Appl. Math. J. Chinese Univ.
2020, 35(2): 220-243

Some recent developments in modeling quantile treatment effects

TANG Sheng-fang

Abstract. This paper provides a selective review of the recent developments on econometric/statistical modeling in quantile treatment effects under both selection on observables and on unobservables. First, we discuss identification, estimation and inference of quantile treatment effects under the framework of selection on observables. Then, we consider the case where the treatment variable is endogenous or self-selected, for which an instrumental variable method provides a powerful tool to tackle this problem. Finally, some extensions are discussed to the data-rich environments, to the regression discontinuity design, and some other approaches to identify quantile treatment effects are also discussed. In particular, some future research works in this area are addressed.

§1 Introduction

In the last three decades, a vast amount of literature in economics and other social sciences has been devoted to seeking to identify the causal effect of a treatment or policy on economic and other social programs and a large fraction of this literature focuses mainly on the mean treatment effects, which include (but not limited to) the average treatment effect (ATE) and average treatment effect on the treated; see, for example, Rosenbaum and Rubin (1983, 1985), Heckman, Ichimura and Todd (1997, 1998), Hahn (1998), Hirano, Imbens and Ridder (2003), Abadie and Imbens (2006, 2016), and among others. Although very important, ATE sometimes reveals only a partial picture of the effects of a treatment. It is common that the treatment may not affect the mean of the outcome distribution but may alter its dispersion or change its shape. Therefore, in applications, the effect of a treatment on the entire distribution of outcome is of interest and importance in both academic and applications. It is well documented in the literature that the distribution of the outcome variable may change in many ways which are not disclosed or are only incompletely disclosed by an investigation of averages. For example,

Received: 2019-12-09. Revised: 2020-03-23.
MR Subject Classification: 62F02, 62G02.
Keywords: average treatment effect, endogeneity, quantile treatment effect, regression discontinuity design.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-020-3980-y.
Supported by the National Natural Science Foundation of China #71631004 (Key Project), the National Science Fund for Distinguished Young Scholars #71625001 and the scholarship from China Scholarship Council (CSC) under the Grant CSC N201806310088.

when evaluating the effects of a policy on income, the income distribution may become more compressed or the lower-tail inequality may decrease while the upper-tail inequality increases. The typical example to this regard is that during the 1992 presidential election, using the yearly Current Population Survey data, the Democrats claimed that "the rich got richer and the poor got poorer" during 1980 to 1992 (the Republican administrations). Indeed, this phenomena could be illustrated by using the so-called quantile treatment effect (QTE) method to support this claim. Hence, instead of considering ATE, applied economists and/or policy makers are more interested in distributional treatment effect or QTE, originally and seminally proposed in the well-known statistics literature by Doksum (1974) and Lehmann (1974). Actually, for any fixed percentile, QTE is defined as the difference between the quantiles of the marginal potential distributions of the treatment and control responses, which offers a great and useful tool to discover the effects on the entire distribution of outcome. In addition, QTE also allows us to investigate numerous interesting hypotheses. For example, if the treatment variable only shifts the location of the distribution of the potential outcome, the QTE is independent of the quantile index. In the special case of a binary treatment, if the distribution of the potential outcome for the treated has first-order stochastic dominance over the distribution of the potential outcome for the control, then QTE function is above the zero line. In summary, the QTE function can provide a powerful tool to summarize the causal effect of a treatment or policy on the marginal distribution of the outcome variable of interest.

Depending on the type of endogeneity of the treatment variable, we can distinguish whether selection to treatment is based on observable variables or on unobservable variables. Selection on observables assumption is often referred to as the unconfoundedness assumption in the statistics literature or as exogenous treatment choice in the econometrics literature, which assumes that given a set of observed covariates, individuals are randomly assigned either to the treatment group or to the control group. In contrast, selection on unobservables is commonly regarded as endogenous treatment choice. In the case where the treatment variable is exogenous conditional on covariates, various approaches have been suggested recently to identify and estimate the conditional or unconditional QTE; see, for example, to name just a few, Firpo (2007), Frölich (2007), Melly (2006), Donald and Hsu (2014), and the references therein. Specifically, our focus in this review is on the weighting estimator of Firpo (2007) and a further discussion of this method is provided in a later section. However, in observational studies, the treatment of interest is often endogenous or self-selected. For example, individuals choose whether to participate in government-subsidized savings plans. Similarly, in trying to estimate the effect of the participation into the 401(K) programs on financial assets holdings, one must face the fact that the unobservable saver heterogeneity may be an unmeasured confounder in the treatment-outcome relationship. Endogeneity of the treatment variable renders conventional quantile regression inconsistent and hence inappropriate for recovering the causal effects of variables on the quantiles of outcomes of interest. In this setup, an instrumental variable (IV) approach provides a powerful tool to address this problem. In this paper, our survey is about the IV approach which is based on the framework developed by Imbens and Angrist (1994) and Abadie (2003). This approach was first proposed by Imbens and Angrist (1994) for average treatment effects and then extended by Imbens and Rubin (1997), Abadie (2002), and Abadie, Angrist and Imbens (2002) to distributional treatment effects. It is well known in the literature that in the setup where both the treatment and the assignment (instrument) are binary, the largest population for which the effects can be point identified is the individuals who respond to a change in the value of the instrument. These individuals are called compliers because they comply with the instrument. Therefore, under this setup, the main estimation is the QTE for the compliers, which is called as the local quantile treatment effect (LQTE) because it corresponds to the QTE for the complier sub-populations. It is worth stressing that this terminology is not related to nonparametric methods that are local with respect to the covariates and it is in analogy to the local average treatment effect introduced by Imbens and Angrist (1994).

It is well documented in the literature that treatment effects can be heterogeneous across different individuals; see, for example, the papers by Heckman and Robb (1985) and Heckman, Smith and Clements (1997). Given the heterogeneity of individual effects, except for the quantile treatment effect for the entire population, researchers may also be of interest to estimate the effect of a treatment or policy on outcome of interest in various sub-populations defined by the possible values of some covariates. For example, researchers may be interested in estimating the effect of maternal smoking during pregnancy on the birth weight of her child under different mother's age or estimating the effect of the participation into the 401(K) programs on financial assets holdings conditional on different ages or income. Recently, Abrevaya, Hsu and Lieli (2015) and Lee, Okui and Whang (2017) considered the conditional average treatment effect model to characterize this heterogeneous effect across different sub-populations. As aforementioned, researchers sometimes may be more interested in the distributional aspects of outcome variable. To this end, a partially conditional quantile treatment effect model (PCQTE) is proposed by Cai, Fang, Lin and Tang (2019a, 2019b) for nonparametric model and Tang, Cai, Fang and Lin (2019) for parametric model to capture the heterogeneously distributional impacts of treatment participation across different sub-populations. Furthermore, to test whether there exits heterogeneity for PCQTE across different sub-populations, Cai, Fang, Lin and Tang (2019a) proposed a consistent test method based on the Cramér-von Mises criterion; see Sections 2.3 and 3.3 for more discussions.

Our main goal in this paper is to provide a selective survey over the methodological advancements for quantile treatment effects. Other recent surveys on the estimation of causal treatment effects for program evaluation from different perspectives and disciplines include, but not limited to, the papers by Angrist and Pischke (2008, 2014), Athey and Imbens (2017), Imbens and Rubin (2015), Imbens and Wooldridge (2009), Lee (2016) and Melly and Wüthrich (2017), Liu, Cai, Fang and Lin (2020), and among many others.

The rest of this paper is organized as follows. Section 2 reviews assumptions, the identification, estimation and inference of the quantile treatment effects under the framework of selection on observables. In Section 3, the review goes to the case where the treatment variable is endogenous. After briefly summarizing some basics, the focus is on the particularities of the quantile estimations and their implications for the identification of the effects in setups without and with covariates. In Section 4, some extensions such as the data-rich environment, the regression discontinuity design, and some other procedures to identify QTE are addressed. Section 5 concludes.

