpha)$ and $c^C(\alpha)$ represent the corresponding critical values at level α implied by the asymptotic null distributions in (i).

Note that the critical values $c^K(\alpha)$ and $c^C(\alpha)$ depend on the true value of $(\gamma', \beta_0', \beta_1')'$ under H_0 and also the distribution of $(X', Z', D)'$, so are nuisance parameter dependent. This motivates us to use the bootstrap to obtain these critical values. Our semiparametric bootstrap procedure is as follows.

Step 1. Draw a bootstrap sample $\{(X_i^*, Z_i^*, D_i^*), 1 \le i \le n\}$ with replacement from the realized values $\{(X_i, Z_i, D_i), 1 \le i \le n\}$. Compute the estimate $\widehat{\gamma}^*$ of γ and get $\widehat{p}_i^* = p(X_i^*, Z_i^*; \widehat{\gamma}^*)$.
Step 2. For every i with $D_i^* = d$, put

$$Y_i^* = \begin{cases} \widehat{F}_{Y|X,p(X,Z),D}^{-1}(U_i^*|X_i^*, \widehat{p}_i^*, d), & \text{if } \widehat{F}_{Y|X,p(X,Z),D}^{-1}(U_i^*|X_i^*, \widehat{p}_i^*, d) \in \mathcal{Y}_d, \\ Y_{i*}, & \text{otherwise,} \end{cases}$$

where Y_{i*} is the Y_i corresponding to X_i^* in the original sample, $\{U_i^*, 1 \le i \le n\}$ is a simulated iid sequence of standard uniformly distributed random variables, and $\widehat{F}_{Y|X,p(X,Z),D}(\cdot|x,p,d)$ takes the form as in (18).
Step 3. Use the bootstrap data $\{(Y_i^*, X_i^*, \widehat{p}_i^*, D_i^*), 1 \le i \le n\}$ to compute estimates $\widehat{H}_n^{d*}$ and $\widehat{F}_n^{d*}$ exactly as in (20) and (21), and compute the corresponding bootstrap realization of the test statistics:

$$T_n^{K^*} = \sqrt{n} \sup_{y_1 \in \mathcal{Y}_1} \left| \widehat{H}_n^{1*}(y_1) - \widehat{F}_n^{1*}(y_1) \right| + \sqrt{n} \sup_{y_0 \in \mathcal{Y}_0} \left| \widehat{H}_n^{0*}(y_0) - \widehat{F}_n^{0*}(y_0) \right|,$$

or

$$T_n^{C*} = n \int_{\mathcal{Y}_1} \left(\widehat{H}_n^{1*}(y_1) - \widehat{F}_n^{1*}(y_1) \right)^2 dw_1(y_1) + n \int_{\mathcal{Y}_0} \left(\widehat{H}_n^{0*}(y_0) - \widehat{F}_n^{0*}(y_0) \right)^2 dw_0(y_0).$$

Step 4. Repeat Step 1-3 B times to get $\left\{T_{nb}^{K^*}\right\}_{b=1}^{B}$ and $\left\{T_{nb}^{C^*}\right\}_{b=1}^{B}$ which approximate the bootstrap distribution of the test statistics, and use the $(1-\alpha)$th empirical quantiles of $\left\{T_{nb}^{K^*}\right\}_{b=1}^{B}$ and $\left\{T_{nb}^{C^*}\right\}_{b=1}^{B}$ to approximate the asymptotic critical values. The bootstrap critical values are denoted as $\widehat{c}_n^K(\alpha)$ and $\widehat{c}_n^C(\alpha)$, respectively.

Since the bootstrap distribution in Step 2 mimics the distribution of the data over $\mathcal{Y}_d$ under the null, our bootstrap procedure is valid even though the data might be generated from an alternative distribution. The following theorem rigorously states this result by extending Corollary 1 of Andrews (1997) and Theorem 3 of Rothe and Wied (2013).

Theorem 6 *Under Condition DR (b) and (c), the following statements hold for any $\alpha \in (0,1)$:*

(i) *Under H_0,*

$$\lim_{n\to\infty} P\left(T_n^K > \widehat{c}_n^K(\alpha)\right) = \alpha \text{ and } \lim_{n\to\infty} P\left(T_n^C > \widehat{c}_n^C(\alpha)\right) = \alpha.$$

(ii) *Under any fixed alternative such that $P(Y \in S_d(\beta_d, \gamma), D = d) > 0$ for all $\beta_d \in \mathcal{B}_d(\mathcal{Y}_d)$, $\gamma \in \Gamma$,*

$$\lim_{n\to\infty} P\left(T_n^K > \widehat{c}_n^K(\alpha)\right) = 1 \text{ and } \lim_{n\to\infty} P\left(T_n^C > \widehat{c}_n^C(\alpha)\right) = 1.$$

(iii) *Under any contiguous local alternative as in Theorem 5(iii),*

$$\lim_{n\to\infty} P\left(T_n^K > \widehat{c}_n^K(\alpha)\right) \geq \alpha \text{ and } \lim_{n\to\infty} P\left(T_n^C > \widehat{c}_n^C(\alpha)\right) \geq \alpha.$$

(i) implies that under H_0, $\widehat{c}_n^K(\alpha) \xrightarrow{p} c^K(\alpha)$ and $\widehat{c}_n^C(\alpha) \xrightarrow{p} c^C(\alpha)$, where the randomness in the probability convergence includes both the randomness of the original sample and the independent randomness of the bootstrap simulations (this also applies to other statements in Theorem 6). (ii) is a corollary of Theorem 5(ii) since $\widehat{c}_n^K(\alpha)$ and $\widehat{c}_n^C(\alpha)$ are bounded in probability under the fixed alternative. (iii) states that T_n^K and T_n^C are *asymptotically locally unbiased*.

We have a few further remarks on Theorem 5 and 6. First, although hard to imagine, it is indeed possible that two marginal distributions match each other although the corresponding conditional distributions are different. As an alternative of our tests which are based on marginal distributions, Andrews (1997) and Rothe and Wied (2013) suggest to test whether the *joint* empirical distribution matches the *joint* distribution implied by the conditional distribution under the null. In our case, we need to test whether the empirical distribution of $(Y_i, X_i, \widehat{p}_i, D_i)$ matches the distribution implied by $F_{Y|X,p(X,Z),D}(y_d|x, \widehat{p}, d) = \Lambda\left(T(x,\widehat{p})'\beta_d(y_d)\right)$ with $\widehat{p}(x,z) = \Lambda\left(R(x,z)'\widehat{\gamma}\right)$. Such tests are technically difficult because the weak limit of the empirical distribution $\{\widehat{p}_i\}_{i=1}^n$ is hard to derive given that $\widehat{p}_i = \Lambda\left(R(X_i,Z_i)'\widehat{\gamma}\right)$ is a *random* transformation of $(X_i', Z_i')'$. Our tests avoid this problem. Also, since the distribution of covariates is usually complicated, our tests are easier to implement than the tests based on the joint distribution given that the distribution of $Y|D$ is usually continuous. Second, as suggested at the end of the last subsection, we can discretize $\mathcal{Y}_d$ in the construction of our test statistics. For example, our test statistics can be

$$T_n^K = \sqrt{n} \sup_{D_i=1, Y_i \in \mathcal{Y}_1} \left| \widehat{H}_n^1(Y_i) - \widehat{F}_n^1(Y_i) \right| + \sqrt{n} \sup_{D_i=0, Y_i \in \mathcal{Y}_0} \left| \widehat{H}_n^0(Y_i) - \widehat{F}_n^0(Y_i) \right|,$$

or

$$T_n^C = \sum_{i=1}^{n} 1(D_i = 1, Y_i \in \mathcal{Y}_1)\left(\widehat{H}_n^1(Y_i) - \widehat{F}_n^1(Y_i)\right)^2 + \sum_{i=1}^{n} 1(D_i = 0, Y_i \in \mathcal{Y}_0)\left(\widehat{H}_n^0(Y_i) - \widehat{F}_n^0(Y_i)\right)^2,$$

where $w_d(y_d)$ in T_n^C corresponds to the empirical distribution of $Y_i \in \mathcal{Y}_d$ such that $D_i = d$. It can be shown that the asymptotic results in Theorem 5 still hold with $w_d(y_d) = P(Y \le y_d, D = d)$ and the bootstrap procedure with corresponding adjustments is still valid. Third, our tests concentrate on testing whether $F_{Y|X,p(X,Z),D}(y_d|x,p,d) = \Lambda\left(T(x,p)'\beta_d(y_d)\right)$, so misspecification in $p(X,Z)$ has only indirect effects on the power of our tests. Nevertheless, before carrying out our tests, we can first use the *conditional Kolmogorov* (CK) test of Andrews (1997) to check whether $p(x,z) = \Lambda\left(R(x,z)'\gamma\right)$ is misspecified; see also McFadden (1974), Horowitz (1985) and Andrews (1988a,b) for related discussions.

5 Counterfactual Analysis

The tools developed in the last section can be used for counterfactual analysis. We first derive the transition matrix of deciles of (marginal) Y_0 to deciles of (marginal) Y_1 as Table 8 of Carneiro et al. (2003), and then estimate a policy relevant treatment parameter under policy invariance. Many other counterfactual analyses can be conducted since the conditional joint distribution of (Y_0, Y_1) given (X, U_D) is known under the RP assumption (A6). So this section provides an alternative way to remove the veil of ignorance in assessing the distributional impacts of social policies as described in Carneiro et al. (2001).

5.1 Derivation of the Transition Matrix

It should be emphasized that we do not impose the unconditional RP assumption in this subsection as in Corollary 2; otherwise, the transition matrix will degenerate to the identity matrix. Actually, only the conditional RP assumption (A6) is imposed.

For each y_i such that $D_i = 0$, we need to derive the corresponding outcome when $D_i = 1$; similarly, for each y_i such that $D_i = 1$, we need to derive the corresponding outcome when $D_i = 0$. Denote the data points with $D_i = 0$ as $\{y_{0i}\}_{i=1}^{n_0}$ and with $D_i = 1$ as $\{y_{1i}\}_{i=1}^{n_1}$, where $n_d = \sum_{i=1}^n 1(D_i = d)$. Assume further that $\{y_{0i}\}_{i=1}^{n_0}$ and $\{y_{1i}\}_{i=1}^{n_1}$ are ordered ascendingly. Now, the counterfactual outcome for y_{0i} is estimated as

$$\widehat{y}_{1i} = \widehat{F}^{-1}_{Y_1|X,U_D}\left(\widehat{F}_{Y_0|X,U_D}(y_{0i}|X_i,\widehat{p}_i)\middle| X_i, \widehat{p}_i\right)$$

and the counterfactural outcome for y_{1i} is estimated as

$$\widehat{y}_{0i} = \widehat{F}^{-1}_{Y_0|X,U_D}\left(\widehat{F}_{Y_1|X,U_D}(y_{1i}|X_i,\widehat{p}_i)\middle| X_i, \widehat{p}_i\right).$$

As a result, the counterfactural samples for Y_0 are $\mathbb{Y}_0 = \{y_{01}, \cdots, y_{0n_0}, \widehat{y}_{01}, \cdots, \widehat{y}_{0n_1}\}$ and the corresponding counterfactual samples for Y_1 are $\mathbb{Y}_1 = \{y_{11}, \cdots, y_{1n_1}, \widehat{y}_{11}, \cdots, \widehat{y}_{1n_0}\}$. Suppose the jth decile of $\mathbb{Y}_0$ is $y_0^{(j)}$ and denote $\mathbb{Y}_0^{(j)} = \left\{y_i \in \mathbb{Y}_0 | y_0^{(j-1)} < y_i \le y_0^{(j)}\right\}$, where $j = 1, \cdots, 10$, $y_0^{(0)} = \min \mathbb{Y}_0$ and $y_0^{(10)} = \max \mathbb{Y}_0$.[20] The set of the corresponding indices in $\mathbb{Y}_0^{(j)}$ is denoted as $\mathbb{I}_0^{(j)} = \left\{i | y_i \in \mathbb{Y}_0^{(j)}\right\}$ and the set of the corresponding

[20] To avoid the identification problem in finite samples, we can assume the counterfactual value of the smallest (say) 0.1% of $\{y_{01}, \cdots, y_{0n_0}\}(\{y_{11}, \cdots, y_{1n_1}\})$ will fall in the first decile of $\mathbb{Y}_1(\mathbb{Y}_0)$ and that of the largest (say) 0.1% of $\{y_{01}, \cdots, y_{0n_0}\}(\{y_{11}, \cdots, y_{1n_1}\})$ will fall in the last decile of $\mathbb{Y}_1(\mathbb{Y}_0)$. Also, the smallest (say) 0.1% of $\{y_{01}, \cdots, y_{0n_0}\}(\{y_{11}, \cdots, y_{1n_1}\})$ will stay in the first decile of $\mathbb{Y}_0(\mathbb{Y}_1)$ and the largest (say) 0.1% of $\{y_{01}, \cdots, y_{0n_0}\}(\{y_{11}, \cdots, y_{1n_1}\})$ will stay in the last decile of $\mathbb{Y}_0(\mathbb{Y}_1)$.

data points in $\mathbb{Y}_1$ is denoted as $\widetilde{\mathbb{Y}}_1^{(j)} = \left\{ y_i \in \mathbb{Y}_1 | i \in \mathbb{I}_0^{(j)} \right\}$. $\mathbb{Y}_1^{(j)}$ and $\mathbb{I}_1^{(j)}$ are similarly defined. Then the transition matrix is estimated as

$$\begin{pmatrix} \frac{\#\left\{\mathbb{I}_0^{(1)} \cap \mathbb{I}_1^{(1)}\right\}}{\#\left\{\mathbb{I}_0^{(1)}\right\}} & \frac{\#\left\{\mathbb{I}_0^{(1)} \cap \mathbb{I}_1^{(2)}\right\}}{\#\left\{\mathbb{I}_0^{(1)}\right\}} & \cdots & \frac{\#\left\{\mathbb{I}_0^{(1)} \cap \mathbb{I}_1^{(10)}\right\}}{\#\left\{\mathbb{I}_0^{(1)}\right\}} \\ \frac{\#\left\{\mathbb{I}_0^{(2)} \cap \mathbb{I}_1^{(1)}\right\}}{\#\left\{\mathbb{I}_0^{(2)}\right\}} & \frac{\#\left\{\mathbb{I}_0^{(2)} \cap \mathbb{I}_1^{(2)}\right\}}{\#\left\{\mathbb{I}_0^{(2)}\right\}} & \cdots & \frac{\#\left\{\mathbb{I}_0^{(2)} \cap \mathbb{I}_1^{(10)}\right\}}{\#\left\{\mathbb{I}_0^{(2)}\right\}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\#\left\{\mathbb{I}_0^{(10)} \cap \mathbb{I}_1^{(1)}\right\}}{\#\left\{\mathbb{I}_0^{(10)}\right\}} & \frac{\#\left\{\mathbb{I}_0^{(10)} \cap \mathbb{I}_1^{(2)}\right\}}{\#\left\{\mathbb{I}_0^{(10)}\right\}} & \cdots & \frac{\#\left\{\mathbb{I}_0^{(10)} \cap \mathbb{I}_1^{(10)}\right\}}{\#\left\{\mathbb{I}_0^{(10)}\right\}} \end{pmatrix},$$

where $\#\{A\}$ for a set A is the number of the elements in A, and $\#\left\{\mathbb{I}_0^{(i)}\right\}$, $i = 1, \cdots, 10$, is roughly $n/10$.

5.2 Relative Marginal Policy Relevant Quantile Treatment Effect

As in Heckman and Vytlacil (2001c) or Section 3 of Heckman and Vytlacil (2005), we postulate a policy question or decision problem of interest and derive the treatment parameter that answer it. We consider a class of policies that affect $p(X, Z)$, but do not affect $F_{Y_d|X,U_D}(y_d|x, u_D)$ or $\Delta_\tau^{MQTE}(x, u_D)$. Usually, we condition on X and study the effect of changing Z. Let a and a' denote two potential policies (without loss of generality, let a be the original observed policy and a' be the hypothetical unrealized policy); then we use subscripts a and a' to distinguish variables under these two policies and assume that the assumptions in Section 2.1 are satisfied under these two policies.

Policy invariance is a key assumption for any study of policy evaluation. It allows analysts to characterize outcomes without specifying how those outcomes are obtained. Policy invariance was first defined and formalized by Marschak (1953) and Hurwicz (1962); see Section 4.6 of Heckman and Vytlacil (2007a) for precise definitions of invariance. For our purpose, we need only the following assumption:

(A7) The distribution of $(Y_{0,a}, Y_{1,a}, U_{D,a})$ conditional on $X_a = x$ is the same as the distribution of $(Y_{0,a'}, Y_{1,a'}, U_{D,a'})$ conditional on $X_{a'} = x$.

This assumption is exactly the assumption (A-7) of Heckman and Vytlacil (2005).[21] To simplify notations, we keep implicit the conditioning on $X_a = x$ and $X_{a'} = x$.

Although in principle, we can estimate the policy relevant quantile treatment effect (PRQTE) as

$$\text{PRQTE}_\tau^{a,a'} = Q_{Y_{a'}}(\tau) - Q_{Y_a}(\tau),^{22}$$

the support of p_a and $p_{a'}$ may be far from $[0, 1]$, where $Q_{Y_{a'}}(\cdot)$ is the inverse function of

$$F_{Y_{a'}}(y) \equiv \int_0^1 \left[\left(1 - F_{p_{a'}}(u_D)\right) F_{Y_{1,a}|U_{D,a}}(y|u_D) + F_{p_{a'}}(u_D) F_{Y_{0,a}|U_{D,a}}(y|u_D) \right] du_D,^{23} \tag{23}$$

and $Q_{Y_a}(\cdot)$ is the inverse function of

$$F_{Y_a}(y) \equiv \int_0^1 \left[\left(1 - F_{p_a}(u_D)\right) F_{Y_{1,a}|U_{D,a}}(y|u_D) + F_{p_a}(u_D) F_{Y_{0,a}|U_{D,a}}(y|u_D) \right] du_D,$$

[21] Policy invariance assumption may not hold in practice. For example, in the education expansion example, increases in college enrollment may affect the spending on each student; as a result, the distribution of $Y_1|(X, U_D)$ will shift left.

[22] Note that we implicitly assume the RP condition holds between the policy regime and the original regime.

[23] See the supplementary materials for its derivation.

$p_a = p_a(Z_a)$ and $p_{a'} = p_{a'}(Z_{a'})$. Extrapolation out of $\mathcal{P}_{x,a}$ and $\mathcal{P}_{x,a'}$ may generate unreliable predictions especially when $\widehat{F}_{Y_a|p_a,D_a}$ is estimated based on a nonparametric procedure. To avoid this problem, Carneiro et al. (2010) consider a marginal version of the policy relevant treatment effect (MPRTE) in the framework of average treatment evaluation.[24] We extend the MPRTE to the marginal policy relevant quantile treatment effect (MPRQTE) below.

To define the MPRQTE, assume $F_{p_{a'}}$ stays in a parametrized space $\{F_\alpha : \alpha \in M, 0 \in M, F_0 = F_{p_a}\}$. Then the MPRQTE is a path derivative along the path $\{F_\alpha : \alpha \in M\}$ of

$$\text{PRQTE}_\tau\left(F_\alpha\right) = \frac{Q_\alpha\left(\tau\right) - Q_0\left(\tau\right)}{E_\alpha[p] - E_0[p]},$$

i.e.,

$$\text{MPRQTE}_\tau\left(\{F_\alpha\}\right) = \lim_{\alpha\to 0} \text{PRQTE}_\tau(F_\alpha)$$

where $E_\alpha[p] = E_{F_\alpha}[p]$, $Q_\alpha\left(\cdot\right)$ is the inverse function of

$$G_\alpha(y) \equiv \int_0^1 \left[\left(1 - F_\alpha(u_D)\right) F_{Y_1|U_D}(y|u_D) + F_\alpha(u_D) F_{Y_0|U_D}(y|u_D)\right] du_D$$

and $Q_0(\cdot)$ is the inverse function of

$$G_0(y) \equiv \int_0^1 \left[\left(1 - F_0(u_D)\right) F_{Y_1|U_D}(y|u_D) + F_0(u_D) F_{Y_0|U_D}(y|u_D)\right] du_D$$

with $F_{Y_d|U_D}(y|u_D) = F_{Y_{d,a}|U_{D,a}}(y|u_D)$. It is straightforward to show that

$$\text{MPRQTE}_\tau\left(\{F_\alpha\}\right) \equiv \lim_{\alpha\to 0} \frac{Q_\alpha\left(\tau\right) - Q_0\left(\tau\right)}{E_\alpha[p] - E_0[p]} = \frac{\int_0^1 \omega(u_D)\left[F_{Y_1|U_D}(Q_0\left(\tau\right)|u_D) - F_{Y_0|U_D}(Q_0\left(\tau\right)|u_D)\right] du_D}{g_0(Q_0\left(\tau\right))},$$

where $g_0(\cdot)$ is the density function corresponding to $G_0(\cdot)$, $\frac{\partial}{\partial\alpha}F_0(u_D)$ is the shorthand expression for $\frac{\partial}{\partial\alpha}F_\alpha(u_D)\big|_{\alpha=0}$, and

$$\omega(u_D) = \frac{\frac{\partial}{\partial\alpha}F_0(u_D)}{\int_0^1 \frac{\partial}{\partial\alpha}F_0(t)dt}$$

is the weight function in Carneiro et al. (2010). Three popular policy changes are $p_\alpha = p + \alpha$, $p_\alpha = p(1+\alpha)$ and $Z_\alpha^k = Z^k + \alpha$, and the corresponding weight functions are $f_p(u_D)$, $\frac{u_D f_p(u_D)}{E_0[p]}$ and $\frac{f_p(u_D) f_V(F_V^{-1}(u_D))}{E[f_V(\mu_D(Z))]}$, where $\mu_D(Z) = Z'\gamma$, and Z^k is the kth component of Z and is continuous; see Table 1 of Carneiro et al. (2011) for a summary. Since $f_p(u_D)$ appears in $\omega(u_D)$, the numerator of $\text{MPRQTE}_\tau\left(\{F_\alpha\}\right)$ can be recovered even if the support of p is a strict subset of $[0,1]$. However, to recover the denominator, we still need the full support condition.