§2 QTE under Selection on Observables

In this section, we review and discuss some recent developments in quantile treatment effects under the identifying restriction that selection to treatment is based on observable characteristics. This setup is about a binary treatment and a scalar outcome. Section 2.1 discusses the model framework, assumptions, and identification of the parameters of interest while Section 2.2 is concerned with estimation and inference. Section 2.3 extends the model framework to estimate the partially conditional quantile treatment effect.

2.1 Model Framework

Here, our presentation is to follow the conventional potential outcome framework initiated by Rubin (1974) in the statistics literature. Specifically, the focus of the analysis is to learn the effect of a binary treatment, represented by the indicator variable D, on some observed outcome variable, denoted by Y. More specifically, $D = 1$ denotes exposure to the treatment, while $D = 0$ denotes lack of exposure to the treatment. Following Rubin (1974), Imbens and Rubin (2015) and others, potential outcomes are defined as follows. Let $Y(1)$ denote the potential outcome under exposure to treatment, and $Y(0)$ stand for the potential outcome under no exposure to treatment. Therefore, the potential outcomes $Y(0)$ and $Y(1)$ have the following relationship with the realized outcome Y; that is, $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$, from which one can see easily that a strong assumption is implicitly needed. Namely, the realized outcome for each particular individual relies only on the value of the treatment variable of that individual, and not on the treatment variable or on outcome values of other individuals. This assumption, first introduced by Cox (1958), is often referred to "Stable Unit Treatment Value Assumption" (SUTVA) by Rubin (1980, 1990). Formally, throughout the paper, the following assumption is imposed.

Assumption 2.1. *(Stable Unit Treatment Value Assumption) The distribution of potential outcomes for one individual is assumed to be independent of potential treatment status of other individuals.*

Remark 2.1: Assumption 2.1 rules out the possibility of interference between individuals and allows the possibility to consider the potential outcomes of one individual to be independent of other individuals' treatment status. SUTVA is a strong assumption and thus may be violated in many practical settings, such as the educational setting with peer effects and the setting where the treatment is a vaccine that immunizes individual against a contagious disease. Recently, several contributions show that particular treatment effects can be point identified under specific relaxations of the SUTVA assumption. For example, Vazquez-Bare (2019a) considered a setting where treatment is randomly assigned and interference can occur within but not between the groups in which individuals are clustered. Furthermore, the author provided conditions under which direct and causal spillover effects can be identified and estimated consistently based on the potential outcomes framework. Similarly, Vazquez-Bare (2019b) analyzed the direct and causal spillover effects within an instrumental variables framework when treatment is not randomly assigned and provided conditions under which parameters of interest can be identified.

See, Vazquez-Bare (2019a, 2019b) and the references therein for more details. In this paper, the SUTVA assumption is imposed as in most treatment effect literature. It is important to note that without Assumption 2.1, the potential outcomes of each individual would depend on the treatments of all individuals in the population. Therefore, it is impossible to identify any treatment effect without imposing some restriction on the dependence between individuals.

In addition to treatment variable and outcome variable of interest, suppose that one also observes a vector of covariates X for each individual, which are predetermined relative to the treatment and oftentimes contain characteristics of the individuals measured before the treatment is known. With the above preparation at hand, one can define the quantile treatment effect and its counterpart for the treated. Formally, for any given quantile level $\tau \in (0,1)$, the quantile treatment effect is defined as

$$\Delta_\tau = q_{1,\tau} - q_{0,\tau}, \tag{1}$$

where $q_{j,\tau} = \inf\{q : P(Y(j) \leq q) \geq \tau\}$ is the τ-th quantile of the distribution of $Y(j)$ for $j = 0$ and 1. Alternatively, one may also be interested in estimating the quantile treatment effect on the treated, given by

$$\Delta_{\tau|D=1} = q_{1,\tau|D=1} - q_{0,\tau|D=1}, \tag{2}$$

where $q_{j,\tau|D=1} = \inf\{q : P(Y(j) \leq q|D = 1) \geq \tau\}$ for $j = 0$ and 1, is the τ-th quantile of the distribution of $Y(j)$ conditional on the treatment state $D = 1$. Due to similarity, in what follows, our presentation is only for QTE.

Generally, the key problem is that for each individual in the population, one observes either $Y(1)$ or $Y(0)$, but never both, so that, without further additional restrictions, the parameters of interest defined in (1) or (2) can not be identified and thus they are not consistently estimable. To solve the identification problem, in this section, it needs to impose the identification restriction that selection to treatment is based on observable variables (exogeneity assumption). In other words, it is assumed that given a set of observed covariates X, individuals are randomly assigned either to the treatment group or to the control group. To be specific, the following strong ignorability assumption as in Rosenbaum and Rubin (1983) is imposed.

Assumption 2.2. *Let* $(Y(1), Y(0), X, D)$ *have a joint distribution. Then, the following two conditions hold:*
(i) (Unconfounded Treatment Assignment): $(Y(1), Y(0))$ *is jointly independent from* D *conditional on covariates* X*.*
(ii) (Common Support): For some $0 < \underline{c} < \bar{c} < 1$, $0 < \underline{c} \leq p(x) = P(D = 1|X = x) \leq \bar{c} < 1$, *where* $p(x)$ *is the propensity score function.*

Remark 2.2: Assumption 2.2 is standard in the treatment effect literature. Part (i) of Assumption 2.2, which was introduced by Rosenbaum and Rubin (1983), states that, conditional on observables X, treatment assignment variable D is independent of potential outcomes $Y(1)$ and $Y(0)$. Although it is a strong assumption, it has been used in many studies on the effect of treatments or programs; see, for example, Abadie and Imbens (2006, 2016), Hirano et al. (2003), Heckman, Ichimura, Smith and Todd (1998), Dehejia and Wahba (1999), Firpo (2007) and among others. Part (ii) of Assumption 2.2 ensures that in the population for all values of X, there are both treatment and control individuals.

Clearly, Assumption 2.2 implies that the distribution of the potential outcomes $Y(1)$ and $Y(0)$ can be identified by

$$F_{Y(1)}(y) = E\big[DI\{Y \le y\}/p(X)\big], \tag{3}$$

and

$$F_{Y(0)}(y) = E\big[(1-D)I\{Y \le y\}/\big(1-p(X)\big)\big]. \tag{4}$$

Consequently, by definition, the QTE parameter Δ_τ can be identified as the difference between two quantile functions:

$$\Delta_\tau = F^{-1}_{Y(1)}(\tau) - F^{-1}_{Y(0)}(\tau), \tag{5}$$

where $F^{-1}_{Y(j)}(\tau) = \inf\{y : F_{Y(j)}(y) \ge \tau\}$ for $j = 0$ and 1. Alternatively, as shown in Firpo (2007), the QTE parameter Δ_τ can be identified directly from Assumption 2.2. Specifically, Firpo (2007) proved that, under Assumption 2.2, the quantile functions of the potential outcome distributions $q_{1,\tau}$ and $q_{0,\tau}$ can be identified by the following moment conditions,

$$E\big[DI\{Y \le q_{1,\tau}\}/p(X) - \tau\big] = 0, \tag{6}$$

and

$$E\big[(1-D)I\{Y \le q_{0,\tau}\}/\big(1-p(X)\big) - \tau\big] = 0. \tag{7}$$

Consequently, the identification of QTE parameter Δ_τ is a straightforward consequence from (6) and (7).