To avoid the full support condition, we can define the relative MPRQTE (RMPRQTE) as

$$\text{RMPRQTE}_\tau\left(\{F_\alpha\}, \{F_\beta\}\right) = \frac{\text{MPRQTE}_\tau\left(\{F_\alpha\}\right)}{\text{MPRQTE}_\tau\left(\{F_\beta\}\right)} = \frac{\int_0^1 \omega_\alpha(u_D)\left[F_{Y_1|U_D}(Q_0\left(\tau\right)|u_D) - F_{Y_0|U_D}(Q_0\left(\tau\right)|u_D)\right] du_D}{\int_0^1 \omega_\beta(u_D)\left[F_{Y_1|U_D}(Q_0\left(\tau\right)|u_D) - F_{Y_0|U_D}(Q_0\left(\tau\right)|u_D)\right] du_D},$$

where $\{F_\beta\}$ is the distribution function sequence associated with another policy, and $\omega_\alpha(u_D)$ and $\omega_\beta(u_D)$ are weight functions associated with $\{F_\alpha\}$ and $\{F_\beta\}$. In practice, there may be more than one policies under con-

[24] Ichimura and Taber (2000) present a discussion of local policy analysis in a model without the MTE structure using a framework developed by Hurwicz (1962).

sideration, and the policy maker needs to choose one among them. In such cases, $\text{RMPRQTE}_\tau(\{F_\alpha\},\{F_\beta\})$ is a useful parameter. Note that identification of $Q_0(\tau)$ still need identification of $G_0(y)$ which requires the full support condition, so we redefine the RMPRQTE as

$$\text{RMPRQTE}_y(\{F_\alpha\},\{F_\beta\}) = \frac{\int_0^1 \omega_\alpha(u_D)\left[F_{Y_1|U_D}(y|u_D) - F_{Y_0|U_D}(y|u_D)\right]du_D}{\int_0^1 \omega_\beta(u_D)\left[F_{Y_1|U_D}(y|u_D) - F_{Y_0|U_D}(y|u_D)\right]du_D}, y \in \mathcal{Y}.$$

Now, this parameter can be identified even if the full support condition does not hold. It is the ratio of the effects induced by two policies on the population with outcome level y and can be estimated by its sample analog, say, $\widehat{\text{RMPRQ}}\text{TE}_y(\{F_\alpha\},\{F_\beta\})$. As in the test of stochastic dominance in Section 4.4, we can test whether $\left|\text{RMPRQTE}_y(\{F_\alpha\},\{F_\beta\})\right| \geq 1$ (or ≤ 1) for $y \in \mathcal{Y}$. The corresponding test statistics are

$$\sqrt{n}\sup_{y\in\mathcal{Y}}\max\left(1-\left|\widehat{\text{RMPRQ}}\text{TE}_y(\{F_\alpha\},\{F_\beta\})\right|,0\right) \text{ and } n\int_{\mathcal{Y}}\left|\max\left(1-\left|\widehat{\text{RMPRQ}}\text{TE}_y(\{F_\alpha\},\{F_\beta\})\right|,0\right)\right|^2 dy.$$

6 Comments on the Literature

In this section, we review three papers on the quantile treatment effect evaluation, pointing out their underlying assumptions, weaknesses, and interactions with our framework.

6.1 Comments on Chernozhukov and Hansen (2005)

Chernozhukov and Hansen (2005) express

$$Y_d = q(d, X, U_d) \text{ with } U_d|X \sim U(0,1)^{25}$$

by the Skorohod representation, where $q(d,x,\tau)$ is the quantile function of Y_d conditional on $X = x$. This representation is essential in developing their identification results. Chernozhukov and Hansen impose the following assumptions on the model:

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

A2. Independence: Conditional on $X = x$, $\{U_d\}$ are independent of Z.

A3. Selection: $D \equiv \delta(Z,X,V)$ for some unknown function δ and random vector V.

A4. Rank Invariance (RI) or Rank Similarity (RS): Conditional on $X = x$, $Z = z$, (a) $\{U_d\}$ are equal to each other; or, more generally, (b) $\{U_d\}$ are identically distributed, conditional on V.

A5. Observed Variables: Observed variables consist of $Y \equiv q(D,X,U_D)$, D, X, and Z.[26]

Some obvious differences between their setup and ours are as follows. A1 restricts Y_d to be continuously distributed, while we do not need this requirement. Their d can be continuous, while we consider only the binary treatment case. To further contrast our model with theirs, we put these two setups side by side for comparison:

[25] Note that U_d may be dependent of X even under unconfoundedness, which generates the usual (parametric) endogeneity. As argued in Section 2, in the usual quantile regression model, $Y = q(X,U)$, U should be expressed as $U(X)$, and $U(X) \sim U(0,1)$ for any value of X. In other words, although the parametric quantile regression is not valid, the nonparametric quantile regression still works. Only if $Y = X'\beta(U)$ with $U \perp X$ and $U \sim U(0,1)$, the parametric quantile regression is valid.

[26] Note that their $U_D = DU_1 + (1-D)U_0$ is different from our U_D.

Chernozhukov and Hansen (2005):	Our Model:
$Y_d = q(d, X, U_d)$ with $U_d \vert X \sim U(0,1)$	$Y_d = q(d, X, V, U_d)$ with $U_d \vert (X, V) \sim U(0,1)$
$D \equiv \delta(Z, X, V)$ with V being a random vector	$D = 1(\mu_D(X, Z) - V \geq 0)$ with $Z \perp V \vert X$
$U_0 \vert (X, Z) = U_1 \vert (X, Z)$ and $Z \perp (U_0, U_1) \vert X$	$U_0 \vert (X, V) = U_1 \vert (X, V)$ and $Z \perp (U_0, U_1) \vert X$

Obviously, Chernozhukov and Hansen put more restrictions on the outcome equation, while we put more restrictions on the choice equation. For example, they assume that the quantile of Y depends on V only through D, while we assume that even given $D = d$, the quantile of Y_d depends on V. On the other hand, they do not impose the index structure on the choice equation. However, under the general setup of outcome equations, the monotonicity assumption, which is implied by the indexed choice model, is hard to relax; see Section 6 of Heckman and Vytlacil (2005) for discussions. Given the indexed structure of D, it is without loss of generality to assume $Z \perp V|X$; see the discussion in Section 2.[27] Also, our RP assumption is different from their RI assumption since the conditional variables are different.[28]

Our target is to identify $q(d, x, v, \tau)$, while Chernozhukov and Hansen's target is to identify $q(d, x, \tau)$.[29] It should be emphasized that Chernozhukov and Hansen do not really impose the RP assumption for their purpose, i.e., they do not use $U_1| (X = x) = U_0| (X = x)$. Otherwise, they do not need to assume A4 since it is implied by $U_1| (X = x) = U_0| (X = x)$. This means that their $q(1, x, \tau) - q(0, x, \tau)$ is only a difference of two quantiles and may not have a causal interpretation (i.e., may not be the difference between two potential outcomes for the same group, say the τth quantile, of individuals). As mentioned in Section 3, this RP assumption is not really used in our partial identification results either, so their identification results are comparable with those in Section 3. As argued in Section 3, $q(d, x, \tau)$ may not be point identified in the general model or even under the nonparametric selection model unless assumptions such as $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$ are satisfied. Of course, as argued in Section 3 of Chernozhukov and Hansen (2005), we can still identify $q(d, x, \tau)$ by expressing $q(d, x, V, U_d)$ in the form of $q(d, x, U_d)$. However, this object is not of practical interest since it is not related to any interpretable parameter of treatment effects unless the RP assumption on $X = x$ is satisfied. To see why, note that

$$q(d, x, \tau) = F^{-1}_{Y_d|X}(\tau|x) \text{ with } F_{Y_d|X}(y|x) = \int q_d^{-1}(y|x, v) dF_{V|X}(v|x),$$

where $q_d^{-1}(y|x, v)$ is the inverse function of $q(d, x, v, \tau)$ with respect to τ.

Their assumption on the outcome equation, combined with A4, imposes strong restrictions on the form of the essential heterogeneity. Consider our outcome equation (1) with additively separable U_d. Under the RS assumption,

$$\begin{aligned}
&Q_\tau(Y_1|X, Z, V) - Q_\tau(Y_0|X, Z, V) \\
&= Q_\tau(\mu_1(X, U_1)|X, Z, V) - Q_\tau(\mu_0(X, U_0)|X, Z, V) \\
&= \mu_1(X) + Q_\tau(U_1|X, Z, V) - \mu_0(X) - Q_\tau(U_0|X, Z, V) \\
&= \mu_1(X) - \mu_0(X)
\end{aligned}$$

does not depend on Z and V so does not depend on D, where the last equality follows from the RS assumption.

[27] As mentioned in Chernozhukov and Hansen (2005), Imbens and Angrist (1994) provide important examples in which Z is assigned depending on D, so it is critical to allow Z and V to be dependent (conditional on X) in their setup.

[28] A4 is also used in Vytlacil and Yildiz (2007) in estimating marginal means of potential outcomes in weakly separable models.

[29] They also mention the "local" identification condition, but the "local" there is different from the "local" in our "local" QTE. Their "local" is local to the vector $\{q(d, x, \tau), d = 0, 1\}$, while our "local" is local to $U_D = u_D$ for a specific u_D.

This is why Theorem 1 of Chernozhukov and Hansen (2005) is stated as $Q_\tau(Y_1|X,Z) - Q_\tau(Y_0|X,Z) = q(1,X,\tau) - q(0,X,\tau)$ and V is not involved.[30] Even in the non-additively separable setup of Y_d, $Y_1 - Y_0 = q(1,X,U) - q(0,X,U)$, where $U|(X,Z) = U_1|(X,Z) = U_0|(X,Z)$ under the RI assumption. $Y_1 - Y_0$ may be correlated with D, but only through the same "one" source as Y_0 is correlated with D. In summary, it is fair to claim that the main problem that Chernozhukov and Hansen target is the selection effect instead of the essential heterogeneity; see Torgovitsky (2011) for related discussions.

If we express our model as $Y = q(D,V,U)$ and $D = 1(\mu_D(Z) - V \geq 0)$, where X is depressed, and $U = DU_1 + (1-D)U_0$, then we can compare our model with those in Chesher (2003), Koenker and Ma (2006), Lee (2007) and Imbens and Newey (2009) where triangular structures are imposed. Their model takes the following form:

$$
\begin{aligned}
Y &= \mu(D,V,U), \\
D &= \varphi(Z,V),
\end{aligned}
$$

where V and U are scalars jointly independent of Z.[31] Similar as our model, their Y is also influenced by two random errors. However, they assume D and Z are both continuous and $\mu(D,V,U)$ is strictly increasing in U and $\varphi(Z,V)$ is strictly increasing in V. So their model does not cover the treatment effect model in this paper. Chesher (2005) provides some partial identification results when D is discrete. However, as mentioned in his Section 5.3, his arguments cannot be extended to the binary endogenous variables case (see Jun et al. (2011) for further discussions). By strengthening D to take a latent index form, we can recover $\mu(D,V,U)$ for V's implied by the D equation.

The general choice equation $D \equiv \delta(Z,X,V)$ of Chernozhukov and Hansen may seem surprising. However, there is a parallel result in the average treatment context without the essential heterogeneity. We state this result formally in the following theorem.

Theorem 7 *Suppose $Y_d = \mu_d(X) + U_d$ with $E[U_d|X] = 0$,[32] $D = \delta(X,Z,V)$, $E[U_0|X,Z] = 0$, $E[U_1 - U_0|X,Z] = 0$ and $(U_1 - U_0) \perp V|\,(X,Z)$.[33] Then $E[Y - \mu_0(X) - D\,(\mu_1(X) - \mu_0(X))\,|X,Z] = 0$. $\mu_1(X) - \mu_0(X)$ can be identified as long as $E[D|X,Z] \neq E[D|X]$ almost surely, which is equivalent to that conditional on X, D is complete for Z. Furthermore, the IV estimator of $\mu_1(X) - \mu_0(X)$ is consistent.*

As in Theorem 4 of Chernozhukov and Hansen (2005), the above theorem shows that the average treatment effect $\mu_1(X) - \mu_0(X)$ can be identified under some completeness assumption. The completeness condition is a global (about Z, not X) condition, while our identification condition in Theorem 1 is a local (abount Z) condition. Due to such kind of completeness conditions, Chernozhukov and Hansen's identification scheme is more convenient to put in the nonparametric IV framework rather than the structural treatment effect model, or their identification scheme is more of "reduced-form" than "structural".

As shown in the proof of Theorem 5, as long as there exist two values of Z, say 0 and 1, such that $E[D|X = x, Z = 0] \neq E[D|X = x, Z = 1]$, $\mu_1(x) - \mu_0(x)$ can be identified. Such an assumption parallels the full rank condition in Theorem 2 of Chernozhukov and Hansen (2005). This full rank condition is equivalent

[30] $E[U_1 - U_0|X,Z,V] = 0$ under the RS assumption, so the usual essential heterogeneity in the average treatment effect evaluation is excluded. The RI assumption $U_1|X,Z = U_0|X,Z$ is somewhat like the common treatment effect assumption.

[31] Chesher (2003) actually uses a local independence condition for local identification. Imbens and Newey (2009) also analyze identification of average derivatives and other functionals of $\mu(D,V,U)$ without the condition that U is a scalar.

[32] In average treatment effect evaluation, the additively separable formulation of Y_d is without loss of generality because we can define the new U_d as $\mu_d(X,U_d) - E\left[\mu_d(X,U_d)|X\right]$.

[33] This implies $(U_1 - U_0) \perp D|(X,Z)$, i.e., the choice is made without knowledge of the idiosyncratic gain after controlling for the observables.

to a monotone likelihood ratio condition,

$$\frac{f_{Y,D|Z}(y_1,1|1)}{f_{Y,D|Z}(y_0,0|1)} > \frac{f_{Y,D|Z}(y_1,1|0)}{f_{Y,D|Z}(y_0,0|0)}$$

for y_1 (and y_0) in a neighborhood of the τth quantile of Y_1 (and Y_0). This condition requires the impact of Z on the joint distribution of (Y,D) to be sufficiently rich, which is in spirit similar to our assumption that $p_x^{\inf}=0$ and $p_x^{\sup}=1$ for identifying the QTE in Section 3, but we need only the impact of Z on D to be sufficiently rich.[34] In our framework,

$$\begin{aligned}\frac{f_{Y,D|Z}(y_1,1|1)}{f_{Y,D|Z}(y_0,0|1)} &= \frac{f_{Y|D,Z}(y_1|1,1)P(D=1|Z=1)}{f_{Y|D,Z}(y_0|0,1)P(D=0|Z=1)} \\ &= \frac{f_{Y_1}(y_1|U_D\le p(1))p(1)}{f_{Y_0}(y_0|U_D>p(1))\,(1-p(1))} = \frac{\displaystyle\int_0^{p(1)} f_{Y_1|U_D}(y_1|u_D)du_D}{\displaystyle\int_{p(1)}^1 f_{Y_0|U_D}(y_0|u_D)du_D}\end{aligned}$$

and similarly

$$\frac{f_{Y,D|Z}(y_1,1|0)}{f_{Y,D|Z}(y_0,0|0)} = \frac{\displaystyle\int_0^{p(0)} f_{Y_1|U_D}(y_1|u_D)du_D}{\displaystyle\int_{p(0)}^1 f_{Y_0|U_D}(y_0|u_D)du_D}.$$

So the monotone likelihood ratio condition requires

$$\frac{\displaystyle\int_0^{p(1)} f_{Y_1|U_D}(y_1|u_D)du_D}{\displaystyle\int_{p(1)}^1 f_{Y_0|U_D}(y_0|u_D)du_D} > \frac{\displaystyle\int_0^{p(0)} f_{Y_1|U_D}(y_1|u_D)du_D}{\displaystyle\int_{p(0)}^1 f_{Y_0|U_D}(y_0|u_D)du_D}. \tag{24}$$

When $\displaystyle\int_{p(0)}^{p(1)} f_{Y_1|U_D}(y_1|u_D)du_D>0$ and $\displaystyle\int_{p(0)}^{p(1)} f_{Y_0|U_D}(y_0|u_D)du_D>0$, i.e., $f_{Y_1|U_D}(y_1|u_D)$ and $f_{Y_0|U_D}(y_0|u_D)du_D$ are not completely zero for $u_D\in(p(0),p(1))$, this condition is equivalent to $p(1)>p(0)$, i.e., Z has a non-trivial impact on D, which is less stringent than the identification condition of the QTE in Section 3, namely, $p(1)=1$ and $p(0)=0$. This is understandable since our model also covers the essential heterogeneity while Chernozhukov and Hansen consider only the selection effect. In the unconfouned case where $f_{Y_1|U_D}(y_1|u_D)$ and $f_{Y_0|U_D}(y_0|u_D)$ does not depend on u_D, (24) reduces to $p(1)/(1-p(1)) > p(0)/(1-p(0))$, which is equivalent to $p(1)>p(0)$. It is easy to see that $p(1)>p(0)$ is also the above identification condition $E[D|X=x,Z=0]\neq E[D|X=x,Z=1]$ in our framework.

Note also that the IV-QRE of Chernozhukov and Hansen (2006) identifies $q(1,X,\tau)-q(0,X,\tau)$ regardless of what Z is used. However, when the essential heterogeneity exists, interpretation of the IV-QRE depends on the specification of Z even if the same set of instruments (among Z) are used in the estimation. This point is emphasized by Angrist et al. (2000) and Heckman and Vytlacil (2005); see also footnote 6 of Carneiro et al. (2011) for an intuitive explanation. Heckman and Vytlacil (2005) also express the usual IV estimator as a weighted average of the MTE in the average treatment context, while the task is quite complicated, even if not impossible, in the quantile treatment environment. Finally, we summarize important literature on the

[34] Their identification conditions do not imply or are implied by our identification assumption that $p_x^{\inf}=0$ and $p_x^{\sup}=1$. Their conditions are not easy to check, while checking of our conditions is quite straightforward, e.g., by drawing histogram of $p(X,Z)$ as in e.g., HIT (1997) and HIST (1998).

average treatment effect and the quantile treatment effect in the following table.

	Average Treatment Effect	Quantile Treatment Effect
Unconfoundedness	Imbens (2004)	Firgo (2007)
Selection Effect Only	Theorem 7 of This Paper*	Chernozhukov and Hansen (2005)
Essential Heterogeneity	$LATE$: Imbens and Angrist (1994) MTE : Heckman and Vytlacil (2005)	$LQTE$: Abadie et al. (2002) $MQTE$: This Paper

Table: Literature of the Average and Quantile Treatment Effect under Various Assumptions

Note: I only provide the most important or summary paper based on my personal judgement.

Note*: Part of this result is scattered in the literature such as Heckman and Robb (1985), but I did not notice the whole result stated explicitly anywhere else.

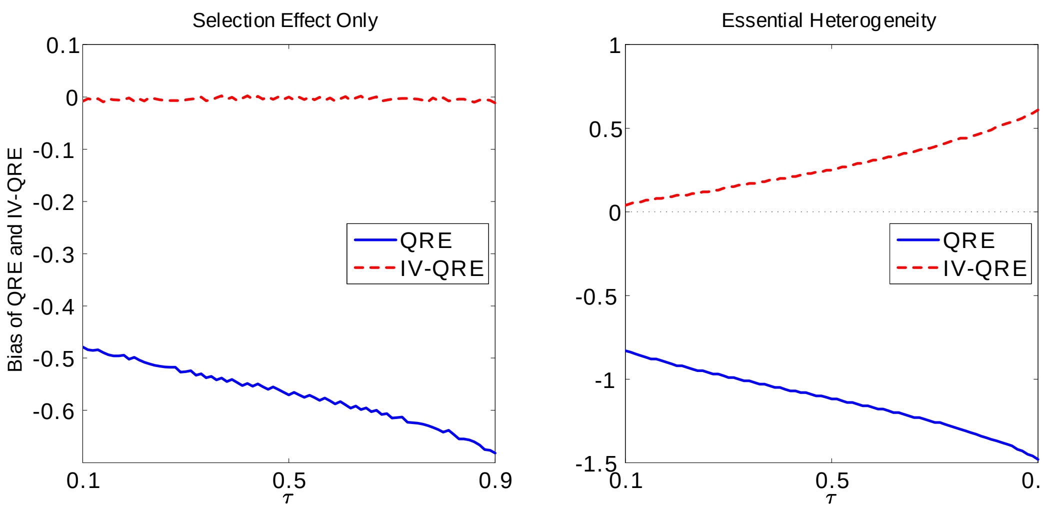

Figure 8: Asymptotic Biases of the QRE and IV-QRE When There is Only Selection Effect and When There is ALSO the Essential Heterogeneity

We close this subsection by a simple example to illustrate the bias of the QRE when the selection effect exists and the bias of the IV-QRE when the essential heterogeneity exists. Assume first only the selection effect exists, $Y_1 = 2U$, $Y_0 = U$ and $D = 1(Z - V > 0)$, where

$$\begin{pmatrix} U \\ V \\ Z \end{pmatrix} \sim N(0, \Sigma) \text{ with } \Sigma = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

When the essential heterogeneity also exists, suppose $Y_1 = V + 2U, Y_0 = 2V + U$, and all other specifications are the same as in the first specification. This specification is the same as that in Section 2.2. The QRE and IV-QRE are solutions to two groups of moment conditions which are developed in the supplementary materials. Figure 8 shows the (asymptotic) biases of the QRE and IV-QRE in the estimation of the QTE under the two specifications. As expected, the QRE is inconsistent as long as D is endogenous, and the IV-QRE is consistent only when the endogeneity comes solely from the selection effect.

6.2 Comments on Abadie et al. (2002)

Identification of $\Delta_\tau^{QTE}(x)$ requires $p_x^{\sup} = 1$ and $p_x^{\inf} = 0$ as shown in Section 3. This hardly holds in practice especially when Z is discrete. A popular alternative is the LQTE of Abadie et al. (2002), which measures the quantile treatment effect only on the compliers.[35] Specifically, pick $z, z' \in \mathcal{Z}_x$; then the LQTE measures the quantile treatment effect for those who change their participation status in response to the change in policy Z from z to z':

$$\Delta_\tau^{LQTE}(x,z,z') = F^{-1}_{Y_1|X,D_z,D_{z'}}(\tau|x,0,1) - F^{-1}_{Y_0|X,D_z,D_{z'}}(\tau|x,0,1), \tag{25}$$

where

$$\begin{aligned} F_{Y_d|X,D_z,D_{z'}}(y_d|x,0,1) &= E\left[1(Y_d \le y_d)|X=x, D_z=0, D_{z'}=1\right] \\ &= E\left[1(Y_d \le y_d)|X=x, p(x,z) < U_D \le p(x,z')\right] \\ &= \tfrac{1}{p(x,z')-p(x,z)} \int_{p(x,z)}^{p(x,z')} F_{Y_d|X,U_D}(y_d|x,u_D)du_D, \end{aligned}$$

and $D_z = 1(p(X,z) - U_D \ge 0)$. When $p(x,z) = 0$ and $p(x,z') = 1$, $\Delta_\tau^{LQTE}(x,z,z') = \Delta_\tau^{QTE}(x)$. Note that $F^{-1}_{Y_1|X,D_z,D_{z'}}(y_1|x,0,1)$ may not be parallel to $F^{-1}_{Y_0|X,D_z,D_{z'}}(y_0|x,0,1)$ as a function of x. However, when $\mathcal{Z}_x = \{0,1\}$, Abadie et al. (2002) assumes[36]

Assumption 3.1: For $\tau \in (0,1)$, there exist $\alpha_\tau \in \mathbb{R}, \beta_\tau \in \mathbb{R}^{d_x}$ such that

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

Assumption 3.1 implicitly assumes rank preservation on the compliers. Their other assumptions as stated in their Assumption 2.1 are parallel to our assumption (A1)-(A5). Different from Chernozhukov and Hansen (2005), Assumption 3.1 is conditioned on D. However, it is easy to see that this assumption implies $P(Y \le \alpha_\tau D + X'\beta_\tau|X, D_1 > D_0) = \tau$.[37] So this assumption is comparable to Theorem 1 of Chernozhukov and Hansen (2005), but as emphasized in the last subsection, Chernozhukov and Hansen (2005) does not consider the essential heterogeneity (and meanwhile consider all persons in the program).

We use (15) to illustrate the key point of Assumption 3.1. Under the normality assumption,

$$\begin{aligned} &F_Y(y|X=x, D=1, D_1 > D_0) = P\left(Y_1 \le y|X=x, V \le \phi + X'\gamma_1 + Z'\gamma_2, \phi + X'\gamma_1 < V \le \phi + X'\gamma_1 + \gamma_2\right) \\ &\quad = P\left(Y_1 \le y|X=x, Z=1, \phi + X'\gamma_1 < V \le \phi + X'\gamma_1 + \gamma_2\right) \\ &\quad = P\left(Y_1 \le y|X=x, \phi + X'\gamma_1 < V \le \phi + X'\gamma_1 + \gamma_2\right) \\ &\quad = \tfrac{1}{p(x,1)-p(x,0)} \int_{p(x,0)}^{p(x,1)} \Phi\left(\tfrac{y-\alpha-\varphi-x'\beta_1-\sigma_{V1}\Phi^{-1}(u_D)}{\sqrt{\sigma_1^2-\sigma_{V1}^2}}\right) du_D, \end{aligned} \tag{26}$$

and similarly,

$$F_Y(y|X=x, D=0, D_1 > D_0) = \frac{1}{p(x,1)-p(x,0)} \int_{p(x,0)}^{p(x,1)} \Phi\left(\frac{y-\alpha-x'\beta_0-\sigma_{V0}\Phi^{-1}(u_D)}{\sqrt{\sigma_0^2-\sigma_{V0}^2}}\right) du_D.$$

[35] It is commonly believed that the LATE is only useful for evaluating the effects of policies in place, but not for forecasting those of new policies. Nevertheless, see Section 6 of Imbens (2010) and Angrist and Fernández-Val (2013) for arguments favorable to the external validity of the LATE estimator.

[36] See also Abadie (2003) for a similar assumption in local average response function estimation.

[37] $P(Y \le \alpha_\tau D + X'\beta_\tau|X, D_1 > D_0) = P(Y \le \alpha_\tau D + X'\beta_\tau|X, D=1, D_1 > D_0)P(D = 1|X, D_1 > D_0) + P(Y \le \alpha_\tau D + X'\beta_\tau|X, D=0, D_1 > D_0)P(D = 0|X, D_1 > D_0) = \tau\left(P(D=1|X, D_1 > D_0) + P(D=0|X, D_1 > D_0)\right) = \tau$. However, $P(Y \le \alpha_\tau D + X'\beta_\tau|X, D_1 > D_0) = \tau$ does not imply $Q_\tau(Y|X,D,D_1 > D_0) = \alpha_\tau D + X'\beta_\tau$ without further assumptions. This is also why Theorem 2 of Chernozhukov and Hansen (2005) is required to identify the QTE when the conditioning set is (X,Z) instead of $(X, D_1 > D_0)$.

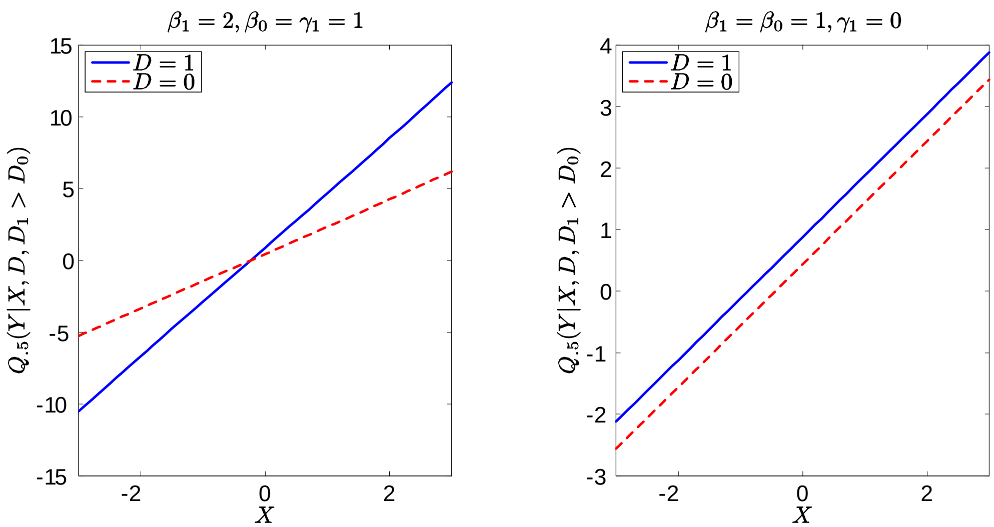

Figure 9: $Q_{.5}(Y|X,D,D_1 > D_0)$ As a Function of D and X: $\sigma_1 = 2$, $\sigma_0 = 1$, $\gamma_2 = 1$

To further simplify our discussion, assume $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_0 & \sigma_1 \\ \sigma_1\sigma_0 & \sigma_0^2 & \sigma_0 \\ \sigma_1 & \sigma_0 & 1 \end{pmatrix}$; in other words, $U_1 = \sigma_1\varepsilon$, $U_0 = \sigma_0\varepsilon$ and $V = \varepsilon$ with $\varepsilon \sim N(0,1)$. In this simple setup, the RP assumption trivially holds. Also, $F_Y(y|X = x, D = d, D_1 > D_0)$ is simplified as

$$\begin{aligned} F_Y(y|X,D=1,D_1>D_0) &= \frac{\max\left\{0, \left[\Phi\left(\min\left\{\frac{y-\alpha-\varphi-X'\beta_1}{\sigma_1}, \phi+X'\gamma_1+\gamma_2\right\}\right) - \Phi\left(\phi+X'\gamma_1\right)\right]\right\}}{\Phi(\phi+X'\gamma_1+\gamma_2)-\Phi(\phi+X'\gamma_1)}, \\ F_Y(y|X,D=0,D_1>D_0) &= \frac{\max\left\{0, \left[\Phi\left(\min\left\{\frac{y-\alpha-X'\beta_0}{\sigma_0}, \phi+X'\gamma_1+\gamma_2\right\}\right) - \Phi\left(\phi+X'\gamma_1\right)\right]\right\}}{\Phi(\phi+X'\gamma_1+\gamma_2)-\Phi(\phi+X'\gamma_1)}. \end{aligned}$$

Assuming the min and max operators do not apply in $F_Y(y|X,D,D_1 > D_0)$, we solve $F_Y(q|X,D,D_1 > D_0) = \tau$ to have

$$\begin{aligned} Q_\tau(Y|X,D=1,D_1>D_0) &= \sigma_1\Phi^{-1}\left\{\tau\left[\Phi\left(\phi+X'\gamma_1+\gamma_2\right)-\Phi\left(\phi+X'\gamma_1\right)\right]+\Phi\left(\phi+X'\gamma_1\right)\right\}+\alpha+\varphi+X'\beta_1, \\ Q_\tau(Y|X,D=0,D_1>D_0) &= \sigma_0\Phi^{-1}\left\{\tau\left[\Phi\left(\phi+X'\gamma_1+\gamma_2\right)-\Phi\left(\phi+X'\gamma_1\right)\right]+\Phi\left(\phi+X'\gamma_1\right)\right\}+\alpha+X'\beta_0. \end{aligned}$$

They are not the same as a function of X. The left graph of Figure 9 shows $Q_\tau(Y|X,D=1,D_1 > D_0)$ and $Q_\tau(Y|X,D=0,D_1 > D_0)$ as a function of X for the case with $\tau = 0.5$, $\sigma_1 = 2$, $\sigma_0 = 1$, $X \sim N(0,1)$, $\beta_1 = 2$, $\beta_0 = \gamma_1 = \gamma_2 = 1$, and $\alpha = \varphi = \phi = 0$. We take $X \in [-3,3]$ which covers most people under consideration. These two functions for other τ's are qualitatively similar, so omitted here. From Figure 9, it is quite clear that $Q_\tau(Y|X,D=1,D_1 > D_0)$ and $Q_\tau(Y|X,D=1,D_1 > D_0)$ are not parallel as a function of X. Anyway, there are special cases where Assumption 3.1 holds: (i) $\beta_1 = \beta_0$ and $\gamma_1 = 0$; (ii) $\gamma_2 = 0$, $\sigma_1\gamma_1 + \beta_1 = \sigma_0\gamma_1 + \beta_0$ (or $\beta_1 - \beta_0 = (\sigma_0 - \sigma_1)\gamma_1$). But these setups are too trivial to happen in practice since either $Y_1 - Y_0$ does not depend on X or Z does not affect D. For comparison, the corresponding $Q_\tau(Y|X,D=1,D_1 > D_0)$ and $Q_\tau(Y|X,D=1,D_1 > D_0)$ for $\tau = 0.5$, $\beta_1 = \beta_0 = 1$ are shown in the right graph of Figure 9.

The example above does not invalidate the analysis in Abadie et al. (2002) completely. As long as we refine their Assumption 3.1 as

$$Q_\tau(Y|X,D,D_1 > D_0) = X'\beta_\tau + D \cdot X'\delta_\tau \tag{27}$$

for $\tau \in (0,1)$, their estimation scheme can still go through with proper adjustment in notations. Note here that X includes a constant. Also, since X can include functions of the original covariates (e.g., polynomials or B-splines), the linear setup of the conditional quantile does not lose generality. Under this setup, $\Delta_\tau^{LQTE}(x,0,1) = x'\delta_\tau$, which depends on x.[38] By adjusting notations in their Theorem 3.1, we can derive the asymptotic distribution of $x'\widehat{\delta}_\tau$ for each $x \in \mathcal{X}$.[39] Of course, given $\widehat{\delta}_\tau$ we can test whether $\underline{\delta}_\tau$'s are equal to zero for a bunch of τ's, where $\underline{\delta}_\tau$ is δ_τ excluding the intercept. However, it is still hard to test $\underline{\delta}_\tau = 0$ for $\tau \in \mathcal{T}$ since the weak limit of $\widehat{\delta}_\tau$, $\tau \in \mathcal{T}$, is unavailable until now.

This misspecification problem also happens in Chernozhukov and Hansen (2006) who assume

$$P(Y \leq D\alpha_\tau + X'\beta_\tau | X, Z) = \tau.$$

In a typical application of their technique, Chernozhukov and Hansen (2004) study the effect of 401(k) participation on savings. Their Table 3 shows that this effect varies with the income level, so the interaction term $D \cdot income$ should be included as a regressor. One main reason for Chernozhukov and Hansen to use this form of conditional quantile is to circumvent the computation problem as noted in Abadie (1997). If we assume

$$P\left(Y \leq X'\beta_\tau + D \cdot X'\delta_\tau\right) | X, Z) = \tau,$$

then the *inverse quantile regression* algorithm of Chernozhukov and Hansen (2006) is not efficient especially when d_X is large. As an alternative, the algorithm of Chen and Pouzo (2012) is still applicable. When D is multi-valued, e.g., $D = S$ is the schooling level as in the application of Chernozhukov and Hansen (2006), it is better to assume $Q_\tau(Y|X,Z) = \alpha_\tau(S) + X'\beta_\tau(S)$ which takes the form of the varying coefficient model (VCM). Section 6 of Hansen et al. (2004) considers the VCM in estimating the return to schooling in a different context.

On the other hand, the distribution regression can be applied to estimate $\Delta_\tau^{LQTE}(x,0,1)$ straightforwardly. From Theorem 1,

$$F_{Y_1|X,D_0,D_1}(y_1|x,0,1) = \frac{P(Y \leq y_1 | X=x, p(X,Z)=p(x,1), D=1)p(x,1) - P(Y \leq y_1 | X=x, p(X,Z)=p(x,0), D=1)p(x,0)}{p(x,1)-p(x,0)},$$

and similarly,

$$F_{Y_0|X,D_0,D_1}(y_0|x,0,1) = \frac{P(Y \leq y_0 | X=x, p(X,Z)=p(x,0), D=0)(1-p(x,0)) - P(Y \leq y_0 | X=x, p(X,Z)=p(x,1), D=0)(1-p(x,1))}{p(x,1)-p(x,0)},$$

so the estimates in (17) can be used to estimate $\Delta_\tau^{LQTE}(x,0,1)$. Specifically,

$$\widehat{F}_{Y_1|X,D_0,D_1}(y_1|x,0,1) = \frac{\widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}(x,1),1)\widehat{p}(x,1) - \widehat{F}_{Y|X,p(X,Z),D}(y_1|x,\widehat{p}(x,0),1)\widehat{p}(x,0)}{\widehat{p}(x,1)-\widehat{p}(x,0)},$$
$$\widehat{F}_{Y_0|X,D_0,D_1}(y_1|x,0,1) = \frac{\widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}(x,0),0)(1-\widehat{p}(x,0)) - \widehat{F}_{Y|X,p(X,Z),D}(y_0|x,\widehat{p}(x,1),0)(1-\widehat{p}(x,1))}{\widehat{p}(x,1)-\widehat{p}(x,0)},$$

and

$$\widehat{\Delta}_{X,z,z'}^{LQTE}(\tau|x,0,1) = \widehat{F}^{-1}_{Y_1|X,D_0,D_1}(\tau|x,0,1) - \widehat{F}^{-1}_{Y_0|X,D_0,D_1}(\tau|x,0,1),$$

where $\widehat{p}(x,z) = \widehat{p}(x,z;\widehat{\gamma}) = \Lambda\left(R(x,z)'\widehat{\gamma}\right)$ and $\widehat{F}_{Y|X,p(X,Z),D}(y|x,p,d) = \Lambda\left(T(x,p)'\,\widehat{\beta}_d(y)\right)$. This procedure

[38] The usual justification of Assumption 3.1 is that α_τ is close to $E[X'\delta_\tau]$; see, e.g., Angrist and Krueger (1999).

[39] To summarize the results, we can integrate the LQTE estimator by $n^{-1}\sum_{i=1}^n X_i'\widehat{\delta}_\tau = \overline{X}\widehat{\delta}_\tau$, which is estimating $E[X'\delta_\tau]$. This estimator is similar to the average derivative estimator in quantile regression; see Chaudhuri et al. (1997). But this estimator is not estimating the ILQTE. In general, $E[X'\delta_\tau] \neq Q_\tau(Y_1|D_1 > D_0) - Q_\tau(Y_0|D_1 > D_0) \equiv \Delta_\tau^{ILQTE}(0,1)$. Actually, $\Delta_\tau^{ILQTE}(0,1)$ is hard to estimate. Also note that when these parameters are used, the results in the application of Abadie et al. (2002) may change.

has two advantages over that of Abadie et al. (2002). First, we do not need to estimate the weights κ or κ_ν as in Abadie et al. (2002), where $\kappa(D,Z,X) = 1 - \frac{D\cdot(1-Z)}{1-P(Z=1|X)} - \frac{(1-D)\cdot Z}{P(Z=1|X)}$ and $\kappa_\nu(Y,D,Z) = E\left[\kappa|Y,D,X\right] = 1 - \frac{D\cdot(1-P(Z=1|Y,D,Z))}{1-P(Z=1|X)} - \frac{(1-D)\cdot P(Z|Y,D,Z)}{P(Z=1|X)}$; rather, only the propensity score is estimated in the first step. Second, our procedure can be applied to continuous, discrete or mixed Y without any special adjustment, while the procedure of Abadie et al. (2002) is only suitable to the continuous Y case (see Abadie (2003) for possible extensions to other data types).

The following theorem states the weak limit of $\widehat{\Delta}^{LQTE}_{X,z,z'}(\tau|x,0,1)$. To ease the statement of the theorem, define $\varphi^{L\prime}_{\gamma,\beta_0(\cdot)}(\eta,\alpha_0)(y_0,x) : \mathbb{R}^{d_\gamma}\times C(\mathcal{Y}_0)^{d_{\beta_0}} \longrightarrow \ell^\infty(\mathcal{Y}_0\mathcal{X})$ as

$$
\begin{aligned}
&\varphi^{L\prime}_{\gamma,\beta_0(\cdot)}(\eta,\alpha)\\
&= \frac{\left[1-p(x,0)\right]\left[F_{10}(y_0,x)-F_{00}(y_0,x)\right]-\left[p(x,1)-p(x,0)\right]\left[1-p(x,0)\right]\lambda_{00}(y_0,x)\frac{\partial T(x,p(x,0))'\beta_0(y_0)}{\partial p}}{\left[p(x,1)-p(x,0)\right]^2}\lambda^p_1(x)R(x,1)'\eta\\
&- \frac{\left[1-p(x,1)\right]\left[F_{10}(y_0,x)-F_{00}(y_0,x)\right]-\left[p(x,1)-p(x,0)\right]\left[1-p(x,0)\right]\lambda_{00}(y_0,x)\frac{\partial T(x,p(x,0))'\beta_0(y_0)}{\partial p}}{\left[p(x,1)-p(x,0)\right]^2}\lambda^p_0(x)R(x,0)'\eta\\
&+ \left[\frac{\left[1-p(x,0)\right]\lambda_{00}(y_0,x)}{p(x,1)-p(x,0)}T(x,p(x,0))' - \frac{\left[1-p(x,1)\right]\lambda_{10}(y_0,x)}{p(x,1)-p(x,0)}T(x,p(x,1)))'\right]\alpha_0(y_0)
\end{aligned}
$$

and $\varphi^{L\prime}_{\gamma,\beta_1(\cdot)}(\eta,\alpha_1)(y_1,x) : \mathbb{R}^{d_\gamma}\times C(\mathcal{Y}_1)^{d_{\beta_1}} \longrightarrow \ell^\infty(\mathcal{Y}_1\mathcal{X})$ as

$$
\begin{aligned}
&\varphi^{L\prime}_{\gamma,\beta_1(\cdot)}(\eta,\alpha)\\
&= \frac{\left[p(x,1)-p(x,0)\right]p(x,1)\lambda_{11}(y_1,x)\frac{\partial T(x,p(x,1))'\beta_1(y_1)}{\partial p} - p(x,0)\left[F_{11}(y_1,x)-F_{01}(y_1,x)\right]}{\left[p(x,1)-p(x,0)\right]^2}\lambda^p_1(x)R(x,1)'\eta\\
&- \frac{\left[p(x,1)-p(x,0)\right]p(x,0)\lambda_{01}(y_1,x)\frac{\partial T(x,p(x,0))'\beta_1(y_1)}{\partial p} - p(x,1)\left[F_{11}(y_1,x)-F_{01}(y_1,x)\right]}{\left[p(x,1)-p(x,0)\right]^2}\lambda^p_0(x)R(x,0)'\eta\\
&+ \left[\frac{p(x,1)\lambda_{11}(y_1,x)}{p(x,1)-p(x,0)}T(x,p(x,1))' - \frac{p(x,0)\lambda_{01}(y_1,x)}{p(x,1)-p(x,0)}T(x,p(x,0)))'\right]\alpha_1(y_1),
\end{aligned}
$$

where $F_{zd}(y,x) = F_{Y|X,p(X,Z),D}(y|x,p(x,z),d)$, $\lambda^p_z(x) = \lambda(R(x,z)'\gamma)$ and $\lambda_{zd}(y,x) = \lambda(T(x,p(x,z)))'\beta_d(y))$.

Theorem 8 *Suppose Assumption DR holds, and $F_{Y_d|X,D_0,D_1}(y|x,0,1)$ admits a positive continuous density $f_{Y_d|X,D_0,D_1}(y_1|x,0,1)$ on an interval $[a,b]$ containing an ϵ-enlargement of the set $\left\{Q_{Y_d|X,D_0,D_1}(\tau|x,0,1)|\tau\in\mathcal{T}\right\}$ for all $x\in\mathcal{X}$. Then*

$$
\begin{aligned}
&\sqrt{n}\left(\widehat{\Delta}^{LQTE}_{X,z,z'}(\tau|x,0,1) - \Delta^{LQTE}_{\tau}(x,0,1)\right)\\
&\rightsquigarrow -\left.\frac{\varphi^{L\prime}_{\gamma,\beta_1(\cdot)}\left(W_\gamma, J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right)}{f_{Y_1|X,D_0,D_1}(y_1|x,0,1)}\right|_{y_1=Q_{Y_1|X,D_0,D_1}(\tau|x,0,1)}\\
&+ \left.\frac{\varphi^{L\prime}_{\gamma,\beta_0(\cdot)}\left(W_\gamma, J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right)}{f_{Y_0|X,D_0,D_1}(y_0|x,0,1)}\right|_{y_0=Q_{Y_0|X,D_0,D_1}(\tau|x,0,1)} \quad \text{in } \ell^\infty(\mathcal{T}\mathcal{X}).
\end{aligned}
$$

The randomness in this weak limit is from three sources, $\widehat{p}$, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, separately, while for the MQTE, $\widehat{p}$ affects $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$ only through $\widehat{\beta}_0$ and $\widehat{\beta}_1$. From the proof of Theorem 8, it is easy to see that the exchangeable bootstrap is valid. As a result, similar inference procedures as in Section 4.4 can be applied so omitted here.

6.3 Comments on Aakvik et al. (2005)

We first consider a simple model to illustrate the limitation of the one-factor model in Aakvik et al. (2005). Suppose

$$\begin{aligned} Y_1^* &= X\beta_1 - U_1, Y_1 = 1(Y_1^* \geq 0), \\ Y_0^* &= X\beta_0 - U_0, Y_0 = 1(Y_1^* \geq 0), \\ D^* &= Z\gamma - V, D = 1(D^* \geq 0), \end{aligned}$$

with $U_1 = \sigma_1(X)\,[-\alpha_1\theta + \varepsilon_1]$, $U_0 = \sigma_0(X)\,[-\alpha_0\theta + \varepsilon_0]$ and $V = -\theta + \varepsilon_D$, where $\sigma_1(X) > 0$ and $\sigma_0(X) > 0$ control the heteroskedasticity, $(X, Z) \perp (\theta, \varepsilon_1, \varepsilon_0, \varepsilon_D)$, $\theta, \varepsilon_1, \varepsilon_0$ and ε_D and independent, and except θ, all others follow $N(0, 1)$. Theoretically,

$$E\left[Y_1 - Y_0 = 1 | X = x, V = v\right] = \frac{\int \Phi\left(\frac{x}{\sigma_1(x)}\beta_1 + \alpha_1\theta\right)\left(1 - \Phi\left(\frac{x}{\sigma_0(x)}\beta_0 + \alpha_0\theta\right)\right)\phi\left(v + \theta\right)dF\left(\theta\right)}{\phi\left(v/\sqrt{2}\right)},$$

where $F(\theta)$ is the CDF of θ. Under the RP assumption,

$$\begin{aligned} E\left[Y_1 - Y_0 = 1 | X = x, V = v\right] &= E[Y_1 = 1 | X = x, V = v] - E[Y_0 = 1 | X = x, V = v] \\ &= \frac{\int \left[\Phi\left(\frac{x}{\sigma_1(x)}\beta_1 + \alpha_1\theta\right) - \Phi\left(\frac{x}{\sigma_0(x)}\beta_0 + \alpha_0\theta\right)\right]\phi\left(v + \theta\right)dF\left(\theta\right)}{\phi\left(v/\sqrt{2}\right)}, \end{aligned}$$

which is smaller than the true value, i.e., the RP assumption does not hold in this example. If we use the model in Aakvik et al. (2005),

$$E\left[Y_1 - Y_0 = 1 | X = x, V = v\right] = \frac{\int \Phi\left(x\beta_1^* + \alpha_1^*\theta\right)\left(1 - \Phi\left(x\beta_0^* + \alpha_0^*\theta\right)\right)\phi\left(v + \theta\right)dF\left(\theta\right)}{\phi\left(v/\sqrt{2}\right)},$$

where $(\gamma^*, \beta_0^*, \beta_1^*)$ and (α_0^*, α_1^*) are the pseudo-true value of $(\gamma, \beta_0, \beta_1)$ and (α_0, α_1), defined as the maximizer of

$$E\left[\ln \int P\left(D, Y | X, Z, \theta\right)dF\left(\theta\right)\right].$$

Here,

$$P\left(D, Y | X, Z, \theta\right) = P(D|Z, \theta)P(Y|D, X, \theta),$$

and

$$\begin{aligned} P(D = 1|Z, \theta) &= \Phi\left(Z\gamma + \theta\right), \\ P(Y = 1|D = 1, X, \theta) &= \Phi\left(X\beta_1 + \alpha_1\theta\right), \\ P(Y = 1|D = 0, X, \theta) &= \Phi\left(X\beta_0 + \alpha_0\theta\right). \end{aligned}$$

Note that D and Y in the expectation of the likelihood function follow the distribution in the true model. To simplify the numerical integrations in the likelihood function, we assume $\alpha_1 = \beta_1 = 2$, $\alpha_0 = \beta_0 = \gamma = 1$, $\sigma_1(X) = 1 + X$, $\sigma_0(X) = 1 + 0.5X$, X and Z are independent and both follow the uniform distribution on two points 0 and 1, and θ can take only three values -1, 0 and 1 with $P(\theta = -1) = P(\theta = 1) = 1/4$

and $P(\theta = 0) = 1/2$.[40] It turns out $(\alpha_0^*, \alpha_1^*, \beta_0^*, \beta_1^*, \gamma^*) = (334.15, 10.03, 176.57, 0.72, 0.85)$, far from the true value $(1, 2, 1, 2, 1)$.