2.2 Estimation Procedures

From the identification results presented above, several procedures have been suggested for estimating Δ_τ. In particular, based on the identification results displayed in (3), (4) and (5), Donald and Hsu (2014) proposed an approach to estimate Δ_τ with three steps. First, obtain the estimator $\widehat{p}(x)$ for the unknown propensity score function $p(x)$ by nonparametric power series estimation. At the second step, obtain the estimators $\widehat{F}_{Y(1)}(y)$ and $\widehat{F}_{Y(0)}(y)$ for the potential outcome distributions $F_{Y(1)}(y)$ and $F_{Y(0)}(y)$. Finally, take inverse and difference to obtain the estimator for Δ_τ. Under some regularity conditions, Donald and Hsu (2014) showed that the proposed estimator converges to a mean zero Gaussian process.

In this section, we describe in detail another estimation approach proposed by Firpo (2007) which is a reweighed version of the procedure proposed by Koenker and Bassett (1978) for a quantile regression setting. Note that using the identification results in (6) and (7), their sample version provides a natural estimator for the quantile parameters $q_{1,\tau}$ and $q_{0,\tau}$. However, this suggestion is infeasible because the propensity score function $p(x)$ is unknown in general but it can be estimated by some standard parametric or nonparametric techniques. Once the estimator $\widehat{p}(x)$ of $p(x)$ is obtained, the sample analogs of (6) and (7) can be used to obtain the estimators for $q_{1,\tau}$ and $q_{0,\tau}$. To be specific, the estimation strategy in Firpo (2007) consists of two steps: (i) nonparametric power series estimation of the propensity score function $p(x)$ and (ii) a weighted quantile regression using the estimated weights from the first step. More specifically, given a sample of n observations on $\{Y_i, X_i, D_i\}$, Firpo (2007) suggested that the QTE parameter Δ_τ can be estimated by $\widehat{\Delta}_\tau = \widehat{q}_{1,\tau} - \widehat{q}_{0,\tau}$, where

$$\widehat{q}_{j,\tau} = \arg\min_q \sum_{i=1}^n \widehat{\omega}_{j,i}\, \rho_\tau(Y_i - q)$$

for $j = 0$ and 1, where $\rho_\tau(\cdot) = u(\tau - I\{u \le 0\})$ is the check function for $0 < \tau < 1$, and the weights $\widehat{\omega}_{j,i}$ are

$$\widehat{\omega}_{1,i} = D_i/\widehat{p}(X_i) \quad \text{and} \quad \widehat{\omega}_{0,i} = (1 - D_i)/[1 - \widehat{p}(X_i)]$$

with $\widehat{p}(x)$ being the nonparametric power series estimator of $p(x)$. Under appropriate regularity conditions, Firpo (2007) showed that the resulting estimator $\widehat{\Delta}_\tau$ is $\sqrt{n}$-consistent and asymptotically normally distributed. Furthermore, Firpo (2007) argued that the asymptotic variance of the estimator $\widehat{\Delta}_\tau$ presented above can attain the semiparametric efficiency bound.

2.3 Partially Conditional Quantile Treatment Effect Model

As discussed in Section 1, in many applications, researchers may be interesting in estimating the effect of a treatment or policy on outcome of interest in various sub-populations defined by the possible values of some component(s) of the pre-treatment variables X. For example, Abrevaya et al. (2015) and Lee et al. (2017) examined the mean effect of maternal smoking during pregnancy on infant birth weights conditional on mother's age. For the partially conditional average treatment effect model, the reader is referred to the papers by Abrevaya et al. (2015) and Lee et al. (2017) for details. Recently, Cai et al. (2019a) extended the approach of Abrevaya et al. (2015) to estimate quantile treatment effect across different sub-populations. Specifically, they proposed the partially conditional quantile treatment effect model to characterize the heterogeneity of a treatment effect across different sub-populations. Therefore, this section reviews this model and its estimation procedure proposed by Cai et al. (2019a).

2.3.1 Model Setup

Let $W \in \mathbb{R}^k$ denote a sub-vector of $X \in \mathbb{R}^p$, $1 \le k < p$. Then, for any given quantile level $\tau \in (0, 1)$, the partially conditional quantile treatment effect (PCQTE) is defined as

$$\Delta_\tau(w) = q_{1,\tau}(w) - q_{0,\tau}(w),$$

where $q_{j,\tau}(w) = \inf\big\{y : P\big(Y(j) \le y \,|\, W = w\big) \ge \tau\big\}$ for $j = 0$ and 1. It should be noted that Assumption 2.2(i) does not hold generally if one only controls the sub-vector W instead of X. Therefore, some new techniques are needed to identify $\Delta_\tau(w)$. As in Cai et al. (2019a), for the sake of parsimony of the exposition, here it is also assumed that the dimension of W is one. As pointed out by Abrevaya et al. (2015), the case for $k = 1$ is the most relevant case in practice, since the resulting estimation can be easily extended to a multivariate case.

2.3.2 Identification and Estimation

For $j = 0$ and 1, let $F_{Y(j)|W}(y|w)$ be the distribution function of $Y(j)$ conditional on $W = w$. Then, under Assumption 2.2, one can show that the conditional distribution function $F_{Y(j)|W}(y|w)$ for $j = 0$ and 1, can be identified from the joint distribution of (Y, X, D), given by

$$F_{Y(0)|W}(y|w) = E\Big((1 - D)I\{Y \le y\}/[1 - p(X)]\Big|W = w\Big), \tag{8}$$

and

$$F_{Y(1)|W}(y|w) = E\Big(DI\{Y \le y\}/p(X)\Big|W = w\Big). \tag{9}$$

Therefore, the parameter of interest $\Delta_\tau(w)$ can be identified by

$$\Delta_\tau(w) = F^{-1}_{Y(1)|W}(\tau|w) - F^{-1}_{Y(0)|W}(\tau|w), \tag{10}$$

where $F^{-1}_{Y(j)|W}(\tau|w) = \inf\{y : F_{Y(j)|W}(y|w) \geq \tau\}$ for $j = 0$ and 1. Thus, if the dimension of W is not very high, following Cai (2002) or Cai and Wang (2008), in view of (8) and (9), one can first estimate easily the conditional CDFs by a nonparametric method, and then compute their inverse to obtain the estimators for the conditional quantiles, so that one can obtain the estimator for $\Delta_\tau(w)$ in (10). If k is large, one can use a dimension reduction approach to estimate the conditional CDFs, such as index method proposed by Hall and Yao (2005).

Alternatively, one can use the estimation procedure proposed by Cai et al. (2019a), which is based on quantile regression technique. Formally, first note that $\Delta_\tau(w)$ can be identified by the following optimization problem

$$\Delta_\tau(w) = q_{1,\tau}(w) - q_{0,\tau}(w), \tag{11}$$

where

$$q_{0,\tau}(w) = \arg\min_q E\Big((1-D)\rho_\tau(Y-q)/(1-p(X))\,\Big|\,W=w\Big), \tag{12}$$

and

$$q_{1,\tau}(w) = \arg\min_q E\Big(D\rho_\tau(Y-q)/p(X)\,\Big|\,W=w\Big). \tag{13}$$

Given the identification results presented in (11), (12) and (13), Cai et al. (2019a) proposed an estimation procedure based on quantile regression in two steps: (i) estimate the unknown propensity score function $p(x)$ parametrically or nonparametrically and (ii) plug the estimator $\widehat{p}(x)$ of $p(x)$ into the sample analogues of (12) and (13) and then obtain the estimators $\widehat{q}_{1,\tau}(w)$ and $\widehat{q}_{0,\tau}(w)$ and thus the estimator $\widehat{\Delta}_\tau(w)$. Specifically, they proposed estimator $\widehat{\Delta}_\tau(w)$ for $\Delta_\tau(w)$, given by

$$\widehat{\Delta}_\tau(w) = \widehat{q}_{1,\tau}(w) - \widehat{q}_{0,\tau}(w), \tag{14}$$

where

$$\widehat{q}_{j,\tau}(w) = \arg\min_q \frac{1}{nh}\sum_{i=1}^{n} \widehat{\omega}_{j,i} K((W_i - w)/h)\rho_\tau(Y_i - q)$$

for $j = 0$ and 1, with $K(\cdot)$ being a kernel function, and h being a bandwidth parameter. Under some regularity conditions, Cai et al. (2019a) showed that the proposed estimator $\widehat{\Delta}_\tau(w)$ given in (14) is consistent and asymptotically normally distributed. The reader is referred to the paper by Cai et al. (2019a) for more details. If k is large, similarly, one can use a dimension reduction approach to estimate the conditional quantiles, such as index method proposed by Li and Lv (2019). Alternatively, one might consider to use a parametric form as in Tang et al. (2019).