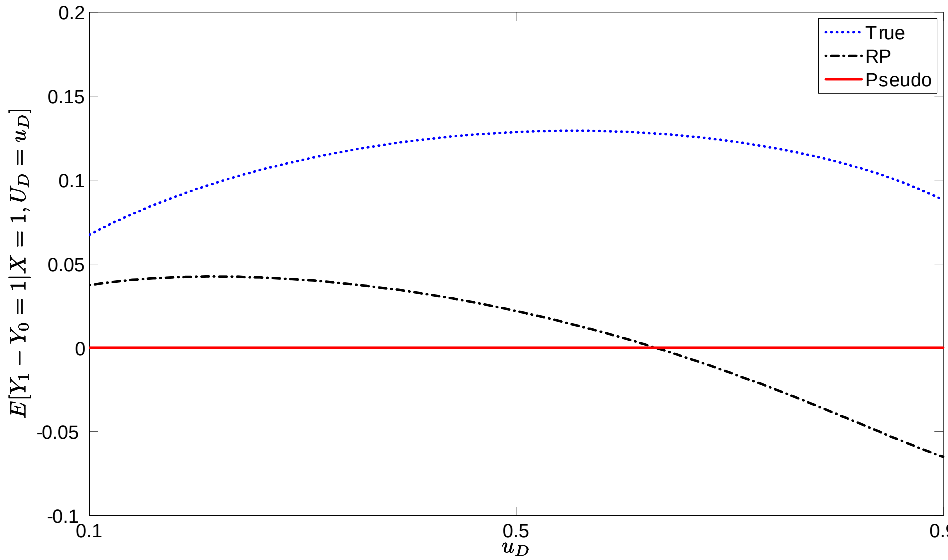

Figure 10: Comparison of $E\,[Y_1 - Y_0 = 1|X = 1, U_D = u_D]$ Under Different Specifications

Figure 10 shows the true $E\,[Y_1 - Y_0 = 1|X = x, V = v]$, $E\,[Y_1 - Y_0 = 1|X = x, V = v]$ under the RP assumption, and $E\,[Y_1 - Y_0 = 1|X = x, V = v]$ with pseudo-true parameter values as a function of $u_D = \Phi(v/\sqrt{2})$ when $x = 1$. As expected, as long as there is misspecification, $E\,[Y_1 - Y_0 = 1|X = x, U_D = u_D]$ cannot be estimated consistently. $E\,[Y_1 - Y_0 = 1|X = 1, U_D = u_D]$ with pseudo-true parameter values may or may not have a smaller bias than $E\,[Y_1 - Y_0 = 1|X = 1, U_D = u_D]$ under the RP assumption, depending on the value of u_D. A striking feature of the one-factor model of Aakvik et al. (2005) in this example is that there is no treatment effect for any $U_D = u_D$, which contrasts the truth.

To compare our model with that of Aakvik et al. (2005), we put them side by side for comparison:

Aakvik et al. (2005):	Our Model:
$Y_d = \mu(\varphi_d(X) + \alpha_d\theta - \varepsilon_d)$	$Y_d = \mu_d(X, V, U_d)$
$D = 1(\mu_D(X, Z) + \theta - \varepsilon_D \geq 0)$	$D = 1(\mu_D(X, Z) - V \geq 0)$ with $Z \perp V\|X$
$(X, Z) \perp (\theta, \varepsilon_0, \varepsilon_0, \varepsilon_D)$ and $(\theta, \varepsilon_0, \varepsilon_0, \varepsilon_D)$ are independent	$U_0\|\,(X, V) = U_1\|\,(X, V)$ and $Z \perp (U_0, U_1)\|X$

where we use a prototype of Aakvik et al. (2005) and its extension. Similar to our setup, their outcome equation also includes two random errors. However, their (Y_0, Y_1) include totally three random errors, $(\theta, \varepsilon_0, \varepsilon_1)$, while our potential outcomes essentially include only two random errors (V, U) with $U|\,(X, V) \equiv U_d|\,(X, V)$. Also, their choice equation includes also two random errors, (θ, ε_D), while ours includes only one, V. On the other hand, our specification of outcome equations is more general in the sense that it allows for heteroskedasticity and nonseparable errors. Also, we do not restrict the dependence between X and (V, U)

[40] The technique used in the assumption of the distribution of θ, i.e., approximating a continuous distribution by a discrete distribution, has been applied to econometric models for duration data by Heckman and Singer (1984).

and among (V, U), while they assume independence between the covariates and the random errors and also among the random errors. To further understand their independence assumption, we represent their model with $\theta, \varepsilon_1, \varepsilon_0$ and ε_D being independent and following $N(0,1)$ as

$$Y_d = \mu(\varphi_d(X) - \alpha_d V - U_d),$$
$$D = 1(\mu_D(X, Z) - V \geq 0),$$

where $U_d \sim N\left(0, 1+\alpha_d^2\right)$, and $|Corr(U_1, U_0)| = \frac{|\alpha_0\alpha_1|}{\sqrt{(1+\alpha_0^2)(1+\alpha_1^2)}} < 1$, $(U_1, U_0, V) \perp (X, Z)$, and $(U_1, U_0) \perp V$. Different from our model, their ranks are only partially maintained since $|Corr(U_1, U_0)| < 1$. Meanwhile, their ranks do not depend on X and V given that $(U_0, U_1) \perp (X, V)$, which implies that conditional ranks and unconditional ranks are the same. Our unconditional ranks are different from conditional ranks; the former are also only partially maintained, while the latter are fully preserved. In summary, these two kinds of models are more complements than substitutes.

An advantage of the factor model over the model with the RP assumption is that it allows for nondegenerate $P(Y_1 > Y_0|Y_0 = y_0, X = x, U_D = u_D)$. In the simple example above,

$$\begin{aligned} P(Y_1 > Y_0|Y_0 = 0, X = x, U_D = u_D) &= P\left(Y_1 = 1|Y_0 = 0, X = x, U_D = u_D\right) \\ &= \frac{\int \Phi\left(\frac{x}{\sigma_1(x)}\beta_1 + \alpha_1\theta\right)\left(1 - \Phi\left(\frac{x}{\sigma_0(x)}\beta_0 + \alpha_0\theta\right)\right)\phi\left(v+\theta\right)dF\left(\theta\right)}{\int\left(1 - \Phi\left(\frac{x}{\sigma_0(x)}\beta_0 + \alpha_0\theta\right)\right)\phi\left(v+\theta\right)dF\left(\theta\right)} \in (0,1)\,. \end{aligned}$$

Nevertheless, merits of the factor model should be determined by how far it is from the true model and how convenient it is to be implemented in practice. A similarity of the factor model with our model is that identification-at-infinity is required to identify various treatment parameters.

7 Application

We use the data of Angrist and Krueger (1991) to illustrate some main points of this paper. Angrist and Krueger (1991) estimate schooling coefficients using quarter of birth as instrument in a sample of 329509 men born 1930-39 from the 1980 census. Quarter of birth is correlated with educational attainment because of a mechanical interaction between compulsory school attendance laws and age at school entry. See the appendix to Angrist and Krueger (1991) for a detailed description of the data. This data set is widely used for various purposes. For example, Bound et al. (1995) use it to illusrate the bias of the IV estimates when the instruments are weak; Chesher (2005) uses it to show the difficulty in estimating identifying nonparametric intervals when the instruments are weak; van der Klaauw (2002) uses it to show that the quarter of birth is not a valid instrument in the conventional sense (independence of the error term in the outcome equation) while can still be used to estimate some treatment effects in the regression discontinuity design framework; Chernozhukov and Hansen (2006) use it to estimate the quantile treatment effects defined and identified in Chernozhukov and Hansen (2005).

Our conditional distribution of the returns to schooling is specified as

$$P(Y \leq y|X, p(X,Z) = p, D = d) = \Lambda\left(\left(p, p^2\right)\alpha_d(y) + X'\beta_d(y)\right),$$ [41]

[41] We also tried cubic polynomials of p, and the results are qualitatively similar. Note here that although we use saturated specification for X given that X is discrete, the specificaiton is not fully saturated to both X and p. The fully saturated model should include the interaction terms of p and X. We neglect such interaction terms because $\mathcal{P}_x$ includes only four points and does not include much variation given that the instruments are relatively weak. Our specification implicitly assumes that

and the treatment status is determined by

$$D = 1(V \leq X'\gamma_1 + Z'\gamma_2),$$

where Y is the log weekly wage, $D = 1(S > 12)$ is the indicator of a high school graduate, and following Angrist and Krueger (1991), X is a vector of covariates consisting of a constant, state and year of birth fixed effects, and Z includes three dummies for the first through third quarter of birth.[42] We specify $\Lambda(\cdot)$ as the CDF of the standard normal and V following a standard normal distribution. Different from Chernozhukov and Hansen (2006), D is binary rather than the years of schooling S to fit in the framework of this paper. Also, we use dummies for the three quarters of birth rather than both the linear projection of S onto X and the three dummies as instruments.

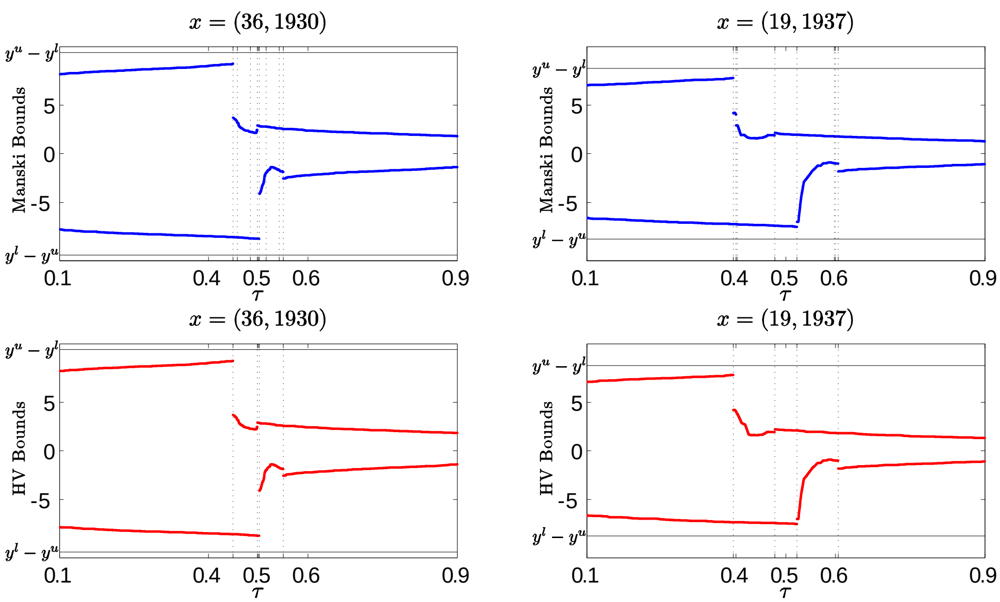

Figure 11: Manski and HV Bounds for $\Delta_\tau^{QTE}(x)$: $x = (36, 1930)$ and $(19, 37)$, $\tau \in [0.1, 0.9]$

We first show the bounds in Section 3 for this data set. Since X and Z are discrete, $p(X, Z)$ can be obtained by its sample analog. Totally, the support of X, $\mathcal{X}$, includes 510 possible values. The minimum number of observations among all $x \in \mathcal{X}$ is 3 and the maximum number is 3203. Averagely, there are 646 data points for each x. For some values of $x \in \mathcal{X}$, not every value of Z is possible. It is rare among all values of $x \in \mathcal{X}$ that $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$, so point identification of $\Delta_\tau^{QTE}(x)$ is almost impossible. There are actually only four x's satisfying $p_x^{\inf} = 0$ and $p_x^{\sup} = 1$, but at most eight data points are available at these x's. $y^l(x)$ and $y^u(x)$ are taken as the minimum and maximum of Y given $X = x$, and the population quantiles in the bounds are estimated by the sample quantiles. Given that $\mathcal{X}$ is large, we only present the bounds at two x's for a taste. Figure 11 shows the bounds with and without Assumption (9) for individuals from state 36 and born in 1930 and individuals from state 19 and born in 1937. These two x's correspond to cases with most observations (3203 data points) and moderate number of observations (666 data points),

$(X, Z) \perp (U_1, U_0, V)$, so the MQTE can be identified over the marginal support of $p(X, Z)$ instead of the conditional support of $p(X, Z)$ given X. Such a strategy is also used in Carneiro et al. (2011).

[42] The dummy for quarter four is omitted to avoid multicollinearity since X includes a constant.

respectively. The bounds in (11) are labeled as "Manski Bounds" and those in (13) are labeled as "HV Bounds". The vertical lines in the upper two figures indicate four possible values of $p(x, z)$ (and $1-p(x, z)$), and the vertical lines in the lower two figures indicate $p_x^{\inf}$, $p_x^{\sup}$, $1-p_x^{\inf}$ and $1-p_x^{\sup}$. From Figure 11, we can draw a few interesting conclusions. First, both Manski bounds and HV bounds are narrower than the trivial bounds $[y^l(x)-y^u(x), y^u(x)-y^l(x)]$ especially when $\tau>0.5$, which indicates that these bounds are indeed informative. Second, $\Delta_\tau^{QTE}(x)=0$ cannot be rejected for all $\tau \in [0.1, 0.9]$. In other words, these bounds cannot determine whether there are quantile treatment effects or not at these two x values. Third, the Manski bounds and the HV bounds are very similar, which indicates that assumption (9) is most likely to hold. Fourth, when $x=(19,1937)$ and τ between 0.4 and 0.45, the Manski upper bound is indeed lower than the HV upper bound. Fifth, for both x values, the range of $p(x, z)$ is quite limited, e.g, between 0.4 and 0.5, which is a sign that Z is weak as already emphasized in the literature.

We now estimate the MQTE for $x=(19,1937)$. The results for $x=(36,1930)$ are qualitatively similar so are not reported here. The bootstrap confidence bands are not calculated since they are too time-consuming. The range of $\widehat{p}_i$ is $[0.2149, 0.6054]$. Out of this range, the estimation shows some abnormality and is not reliable. $\widehat{F}_{Y_d|X,U_D}(y|x,u_D)$ for $u_D=0.3$ and 0.5 and the MQTEs for $\tau=0.1, 0.25, 0.5, 0.75$ and 0.9 when $u_D \in [0.2, 0.6]$ are shown in Figure 12. From Figure 12, a few results of interest are as follows. First, as expected, $\widehat{F}_{Y_1|X,U_D}(y|x,u_D)$ stochastically dominates $\widehat{F}_{Y_0|X,U_D}(y|x,u_D)$ for the two u_D values. Second, as expected, $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$ is a decreasing function of u_D for all τ's, which implies that for individuals with any level of income, those who are reluctant to attend college have lower returns to college. Third, the return to college is decreasing in τ and becomes stable when τ gets large for all values of u_D. Chernozhukov and Hansen (2006) interpret τ as an ability index, so interpret this result as that people with high ability will generate high earnings regardless of their education level, while those with lower ability gain more from the training provided by formal education. Fourth, the variation of the return to college gets larger when u_D gets larger. Fifth, for individuals with any level of income, the counterfactual income when $D=0$ is increasing in u_D. This means that when $D=0$, individuals with relatively high income tend not to attend the college. The counterfactual income when $D=1$ is different. For high-income individuals, the counterfactual income of attending college is decreasing in u_D, while the low-income individuals have an increasing counterfactual income in u_D. This means that for high-income individuals, they attend college because they can potentially have a higher income (check $\widehat{F}^{-1}_{Y_0|X,U_D}(\tau|x,u_D)$ and $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$), while for low-income individuals, they attend college not because their potential income of attending college would be higher but because their potential income of not attending college would be even lower. Dependence of $\widehat{F}^{-1}_{Y_d|X,U_D}(\tau|x,u_D)$ on u_D indicates the invalidity of unconfoundedness. Sixth, the variation of potential income when $D=0$ is increasing in u_D, while the variation when $D=1$ is decreasing. This means that those who are eager to attend college have much more chances if attending college than not, while the converse result is more suitable to those who are reluctant to attend college. All these results are intuitively understandable in reality.

We next analyze the RMPRQTE between the constant shifts $p_\alpha=p+\alpha$ (Policy I) and the proportional shifts $p_\alpha=p(1+\alpha)$ (Policy II) when $x=(19,1937)$. We will not check the policy effect of the marginal change in Z since it is discrete. $\widehat{\text{RMPRQTE}}$ as a function of y is shown in the upper right panel of Figure 12, where $\mathcal{Y}=\mathcal{Y}_0 \cap \mathcal{Y}_1$ with $\mathcal{Y}_d=\left[\widehat{Q}_{Y_d}(0.01), \widehat{Q}_{Y_d}(0.99)\right]$ and $\widehat{Q}_{Y_d}(\tau)$ being the τth sample quantile of Y_d. As mentioned above, $p(X,Z)|X$ can take only four values, which makes the estimation of $f_{p|X}(u_D|x)$ in the weight function $\omega(u_D|x)$ impossible. To avoid this problem, we assume the distribution of $p(X,Z)|X$ does not depend on X, so the marginal density $f_p(u_D)$ is used in the weight function. This assumption is roughly valid since the quarter of birth seems uncorrelated with the state or year of birth. In our estimation, $f_p(u_D)$ is approximated by 20 discrete point masses. From Figure 12, we can see that Policy I has a larger effect than Policy II for most income levels. Note that both the numerator and denominator of $\widehat{\text{RMPRQTE}}$ are negative. This is because as education expands, more people are enrolled in college. These marginally enrolled people are less eager to attend college since the return is relatively low. This is somewhat like the scenario that increases in college enrollment will deteriorate the quality of college graduate and make the average wage level decrease.

We finally estimate the transition matrix defined in Section 5.1. It can be shown that the transition matrix equals

$$\begin{pmatrix}
\boxed{0.930} & 0.074 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.088 & \boxed{0.707} & 0.142 & 0.052 & 0.009 & 0.002 & 0 & 0 & 0 & 0 \\
0 & 0.191 & \boxed{0.324} & 0.166 & 0.084 & 0.092 & 0.086 & 0.041 & 0.014 & 0.003 \\
0 & 0.048 & \boxed{0.311} & \mathbf{0.226} & 0.077 & 0.053 & 0.062 & 0.076 & 0.078 & 0.073 \\
0 & 0.012 & 0.088 & \boxed{0.311} & \mathbf{0.193} & 0.119 & 0.072 & 0.052 & 0.063 & 0.107 \\
0 & 0 & 0.036 & 0.076 & \boxed{0.230} & \mathbf{0.191} & 0.171 & 0.086 & 0.034 & 0.162 \\
0 & 0 & 0.033 & 0.039 & 0.195 & \boxed{0.208} & \mathbf{0.162} & 0.131 & 0.046 & 0.187 \\
0 & 0 & 0.019 & 0.117 & 0.187 & \boxed{0.225} & 0.203 & \mathbf{0.149} & 0.079 & 0.097 \\
0 & 0 & 0.001 & 0.015 & 0.028 & 0.100 & 0.230 & \boxed{0.376} & \mathbf{0.156} & 0.023 \\
0 & 0 & 0 & 0 & 0 & 0.012 & 0.016 & 0.093 & \boxed{0.532} & \mathbf{0.350}
\end{pmatrix}$$

So unconditionally, the RP assumption does not hold since this transition matrix is far from the identify matrix although there is a strong positive dependence between the two potential outcomes. The transition matrix also shows that the independence assumption across counterfactual outcomes, which is the Veil of Ignorance assumption used in applied welfare theory (see, e.g., Sen (1997)) or in aggregate income inequality decomposition (see, e.g., DiNardo et al. (1996)), does not hold either. Another interesting phenomenon is that the diagonal element is not the largest in seven out of ten rows and many individuals from the low-income stratum may jump to the high-income stratum if attending college. This means that education indeed changes people's income strata.

8 Conclusion

This paper studies identification, estimation and inference of quantile treatment effects which are useful in economic policy analysis. We use the MQTE to unify the literature and organize the whole paper. The contributions of this paper can be summarized in four aspects. First, we clarify some key concepts in quantile treatment effect evaluation. For example, what is the meaning of the rank preservation assumption and why is it important? what is the difference between quantile of difference and difference of quantiles? what is the relationship between the MQTE and other parameters of quantile treatment effects? Second, we estimate various quantile treatment effects based on the distribution regression, derive their weak limits and show the validity of the bootstrap inferences. These quantile treatment parameters include the MQTE, the QTE, the IQTE and the LQTE. Third, we conduct two counterfactual analyses, namely, deriving the transition matrix and developing the RMPRQTE parameter. These two tools are useful to remove the veil of ignorance in assessing the distributional impacts of social policies. Fourth, we develop sharp bounds for the QTE and provide sufficient and necessary conditions for point identification with and without the monotonicity assumption. These results are useful to clarify the difference between this paper (which considers both the selection effect and the essential heterogeneity) and the literature such as Chernozhukov and Hansen (2005) (who consider only the selection effect). For example, under the monotonicity assumption, point identification

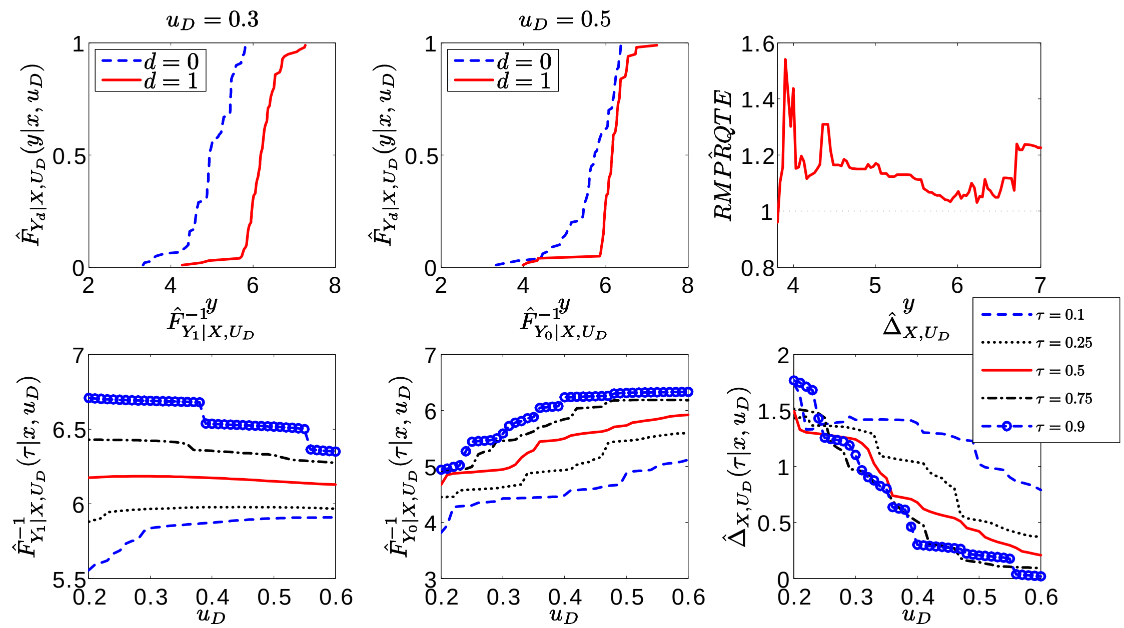

Figure 12: $\widehat{F}_{Y_d|X,U_D}(y|x,u_D)$, $\widehat{F}^{-1}_{Y_d|X,U_D}(\tau|x,u_D)$, $\widehat{\Delta}_{X,U_D}(\tau|x,u_D)$ and $\widehat{\text{RMPRQTE}}$: $x=(19,1937)$

of the QTE in the framework of Chernozhukov and Hansen (2005) only requires the instrument to have a nontrivial effect on the propensity score, while in our framework, point identification requires the effect of the instrument on the propensity score to be not only nontrivial but extremal.

There are many interesting problems which are unsolved in this paper and will be considered in our future research. First, our propensity score estimator is parametric or semiparametric, while a nonparametric estimator is also popular, see, e.g., Abadie (2003), Hirano et al. (2003), Carneiro and Lee (2009) and Cattaneo (2010) among others. When the second-step estimator is only finite-dimensional, Newey (1994), Chen et al. (2003) and Ichimura and Lee (2010) derive its asymptotic distribution and prove the bootstrap validity. However, when the second-step estimator is infinite-dimensional as in our case, its week limit and bootstrap validity involve nontrivial technical complications and will be pursued in a separate paper. Nevertheless, our semiparametric estimator (combined with the goodness-of-fit test in Section 4.5) is enough for any practical purpose. Second, our estimation scheme can be easily extended to multi-valued treatment case. We refer to Cattaneo (2010) for the relevant literature on this topic (especially under unconfoundedness), and Heckman and Vytlacil (2007b) and Heckman et al. (2006, 2008) for the MTE estimation. However, extension to the case with continuous treatment is not trivial; see Florens et al. (2008) for some identification results in the average treatment scenario. Third, a formal solution to the testing and inference problems at the end of Section 3.3 is desirable. Finally, a key assumption in this paper is the RP condition (A6). When this assumption does not hold, even the MQTE cannot be point identified. There are two responses to the relaxation of the RP assumption. First, we can test whether the RP assumption holds. Second, we can construct bounds for the MQTE (and related parameters such as the QTE and IQTE) and conduct inferences on the identified sets; see Kitagawa (2009), Fan and Zhu (2009), Fan and Park (2009, 2010, 2012), Fan and Wu (2010) and Kim (2013) for some related recent developments in different contexts.

References

Aakvik, A., J.J. Heckman and E. Vytlacil, 1999, Semiparametric Program Evaluation Lessons from an Evaluation of a Norwegian Training Program, mimeo, Department of Economics, University of Chicago.

Aakvik, A., J.J. Heckman and E. Vytlacil, 2005, Estimating Treatment Effects for Discrete Outcomes When Responses to Treatment Vary: An Application to Norweigian Vocational Rehabilitation Programs, *Journal of Econometrics*, 125, 15-51.

Abadie, A., 1997, Changes in Spanish Labor Income Structure During the 1980's: A Quantile Regression Approach, *Investigaciones Economicas*, 21, 253-272.

Abadie, A., 2002, Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models, *Journal of the American Statistical Association*, 97, 284-292.

Abadie, A., 2003, Semiparametric Instrumental Variable Estimation of Treatment Response Models, *Journal of Econometrics*, 113, 231-263.

Abadie, A., J.D. Angrist and G.W. Imbens, 2002, Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings, *Econometrica*, 70, 91-117.

Abbring, J.H. and J.J. Heckman, 2007, Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier Science B.V., Ch. 72, 5145-5303.

Anderson, G., 1996, Nonparametric Tests of Stochastic Dominance in Income Distributions, *Econometrica*, 64, 1183-1193.

Andrews, D.W.K., 1988a, Chi-Square Diagnostic Tests for Econometric Models: Theory, *Econometrica*, 56, 1419-1453.

Andrews, D.W.K., 1988b, Chi-square Diagnostic Tests for Econometric Models: Introduction and Applications, *Journal of Econometrics*, 37, 135-156.

Andrews, D.W.K., 1997, A Conditional Kolmogorov Test, *Econometrica*, 65, 1097-1128.

Angrist, J.D. and I. Fernández-Val, 2013, ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework, *Advances in Economics and Ecnometrics*, Tenth World Congress, Vol III, D. Acemoglu, M. Arellano and E. Dekel, eds., Cambridge: Cambridge University Press.

Angrist, J.D., K. Graddy and G.W. Imbens, 2000, The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish, *Review of Economic Studies*, 67, 499-527.

Angrist, J.D. and A.B. Krueger, 1991, Does Compulsory School Attendance Affect Schooling and Earnings, *Quarterly Journal of Economics*, 106, 979-1014.

Angrist, J.D. and A.B. Krueger, 1999, Empirical Strategies in Labor Economics, *Handbook of Labor Economics*, Vol. 3A, O. Ashenfelter and D. Card, eds., Amsterdam: North-Holland, Ch. 23, 1277-1366.

Athey, S. and G.W. Imbens, 2006, Identification and Inference in Nonlinear Difference-In-Differences Models, *Econometrica*, 74, 431–497.

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

Barrett, G. and S. Donald, 2003, Consistent Tests for Stochastic Dominance, *Econometrica*, 71, 71-104.

Barrett, G. and S. Donald, 2009, Statistical Inference with Generalized Gini Indices of Inequality, Poverty, and Welfare, *Journal of Business and Economic Statistics*, 27, 1-17.

Bhattacharya, D., 2007, Inference on Inequality from Household Survey Data, *Journal of Econometrics*, 137, 674-707.

Bickel, P. and J.-J. Ren, 2001, The Bootstrap in Hypothesis Testings, *Lecture Notes - Monograph Series*, 36, 91-112.

Bitler, M.P., J.B. Gelbach and H.W. Hoynes, 2006, What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments, *American Economic Review*, 96, 988-1012.

Björklund, A. and R. Moffitt, 1987, The Estimation of Wage Gains and Welfare Gains in Self-Selection Models, *Review of Economics and Statistics*, 69, 42-49.

Bound, J., D.A. Jaeger and R.M. Baker, 1995, Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak, *Journal of the American Statistical Association*, 90, 443-450.

Card, D., 1996, The Effect of Unions on the Structure of Wages: A Longitudinal Analysis, *Econometrica*, 64, 957-979.

Carneiro, P., K.T. Hansen and J.J. Heckman, 2001, Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies, *Swedish Economic Policy Review*, 8, 273-301.

Carneiro, P., K.T. Hansen and J.J. Heckman, 2003, Estimating Distributions of Treatment Effects with An Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice, *International Economic Review*, 44, 361-422.

Carneiro, P., J.J. Heckman and E. Vytlacil, 2010, Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin, *Econometrica*, 78, 377-394.

Carneiro, P., J.J. Heckman and E. Vytlacil, 2011, Estimating Marginal Returns to Education, *American Economic Review*, 101, 2754-2781.

Carneiro, P. and S. Lee, 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, 191-209.

Cattaneo, M.D., 2010, Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability, *Journal of Econometrics*, 155, 138-154.

Chaudhuri, P., K. Doksum and A. Samarov, 1997, On Average Derivative Quantile Regression, *Annals of Statistics*, 25, 715-744.

Chen, X., 2007, Large Sample Sieve Estimation of Semi-Nonparametric Models, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Elsevier Science B.V., Ch. 76, 5549-5632.

Chen, X., O. Linton and I. van Keilegom, 2003, Estimation of Semiparametric Models When the Criterion Function is Not Smooth, *Econometrica*, 71, 1591-1608.

Chen, X. and D. Pouzo, 2012, Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals, *Econometrica*, 80, 277-321.

Chernozhukov, V. and C. Hansen, 2004, The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis, *Review of Economics and Statistics*, 86, 735-751.

Chernozhukov, V. and C. Hansen, 2005, An IV Model of Quantile Treatment Effects, *Econometrica*, 73, 245-261.

Chernozhukov, V. and C. Hansen, 2006, Instrumental Quantile Regression Inference for Structural and Treatment Effect Models, *Journal of Econometrics*, 132, 491-525.

Chernozhukov, V. and C. Hansen, 2013, Quantile Models with Endogeneity, *Annual Review of Economics*, 5, 57-81.

Chernozhukov, V., I. Fernández-Val and A. Galichon, 2010, Quantile and Probability Curves without Crossing, *Econometrica*, 78, 1093-1125.

Chernozhukov, V., I. Fernández-Val and B. Melly, 2013, Inference on Counterfactual Distributions, *Econometrica*, 81, 2205–2268.

Chernozhukov, V., G.W. Imbens and W.K. Newey, 2007, Instrumental Variable Estimation of Nonseparable Models, *Journal of Econometrics*, 139, 4-14.

Chernozhukov, V., S. Lee and A. Rosen, 2013, Intersection Bounds: Estimation and Inference, *Econometrica*, 667-737.

Chesher, A., 2003, Identification in Nonseparable Models, *Econometrica*, 71, 1405-1441.

Chesher, A., 2005, Nonparametric Identification Under Discrete Variation, *Econometrica*, 73, 1525-1550.

Cox, D.R., 1972, Regression Models and Life Tables (with discussion), *Journal of the Royal Statistical Society*, Series B, 34, 187-220.

Das, M., W.K. Newey and F. Vella, 2003, Nonparametric Estimation of Sample Selection Models, *Review of Economic Studies*, 70, 33-58.

Davidson, R. and J. Duclos, 2000, Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality, *Econometrica*, 68, 1435-1464.

DiNardo, J., N.M. Fortin and T. Lemieux, 1996, Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach, *Econometrica*, 64, 1001-1044.

Doksum, K., 1974, Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case, *Annals of Statistics*, 2, 267-277.

Durbin, J., 1973, *Distribution Theory for Tests Based on the Sample Distribution Function*, Philadelphia: SIAM.

Fan, J. and I. Gijbels, 1996, *Local Polynomial Modelling and its Applications*, London: Chapman & Hall.

Fan, Y. and S.S. Park, 2009, Partial Identification of The Distribution of Treatment Effects and Its Confidence Sets, *Advances in Econometrics*, 25, 3–70.

Fan, Y. and S.S. Park, 2010, Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference, *Econometric Theory*, 26, 931-951.

Fan, Y. and S.S. Park, 2012, Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments, *Journal of Econometrics*, 167, 330-344.

Fan, Y. and J. Wu, 2010, Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and its Confidence Sets, *Review of Economic Studies*, 77, 1002-1041.

Fan, Y. and D. Zhu, 2009, Partial Identification and Confidence Sets for Functionals of the Joint Distribution of Potential Outcomes, mimeo, Vanderbilt University.

Firpo, S., 2007, Efficient Semiparametric Estimation of Quantile Treatment Effects, *Econometrica*, 259-276.

Florens, J.P., J.J. Heckman, C. Meghir and E.J. Vytlacil, 2008, Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects, *Econometrica*, 76, 1191-1206.

Foresi, S. and F. Peracchi, 1995, The Conditional Distribution of Excess Returns: An Empirical Analysis, *Journal of the American Statistical Association*, 90, 451-466.

Fréchet, M., 1951, Sur Les Tableaux de Corrélation Dont Les Marges Sont Données, *Annals University Lyon*, Series A, 14, 53-77.

Gagliardini, P. and O. Scaillet, 2012, Nonparametric Instrumental Variable Estimation of Structural Quantile Effects, *Econometrica*, 80, 1533-1562.

Hahn, J., 1998, On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects, *Econometrica*, 66, 315-331.

Hansen, K.T., J.J. Heckman and K.J. Mullen, 2004, The Effect of Schooling and Ability on Achievement Test Scores, *Journal of Econometrics*, 121, 39-98.

Heckman, J.J., 1984, The χ^2 Goodness of Fit Statistic for Models with Parameters Estimated from Microdata, *Econometrica*, 52, 1543-1547.

Heckman, J.J., 1992, Randomization and Social Program Evaluation, *Evaluating Welfare and Training Program*, C.F. Manski and I. Garfinkle, eds., Cambridge, MA: Harvard University Press, 201-230.

Heckman, J.J., 1997, Instrumental Variables: A Study of the Implicit Assumptions Underlying One Widely Used Estimator for Program evaluations, *Journal of Human Resources*, 32, 441-462.

Heckman, J.J., H. Ichimura and P.E. Todd (HIT), 1997, Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Programme, *The Review of Economic Studies*, 64, 605–654.

Heckman, J.J., H. Ichimura, J. Smith and P.E. Todd (HIST), 1998, Characterizing Selection Bias Using Experimental Data, *Econometrica*, 66, 1017-1098.

Heckman, J.J. and R. Robb, Jr., 1985, Alternative Methods for Evaluating the Impact of Interventions, *Longitudinal Analysis of Labor Market Data*, J.J Heckman and B. Singer, eds., New York: Cambridge University Press, 156-245.

Heckman, J.J. and B. Singer, 1984, A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data, *Econometrica*, 52, 271-320.

Heckman, J.J. and J.A. Smith, 1993, Assessing the Case for Randomized Evaluation of Social Programs, in *Measuring Labour Market Measures: Evaluating he Effects of Active Labour Market Policies*, K. Jensen and P.K. Madsen, eds., Copenhagen: Danish Ministry of Labor, 35-96.

Heckman, J.J. and J.A. Smith, 1998, Evaluating the Welfare State, *Econometric and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, S. Strøm, eds., Cambridge: Cambridge University Press, 241-318.

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

Heckman, J.J., S. Urzua and E.J. Vytlacil, 2006, Understanding Instrumental Variables in Models with Essential Heterogeneity, *Review of Economics and Statistics*, 88, 389-432.

Heckman, J.J., S. Urzua and E.J. Vytlacil, 2008, Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case, *Annales d'Economie et de Satistique*, 91-92, 151-174.

Heckman, J.J. and E.J. Vytlacil, 1999, Local Instrumental Variables and Latent Variable Models for Identifying the Bounding Treatment Effects, *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

Heckman, J.J. and E.J. Vytlacil, 2001a, Local Instrumental Variables, *Nonlinear Statistical Inference*, C. Hsiao, K. Morimune, and J.L. Powell, eds., Cambridge: Cambridge University Press, 1-46.

Heckman, J.J. and E.J. Vytlacil, 2001b, Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect, *Econometric Evaluation of Labour Market Policies*, M. Lechner and M. Pfeiffer, eds., Heidelberg: Physica-Verlag, 1-15.

Heckman, J.J. and E.J. Vytlacil, 2001c, Policy-Relevant Treatment Effects, *American Economic Review*, 91, 107-111.

Heckman, J.J. and E.J. Vytlacil, 2005, Structural Equations, Treatment Effects, and Econometric Policy Evaluation, *Econometrica*, 73, 669-738.

Heckman, J.J. and E.J. Vytlacil, 2007a, Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier Science B.V., Ch. 70, 4779-4874.

Heckman, J.J. and E.J. Vytlacil, 2007b, Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments, *Handbook of Econometrics*, Vol. 6B, J.J. Heckman and E.E. Leamer, eds., Amsterdam: Elsevier Science B.V., Ch. 71, 4875-5143.

Hirano, K., G.W. Imbens and G. Ridder, 2003, Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score, *Econometrica*, 71, 1161–1189.

Hoeffding, W., 1940, Masstabinvariate Korrelationstheorie, *Schriften des Mathematischen Instituts und des Instituts Für Angewandte Mathematik der Universität Berlin*, 5, 179-233.

Holland, P., 1986, Statistics and Causal Inference (with discussion), *Journal of the American Statistical Association*, 81, 945–970.

Hong, H. and E. Tamer, 2003, Inference in Censored Models with Endogenous Regressors, *Econometrica*, 71, 905-932.

Honoré, B. and L. Hu, 2004, On the Performance of Some Robust Instrumental Variables Estimators, *Journal of Business and Economic Statistics*, 22, 30-39.

Horowitz, J.L., 1985, Testing Probabilistic Discrete Choice Models of Travel Demand by Comparing Predicted and Observed Aggregate Choice Shares, *Transportation Research-B*, 19B, 17-38.

Horowitz, J.L. and S. Lee, 2007, Nonparametric Instrumental Variables Estimation of a Quantile Regression Model, *Econometrica*, 75, 1191-1208.

Hurwicz, L., 1962, On the Structural Form of Interdependent Systems, *Logic, Methodology and Philosophy of Science*, E. Nagel, P. Suppes, and A. Tarski, eds., Stanford, CA: Stanford University Press, 232-239.

Ibragimov, I. and R. Has'minskii, 1981, *Statistical Estimation: Asymptotic Theory*, New York: Springer-Verlag.

Ichimura, H. and S. Lee, 2010, Characterization of the Asymptotic Distribution of Semiparametric M-estimators, *Journal of Econometrics*, 159, 252-266.

Ichimura, H. and C. Taber, 2000, Direct Estimation of Policy Impacts, Technical Working Paper 254, NBER.

Imbens, G.W., 2004, Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86, 4-29.

Imbens, G.W., 2010, Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009), *Journal of Economic Literature*, 48, 399-423.

Imbens, G.W. and J.D. Angrist, 1994, Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, 467-475.

Imbens, G.W. and C. Manski, 2004, Confidence Intervals for Partially Identified Parameters, *Econometrica*, 72, 1845-1857.

Imbens, G.W. and W.K. Newey, 2009, Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity, *Econometrica*, 77, 1481-1512.

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

Imbens, G.W. and J.M. Wooldridge, 2009, Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, 47, 5-86.

Jun, S.J., 2008, Weak Identification Robust Tests in an Instrumental Quantile Model, *Journal of Econometrics*, 144, 118-138.

Jun, S.J., J. Pinkse and H. Xu, 2011, Tighter Bounds in Triangular Systems, *Journal of Econometrics*, 161, 122-128.

Kato, K., 2011, A Note on Moment Convergence of Bootstrap M-estimators, *Statistics and Risk Modeling*, 28, 51-61.

Kim, J.H., 2013, Identifying the Distribution of Treatment Effects under Support Restrictions, mimeo, Columbia University.

Kitagawa, T., 2009, Identification Region of the Potential Outcome Distributions Under Instrument Independence, mimeo, UCL.

Koenker, R., 2005, *Quantile Regression*, New York: Cambridge University Press.

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

Koenker, R. and J. Yoon, 2009, Parametric Links for Binary Choice Models: A Fisherian-Bayesian Colloquy, *Journal of Econometrics*, 152, 120-130.

Lee, S., 2007, Endogeneity in Quantile Regression Models: A Control Function Approach, *Journal of Econometrics*, 141, 1131–1158.

Lehmann, E.L., 1974, *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

Linton, O., E. Maasoumi and Y. Whang, 2005, Consistent Testing for Stochastic Dominance under General Sampling Schemes, *Review of Economic Studies*, 72, 735-765.

Ma, L. and R.W. Koenker, 2006, Quantile Regression Methods for Recursive Structural Equation Models, *Journal of Econometrics*, 134, 471-506.

Manski, C.F., 1994, The Selection Problem, *Advances in Econometrics: Sixth World Congress*, C. Sims, eds., Cambridge: Cambridge University Press.

Manski, C.F., 1995, *Identification Problems in the Social Sciences*, Cambridge and London: Harvard University Press.

Manski, C.F. 1996, Learning About Treatment Effects from Experiments with Random Assignment of Treatments, *Journal of Human Resources*, 31, 709-733.

Manski, C.F., 1997, Monotone Treatment Response, *Econometrica*, 65, 25-37.

Manski, C.F., 2003, *Partial Identification of Probabilities Distributions*, New York and Heidelberg: Springer.

Marschak, J., 1953, Economic Measurements for Policy and Predictions, *Studies in Econometric Method*, W.C. Hood and T.C. Koopmans, eds., New York: Wiley, 1-26.

McFadden, D., 1974, Conditional Logit Analysis of Qualitative Choice Behavior, *Frontiers of Econometrics*, P. Zarembka, eds., New York: Academic Press, Ch. 4, 105-142.

McFadden, D., 1989, Testing for Stochastic Dominance, *Studies in the Economics of Uncertainty (in Honor of J. Hadar)*, T.B. Fomby and T.K. Seo, eds., New York: Springer-Verlag, 113-134.

Newey, W.K., 1984, A Method of Moment Interpretation of Sequential Estimators, *Economic Letters*, 14, 201-206.

Newey, W.K., 1994, The Asymptotic Variance of Semiparametric Estimators, *Econometrica*, 62, 1349-1382.

Powell, J., 1984, Least Absolute Deviation Estimation for the Censored Regression Model, *Journal of Econometrics*, 25, 303-325.

Powell, J., 1986, Censored Regression Quantiles, *Journal of Econometrics,* 32, 143-155.

Rothe, C. and D. Wied, 2013, Misspecification Testing in a Class of Conditional Distributional Models, *Journal of the American Statistical Association*, 108, 314-324.

Sakata, S., 2007, Instrumental Variable Estimator Based on Conditional Median Restriction, *Journal of Econometrics*, 141, 350-382.

Sen, A., 1997, *On Economic Inequality*, 2nd edition, Oxford: Clarendon Press.

Sen, A., 2000, Social Justice and the Distribution of Income, *Handbook of Income Distribution*, Vol. 1, A.B. Atkinson and F. Bourguignon, eds., Amsterdam: North Holland, Ch. 1, 59-85.

Stoye, J., 2009, More on Confidence Intervals for Partially Identified Parameters, *Econometrica*, 77, 1299-1315.

Torgovitsky, A., 2011, Identification of Nonseparable Models with General Instruments, mimeo, Department of Economics, PSU.

van der Klaauw, W., 2002, Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach, *International Economic Review*, 43, 1249-1287.

van der Vaart, A.W. and J.A. Wellner, 1996, *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.

Vytlacil, E.J., 2002, Independence, Monotonicity, and Latent Index Models: An Equivalence Result, *Econometrica*, 70, 331-341.

Vytlacil, E.J., 2006, A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results, *Oxford Bulletin of Economics and Statistics*, 68, 515-518.

Vytlacil, E.J., and N. Yildiz, 2007, Dummy Endogenous Variables in Weakly Separable Models, *Econometrica*, 75, 757-779.

Appendix: Proofs

Proof of Theorem 2. We concentrate on the point identification of $Q_\tau(Y_1|X=x)$. Note that

$$\begin{aligned}
\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z) &\le \sup_{z\in\mathcal{Z}_x} Q_{Y|X,Z,D}\left(1-\frac{1-\tau}{p_x^{\sup}}\middle| x,z,1\right)1\left(p_x^{\sup}>1-\tau\right)+y^l(x)1\left(p_x^{\sup}\le 1-\tau\right)\\
&\le y^u(x)1\left(p_x^{\sup}>1-\tau\right)+y^l(x)1\left(p_x^{\sup}\le 1-\tau\right),\\
\inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z) &\ge \inf_{z\in\mathcal{Z}_x} Q_{Y|X,Z,D}\left(\frac{\tau}{p_x^{\sup}}\middle| x,z,1\right)1\left(p_x^{\sup}\ge\tau\right)+y^u(x)1\left(p_x^{\sup}<\tau\right)\\
&\ge y^l(x)1\left(p_x^{\sup}\ge\tau\right)+y^u(x)1\left(p_x^{\sup}<\tau\right).
\end{aligned}$$

When $0\le p_x^{\sup}<\min\{\tau,1-\tau\}$, $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)\le y^l(x)<y^u(x)\le \inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$, so $Q_{Y_1|X}(\tau|x)$ cannot be point identified. So it is necessary for $p_x^{\sup}\ge\min\{\tau,1-\tau\}$ to point identify $Q_{Y_1|X}(\tau|x)$.

Assume $p_x^{\sup}$ is achieved at some value that Z can take, say, $z^{\sup}$. If $p_x^{\sup}=1$, $Q_{Y|X,Z,D}(\tau|x,z^{\sup},1)\le \sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)\le \inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)\le Q_{Y|X,Z,D}(\tau|x,z^{\sup},1)$, so $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)=\inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$.