2.3.3 Specification Test

In some real applications, researchers may be interested in the effects of programs beyond point estimates of the treatment effects in various sub-populations. For example, policymakers may be of interest to investigate whether the effect of the treatment is zero or a constant for all sub-populations defined by covariates. To this end, Crump, Hotz, Imbens and Mitnik (2008) developed two nonparametric tests based on series approach in which the first test is to test whether the treatment has a zero average effect for all sub-populations defined by covariates and the second is to test whether the average effect conditional on the covariates is identical for all sub-populations, in other words, whether there is heterogeneity in average treatment effects by covariates. The reader is referred to the papers by Crump et al. (2008) for details.

To test whether there is heterogeneity in quantile treatment effect across different sub-populations, Cai et al. (2019a) considered the following testing hypothesis

$$H_0 : \Delta_\tau(w) = \Delta_\tau, \quad \text{for all } \ w \in \mathcal{W} \quad \text{versus} \quad H_1 : \Delta_\tau(w) \neq \Delta_\tau, \tag{15}$$

where $\tau \in (0, 1)$ is the quantile level, Δ_τ is the τ-th unconditional quantile treatment effect defined in Section 2.1 and $\mathcal{W}$ denotes the support of W. Under the null hypothesis, PCQTE is equal to the corresponding unconditional QTE, whereas, under the alternative, there are some values of the covariate W for which the quantile effect of the treatment differs from the unconditional quantile treatment effect Δ_τ. To test the hypothesis displayed in (15), Cai et al. (2019a) proposed the following test statistics based on Cramér-von Mises criterion, given by

$$J_n = \int \left(\widehat{\Delta}_\tau(w) - \widehat{\Delta}_\tau\right)^2 \lambda(w) dw,$$

where $\lambda(w)$ is a non-negative weighting function defined on the support of W, $\widehat{\Delta}_\tau(w)$ is the semiparametric estimator of $\Delta_\tau(w)$ in (14) and $\widehat{\Delta}_\tau$ is a $\sqrt{n}$-consistent estimator for Δ_τ. Under some regularity conditions, they established the asymptotic properties of the test statistic J_n, including consistency and asymptotic normality; see Cai et al. (2019a) for details.

Remark 2.3: Except for the testing issue displayed in (15), one may be interested in testing whether the partially conditional quantile treatment effect model is correctly specified; that is, the more general interest than testing (15) is to consider the hypothesis testing problem

$$H_0 : \Delta_\tau(w) = \Delta_{\tau,0}(w; \theta_\tau) \ \text{ for all } \ w,$$

where $\Delta_{\tau,0}(\cdot)$ is a known function with unknown parameter θ_τ. In particular, one might have an interest in testing

$$H_0 : \Delta_\tau(w) \leq 0 \ \text{ or } \ \geq 0 \text{ for all } \ w,$$

which leads to studying the stochastic dominance such as $Y(1) \leq Y(0)$ or $Y(1) \geq Y(0)$ for all $w \in \mathcal{W}$. These extensions are beyond the scope of this paper but certainly worth pursuing in future research.

§3 QTE under Selection on Unobservables

In the preceding section, the setting is on where treatment assignment is confounded but there exist a set of observed covariates X, such that treatment assignment becomes unconfounded conditional on X. In some applications, however, treatment variable is self-selected and potentially endogenous, as stated in Section 1. Under these settings, an instrumental variable approach provides a powerful tool to address this problem. Over the last three decades, several instrumental variable approaches have been proposed to identify and estimate treatment effect parameters of interest. In this section, our focus is on the procedures developed by Imbens and Angrist (1994) to achieve identification through a monotonicity assumption in the treatment choice equation. But the limitation is to consider models with binary endogenous variables. Unfortunately, so far, there have been almost no results available for multi-valued or continuous treatments under this framework.

3.1 Models Without Covariates

3.1.1 Model Setup

Again, in this section, our discussion is on the potential outcome framework and the consideration is limited to the simplest setup where both the treatment and the assignment are binary. Formally, let $Y(1)$ and $Y(0)$ be the potential outcomes that an individual would attain with and without being exposed to a treatment. Define D to be the binary treatment variable which equals to one when individual has been exposed to the treatment and equals to zero otherwise. Similarly, let Z be a binary instrumental variable which is independent of the potential outcomes $Y(1)$ and $Y(0)$ but which is correlated with the treatment variable D in the population. Denote $D(z)$ the potential treatment status when Z is set exogenously to 0 or 1. Clearly, one cannot observe both potential treatment status $D(1)$ and $D(0)$. Instead, the realized treatment status $D = ZD(1) + (1 - Z)D(0)$ is observed. Again, note that without Assumption 2.1, the potential outcomes of each individual would rely on the assignments and treatments of all individuals in the population. It is impossible to identify any treatment effect without imposing some restrictions on the dependence between individuals.

With treatment D and instrument Z being binary, it is easy to see that the population can be partitioned into four types $\mathcal{T}$ defined in terms of the values of $D(1)$ and $D(0)$, presented in Table 1.

Table 1: Definition of Types

Types ($\mathcal{T}$)	$D(1)$	$D(0)$	Notion
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	0	0	Never takers

One can see easily from Table 1 that the compliers are those experimental individuals that comply with the treatment assignment in both treatment arms while the defiers are affected by the assignment in the opposite direction: they take the treatment when they are assigned to the control group but refuse to take it when they are assigned to the treatment group. The never-takers and the always-takers are not affected by the assignment variable Z, so there is no source of random variation for these types in the data, which implies that the treatment effects for these never-takers and always-takers can not be identified if the treatment effect is allowed to be arbitrarily heterogeneous. This also implies that it is impossible to identify the treatment effects for the whole population or for the treated sub-population. It is only able to identify the causal effects for the individuals who respond to a change in the value of the instrument; that is, the compliers or defiers.

As discussed in Angrist, Imbens and Rubin (1996), since the defiers and the compliers react to a change in the assignment in opposite directions, the intention-to-treat effect is a weighted average of the individual treatment effect with negative weights for the defiers, positive weights for the compliers and zero weights for both other types. Therefore, even though there could be a positive treatment effect for all the individuals considered, the intention-to-treat effect may be 0. This is sufficient to show that, in the case of arbitrarily heterogeneous treatment, the causal effect can not be identified. Because we can only observe $D(1)$ or $D(0)$ for each individual, we are unable to determine the type of each individual to separate compliers from defiers. Several procedures have been suggested to solve this mixture problem and two of them are commonly adopted in the treatment effect literature: either it is assumed that there are no defiers (homogeneity assumption for the assignment effect) or the outcome is restricted to be the same for the compliers and the defiers (homogeneity assumption for the treatment effect). In this section, we review the approach to instrumental variables which follows the first assumption. Formally, to identify meaningful treatment effect parameters, the following identification assumption is needed and stated here.