If $(Y_1,Y_0)\perp D|X,Z$, then under the assumption (8),

$$\begin{aligned}
\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z) &= \sup_{z\in\mathcal{Z}_x}\left\{Q_{Y|X}\left(1-\frac{1-\tau}{p(x,z)}\middle| x\right)1(p(x,z)>1-\tau)+y^l(x)1\left(p(x,z)\le 1-\tau\right)\right\}\\
&\le Q_{Y|X}\left(1-\frac{1-\tau}{p_x^{\sup}}\middle| x\right)1(p_x^{\sup}>1-\tau)+y^l(x)1\left(p_x^{\sup}\le 1-\tau\right),\\
\inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z) &= \inf_{z\in\mathcal{Z}_x}\left\{Q_{Y|X}\left(\frac{\tau}{p(x,z)}\middle| x\right)1(p(x,z)\ge\tau)+y^u(x)1\left(p(x,z)>\tau\right)\right\}\\
&\ge Q_{Y|X}\left(\frac{\tau}{p_x^{\sup}}\middle| x\right)1\left(p_x^{\sup}\ge\tau\right)+y^u(x)1\left(p_x^{\sup}<\tau\right).
\end{aligned}$$

When $p_x^{\sup}=1$, $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)=Q_{Y|X}(\tau|x)=\inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$, so $p_x^{\sup}=1$ is sufficient for point identification of $Q_{Y_1|X}(\tau|x)$. Suppose $Y_1|X=x$ is continuously distributed with a positive density on $(y^l(x),y^u(x))$. When $1>p_x^{\sup}>\max\{\tau,1-\tau\}$, $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)\le Q_{Y|X}\left(1-\frac{1-\tau}{p_x^{\sup}}\middle| x\right)<Q_{Y|X}\left(\frac{\tau}{p_x^{\sup}}\middle| x\right)=\inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$. When $\min\{\tau,1-\tau\}\le p_x^{\sup}\le\max\{\tau,1-\tau\}$, either $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)\le y^l(x)<Q_{Y|X}\left(\frac{\tau}{p_x^{\sup}}\middle| x\right)\le \inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$ or $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)\le Q_{Y|X}\left(1-\frac{1-\tau}{p_x^{\sup}}\middle| x\right)<y^u(x)\le \inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$ or $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)\le y^l(x)<y^u(x)\inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$ or $\sup_{z\in\mathcal{Z}_x} L^1_\tau(x,z)\le Q_{Y|X}\left(1-\frac{1-\tau}{p_x^{\sup}}\middle| x\right)<Q_{Y|X}\left(\frac{\tau}{p_x^{\sup}}\middle| x\right)\le \inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$. So $p_x^{\sup}=1$ is also necessary for point identification of $Q_{Y_1|X}(\tau|x)$. ■

Proof of Theorem 3. Depress the conditioning on $X=x$ to simplify notations. From Theorem 1, $\int_0^{p_x^{\sup}} F_{Y_1|U_D}(y|u_D)du_D$ and $\int_{p_x^{\inf}}^1 F_{Y_0|U_D}(y|u_D)du_D$ can be identified, but the distribution of (D,Y,X,Z) contains no information on $\int_{p_x^{\sup}}^1 F_{Y_1|U_D}(y|u_D)du_D$ and $\int_0^{p_x^{\inf}} F_{Y_0|U_D}(y|u_D)du_D$. Nevertheless, note that

$$\begin{aligned}
P\left(Y\le y|p(Z)=p_x^{\sup},D=1\right)p_x^{\sup} \le P(Y_1\le y) &= \int_0^{p_x^{\sup}} F_{Y_1|U_D}(y|u_D)du_D+\int_{p_x^{\sup}}^1 F_{Y_1|U_D}(y|u_D)du_D \quad (28)\\
&\le P\left(Y\le y|p(Z)=p_x^{\sup},D=1\right)p_x^{\sup}+(1-p_x^{\sup}),
\end{aligned}$$

and

$$P\left(Y \leq y|p(Z) = p_x^{\inf}, D=0\right)\left(1-p_x^{\inf}\right) \leq P(Y_0 \leq y) = \int_{p_x^{\inf}}^{1} F_{Y_1|U_D}(y|u_D)du_D + \int_{0}^{p_x^{\inf}} F_{Y_0|U_D}(y|u_D)du_D$$
$$\leq P\left(Y \leq y|p(Z) = p_x^{\inf}, D=0\right)\left(1-p_x^{\inf}\right) + p_x^{\inf}.$$

We concentrate on bounding $Q_{Y_1}(\tau)$ since the results for $Q_{Y_0}(\tau)$ can be similarly derived. We derive the proof into six steps. The first four steps are similar to the proof of Proposition 2 in Manski (1994).

Step 1: $R_\tau^1(x)$ is an upper bound for $Q_{Y_1}(\tau)$.

By (28),

$$P\left(Y \leq y|p(Z) = p_x^{\sup}, D=1\right)p_x^{\sup} \geq \tau \Longrightarrow P(Y_1 \leq y) \geq \tau. \tag{29}$$

The premise of (29) is empty if that $p_x^{\sup} < \tau$. Suppose that $p_x^{\sup} \geq \tau$. Then the definition of $R_\tau^1(x)$ states that

$$R_\tau^1(x) \equiv \min\left\{t : P\left(Y \leq t|p(Z) = p_x^{\sup}, D=1\right) \geq \tau/p_x^{\sup}\right\}.$$

It follows that $P(Y_1 \leq R_\tau^1(x)) \geq \tau$. Hence $Q_{Y_1}(\tau) \leq R_\tau^1(x)$.

Step 2: $L_\tau^1(x)$ is a lower bound for $Q_{Y_1}(\tau)$.

By (28),

$$P\left(Y \leq y|p(Z) = p_x^{\sup}, D=1\right)p_x^{\sup} + (1-p_x^{\sup}) < \tau \Longrightarrow P(Y_1 \leq y) < \tau,$$

which can be rewritten as

$$P\left(Y \leq y|p(Z) = p_x^{\sup}, D=1\right) < 1 - \frac{1-\tau}{p_x^{\sup}} \Longrightarrow P(Y_1 \leq y) < \tau. \tag{30}$$

The premise of (30) is empty if $p_x^{\sup} \leq 1-\tau$. Suppose that $p_x^{\sup} > 1-\tau$. Then the definition of $L_\tau^1(x)$ states that

$$L_\tau^1(x) \equiv \min\left\{t : P\left(Y \leq t|p(Z) = p_x^{\sup}, D=1\right) \geq 1 - \frac{1-\tau}{p_x^{\sup}}\right\}.$$

It follows that, for all $\eta > 0$, $P\left(Y_1 \leq L_\tau^1(x) - \eta\right) < \tau$. Hence $Q_{Y_1}(\tau) \geq L_\tau^1(x)$.

Step 3: $R_\tau^1(x)$ is the least upper bound for $Q_{Y_1}(\tau)$.

First let $p_x^{\sup} \geq \tau$. For any $\lambda > 0$,

$$P\left(Y_1 \leq R_\tau^1(x) - \lambda\right) = P\left(Y \leq R_\tau^1(x) - \lambda|p(Z) = p_x^{\sup}, D=1\right)p_x^{\sup} + \int_{p_x^{\sup}}^{1} F_{Y_1|U_D}(R_\tau^1(x) - \lambda|u_D)du_D.$$

Suppose $F_{Y_1|U_D}(R_\tau^1(x) - \lambda|u_D) = 0$ for $u_D \in (p_x^{\sup}, 1]$, as is possible in the absence of other information. Then the definition of $R_\tau^1(x)$ implies that

$$P\left(Y_1 \leq R_\tau^1(x) - \lambda\right) = P\left(Y \leq R_\tau^1(x) - \lambda|p(Z) = p_x^{\sup}, D=1\right)p_x^{\sup} < \tau.$$

Hence $Q_{Y_1}(\tau) > R_\tau^1(x) - \lambda$. Now let $p_x^{\sup} < \tau$. For any $t < y^u(x)$,

$$P(Y_1 \leq t) = P\left(Y \leq t|p(Z) = p_x^{\sup}, D=1\right)p_x^{\sup} + \int_{p_x^{\sup}}^{1} F_{Y_1|U_D}(t|u_D)du_D. \tag{31}$$

Suppose that $F_{Y_1|U_D}(t|u_D) = 0$ for $u_D \in (p_x^{\sup}, 1]$. Then

$$P(Y_1 \leq t) = P\left(Y \leq t|p(Z) = p_x^{\sup}, D=1\right)p_x^{\sup} < \tau.$$

Hence $Q_{Y_1}(\tau) > t$.

Step 4: $L^1_\tau(x)$ is the greatest lower bound for $Q_{Y_1}(\tau)$.

First let $p_x^{\sup} > 1-\tau$. For any $\lambda > 0$,

$$P\left(Y_1 \le L^1_\tau(x)+\lambda\right) = P\left(Y \le L^1_\tau(x)+\lambda | p(Z) = p_x^{\sup}, D=1\right) p_x^{\sup} + \int_{p_x^{\sup}}^1 F_{Y_1|U_D}(L^1_\tau(x)+\lambda|u_D)du_D.$$

Suppose that $F_{Y_1|U_D}(L^1_\tau(x)+\lambda|u_D) = 1$ for $u_D \in (p_x^{\sup}, 1]$, as is possible in the absence of other information. Then the definition of $L^1_\tau(x)$ implies that

$$P\left(Y_1 \le L^1_\tau(x)+\lambda\right) = P\left(Y \le L^1_\tau(x)+\lambda | p(Z) = p_x^{\sup}, D=1\right) p_x^{\sup} + (1-p_x^{\sup}) \ge \tau.$$

Hence $Q_{Y_1}(\tau) \le L^1_\tau(x)+\lambda$. Now, let $p_x^{\sup} \le 1-\tau$. Let $t > y^l(x)$ and suppose that $F_{Y_1|U_D}(t|u_D) = 1$. Then by (31),

$$P(Y_1 \le t) = P\left(Y \le t | p(Z) = p_x^{\sup}, D=1\right) p_x^{\sup} + (1-p_x^{\sup}) \ge \tau.$$

Hence $Q_{Y_1}(\tau) \le t$.

Step 5: $R^1_\tau(x) = L^1_\tau(x)$ if $p_x^{\sup} = 1$; $p_x^{\sup} = 1$ if $R^1_\tau(x) = L^1_\tau(x)$ when $Y|X = x, p(X,Z) = p_x^{\sup}, D = 1$ is continuously distributed with a positive density on $(y^l(x), y^u(x))$.

Suppose $p_x^{\sup} = 1$, then $L^1_\tau(x) = Q_{Y|X,p(X,Z),D}(\tau|x, p_x^{\sup}, 1) = R^1_\tau(x)$.

Fix $\tau \in (0, 1/2]$. When $0 \le p_x^{\sup} < \tau$, $L^1_\tau(x) = y^l(x) < y^u(x) = R^1_\tau(x)$. When $1 > p_x^{\sup} > 1-\tau$, $L^1_\tau(x) = Q_{Y|X,p(X,Z),D}\left(1 - \frac{1-\tau}{p_x^{\sup}}\middle| x, p_x^{\sup}, 1\right) < Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x^{\sup}}\middle| x, p_x^{\sup}, 1\right) = R^1_\tau(x)$. When $1-\tau \ge p_x^{\sup} \ge \tau$, $L^1_\tau(x) = y^l(x) < Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x^{\sup}}\middle| x, p_x^{\sup}, 1\right)$. So when $\tau \le 1/2$, $Q_\tau(Y_1|X=x)$ cannot be point identified unless $p_x^{\sup} = 1$. Similarly, when $1 > \tau > 1/2$, $Q_{Y_1|X}(\tau|x)$ cannot be point identified unless $p_x^{\sup} = 1$.

Step 6: $\sup\limits_{z\in\mathcal{Z}_x} L^1_\tau(x,z) = L^1_\tau(x)$ and $\inf\limits_{z\in\mathcal{Z}_x} R^1_\tau(x,z) = R^1_\tau(x)$ under assumption (9).

Note that

$$\begin{aligned}
\inf_{z\in\mathcal{Z}_x} R^1_\tau(x,z) &= \inf_{z\in\mathcal{Z}_x}\left\{Q_{Y|X,Z,D}\left(\frac{\tau}{p(x,z)}\middle| x,z,1\right)1(p(x,z)\ge\tau) + y^u(x)1(p(x,z)<\tau)\right\} \\
&= \inf_{p_x\in\mathcal{P}_x}\left\{Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x}\middle| x,p_x,1\right)1(p_x\ge\tau) + y^u(x)1(p_x<\tau)\right\}
\end{aligned}$$

and

$$R^1_\tau(x) = Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x^{\sup}}\middle| x, p_x^{\sup}, 1\right)1(p_x^{\sup}\ge\tau) + y^u(x)1(p_x^{\sup}<\tau).$$

If $p_x^{\sup} < \tau$, the result is trivial, so assume $p_x^{\sup} \ge \tau$. First, $\inf\limits_{z\in\mathcal{Z}_x} R^1_\tau(x,z) \le R^1_\tau(x)$, so we need only show $\inf\limits_{z\in\mathcal{Z}_x} R^1_\tau(x,z) \ge R^1_\tau(x)$. Since $\inf\limits_{z\in\mathcal{Z}_x} R^1_\tau(x,z)$ must be achieved at some $p_x \in \mathcal{P}_x$ such that $p_x \ge \tau$, we need only show

$$r^1_\tau(x,p_x) \equiv Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x}\middle| x,p_x,1\right) \ge Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p_x^{\sup}}\middle| x,p_x^{\sup},1\right) \equiv r^1_\tau(x)$$

for any $p_x \in \mathcal{P}_x$ and $p_x \ge \tau$. From the definition of $r^1_\tau(x,p_x)$, $r^1_\tau(x,p_x)$ is the infimum of r such that $P(Y \le r|X=x, p(Z)=p_x, D=1) \ge \tau/p_x$ or $\int_0^{p_x} F_{Y_1|X,U_D}(r|x,u_D)du_D \ge \tau$. Since $F_{Y_1|X,U_D}(r|x,u_D) \ge 0$, $\int_0^{p_x^{\sup}} F_{Y_1|X,U_D}(r^1_\tau(x,p_x)|x,u_D)du_D \ge \tau$ or $P\left(Y \le r^1_\tau(x,p_x)|X=x, p(Z)=p_x^{\sup}, D=1\right) \ge \tau/p_x^{\sup}$. Since

$r_\tau^1(x)$ is the infimum of r satisfying $P\left(Y \leq r | X = x, p(Z) = p_x^{\max}, D = 1\right) \geq \tau / p_x^{\sup}$, $r_\tau^1(x, p_x) \geq r_\tau^1(x)$. Similarly, we can show $\sup_{z \in \mathcal{Z}_x} L_\tau^1(x, z) = L_\tau^1(x)$. ■

Proof of Theorem 4. The following theorem applies to the case where $\mathcal{Y}_d$ is a compact interval of $\mathbb{R}$. The case where Y_d is discrete is simpler. Lemma E.1 and E.2 in the following proof are referred to the corresponding lemmas in Appendix E of Chernozhukov et al. (2013).

Step 1: We apply the proof idea of Theorem 5.2 of Chernozhukov et al. (2013) to derive the weak limit of $\widehat{\beta}_d(y_d)$.

If we use the notation of Chernozhukov et al. (2013), $u = y = (y_0, y_1)'$, $\theta(u) = (\gamma, \beta_0(y)', \beta_1(y)')' \equiv (\gamma, \beta(y)')' \in \mathbb{R}^{d_\theta}$, and $\mathcal{U} = \mathcal{Y} \equiv \mathcal{Y}_0\mathcal{Y}_1 = \{(y_0, y_1) | y_0 \in \mathcal{Y}_0, y_1 \in \mathcal{Y}_1\}$. Note that the first element of $\theta(u)$ does not depend on y. Let

$$\varphi_{y,\theta}(D, Y, X, Z) = \left(\begin{array}{l} (\widetilde{p} - D)\, H(R'\gamma) R, \\ (1 - D)\left[\Lambda(T'\beta_0) - 1(Y \leq y_0)\right] H(T'\beta_0) T, \\ D\left[\Lambda(T'\beta_1) - 1(Y \leq y_1)\right] H(T'\beta_1) T, \end{array} \right)$$

where $H(\cdot) = \lambda(\cdot) / \{\Lambda(\cdot)\left[1 - \Lambda(\cdot)\right]\}$. Let $\Psi(\theta, y) = P\left(\varphi_{y,\theta}\right)$ and $\widehat{\Psi}(\theta, y) = P_n\left(\varphi_{y,\theta}\right)$, where P_n is the empirical measure and P is the corresponding probability measure. From the first order conditions, distribution regression in the sample obeys $\widehat{\theta}(y) = \phi(\widehat{\Psi}(\theta, y), 0)$ for each $y \in \mathcal{Y}$, where ϕ is the Z-map defined in Appendix E.1. Then, by Step 3 below

$$\sqrt{n}\left(\widehat{\Psi} - \Psi\right) \rightsquigarrow W \equiv (W_\gamma, W_0(y_0), W_1(y_1)) \text{ in } \ell^\infty\left(\mathbb{R}^{d_\theta} \times \mathcal{Y}\right),\ W(y, \theta) = \mathbb{G}\left(\varphi_{y,\theta}\right),$$

where W has continuous paths a.s. with three components being independent of each other. Step 4 verifies the Conditions of Lemma E.1 for

$$\dot{\Psi}_{\theta(u),u} = \left(\begin{array}{ccc} J_p & \mathbf{0} & \mathbf{0} \\ J_{0p}(y_0) & J_0(y_0) & \mathbf{0} \\ J_{1p}(y_1) & \mathbf{0} & J_1(y_1) \end{array} \right) \equiv J(y),$$

where J_p, $J_d(y_d)$ and $J_{dp}(y_d)$ are defined in the main text, which also implies $y \longmapsto \beta(y)$ is continuously differentiable on $\mathcal{Y}$. Then, by Lemma E.2, the map ϕ is Hadamard-differentiable with the derivative map $w \longmapsto -J^{-1}w$ at $(\Psi, 0)$. Therefore, we can conclude by the Functional Delta Method that

$$\sqrt{n}\left(\widehat{\theta}(\cdot) - \theta(\cdot)\right) \rightsquigarrow -J^{-1}(\cdot) W(\theta(\cdot), \cdot) \text{ in } \ell^\infty\left(\mathcal{Y}\right)^{d_\theta}.$$

This further implies that

$$\sqrt{n}\left(\widehat{\beta}(y) - \beta(y)\right) \rightsquigarrow -\left(\begin{array}{c} J_0(y_0)^{-1}\left(J_{0p}(y_0) J_p^{-1} W_\gamma + W_0(y_0)\right) \\ J_1(y_1)^{-1}\left(J_{1p}(y_1) J_p^{-1} W_\gamma + W_1(y_1)\right) \end{array} \right) \text{ in } \ell^\infty\left(\mathcal{Y}\right)^{d_\beta}.$$

Note that the weak convergence in this step does not reply on compactness of $\mathcal{Y}$.

Step 2: We derive the results for conditional CDFs and the MQTE. [Here we shall rely on compactness of $\mathcal{YXP}$. Then $\mathcal{Y}_0$ and $\mathcal{Y}_1$ are closed interval of $\mathbb{R}$.]

Consider the mapping $\varphi : \mathbb{D}_\varphi \subset \ell^\infty\left(\mathcal{Y}\right)^{d_\beta} \longmapsto \ell^\infty\left(\mathcal{YXP}\right)^2$, defined as $b \longmapsto \varphi(b)$,

$$\varphi(b)(y, p, x) = \left(\begin{array}{c} \Lambda(T(x,p))' b_0(y_0)) - (1 - p)\frac{\partial T(x,p)'}{\partial p} b_0(y_0) \cdot \lambda\left(T\left(x, p\right)' b_0(y_0)\right), \\ \Lambda(T(x,p))' b_1(y_1)) + p\frac{\partial T(x,p)'}{\partial p} b_1(y_1) \cdot \lambda\left(T\left(x, p\right)' b_1(y_1)\right). \end{array} \right)$$

It is straightforward to deduce that this map is Hadamard differentiable at $b(\cdot) = \beta(\cdot)$ tangentially to $C(\mathcal{Y})^{d_\beta}$ with the derivative map given by: $\alpha \longmapsto \varphi'_{\beta(\cdot)}(\alpha)$,

$$\varphi'_{\beta(\cdot)}(\alpha)(y,x,p) = \begin{pmatrix} \varphi'_{\beta_0(\cdot)}(\alpha)(y,x,p) \\ \varphi'_{\beta_1(\cdot)}(\alpha)(y,x,p) \end{pmatrix},$$

where $\varphi'_{\beta_0(\cdot)}$ and $\varphi'_{\beta_1(\cdot)}$ are defined in the main text. Since

$$\widehat{F}_{Y_0|X,U_D}(y_0|x,p) = \Lambda(T(x,p))'\widehat{\beta}_0(y_0)) - (1-p)\frac{\partial T(x,p)'}{\partial p}\widehat{\beta}_0(y)\cdot\lambda\left(T(x,p)'\widehat{\beta}_0(y_0)\right),$$
$$\widehat{F}_{Y_1|X,U_D}(y_1|x,p) = \Lambda(T(x,p))'\widehat{\beta}_1(y_1)) + p\frac{\partial T(x,p)'}{\partial p}\widehat{\beta}_1(y)\cdot\lambda\left(T(x,p)'\widehat{\beta}_1(y_1)\right),$$

by the delta method it follows that

$$\sqrt{n}\begin{pmatrix} \widehat{F}_{Y_0|X,U_D}(y_1|x,p) - F_{Y_0|X,U_D}(y_1|x,p) \\ \widehat{F}_{Y_1|X,U_D}(y_1|x,p) - F_{Y_1|X,U_D}(y_1|x,p) \end{pmatrix} \rightsquigarrow -\begin{pmatrix} \varphi'_{\beta_0(\cdot)}\left(J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right) \\ \varphi'_{\beta_1(\cdot)}\left(J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right) \end{pmatrix} \text{ in } \ell^\infty(\mathcal{YXP})^2,$$

and

$$\sqrt{n}\begin{pmatrix} \widehat{F}^*_{Y_0|X,U_D}(y_1|x,p) - \widehat{F}_{Y_0|X,U_D}(y_1|x,p) \\ \widehat{F}^*_{Y_1|X,U_D}(y_1|x,p) - \widehat{F}_{Y_1|X,U_D}(y_1|x,p) \end{pmatrix} \overset{*}{\rightsquigarrow} -\begin{pmatrix} \varphi'_{\beta_0(\cdot)}\left(J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right) \\ \varphi'_{\beta_1(\cdot)}\left(J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right) \end{pmatrix} \text{ in } \ell^\infty(\mathcal{YXP})^2.$$

Also, $\Delta_\tau^{MQTE}(x,u_D)$ is estimated as

$$\widehat{\Delta}_{X,U_D}(\tau|x,u_D) = \widehat{F}^{-1}_{Y_1|X,U_D}(\tau|x,u_D) - \widehat{F}^{-1}_{Y_0|X,U_D}(\tau|x,u_D).$$

By the proof of Theorem 4.1(2) of Chernozhukov et al. (2013), the results in the theorem follow.
Step 3: We verify that $\left\{\varphi_{y,\theta}(Y,X,Z)|(y,\theta)\in\mathcal{Y}\times\mathbb{R}^{d_\theta}\right\}$ is P-Donsker with a square integrable envelope.

We only analyze $D\left[\Lambda(T'\beta_1) - 1(Y\le y_1)\right]H(T'\beta_1)T$ for illustration. The function classes $\mathcal{F}_1 = \left\{T'\beta_1|\beta_1\in\mathbb{R}^{d_T}\right\}$, $\mathcal{F}_2 = \{1(Y\le y_1)|y_1\in\mathcal{Y}_1\}$, and $\{DT_q|q=1,\cdots,d_T\}$ are VC classes of functions. The final class $\mathcal{G} = \{(\Lambda(\mathcal{F}_1)-\mathcal{F}_2)H(\mathcal{F}_1)DT_q|q=1,\cdots,d_T\}$ is a Lipschitz transformation of VC classes with Lipschitz coefficient bounded by const$\cdot\|T\|$ and envelope function const$\cdot\|T\|$, which is square-integrable. Hence $\mathcal{G}$ is Donsker by Example 9.19 in van der Vaart (1998). Finally, the map $(\beta_1,y_1)\longmapsto D\left[\Lambda(T'\beta_1)-1(Y\le y_1)\right]H(T'\beta_1)T$ is continuous at each $(\beta_1,y_1)\in\mathbb{R}^{d_{\beta_1}}\times\mathcal{Y}_1$ with probability one by the absolute continuity of the conditional distribution of Y (when $\mathcal{Y}$ is not finite).
Step 4: We verify conditions (a)-(c) of Lemma E.1.

Condition (a) and (b) are immediate by the assumption. To verify (c), a straightforward computation gives that for (θ,y) in the neighborhood of $(\theta(y),y)$,

$$\frac{\partial}{\partial\theta'}\Psi(\theta,y) = \tag{32}$$
$$\begin{pmatrix} E\left[\{h[p-D]+H\lambda\}RR'\right] & \mathbf{0} & \mathbf{0} \\ A_{0\gamma} & E\left[(1-D)\{h_0[\Lambda_0-1(Y\le y_0)]+H_0\lambda_0\}TT'\right] & \mathbf{0} \\ A_{1\gamma} & \mathbf{0} & E\left[D\{h_1[\Lambda_1-1(Y\le y_1)]+H_1\lambda_1\}TT'\right] \end{pmatrix},$$

and

$$\frac{\partial}{\partial y'}\Psi(\theta,y)=\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ -E\left[(1-D)f_{Y|X,Z,D}(y|X,Z,D)H_0T\right] & \mathbf{0}\\ \mathbf{0} & -E\left[Df_{Y|X,Z,D}(y|X,Z,D)H_1T\right]\end{pmatrix}$$
$$=\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ -E\left[(1-\widetilde{p})f_{Y_0|X,Z}(y|X,Z)H_0T\right] & \mathbf{0}\\ \mathbf{0} & -E\left[\widetilde{p}f_{Y_1|X,Z}(y|X,Z)H_1T\right]\end{pmatrix},$$

where

$$A_{0\gamma}=E\left[(1-D)\lambda\left\{\frac{\partial T(X,p)'\beta_0}{\partial p}\left[h_0\left(\Lambda_0-1(Y\le y_0)\right)+\lambda_0H_0\right]T+H_0\left[\Lambda_0-1(Y\le y_0)\right]\frac{\partial T(X,p)}{\partial p}\right\}R'\right],$$
$$A_{1\gamma}=E\left[D\lambda\left\{\frac{\partial T(X,p)'\beta_1}{\partial p}\left[h_1\left(\Lambda_1-1(Y\le y_1)\right)+\lambda_1H_1\right]T+H_1\left[\Lambda_1-1(Y\le y_1)\right]\frac{\partial T(X,p)}{\partial p}\right\}R'\right]$$

and $T=T(X,p)$, $R=R(X,Z)$, $H=H(R'\gamma)$, $p=\Lambda(R'\gamma)$, $h=h(R'\gamma)$, $\lambda=\lambda(R'\gamma)$, $H_d=H(T'\beta_d)$, $h_d=h(T'\beta_d)$, $\Lambda_d=\Lambda(T'\beta_d)$, $\lambda_d=\lambda(T'\beta_d)$. Both terms are continuous in (θ,y) at $(\theta(y),y)$ for each $y\in\mathcal{Y}$. The computation above as well as verification of continuity follows from using the dominated convergence theorem, and the following ingredients: (1) a.s. continuity of the map $(\theta,y)\longmapsto\frac{\partial}{\partial\theta'}\varphi_{y,\theta}(Y,X,Z)$, (2) domination of $\left\|\frac{\partial}{\partial\theta'}\varphi_{y,\theta}(Y,X,Z)\right\|$ by a square-integrable function const$\cdot\|(R,T)\|$, (3) a.s. continuity and uniform boundedness of the conditional density function $y\longmapsto f_{Y|X,Z,D}(y|X,Z,d)$, and (4) $H(\cdot)$ is bounded uniformly on $\mathbb{R}$, a.s. By assumption and the lower-triangular form of $\frac{\partial}{\partial\theta'}\Psi(\theta,y)$, $\frac{\partial}{\partial\theta'}\Psi(\theta(y),y)$ is positive-definite uniformly in $y\in\mathcal{Y}$. ∎

Proof of Theorem 5. Since T_n^K and T_n^C are continuous functionals of $\widehat{H}_n^d(y)-\widehat{F}_n^d(y)$, $y\in\mathcal{Y}_d$, we can apply the continuous mapping theorem to get their asymptotic distributions as long as the weak limits of $\widehat{H}_n^d(y)-\widehat{F}_n^d(y)$, $y\in\mathcal{Y}_d$, are derived. For this purpose, let

$$\varphi_{y,\theta}(D,Y,X,Z)=\begin{pmatrix}(\widetilde{p}-D)\,H(R'\gamma)R,\\ (1-D)\left[\Lambda(T'\beta_0)-1(Y\le y_0)\right]H(T'\beta_0)T,\\ D\left[\Lambda(T'\beta_1)-1(Y\le y_1)\right]H(T'\beta_1)T,\\ (1-D)\cdot 1(Y\le y_0)-H^0(y_0),\\ D\cdot 1(Y\le y_1)-H^1(y_1),\\ (1-D)\cdot\Lambda(T'\beta_0)-F^0(y_0),\\ D\cdot\Lambda(T'\beta_1)-F^1(y_1),\end{pmatrix}$$

and $\Psi(\theta,y)=P(\varphi_{y,\theta})$, where $\theta(y)=\left(\gamma,\beta_0(y_0)',\beta_1(y_1)',H^0(y_0),H^1(y_1),F^0(y_0),F^1(y_1)\right)'$ and other notations are the same as in the proof of Theorem 4. It can be shown that

$$\dot{\Psi}_{\theta(y),y}=\begin{pmatrix}J_p & \mathbf{0} & \mathbf{0} & 0 & 0 & 0 & 0\\ J_{0p}(y_0) & J_0(y_0) & \mathbf{0} & 0 & 0 & 0 & 0\\ J_{1p}(y_1) & \mathbf{0} & J_1(y_1) & 0 & 0 & 0 & 0\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -1 & 0 & 0 & 0\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & -1 & 0 & 0\\ A_\gamma^0 & A_0^0 & \mathbf{0} & 0 & 0 & -1 & 0\\ A_\gamma^1 & \mathbf{0} & A_1^1 & 0 & 0 & 0 & -1\end{pmatrix},$$

where A_γ^d and A_d^d are defined in the main text. By the Functional Delta Method, it follows that

$$\sqrt{n}\begin{pmatrix} \widehat{H}_n^0(y_0)-H^0(y_0) \\ \widehat{H}_n^1(y_1)-H^1(y_1) \\ \widehat{F}_n^0(y_0)-F^0(y_0) \\ \widehat{F}_n^1(y_1)-F^1(y_1) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_H^0(y_0) \\ \mathbb{G}_H^1(y_1) \\ \mathbb{G}_F^0(y_0) \\ \mathbb{G}_F^1(y_1) \end{pmatrix} \text{ in } \ell^\infty(\mathcal{Y})^4, \tag{33}$$

where $\mathbb{G}_H^d(y_d)$ and $\mathbb{G}_F^d(y_d)$ are defined in the main text. By the continuous mapping theorem, the asymptotic null distributions in (i) follow.

Under H_1, if $P(Y \in S_d(\beta_d, \gamma), D = d) > 0$,

$$\begin{aligned} T_n^K &= \sqrt{n}\sup_{y\in\mathcal{Y}_1}\left|\widehat{H}_n^1(y)-H^1(y)-\left(\widehat{F}_n^1(y)-\overline{F}^1(y)\right)+\left(H^1(y)-\overline{F}^1(y)\right)\right| \\ &+\sqrt{n}\sup_{y\in\mathcal{Y}_0}\left|\widehat{H}_n^0(y)-H^0(y)-\left(\widehat{F}_n^0(y)-\overline{F}^0(y)\right)+\left(H^0(y)-\overline{F}^0(y)\right)\right| \\ &= O_p\left(\sqrt{n}\right), \end{aligned}$$

and similarly, $T_n^C = O_p(n)$, which implies that for any sequences of random variables $\{c_n : n \geq 1\}$ with $c_n = O_p(1)$,

$$\lim_{n\to\infty} P\left(T_n^K > c_n\right) = 1 \text{ and } \lim_{n\to\infty} P\left(T_n^C > c_n\right) = 1,$$

where $\overline{F}^d(y) = E\left[1(D=d)\Lambda(T(X,\Lambda(R(X,Z)'\overline{\gamma}))'\overline{\beta}_d)\right]$ with $\overline{\beta}_d \in \mathcal{B}_d(\mathcal{Y}_d)$ and $\overline{\gamma} \in \Gamma$ being pseudo-true values of β_d and γ. Note here that although we can show a similar result as in (33), J_p, $J_d(\cdot)$ and $J_{dp}(\cdot)$ cannot be simplified under H_1 and must take the form as in (32). Nevertheless, $\sqrt{n}\sup_{y\in\mathcal{Y}_d}\left|\widehat{H}_n^d(y)-H^d(y)-\left(\widehat{F}_n^d(y)-\overline{F}^d(y)\right)\right| = O_p(1)$ and is dominated by $\sqrt{n}\sup_{y\in\mathcal{Y}_d}\left|\left(H^d(y)-\overline{F}^d(y)\right)\right|$.

From Lemma 2.8.7 in Van der Vaart and Wellner (1996, p. 174), $\sqrt{n}\left(\widehat{H}_n^d(y_d)-H_n^d(y_d)\right)$ and $\sqrt{n}\left(\widehat{F}_n^d(y_d)-F_n^d(y_d)\right)$ under the local alternative have the same weak limit as $\sqrt{n}\left(\widehat{H}_n^d(y_d)-H^d(y_d)\right)$ and $\sqrt{n}\left(\widehat{F}_n^d(y_d)-F^d(y_d)\right)$ under the null (associated with p_* and F_*^d), where $H_n^d(y_d)$ and $F_n^d(y_d)$ are the counterparts of $H^d(y_d)$ and $F^d(y_d)$ under the local alternative. So it remains to find the limits of $\sqrt{n}\left(H_n^d(y_d)-H^d(y_d)\right)$ and $\sqrt{n}\left(F_n^d(y_d)-F^d(y_d)\right)$.

$$\begin{aligned} \sqrt{n}\left(H_n^d(y_d)-H^d(y_d)\right) &= \delta_d E\left[\left(Q^d(y_d|X,q_n(X,Z))-F_*^d(y_d|X,q_n(X,Z))\right)1(D=d)\right] \\ &+\delta_d E\left[\left(F_*^d(y_d|X,q_n(X,Z))-F_*^d(y_d|X,p_*(X,Z))\right)1(D=d)\right] \\ &\longrightarrow \delta_d E\left[\left(Q^d(y_d|X,p_*(X,Z))-F_*^d(y_d|X,p_*(X,Z))\right)1(D=d)\right] \end{aligned}$$

by the bounded convergence theorem. Since

$$\sqrt{n}\left(F_n^d(y_d)-F^d(y_d)\right) = \sqrt{n}\left(E\left[\Lambda(T_n'\beta_{dn})\cdot 1(D=d)\right]-E\left[\Lambda(T'\beta_d)\cdot 1(D=d)\right]\right)$$

where γ_n in T_n and β_{dn} are the true values of γ and β_d under local alternative, we have

$$\begin{aligned} \sqrt{n}\left(F_n^d(y_d)-F^d(y_d)\right) &\longrightarrow -\delta_\gamma\left[A_\gamma^d + A_d^d J_d(y_d)^{-1}J_{dp}(y_d)\right]J_p^{-1}E\left[(q-D)\,H(R'\gamma)R\right] \\ &- \delta_d A_d^d J_d(y_d)^{-1}E\left[\left[\Lambda(T'\beta_d)-Q^d(y_d|X,\widetilde{p})\right]H(T'\beta_d)T\cdot 1(D=d)\right], \end{aligned}$$

by a similar argument as in Step 1 of Theorem 4. ■

Proof of Theorem 6. We take T_n^K as an example since the proof for T_n^C is similar.

Because $\left(\widehat{H}_n^0(y_0), \widehat{H}_n^1(y_1), \widehat{F}_n^0(y_0), \widehat{F}_n^1(y_1)\right) = \phi(\widehat{\Psi}(\theta, y))$ and $\phi(\cdot)$ is Hadamard differentiable at $\Psi(\theta, y)$, where $\widehat{\Psi}(\theta, y) = P_n(\varphi_{y,\theta})$ and $\Psi(\theta, y) = P(\varphi_{y,\theta})$ with $\varphi_{y,\theta}$ defined in Theorem 5, by the functional delta method for the bootstrap (Van der Vaart and Wellner (1996), Theorem 3.9.11, p. 378), it holds under either the null or a fixed alternative that $\sqrt{n}\left(\widehat{H}_n^{0*}(y_0) - \widehat{F}_n^0(y_0), \widehat{H}_n^{1*}(y_1) - \widehat{F}_n^1(y_1), \widehat{F}_n^{0*}(y_0) - \widehat{F}_n^0(y_0), \widehat{F}_n^{1*}(y_1) - \widehat{F}_n^1(y_1)\right)$ weakly converges to a tight mean zero Gaussian process whose distribution coincides with that of the limit process in (33) under H_0.

Under H_0, it follows from the above result that $\widehat{c}_n(\alpha) = c(\alpha) + o_p(1)$ under H_0. This implies that T_n and $T_n - (\widehat{c}_n(\alpha) - c(\alpha))$ converges to the same limiting distribution as $n \to \infty$, and hence we have that $P(T_n > \widehat{c}_n(\alpha)) = \alpha + o(1)$ under the null.

Under a fixed alternative, $\widehat{c}_n(\alpha)$ is bounded in probability from the above result. Thus, for any $\epsilon > 0$, there exists a constant M such that $P(\widehat{c}_n(\alpha) > M) < \epsilon + o(1)$. Using elementary inequalities, we also have that

$$P(T_n \le \widehat{c}_n(\alpha)) = P(T_n \le \widehat{c}_n(\alpha), \widehat{c}_n(\alpha) \le M) + P(T_n \le \widehat{c}_n(\alpha), \widehat{c}_n(\alpha) > M) \le P(T_n \le M) + P(\widehat{c}_n(\alpha) > M).$$

From Theorem 5(ii), we know that $P(T_n \le M) = o(1)$, and thus $P(T_n \le \widehat{c}_n(\alpha)) < \epsilon + o(1)$, which implies the statement of the theorem since ϵ can be chosen arbitrarily small.

As to part (iii), we can use Anderson's Lemma (e.g., see Ibragimov and Has'minski (1981), Lemma 10.1, p. 155) and similar arguments of Andrews (1997, p. 1114) to show that $\lim_{n\to\infty} P(T_n > c(\alpha)) \ge \alpha$. Furthermore, we have already shown in part (i) that $P(T_n > \widehat{c}_n(\alpha)) = P(T_n > c(\alpha)) + o(1)$ under the null. By using contiguity arguments, this can also shown to be true under the local alternative; see, for example, the proof of Corollary 2.1 in Bickel and Ren (2001, p. 97). ■

Proof of Theorem 7. Note that

$$\begin{aligned}
& E[Y - \mu_0(X) - D(\mu_1(X) - \mu_0(X))\,|X, Z] \\
&= E[U_0 + D(U_1 - U_0)\,|X, Z] \\
&= E[\delta(X, Z, V)(U_1 - U_0)\,|X, Z] \\
&= E[\delta(X, Z, V)|X, Z]\, E[(U_1 - U_0)\,|X, Z] \\
&= 0,
\end{aligned}$$

where the second equality uses $E[U_0|X, Z] = 0$, and the third equality uses $(U_1 - U_0) \perp V|\,(X, Z)$, and the fourth equality uses $E[U_1 - U_0|X, Z] = 0$. This is the parallel result of Theorem 1 of Chernozhukov and Hansen (2005) in the average treatment effect evaluation.

To identify $\mu_1(X) - \mu_0(X)$, we need for all $\lambda_0(X)$ and $\lambda_1(X)$ with finite expectation,

$$E[\lambda_0(X) + D\lambda_1(X)|X, Z] = \lambda_0(X) + E[D|X, Z]\lambda_1(X) = 0 \Longrightarrow \lambda_1(X) = 0. \tag{34}$$

If $E[D|X, Z] > 0$, then $\lambda_1(X) = 0$ is equivalent to $\lambda_0(X) = \lambda_1(X) = 0$. If Z can take only one value, obviously, (34) cannot be satisfied. Suppose Z can take two values, z_1 and z_2, then $E[\lambda_0(X) + D\lambda_1(X)|X, Z] = 0$ implies

$$\begin{aligned}
\lambda_0 + E[D|Z = z_1]\lambda_1 = 0, \\
\lambda_0 + E[D|Z = z_2]\lambda_1 = 0,
\end{aligned}$$

i.e.,

$$\begin{bmatrix} 1 & E[D|Z=z_1] \\ 1 & E[D|Z=z_2] \end{bmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where we depress the conditioning on X to simplify the notations. As long as $E[D|Z=z_1] \neq E[D|Z=z_2]$, we must have $\lambda_0 = \lambda_1 = 0$. Interestingly, we do not require $E[D|Z] \in (0,1)$ as long as Z has a nontrivial impact on D. Also, this condition is equivalent to that D is complete for the "parameter" Z in the family of $F_{D|Z}$. To see why, note that for any $\lambda(D)$ with finite expectation (which is equivalent to $|\lambda(0)| < \infty$ and $|\lambda(1)| < \infty$), $E[\lambda(D)|Z] = 0$ implies

$$\begin{aligned} \lambda(0)P_{D|Z}(0|Z) + \lambda(1)P_{D|Z}(1|Z) &= \lambda(0)(1 - E[D|Z]) + \lambda(1)E[D|Z] \\ &= \lambda(0) + (\lambda(1) - \lambda(0))\, E[D|Z] = 0. \end{aligned}$$

If $\lambda_0 + \lambda_1 E[D|Z] = 0$ for any λ_0 and $\lambda_1 \Longrightarrow \lambda_0 = \lambda_1 = 0$; then $\lambda(0) + (\lambda(1) - \lambda(0))\, E[D|Z] = 0$ for any $\lambda(0)$ and $\lambda(1) \Longrightarrow \lambda(0) = \lambda(1) = 0$.

The usual IV estimator of β in the regression $Y = \alpha + D\beta + \varepsilon$ with $\varepsilon = U_0 + D\,(U_1 - U_0)$ is

$$\widehat{\beta}_{IV} = \frac{\widehat{Cov}(Z,Y)}{\widehat{Cov}(Z,D)},$$

where $\widehat{Cov}$ is the covariance estimate (conditional on X). It is easy to see that

$$\widehat{\beta}_{IV} \xrightarrow{p} \beta + \frac{Cov(Z,\varepsilon)}{Cov(Z,D)},$$

where $Cov(\cdot,\cdot)$ is understood as the conditional covariance given X.

$$\begin{aligned} Cov(Z,\varepsilon) &= Cov(Z, U_0 + D\,(U_1 - U_0)) \\ &= Cov(Z, U_0) + Cov(Z, D\,(U_1 - U_0)) \\ &= Cov(Z, D\,(U_1 - U_0)), \end{aligned}$$

where the last equality is implied by the assumption $E[U_0|X,Z] = 0$ and $E[U_0|X] = 0$.

$$\begin{aligned} Cov(Z, D\,(U_1 - U_0)) &= E\left[ZD\,(U_1 - U_0)\right] - E\left[Z\right] E\left[D\,(U_1 - U_0)\right] \\ &= E\left[ZE\left[D|Z\right] E[U_1 - U_0|Z]\right] - E\left[Z\right] E\left[E\left[D|Z\right] E\left[U_1 - U_0|Z\right]\right] \\ &= 0, \end{aligned}$$

where $E[\cdot|\cdot]$ is understood as conditional on X, the second equality is from $(U_1 - U_0) \perp V|\,(X,Z)$ and the law of iterated expectation, and the last equality is from $E[U_1 - U_0|X,Z] = 0$. This derivation is similar as that in footnote 5 of Heckman et al. (2006).