Assumption 3.1.
(i) Independence of the instrument: the random vector $(Y(1), Y(0), D(1), D(0))$ *is independent of* Z*.*
(ii) First stage: $0 < P(Z = 1) < 1$ *and* $P(D(1) = 1) > P(D(0) = 1)$*.*
(iii) Monotonicity: $P(D(1) \geq D(0)) = 1$*.*

Remark 3.1: This assumption is commonly made in the literature, which was also assumed in Imbens and Angrist (1994) and Abadie (2002). First, Assumption 3.1(i) is an unconfounded instrument restriction, which means that Z is “as good as randomly assigned”. It is mechanically satisfied if the assignment has been randomized. Second, Assumption 3.1(ii) assumes that D and Z are correlated. Indeed, the strength of this correlation can be measured by the ratio of the compliers $P(\mathcal{T} = c) = E(D|Z = 1) - E(D|Z = 0)$; see Lemma 2.1 in Abadie (2003). Finally, Assumption 3.1(iii) says that the potential treatment state of any individual does not decrease in the instrument. It rules out the existence of defiers (type $\mathcal{T} = d$), because for the defiers, $D(1) < D(0)$. As a consequence, always takers, never takes and compliers exhaustively partition the whole population.

In summary, under Assumption 3.1, one can identify effects for the individuals that respond to a change in the value of the instrument, that is, the compliers group. For the easy exposition, $\mathcal{C}$ is used to denote the compliers type $\mathcal{T} = c$ in the following. Based on the aforementioned discussions, the main estimation is the QTE for the compliers, which is the so-called the local quantile treatment effect (LQTE) because it corresponds to the QTE for a subpopulation, defined as

$$\Delta_{\tau|\mathcal{C}} = q_{1,\tau|\mathcal{C}} - q_{0,\tau|\mathcal{C}}, \tag{16}$$

where $q_{j,\tau|\mathcal{C}} = \inf\left\{y : P(Y(j) \leq y|\mathcal{T} = c) \geq \tau\right\}$ for $j = 0$ and 1, and $\tau \in (0, 1)$ is the quantile level.

3.1.2 Identification and Estimation

To discuss the identification of the LQTE parameter $\Delta_{\tau|\mathcal{C}}$, it is convenient to introduce some notations. Let $F_{Y(1)|\mathcal{C}}(y)$ and $F_{Y(0)|\mathcal{C}}(y)$ be the distribution functions of $Y(1)$ and $Y(0)$ conditional on compliers, respectively. Then, by the definition of the LQTE parameter $\Delta_{\tau|\mathcal{C}}$, to identify $\Delta_{\tau|\mathcal{C}}$, one should first identify the conditional distributions $F_{Y(1)|\mathcal{C}}(y)$ and $F_{Y(0)|\mathcal{C}}(y)$. Several approaches have been proposed to identify them in the literature. For example, Imbens and Rubin (1997) showed that $F_{Y(1)|\mathcal{C}}(y)$ and $F_{Y(0)|\mathcal{C}}(y)$ can be identified by working directly with their densities. In this section, the review is typically devoted to the method proposed by Abadie (2002). Specifically, as in Lemma 2.1 of Abadie (2002), under Assumption 3.1, the conditional distribution functions $F_{Y(j)|\mathcal{C}}(y)$ for $j = 0$ and 1 can be identified from the joint distribution of (Y, D, Z), which is given by, respectively,

$$F_{Y(0)|\mathcal{C}}(y) = \frac{E\big(I\{Y \le y\}(1-D)|Z=1\big) - E\big(I\{Y \le y\}(1-D)|Z=0\big)}{E\big(1-D|Z=1\big) - E\big(1-D|Z=0\big)} \tag{17}$$

and

$$F_{Y(1)|\mathcal{C}}(y) = \frac{E\big(I\{Y \le y\}D|Z=1\big) - E\big(I\{Y \le y\}D|Z=0\big)}{E\big(D|Z=1\big) - E\big(D|Z=0\big)}. \tag{18}$$

Consequently, the LQTE parameter $\Delta_{\tau|\mathcal{C}}$ defined in (16) is identified as

$$\Delta_{\tau|\mathcal{C}} = F^{-1}_{Y(1)|\mathcal{C}}(\tau) - F^{-1}_{Y(0)|\mathcal{C}}(\tau),$$

where $F^{-1}_{Y(j)|\mathcal{C}}(\tau) = \inf\{y : F_{Y(j)|\mathcal{C}}(y) \ge \tau\}$ for $j = 0$ and 1.

Given a sample of n observations on $\{Y_i, D_i, Z_i\}_{i=1}^n$, based on the identification result in (17) and (18), $F_{Y(j)|\mathcal{C}}(y)$ can be estimated via two-stage least-square (2SLS) with dependent variable $I\{Y_i \le y\}D_i$, endogenous regressor D_i and instrument Z_i. Although the estimated cumulative distribution function should necessarily be non-monotone, it can be monotonized using the rearrangement method of Chernozhukov, Fernández-Val and Galichon (2010). Finally, the estimation of $\Delta_{\tau|\mathcal{C}}$ can be obtained by inverting the empirical potential outcome CDFs and then taking difference. Under some regularity conditions, one can show that the resulting estimator of $\Delta_{\tau|\mathcal{C}}$ is consistent and asymptotically normally distributed if the densities of the potential outcomes among compliers are strictly positive, namely, $f_{Y(d)|\mathcal{C}}(y) > 0$ for $d \in \{0, 1\}$. The reader is referred to the paper by Chernozhukov et al. (2010) for details.

3.2 Model With Covariates

So far, identification and estimation of LQTE are without covariates. But, this may be commonly the case in observational data in which the instrument variable is typically not explicitly randomized like in an experiment. Hence, one may be interested in including a vector of covariates in the estimation for two reasons. First, the validity of Assumption 3.1 may be plausible only after conditioning on observable covariates. Examples include the stratified randomized experiment where the assignment probabilities are different across the strata and the case in observational studies in which the instrument has not been randomized. Another reason incorporating covariates in the estimation is for efficiency, as shown in Frölich and Melly (2013), even though Assumption 3.1 is valid unconditionally; see, for example, Frölich and Melly (2013) and Section 10.2 in Melly and Wüthrich (2017) for more discussions. Therefore, in this section, our setup goes to the case that the instrument is valid after conditioning on a vector of covariates, denoted by X, which also covers randomized instrument as a special case.

3.2.1 Conditional LQTE

In applied fields, researchers may be interested in estimating conditional LQTE, because conditional LQTE allows analyzing the heterogeneity of the effects with respect to the observable covariates. In addition, conditional LQTE can be used to decompose the total variance within and between components. A pioneering application of conditional LQTE can be found in the work by Abadie et al. (2002) to seek to investigation of the effects of training program provided under Job Training Partnership Act on the distribution of earnings. Formally, conditional LQTE is defined as the quantile treatment effect for the compliers conditional on covariates X, which is given by

$$\Delta_{\tau|\mathcal{C}}(x) = q_{1,\tau|\mathcal{C}}(x) - q_{0,\tau|\mathcal{C}}(x), \tag{19}$$

where $q_{j,\tau|\mathcal{C}}(x) = \inf\big\{y : P(Y(j) \leq y|\mathcal{T} = c, X = x) \geq \tau\big\}$ for $j = 0$ and 1. Let $F_{Y(j)|\mathcal{C}}(y|x)$ for $j = 0$ and 1, be the distribution function of $Y(j)$ for compliers conditional on $X = x$. Then, it is easy to see from (19) that, to identify the parameter of interest $\Delta_{\tau|\mathcal{C}}(x)$, one should first identify $F_{Y(j)|\mathcal{C}}(y|x)$ for $j = 0$ and 1 and then compute their inverse and finally, take difference to obtain $\Delta_{\tau|\mathcal{C}}(x)$. To obtain the estimator for $\Delta_{\tau|\mathcal{C}}(x)$, the following identifying assumption, similar to Assumption 3.1, is needed.

Assumption 3.2. *For all x in the support of X:*
(i) Independence of the instrument: the random vector $(Y(1), Y(0), D(1), D(0))$ is jointly independent of Z conditional on $X = x$.
(ii) Relevant instrument: $P\big(D(1) = 1|X = x\big) > P\big(D(0) = 1|X = x\big)$ and $P(Z = 1|X = x) \in (0, 1)$.
(iii) Monotonicity: $P\big(D(1) \geq D(0)|X = x\big) = 1$.

Remark 3.2: Assumption 3.2 indeed is simply the conditional version of Assumption 3.1 and it is also assumed in Abadie et al. (2002) and Abadie (2003). Actually, Assumption 3.2(i) is weaker than Assumption 3.1(i), because independence now is only required to hold among units with the same values of X rather than unconditionally, implying that Z is as good as randomly assigned given X. Furthermore, Assumption 3.2(ii) implies that compliers exist for every value of X in its support and requires the support of X to be identical in the $Z = 0$ and $Z = 1$ sub-populations. Finally, Assumption 3.2(iii) requires that defiers do not exist for every value of X.

Using analogous arguments as in previous section, the distribution functions of the potential outcomes $Y(1)$ and $Y(0)$ for the compliers given $X = x$ can be identified under Assumption 3.2 by

$$F_{Y(0)|\mathcal{C}}(y|x) = \frac{E\big(I\{Y \leq y\}(1-D)|X = x, Z = 1\big) - E\big(I\{Y \leq y\}(1-D)|X = x, Z = 0\big)}{E\big(1 - D|X = x, Z = 1\big) - E\big(1 - D|X = x, Z = 0\big)} \tag{20}$$

and

$$F_{Y(1)|\mathcal{C}}(y|x) = \frac{E\big(I\{Y \le y\}D|X=x, Z=1\big) - E\big(I\{Y \le y\}D|X=x, Z=0\big)}{E\big(D|X=x, Z=1\big) - E\big(D|X=x, Z=0\big)}, \tag{21}$$

respectively. It is easy to see that the nonparametric estimation of these conditional distributions may suffer from the curse of dimensionality if the dimension of X is high. For this reason, Abadie et al. (2002) imposed a linear restriction for the conditional quantile functions, given by

$$q_{1,\tau|\mathcal{C}}(x) = \delta_{\tau|\mathcal{C}} + \beta_\tau' x, \quad \text{and} \quad q_{0,\tau|\mathcal{C}}(x) = \beta_\tau' x.$$

In this model, the parameter of primary interest is $\delta_{\tau|\mathcal{C}}$, which gives the difference in the conditional τ-th quantiles of $Y(1)$ and $Y(0)$ and thus the conditional LQTE. It is important to note that although both the conditional quantile functions of the potential outcomes $Y(1)$ and $Y(0)$ for the compliers conditional on $X = x$ are functions of $X = x$, the conditional LQTE parameter of interest $\delta_{\tau|\mathcal{C}}$ does not rely on the covariates X. This is because the model above assumes that the two conditional quantile functions have the same coefficients β_τ. The particular linear form of the conditional quantile functions makes analysis simpler. Of course, one can easily introduce a more complex form, including interactions between the treatment variable and the control variables or a nonparametric form, which allows researchers to investigate the heterogeneous treatment effects with respect to the observable covariates. This is warranted as a future research topic.

To estimate $\delta_{\tau|\mathcal{C}}$ and β_τ, Abadie et al. (2002), applying a result in Abadie (2003), showed that the parameter of interest $\delta_{\tau|\mathcal{C}}$ is identified by α_τ, which solves the following weighted quantile regression problem

$$(\alpha_\tau, \beta_\tau) = \arg\min_{\alpha,\beta} E\Big(W^{AAI} \cdot \rho_\tau\big(Y - \alpha D - X'\beta\big)\Big), \tag{22}$$

where

$$W^{AAI} = 1 - D(1-Z)/[1 - P(Z=1|X)] - (1-D)Z/P(Z=1|X).$$

Following the analogy principle, a natural estimator of $(\alpha_\tau, \beta_\tau)$ is the sample counterpart of (22). However, it is important to note that although the population objective function (22) is globally convex in (α, β), its sample counterpart is typically not because W^{AAI} is negative when $D \ne Z$, implying that the objective function has many local minima. To this end, Abadie et al. (2002) suggested replacing the weight W^{AAI} with its projection on (Y, D, X), which can be shown to be always positive. Therefore, their estimation strategy consists of three steps: (i) nonparametric power series estimation of the instrument propensity score function $P(Z=1|X)$; (ii) nonparametric power series estimation of the positive weight; and (iii) a weighted quantile regression using the estimated weight from the second step. Under appropriate regularity conditions, the resulting estimators $\widehat{\alpha}_\tau$ and $\widehat{\beta}_\tau$ are $\sqrt{n}$-consistent and asymptotically normally distributed; see Abadie et al. (2002) for details.

3.2.2 Unconditional LQTE

The previous section is about conditional quantile treatment effect; that is, the quantile treatment effect is defined within the sub-population with the same covariates X. In this section, the discussion is on unconditional quantile treatment effect in the presence of covariates. The distinction between conditional quantile treatment effect and unconditional quantile treatment effect is important because of the definition of quantiles. For a more detailed discussion about the difference between conditional and unconditional quantile treatment effect, the reader is referred to the paper by Frölich and Melly (2013).

Concerning unconditional LQTE estimation with covariates, first noting that from the representation of $F_{Y(1)|\mathcal{C}}(y|x)$ in (21) and the fact that $P(\mathcal{T}=c|X=x)=E(D|X=x,Z=1)-E(D|X=x,Z=0)$, it is easy to know that the distribution function of the potential outcome $Y(1)$ for the compliers can be identified as

$$\begin{aligned} F_{Y(1)|\mathcal{C}}(y)=&\int F_{Y(1)|\mathcal{C}}(y|x)dF_{X|\mathcal{C}}(x)=\int F_{Y(1)|\mathcal{C}}(y|x)\frac{P(\mathcal{T}=c|X=x)}{P(\mathcal{T}=c)}dF_X(x)\\ =&\frac{\int\big(E[I\{Y\le y\}D|X=x,Z=1]-E[I\{Y\le y\}D|X=x,Z=1]\big)dF_X(x)}{\int\big(E[D|X=x,Z=1]-E[D|X=x,Z=0]\big)dF_X(x)}, \end{aligned} \tag{23}$$

where the first equality is an application of the law of total expectation and the second equality follows from the Bayes' law. Similarly, one can obtain the identifying result for $F_{Y(0)|\mathcal{C}}(y)$. Alternatively, by Parts (b) and (c) of Theorem 3.1 in Abadie (2003), one can obtain a weighted representation for $F_{Y(1)|\mathcal{C}}(y)$ and $F_{Y(0)|\mathcal{C}}(y)$ (see Frölich and Melly (2013)):

$$F_{Y(1)|\mathcal{C}}(y)=\frac{E\big(I\{Y\le y\}DW^{FM}\big)}{E\big(DW^{FM}\big)},\quad\text{and}\quad F_{Y(0)|\mathcal{C}}(y)=\frac{E\big(I\{Y\le y\}(1-D)W^{FM}\big)}{E\big(DW^{FM}\big)}, \tag{24}$$

where

$$W^{FM}=[Z-E(Z|X)](2D-1)/[E(Z|X)(1-E(Z|X))].$$

This identifies the unconditional LQTE as the difference between the quantiles:

$$\Delta_{\tau|\mathcal{C}}=F^{-1}_{Y(1)|\mathcal{C}}(\tau)-F^{-1}_{Y(0)|\mathcal{C}}(\tau),$$