All the elegancy in the identification of $\mu_1(X) - \mu_0(X)$ and the consistency of $\widehat{\beta}_{IV}$ hinges on the linear structure of $\mu_0(X) + D\,(\mu_1(X) - \mu_0(X))$ in D. ■

Proof of Theorem 8. Consider the mapping $\varphi : \mathbb{D}_\varphi \subset \mathbb{R}^{d_\gamma} \times \ell^\infty(\mathcal{Y})^{d_\beta} \longmapsto \mathbb{R}^2 \times \ell^\infty(\mathcal{Y}\mathcal{X})^4$, defined as

$(a,b) \longmapsto \varphi(a,b)$,

$$\varphi(a,b)(x,y) = \begin{pmatrix} \Lambda(R(x,1)'a) \\ \Lambda(R(x,0)'a) \\ \Lambda(T(x,p(x,1;a))'b_1(y_1)) \\ \Lambda(T(x,p(x,0;a))'b_1(y_1)) \\ \Lambda(T(x,p(x,1;a))'b_0(y_0)) \\ \Lambda(T(x,p(x,0;a))'b_0(y_0)) \end{pmatrix}.$$

It is straightforward to deduce that this map is Hadamard differentiable at $(a,b(\cdot)) = (\gamma,\beta(\cdot))$ tangentially to $\mathbb{R}^{d_\gamma} \times C(\mathcal{Y})^{d_\beta}$ with the derivative map given by: $(\eta,\alpha) \longmapsto \varphi'_{\gamma,\beta(\cdot)}(\eta,\alpha)$,

$$\varphi'_{\gamma,\beta(\cdot)}(\eta,\alpha)(y,x) = \begin{pmatrix} \lambda(R(x,1)'\gamma)R(x,1)'\eta \\ \lambda(R(x,0)'\gamma)R(x,0)'\eta \\ \lambda(T(x,p(x,1;\gamma))'\beta_1(y_1)) \left[\frac{\partial T(x,p(x,1;\gamma))'\beta_1(y_1)}{\partial p}\lambda(R(x,1)'\gamma)R(x,1)'\eta + T(x,p(x,1;\gamma)))'\alpha_1(y_1)\right] \\ \lambda(T(x,p(x,0;\gamma))'\beta_1(y_1)) \left[\frac{\partial T(x,p(x,0;\gamma))'\beta_1(y_1)}{\partial p}\lambda(R(x,0)'\gamma)R(x,0)'\eta + T(x,p(x,0;\gamma)))'\alpha_1(y_1)\right] \\ \lambda(T(x,p(x,1;\gamma))'\beta_0(y_0)) \left[\frac{\partial T(x,p(x,1;\gamma))'\beta_0(y_0)}{\partial p}\lambda(R(x,1)'\gamma)R(x,1)'\eta + T(x,p(x,1;\gamma)))'\alpha_0(y_0)\right] \\ \lambda(T(x,p(x,0;\gamma))'\beta_0(y_0)) \left[\frac{\partial T(x,p(x,0;\gamma))'\beta_0(y_0)}{\partial p}\lambda(R(x,0)'\gamma)R(x,0)'\eta + T(x,p(x,0;\gamma)))'\alpha_0(y_0)\right] \end{pmatrix}.$$

Since

$$\widehat{F}_{Y_1|X,D_0,D_1}(y_1|x,0,1) = \frac{\Lambda(T(x,\Lambda(R(x,1)'\widehat{\gamma}))'\widehat{\beta}_1(y_1))\Lambda(R(x,1)'\widehat{\gamma}) - \Lambda(T(x,\Lambda(R(x,0)'\widehat{\gamma}))'\widehat{\beta}_1(y_1))\Lambda(R(x,0)'\widehat{\gamma})}{\Lambda(R(x,1)'\widehat{\gamma}) - \Lambda(R(x,0)'\widehat{\gamma})},$$

$$\widehat{F}_{Y_0|X,D_0,D_1}(y_0|x,0,1) = \frac{\Lambda(T(x,\Lambda(R(x,0)'\widehat{\gamma}))'\widehat{\beta}_0(y_0))\left(1-\Lambda(R(x,0)'\widehat{\gamma})\right) - \Lambda(T(x,\Lambda(R(x,1)'\widehat{\gamma}))'\widehat{\beta}_0(y_0))\left(1-\Lambda(R(x,1)'\widehat{\gamma})\right)}{\Lambda(R(x,1)'\widehat{\gamma}) - \Lambda(R(x,0)'\widehat{\gamma})},$$

by the delta method it follows that

$$\sqrt{n}\begin{pmatrix} \widehat{F}_{Y_1|X,D_0,D_1}(y_1|x,0,1) - F_{Y_1|X,D_0,D_1}(y_1|x,0,1) \\ \widehat{F}_{Y_0|X,D_0,D_1}(y_0|x,0,1) - F_{Y_0|X,D_0,D_1}(y_0|x,0,1) \end{pmatrix}$$
$$\rightsquigarrow \begin{pmatrix} \varphi^{L\prime}_{\gamma,\beta_1(\cdot)}\left(W_\gamma, -J_1(y_1)^{-1}\left(J_{1p}(y_1)J_p^{-1}W_\gamma + W_1(y_1)\right)\right) \\ \varphi^{L\prime}_{\gamma,\beta_0(\cdot)}\left(W_\gamma, -J_0(y_0)^{-1}\left(J_{0p}(y_0)J_p^{-1}W_\gamma + W_0(y_0)\right)\right) \end{pmatrix} \text{ in } \ell^\infty(\mathcal{YX})^2,$$

where

$$\begin{aligned}
&\varphi^{L\prime}_{\gamma,\beta_1(\cdot)}(\eta,\alpha_1) \\
&= \frac{\begin{array}{c} F_{Y|X,p(X,Z),D}(y_1|x,p(x,1),1)\lambda(R(x,1)'\gamma)R(x,1)' - F_{Y|X,p(X,Z),D}(y_1|x,p(x,0),1)\lambda(R(x,0)'\gamma)R(x,0)' \\ +p(x,1)\lambda(T(x,p(x,1)))'\beta_1(y_1))\frac{\partial T(x,p(x,1))'\beta_1(y_1)}{\partial p}\lambda(R(x,1)'\gamma)R(x,1)' \\ -p(x,0)\lambda(T(x,p(x,0)))'\beta_1(y_1))\frac{\partial T(x,p(x,0))'\beta_1(y_1)}{\partial p}\lambda(R(x,0)'\gamma)R(x,0)' \end{array}}{p(x,1)-p(x,0)}\eta \\
&- \frac{F_{Y|X,p(X,Z),D}(y_1|x,p(x,1),1)p(x,1) - F_{Y|X,p(X,Z),D}(y_1|x,p(x,0),1)p(x,0)}{[p(x,1)-p(x,0)]^2}\left[\lambda(R(x,1)'\gamma)R(x,1)' - \lambda(R(x,0)'\gamma)R(x,0)'\right]\eta \\
&+ \left[\frac{p(x,1)\lambda(T(x,p(x,1))'\beta_1(y_1))}{p(x,1)-p(x,0)}T(x,p(x,1))' - \frac{p(x,0)\lambda(T(x,p(x,0))'\beta_1(y_1))}{p(x,1)-p(x,0)}T(x,p(x,0)))'\right]\alpha_1(y_1),
\end{aligned}$$

and

$$
\begin{aligned}
&\varphi^{L\prime}_{\gamma,\beta_0(\cdot)}(\eta,\alpha_0)\\
&= \frac{\begin{array}{c} F_{Y|X,p(X,Z),D}(y_0|x,p(x,1),0)\lambda(R(x,1)'\gamma)R(x,1)' - F_{Y|X,p(X,Z),D}(y_0|x,p(x,0),0)\lambda(R(x,0)'\gamma)R(x,0)' \\ +\left[1-p(x,0)\right]\lambda(T(x,p(x,0))'\beta_0(y_0))\frac{\partial T(x,p(x,0))'\beta_0(y_0)}{\partial p}\lambda(R(x,0)'\gamma)R(x,0)' \\ -\left[1-p(x,1)\right]\lambda(T(x,p(x,1))'\beta_0(y_0))\frac{\partial T(x,p(x,1))'\beta_0(y_0)}{\partial p}\lambda(R(x,1)'\gamma)R(x,1)' \end{array}}{p(x,1)-p(x,0)}\eta\\
&- \frac{F_{Y|X,p(X,Z),D}(y_0|x,p(x,0),0)\left[1-p(x,0)\right] - F_{Y|X,p(X,Z),D}(y_0|x,p(x,1),0)\left[1-p(x,1)\right]}{\left[p(x,1)-p(x,0)\right]^2}\\
&\qquad\qquad \cdot\left[\lambda(R(x,1)'\gamma)R(x,1)' - \lambda(R(x,0)'\gamma)R(x,0)'\right]\eta\\
&+ \left[\frac{\left[1-p(x,0)\right]\lambda(T(x,p(x,0))'\beta_0(y_0))}{p(x,1)-p(x,0)}T(x,p(x,0))' - \frac{\left[1-p(x,1)\right]\lambda(T(x,p(x,1))'\beta_0(y_0))}{p(x,1)-p(x,0)}T(x,p(x,1)))'\right]\alpha_0(y_0),
\end{aligned}
$$

which can be simplified to the form in the main text. By the proof of Theorem 4.1(2) of Chernozhukov et al. (2013), the results in the theorem follow. ■

Supplementary Materials

1. Derivation of Weights for the QTT and QTUT

The following derivation is similar as that in Section 4 of Heckman and Vytlacil (2001a). We start from $F_{Y_d|X,U_D}(y_d|x,u_D)$ since it can be identified from our data directly. First,

$$F_{Y_d|X,U_D,D}(y_d|x,p,1) = \frac{1}{p}\int_0^p F_{Y_d|X,U_D}(y_d|x,u_D)du_D.$$

Second,

$$F_{Y_d|X,D}(y_d|x,1) = \int_0^1 F_{Y_d|X,U_D,D}(y_d|x,p,1)dF_{p(X,Z)|X,D}(p|x,1).$$

Using Bayes' rule, it follows that

$$dF_{p(X,Z)|X,D}(p|x,1) = \frac{P(D=1|X=x,p(X,Z)=p)}{P(D=1|X=x)}dF_{p(X,Z)|X}(p|x).$$

Since $P(D=1|X=x,p(X,Z)=p)=p$, it follows that

$$F_{Y_d|X,D}(y_d|x,1) = \frac{1}{P(D=1|X=x)}\int_0^1\left[\int_0^p F_{Y_d|X,U_D}(y_d|x,u_D)du_D\right]dF_{p(X,Z)|X}(p|x).$$

Note further that since $P(D=1|X=x) = E\left[p(X,Z)|X=x\right] = \displaystyle\int_0^1 \left(1-F_{p(X,Z)|X}(t|x)\right)dt$,

$$\begin{aligned}
F_{Y_d|X,D}(y_d|x,1) &= \frac{1}{P(D=1|X=x)}\int_0^1\left[\int_0^1 1(u_D\le p)F_{Y_d|X,U_D}(y_d|x,u_D)du_D\right]dF_{p(X,Z)|X}(p|x)\\
&= \frac{1}{\displaystyle\int_0^1\left(1-F_{p(X,Z)|X}(t|x)\right)dt}\int_0^1\left[\int_0^1 1(u_D\le p)dF_{p(X,Z)|X}(p|x)\right]F_{Y_d|X,U_D}(y_d|x,u_D)du_D\\
&= \int_0^1 F_{Y_d|X,U_D}(y_d|x,u_D)\frac{1-F_{p(X,Z)|X}(u_D|x)}{\displaystyle\int_0^1\left(1-F_{p(X,Z)|X}(t|x)\right)dt}du_D\\
&= \int_0^1 F_{Y_d|X,U_D}(y_d|x,u_D)h_{TT}(x,u_D)du_D,
\end{aligned}$$

where $h_{TT}(x,u_D) = \left(1-F_{p(X,Z)|X}(u_D|x)\right)/E[p(X,Z)|X=x]$ is the same weight as in Heckman and Vytlacil (2001a). Similarly,

$$F_{Y_d|X,D}(y_d|x,0) = \int_0^1 F_{Y_d|X,U_D}(y_d|x,u_D)h_{TUT}(x,u_D)du_D,$$

where $h_{TUT}(x,u_D) = F_{p(X,Z)|X}(u_D|x)/E[1-p(X,Z)|X=x]$. Third, from $F_{Y_d|X,D}(y_d|x,1)$ and $F_{Y_d|X,D}(y_d|x,0)$, we can identify

$$\Delta_\tau^{QTT}(x) = F^{-1}_{Y_1|X,D}(\tau|x,1) - F^{-1}_{Y_0|X,D}(\tau|x,1)$$

and

$$\Delta_\tau^{QTUT}(x) = F^{-1}_{Y_1|X,D}(\tau|x,0) - F^{-1}_{Y_0|X,D}(\tau|x,0).$$

2. Another Counterexample for Section 3.3

Example 3 *Suppose* $Y_d \in [0, y^u(x)]$ *conditional on* $X = x$. $p^{\sup}_{x1} \equiv P(Y = 0|X = x, p(X,Z) = p^{\sup}_x, D = 1) \in (0,1)$ *and* $p^{\inf}_{x0} \equiv P(Y = 0|X = x, p(X,Z) = p^{\inf}_x, D = 0) \in (0,1)$. $Y|X = x, p(X,Z) = p^{\sup}_x, D = 1$ *and* $Y|X = x, p(X,Z) = p^{\inf}_x, D = 0$ *are continuously distributed with a positive density on* $(0, y^u(x))$. *First check the bound for* $Q_{Y_1|X}(\tau|x)$*:*

$$L^1_\tau(x) = \begin{cases} Q_{Y|X,p(X,Z),D}\left(1 - \frac{1-\tau}{p^{\sup}_x}\middle| x, p^{\sup}_x, 1\right), & \text{if } p^{\sup}_x > 1-\tau \text{ and } 1 - \frac{1-\tau}{p^{\sup}_x} > p^{\sup}_{x1}, \\ 0, & \text{if } p^{\sup}_x \le 1-\tau \text{ or } [p^{\sup}_x > 1-\tau \text{ and } 1 - \frac{1-\tau}{p^{\sup}_x} \le p^{\sup}_{x1}], \end{cases}$$

$$R^1_\tau(x) = \begin{cases} 0, & \text{if } p^{\sup}_x \ge \tau \text{ and } \frac{\tau}{p^{\sup}_x} \le p^{\sup}_{x1}, \\ Q_{Y|X,p(X,Z),D}\left(\frac{\tau}{p^{\sup}_x}\middle| x, p^{\sup}_x, 1\right) & \text{if } p^{\sup}_x \ge \tau \text{ and } \frac{\tau}{p^{\sup}_x} > p^{\sup}_{x1}, \\ y^u(x), & \text{if } p^{\sup}_x < \tau. \end{cases}$$

When $\max\left\{1-\tau, \frac{\tau}{p^{\sup}_{x1}}\right\} < p^{\sup}_x \le \frac{1-\tau}{1-p^{\sup}_{x1}}$ *or* $\frac{\tau}{p^{\sup}_{x1}} \le p^{\sup}_x \le 1-\tau$, $L^1_\tau(x) = R^1_\tau(x) = 0$. *Otherwise,* $L^1_\tau(x) < R^1_\tau(x)$ *unless* $p^{\sup}_x = 1$. *Similarly, when* $1 - \frac{1-\tau}{1-p^{\inf}_{x0}} \le p^{\inf}_x < \min\left\{\tau, 1 - \frac{\tau}{p^{\inf}_{x0}}\right\}$ *or* $\tau \le p^{\inf}_x \le 1 - \frac{\tau}{p^{\inf}_{x0}}$, $L^0_\tau(x) = R^0_\tau(x) = 0$. *Otherwise,* $L^0_\tau(x) < R^0_\tau(x)$ *unless* $p^{\inf}_x = 0$. *In Figure 6, only the blue area augmented by* $\{p^{\sup}_x = 1\}$ $(\{p^{\inf}_x = 0\})$ *is the combination of* $p^{\sup}_x$ $(p^{\inf}_x)$ *and* $p^{\sup}_{x1}$ $(p^{\inf}_{x0})$ *for point identification of* $Q_{Y_1|X}(\tau|x)$ $(Q_{Y_0|X}(\tau|x))$ *when* $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$. *Obviously,* $p^{\inf}_x = 0$ *and* $p^{\sup}_x = 1$ *are not necessary for point identification of* $\Delta^{QTE}_\tau(x)$. *Only if* $p^{\sup}_{x1} \le \tau$ *and* $p^{\inf}_{x0} \le \tau$, $p^{\inf}_x = 0$ *and* $p^{\sup}_x = 1$ *are necessary.* $\square$

3. Derivation of Equation (23)

The following derivation is similar as that in Appendix B of Heckman and Vytlacil (2005).

$$\begin{aligned} F_{Y_{a'}|X_{a'}}(y_d|x) &= \int_0^1 F_{Y_{a'}|X_{a'},p_{a'}}(y_d|x,p)dF_{p_{a'}|X_{a'}}(p|x) \\ &= \int_0^1 \left[\int_0^1 1_{[0,p]}(u_D)F_{Y_{1,a'}|X_{a'},U_{D,a'}}(y_d|x,u_D) + 1_{(p,1]}(u_D)F_{Y_{0,a'}|X_{a'},U_{D,a'}}(y_d|x,u_D)du_D\right] dF_{p_{a'}|X_{a'}}(p|x) \\ &= \int_0^1 \left[\int_0^1 1_{[u_D,1]}(p)F_{Y_{1,a'}|X_{a'},U_{D,a'}}(y_d|x,u_D) + 1_{(0,u_D]}(p)F_{Y_{0,a'}|X_{a'},U_{D,a'}}(y_d|x,u_D)dF_{p_{a'}|X_{a'}}(p|x)\right] du_D \\ &= \int_0^1 \left[\left(1 - F_{p_{a'}|X_{a'}}(u_D|x)\right) F_{Y_{1,a'}|X_{a'},U_{D,a'}}(y_d|x,u_D) + F_{p_{a'}|X_{a'}}(u_D|x)F_{Y_{0,a'}|X_{a'},U_{D,a'}}(y_d|x,u_D)\right] du_D \\ &= \int_0^1 \left[\left(1 - F_{p_{a'}|X_{a'}}(u_D|x)\right) F_{Y_{1,a}|X_a,U_{D,a}}(y_d|x,u_D) + F_{p_{a'}|X_{a'}}(u_D|x)F_{Y_{0,a}|X_a,U_{D,a}}(y_d|x,u_D)\right] du_D \end{aligned}$$

where $p_{a'} = p_{a'}(X_{a'}, Z_{a'})$, the third equality is from Fubini's theorem, and the last equality is from assumption (A7).

4. Moment Conditions for the QRE and IV-QRE

If we use the QRE to estimate the QTE, $\Delta_\tau^{QTE}(x)$ is the β such that for some α, $Q_\tau\left(Y-\alpha-D\beta|X=x\right)=0$. The moment condition to identify (α,β) is

$$E\left[\left.\begin{pmatrix}1\\ D\end{pmatrix}(\tau-1(Y\le\alpha+D\beta))\right|X=x\right]=0,$$

i.e.,

$$\begin{pmatrix}\tau\\ \tau E[p(X,Z)|X=x]\end{pmatrix}=\begin{pmatrix}P\left(Y\le\alpha+D\beta|X=x\right)\\ E[D1(Y\le\alpha+D\beta)|X=x]\end{pmatrix}.$$

We calculate these moment conditions in our framework. First,

$$P\left(Y\le\alpha+D\beta|X=x\right)=\int P\left(Y\le\alpha+D\beta|X=x,p(X,Z)=p\right)dF_{p(X,Z)|X}(p|x),$$

where

$$\begin{aligned}&P\left(Y\le\alpha+D\beta|X=x,p(X,Z)=p\right)\\ &=P\left(Y\le\alpha+D\beta|X=x,p(X,Z)=p,D=1\right)P(D=1|X=x,p(X,Z)=p)\\ &\quad+P\left(Y\le\alpha+D\beta|X=x,p(X,Z)=p,D=0\right)P(D=0|X=x,p(X,Z)=p)\\ &=P\left(Y_1\le\alpha+\beta|X=x,U_D\le p\right)p+P\left(Y_0\le\alpha|X=x,U_D>p\right)(1-p)\\ &=\int_0^p F_{Y_1|X,U_D}(\alpha+\beta|x,u_D)du_D+\int_p^1 F_{Y_0|X,U_D}(\alpha|x,u_D)du_D.\end{aligned}$$

Second,

$$E[D1(Y\le\alpha+D\beta)|X=x]=\int E[D1(Y\le\alpha+D\beta)|X=x,p(X,Z)=p]dF_{p(X,Z)|X}(p|x),$$

where

$$\begin{aligned}&E[D1(Y\le\alpha+D\beta)|X=x,p(X,Z)=p]\\ &=E[D1(Y\le\alpha+D\beta)|X=x,p(X,Z)=p,D=1]P(D=1|X=x,p(X,Z)=p)\\ &\quad+E[D1(Y\le\alpha+D\beta)|X=x,p(X,Z)=p,D=0]P(D=0|X=x,p(X,Z)=p)\\ &=P\left(Y_1\le\alpha+\beta|X=x,U_D\le p\right)p\\ &=\int_0^p F_{Y_1|X,U_D}(\alpha+\beta|x,u_D)du_D\end{aligned}$$

In summary,

$$\begin{pmatrix}\tau\int(1-p)\,dF_{p(X,Z)|X}(p|x)\\ \tau\int p dF_{p(X,Z)|X}(p|x)\end{pmatrix}=\begin{pmatrix}\int\left[\int_p^1 F_{Y_0|X,U_D}(\alpha|x,u_D)du_D\right]dF_{p(X,Z)|X}(p|x)\\ \int\left[\int_0^p F_{Y_1|X,U_D}(\alpha+\beta|x,u_D)du_D\right]dF_{p(X,Z)|X}(p|x)\end{pmatrix},$$

which is a very nonlinear system of functions. If $F_{Y_0|X,U_D}(y_0|x,u_D)$ and $F_{Y_1|X,U_D}(y_1|x,u_D)$ do not depend on u_D, or U_D is independent of (Y_0,Y_1) given X, or the unconfoundedness assumption holds, then β will identify $\Delta_\tau^{QTE}(x)$; otherwise, won't.

The IV-QRE of Chernozhukov and Hansen (2006) is estimating $q(1,x,\tau)-q(0,x,\tau)$, where $q(d,x,\tau)$ is defined by $E\left[1\left(Y\le q(D,X,\tau)\right)-\tau|X=x,Z\right]=0$. If use $p(X,Z)$ as the instrument, then the corresponding moment equations are

$$\begin{pmatrix} \tau \\ \tau E[p(X,Z)|X=x] \end{pmatrix} = \begin{pmatrix} \int \left[\int_0^p F_{Y_1|X,U_D}(\alpha+\beta|x,u_D)du_D + \int_p^1 F_{Y_0|X,U_D}(\alpha|x,u_D)du_D \right] dF_{p(X,Z)|X}(p|x) \\ \int p \left[\int_0^p F_{Y_1|X,U_D}(\alpha+\beta|x,u_D)du_D + \int_p^1 F_{Y_0|X,U_D}(\alpha|x,u_D)du_D \right] dF_{p(X,Z)|X}(p|x) \end{pmatrix}$$

where the first moment condition is from above, and the second moment condition is from

$$E[p(X,Z)1(Y \leq \alpha + D\beta)|X=x] = \int E[p(X,Z)1(Y \leq \alpha + D\beta)|X=x, p(X,Z)=p] dF_{p(X,Z)|X}(p|x),$$

and

$$\begin{aligned} & E[p(X,Z)1(Y \leq \alpha + D\beta)|X=x, p(X,Z)=p] \\ = {} & pE[1(Y \leq \alpha + D\beta)|X=x, p(X,Z)=p, D=1]P(D=1|X=x, p(X,Z)=p) \\ & + pE[1(Y \leq \alpha + D\beta)|X=x, p(X,Z)=p, D=0]P(D=0|X=x, p(X,Z)=p) \\ = {} & pE[1(Y_1 \leq \alpha + \beta)|X=x, p(X,Z)=p, U_D \leq p]p + pE[1(Y_0 \leq \alpha)|X=x, p(X,Z)=p, U_D > p](1-p) \\ = {} & p \left[\int_0^p F_{Y_1|X,U_D}(\alpha+\beta|x,u_D)du_D + \int_p^1 F_{Y_0|X,U_D}(\alpha|x,u_D)du_D \right]. \end{aligned}$$

In the unconfoundedness case, the system reduces to

$$\begin{pmatrix} \tau \\ \tau E[p(X,Z)|X=x] \end{pmatrix} = \begin{pmatrix} \int \left[pF_{Y_1|X}(\alpha+\beta|x) + (1-p)F_{Y_0|X}(\alpha|x) \right] dF_{p(X,Z)|X}(p|x) \\ \int p \left[pF_{Y_1|X}(\alpha+\beta|x) + (1-p)F_{Y_0|X}(\alpha|x) \right] dF_{p(X,Z)|X}(p|x) \end{pmatrix}$$

which can be satisfies by $\alpha = q(0,x,\tau)$ and $\alpha + \beta = q(1,x,\tau)$.

In the example of the main context, we assume that α is known and numerically solve the moment conditions. Specifically, for the first specification where only the selection effect exists, the moment condition for the QRE is

$$\frac{\tau}{2} = \int \left[\int_0^p \Phi \left(\frac{\frac{\Phi^{-1}(\tau)+\beta}{2} - 0.5\Phi^{-1}(u_D)}{\sqrt{0.75}} \right) du_D \right] dp,$$

and for the IV-QRE is

$$\frac{\tau}{2} = \int p \left[\int_0^p \Phi \left(\frac{\frac{\Phi^{-1}(\tau)+\beta}{2} - 0.5\Phi^{-1}(u_D)}{\sqrt{0.75}} \right) du_D + \int_p^1 \Phi \left(\frac{\Phi^{-1}(\tau) - 0.5\Phi^{-1}(u_D)}{\sqrt{0.75}} \right) du_D \right] dp;$$

for the second specification where only the essential heterogeneity also exists, the moment condition for the QRE is

$$\frac{\tau}{2} = \int \left[\int_0^p \Phi \left(\frac{\frac{\sqrt{7}\Phi^{-1}(\tau)+\beta-\Phi^{-1}(u_D)}{2} - 0.5\Phi^{-1}(u_D)}{\sqrt{0.75}} \right) du_D \right] dp,$$

and for the IV-QRE is

$$\frac{\tau}{2} = \int p \left[\int_0^p \Phi \left(\frac{\frac{\sqrt{7}\Phi^{-1}(\tau)+\beta-\Phi^{-1}(u_D)}{2} - 0.5\Phi^{-1}(u_D)}{\sqrt{0.75}} \right) du_D + \int_p^1 \Phi \left(\frac{\sqrt{7}\Phi^{-1}(\tau) - 2\Phi^{-1}(u_D) - 0.5\Phi^{-1}(u_D)}{\sqrt{0.75}} \right) du_D \right] dp.$$