where $F^{-1}_{Y(j)|\mathcal{C}}(\tau)=\inf\{y:F_{Y(j)|\mathcal{C}}(y)\ge\tau\}$ for $j=0$ and 1. In addition, as in Frölich and Melly (2013), the unconditional LQTE parameter $\Delta_{\tau|\mathcal{C}}$ can also be identified directly from the following weighted quantile regression problem:

$$(q_{0,\tau|\mathcal{C}},\Delta_{\tau|\mathcal{C}})=\arg\min_{\alpha,\beta}E\big(\rho_\tau(Y-\alpha-\beta D)\cdot W^{FM}\big), \tag{25}$$

where $q_{0,\tau|\mathcal{C}}$ is the τ-th quantile of the conditional distribution function $F_{Y(0)|\mathcal{C}}(y)$. This optimization problem is non-convex since W^{FM} is negative for $Z\ne D$. This complicates the optimization problem because local optima could exist, but this problem can be solved relatively easily by rewriting (25) as two one-dimensional optimization problems. In other words, (25) can be equivalently written as

$$(q_{1,\tau|\mathcal{C}},q_{0,\tau|\mathcal{C}})=\Big(\arg\min_{q_1}E\big[\rho_\tau(Y-q_1)\cdot W^{FM}|D=1\big],\ \arg\min_{q_0}E\big[\rho_\tau(Y-q_0)\cdot W^{FM}|D=0\big]\Big),$$

where $q_{1,\tau|\mathcal{C}}$ is the τ-th quantile of the conditional distribution function $F_{Y(1)|\mathcal{C}}(y)$, which are two separate one-dimensional optimization problems in the $D=1$ and $D=0$ sub-populations such that one can easily use grid-search procedures supported by visual inspection of the objective function for local minima.

Given a sample of n observations $\{Y_i,X_i,D_i,Z_i\}_{i=1}^n$, estimators based on all three representations in (23), (24) and (25) can be obatined. Specifically, Hsu, Lai and Lieli (2015) proposed a weighted CDF estimator based on (24) and established the asymptotic distribution for the whole LQTE process, while Frölich and Melly (2013) analyzed a weighted quantile regression estimator based on (25). Furthermore, under some regularity conditions, they proved that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normally distributed, and achieves the semiparametric efficiency bound.

3.3 Partially Conditional Local Quantile Treatment Effects

Again, in the setting where selection to treatment is based on unobservables, policy-makers may also be interested in estimating the quantile treatment effect in various sub-populations defined by the possible values of a subset of the pre-treatment variables X. For example, researchers may have an interest to estimate the effect of the participation into the 401(K) programs on financial assets holdings conditional on ages or income. Indeed, Cai et al. (2019c) proposed a method to test if the conditional independence holds for the 401(K) programs and the testing result is that the conditional independence does not hold for this example. Therefore, to characterize the heterogeneity of the effect of an endogenous treatment along the outcome distribution across different sub-populations, Cai et al. (2019b) considered the partially conditional local quantile treatment effect (PCLQTE), which is defined as

$$\Delta_{\tau|\mathcal{C}}(w) = q_{1,\tau|\mathcal{C}}(w) - q_{0,\tau|\mathcal{C}}(w),$$

where $q_{j,\tau|\mathcal{C}}(w) = \inf\big\{y : P\big(Y(j) \leq y\,|\,W = w, \mathcal{T} = c\big) \geq \tau\big\}$ for $j = 0$ and 1, and similar to Section 2.3.1, W is a subset of X. Similar to Section 2.3, the dimension of W is assumed to be one. Again, it is worth stressing that Assumption 3.2 does not hold generally if one only controls the subset W instead of X. Similarly, we also can consider the partially conditional local quantile treatment effect for the treated, which is given by

$$\Delta_{\tau|\mathcal{C},D=1}(w) = q_{1,\tau|\mathcal{C},D=1}(w) - q_{0,\tau|\mathcal{C},D=1}(w),$$

where $q_{j,\tau|\mathcal{C},D=1}(w) = \inf\big\{y : P\big(Y(j) \leq y|W = w, \mathcal{T} = c, D = 1\big) \geq \tau\big\}$ for $j = 0$ and 1. For $j = 0$ and 1, let $F_{Y(j)|\mathcal{C}}(y|w)$ be the distribution function of $Y(j)$ for compliers conditional on $W = w$. Under Assumption 3.2, Cai et al. (2019b) showed that the conditional distribution functions $F_{Y(1)|\mathcal{C}}(y|w)$ and $F_{Y(0)|\mathcal{C}}(y|w)$ can be identified by

$$F_{Y(0)|\mathcal{C}}(y|w) = \frac{1}{\gamma_0(w)} E\left[\frac{(Z - \pi(X))(D-1)}{\pi(X)(1-\pi(X))} I\{Y \leq y\}\middle| W = w\right], \tag{26}$$

and

$$F_{Y(1)|\mathcal{C}}(y|w) = \frac{1}{\gamma_1(w)} E\left[\frac{(Z - \pi(X))D}{\pi(X)(1-\pi(X))} I\{Y \leq y\}\middle| W = w\right], \tag{27}$$

respectively, where $\pi(X) = P(Z = 1|X)$ is the instrument propensity score function,

$$\gamma_1(w) = E\left[\frac{(Z - \pi(X))D}{\pi(X)(1-\pi(X))}\middle| W = w\right], \quad \text{and} \quad \gamma_0(w) = E\left[\frac{(Z - \pi(X))(D-1)}{\pi(X)(1-\pi(X))}\middle| W = w\right].$$

Consequently, the partially conditional local quantile treatment effect parameter $\Delta_{\tau|\mathcal{C}}(w)$ is identified as

$$\Delta_{\tau|\mathcal{C}}(w) = F^{-1}_{Y(1)|\mathcal{C}}(\tau|w) - F^{-1}_{Y(0)|\mathcal{C}}(\tau|w), \tag{28}$$

where $F^{-1}_{Y(j)|\mathcal{C}}(\tau|w) = \inf\{y : F_{Y(j)|\mathcal{C}}(y|w) \geq \tau\}$ for $j = 0$ and 1, and $\tau \in (0, 1)$ is the quantile level. Similarly, one can obtain the identifying result for the parameter of interest $\Delta_{\tau|\mathcal{C},D=1}(w)$. For more details, see the paper by Cai et al. (2019b).

Based on the identifying results in (27), (26) and (28), one can first estimate the two conditional CDFs and then compute the inverse to obtain the conditional quantiles so that one can obtain an estimator for $\Delta_{\tau|\mathcal{C}}(w)$. Alternatively, Cai et al. (2019b) proposed an estimator for $\Delta_{\tau|\mathcal{C}}(w)$ based on quantile regression method, given by

$$\widehat{\Delta}_{\tau|\mathcal{C}}(w) = \widehat{q}_{1,\tau|\mathcal{C}}(w) - \widehat{q}_{0,\tau|\mathcal{C}}(w),$$

where

$$\widehat{q}_{j,\tau|\mathcal{C}}(w) = \arg\min_{q} \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{W_i - w}{h}\Big)\widehat{\omega}_{j,i}(Z_i)\rho_\tau(Y_i - q)$$

with

$$\widehat{\omega}_{1,i}(Z_i) = \frac{\big(Z_i - \widehat{\pi}(X_i)\big)D_i}{\widehat{\pi}(X_i)\big(1-\widehat{\pi}(X_i)\big)} \quad \text{and} \quad \widehat{\omega}_{0,i}(Z_i) = \frac{\big(Z_i - \widehat{\pi}(X_i)\big)(1-D_i)}{\widehat{\pi}(X_i)\big(1-\widehat{\pi}(X_i)\big)}.$$

Here, $\widehat{\pi}(x)$ is the parametric or nonparametric estimator of $\pi(x)$, $K(\cdot)$ is a kernel function, and h is the bandwidth parameter. Similar estimating result can be obtained for $\Delta_{\tau|\mathcal{C},D=1}(w)$. Under some regularity conditions, Cai et al. (2019b) showed that the proposed estimators $\widehat{\Delta}_{\tau|\mathcal{C}}(w)$ and $\widehat{\Delta}_{\tau|\mathcal{C},D=1}(w)$ are consistent and asymptotically normally distributed.

§4 Some Recent Extensions

4.1 A Big Data Era

All the models and their estimation methods discussed in Sections 2 and 3 assume basically that the dimension of covariates X is small relative to the sample size, but in practice researchers often use high-dimensional models. The reason for the presence of high-dimensional X is either because many variables are available in the raw data set or because we want to include interactions and other transformations of the control variables, such as dummy variable expansions of categorical variables, power or other series expansions of continuous variables, higher-order interactions, or other technical regressors generated from the original available variables. The presence of high-dimensional X renders conventional theory (where the dimension of X is taken to be small) invalid for estimating the treatment effect parameters and has motivated researchers to develop new program evaluation inference methods to account for high-dimensional X.

Recently, Belloni, Chernozhukov, Fernández-Val and Hansen (2017) provided an estimator for LQTE based on (23) employing machine learning technique from the high-dimensional literature in statistics. This methodology, which allows for very large dimension of X (much larger than the sample size), such as some variables from the online data, consists of two steps. First, use model selection method, for example, least absolute shrinkage and selection operator (LASSO) or other methods, to automatically select the relevant covariates from the large set of pre-treatment covariates. Then, construct estimators for the treatment effect parameters of interest using only the small (usually much smaller than the sample size) set of selected covariates. By assuming that key reduced-form predictive relationships are approximately sparse, Belloni et al. (2017) showed that valid inference can be performed after data-driven selection of control variables. Furthermore, they established the limiting laws of the estimators of the whole LQTE process as a function of the quantile index. This allows us to construct confidence band for the LQTE function over a quantile continuum and to test functional hypotheses as well as to test dominance relations between the potential outcomes. More general methods and a review of the literature on how to incorporate some big data techniques into studying the quantile (average) treatment effects can found partially in Belloni et al. (2017) and Athey and Imbens (2017), and among others.

4.2 Regression Discontinuity Design

In this section, the quantile treatment effect is reviewed under the framework of regression discontinuity design (RDD) which is related to the quantile treatment effect under endogeneity considered above. Actually, in causal analysis, RDD is split into two categories, the sharp and fuzzy RDD. This categorization is based on how the treatment assignment is determined by a running variable R. For the fuzzy design, the treatment assignment probability is partly determined by the running variable R at the threshold r_0, while for the sharp design, this probability jumps from 0 to 1 at the threshold. Therefore, the fuzzy RDD covers the sharp RDD as a special case. As discussed in Imbens and Lemieux (2008), if the distribution of the potential outcomes is continuous in the running variable R, then, the discontinuity becomes a valid instrumental variable for the treatment, where the instrument is an indicator for exceeding the threshold r_0 in the running variable R. In this context, compliers are the individuals that receive the treatment when the running variable approaches the threshold from above, but not when it approaches from below, and monotonicity means that all individuals that switch treatment at the discontinuity do so in the same direction. Based on the discussions above, it is easy to see that this design fits in the framework outlined in previous section, with the exception that the identification is local at the threshold r_0, which requires nonparametric estimation; see Imbens and Lemieux (2008) for more discussions.

Under some Assumptions, Frandsen, Frölish and Melly (2012) showed that the distribution functions of the potential outcomes for the compliers at the threshold r_0 are identified for any y in $\mathbb{R}$ from the joint distribution of (Y, D, R), which are given by, respectively,

$$F_{Y(0)|\mathcal{C},R=r_0}(y) = \frac{\lim\limits_{r\to r_0^+} E\big[I\{Y \le y\}(1-D)|R=r\big] - \lim\limits_{r\to r_0^-} E\big[I\{Y \le y\}(1-D)|R=r\big]}{\lim\limits_{r\to r_0^+} E\big[1-D|R=r\big] - \lim\limits_{r\to r_0^-} E\big[1-D|R=r\big]},$$

and

$$F_{Y(1)|\mathcal{C},R=r_0}(y) = \frac{\lim\limits_{r\to r_0^+} E\big[I\{Y \le y\}D|R=r\big] - \lim\limits_{r\to r_0^-} E\big[I\{Y \le y\}D|R=r\big]}{\lim\limits_{r\to r_0^+} E\big[D|R=r\big] - \lim\limits_{r\to r_0^-} E\big[D|R=r\big]}.$$

It is important to note that the distributions are identified only for the local compliers; that is, the compliers whose running variable is arbitrarily close to the threshold r_0. In this setting, a natural estimation to consider is the quantile treatment effect for the local compliers:

$$\Delta_{\tau|\mathcal{C},R=r_0} = q_{1,\tau|\mathcal{C},R=r_0} - q_{0,\tau|\mathcal{C},R=r_0},$$

where $q_{j,\tau|\mathcal{C},R=r_0}$ is the τ-th conditional quantile of the distribution function of the potential outcome $Y(j)$ for the local compliers, namely, $q_{j,\tau|\mathcal{C},R=r_0} = \inf\big\{y : F_{Y(j)|\mathcal{C},R=r_0}(y) \ge \tau\big\}$ for $j = 0$ and 1.

It is clear that both the representations of $F_{Y(1)|\mathcal{C},R=r_0}(y)$ and $F_{Y(0)|\mathcal{C},R=r_0}(y)$ are functions of four conditional means at boundary points. To estimate these four conditional means, Frandsen et al. (2012) considered using local linear techniques because it can automatically correct bias at boundaries. The estimated conditional CDFs are subsequently inverted to obtain an estimation for $\Delta_{\tau|\mathcal{C},R=r_0}$. Moreover, they established the uniform consistency and asymptotic normality for the proposed estimator; see Frandsen et al. (2012) for details.

Recently, Shen and Zhang (2016) suggested various consistent uniform tests for examining whether a treatment or policy intervention can unambiguously improve or deteriorate the outcome of interest and whether a treatment or policy intervention has an effect on any part of the outcome distribution under the RDD framework, which complements the work by Frandsen et al. (2012). Furthermore, they showed under some regularity conditions that the proposed tests are distribution free, meaning that the critical values and p-values can be easily tabulated; see Shen and Zhang (2016) for more discussions.

4.3 Other Approaches to Identify QTE

So far, we have reviewed instrumental variable methods to estimate conditional or unconditional QTEs under the framework developed by Imbens and Angrist (1994) and Abadie (2003). Another most-used approach to instrumental variable identification and estimation of QTE is the instrumental variable quantile regression (IVQR) model developed by Chernozhukov and Hansen (2005, 2006). In this setup, instead of restricting heterogeneity in the treatment choice equation, it restricts heterogeneity in the outcome equation by imposing a stochastic rank preservation condition, a restriction on the evolution of individual ranks across treatment states, under which QTE can be identified. On the surface, those two models do not seem to be connected and neither model is more general than the other. In a recent paper, Wüthrich (2019) discussed the relationship between the IVQR model and the LQTE model, finding that the two models are closely-connected; see Wüthrich (2019) for more discussions. Some extensions to setups with nonbinary instruments while maintaining the assumption that the treatment is binary also have been considered in the treatment effect literature. First, if the instrument is multi-valued (or there are several instruments), it is easy to see that by the similar arguments, one can identify an LQTE with respect to any pair of distinct values of instrument variable Z satisfying Assumption 3.1. When the instrument is continuous, it is possible to identify a continuum of treatment effects, which were outlined in Heckman and Vytlacil (2001) and Heckman and Vytlacil (2005), who developed this approach for average treatment effects and called the resulting parameter based on an infinitesimal change in the instrument the marginal treatment effect. Using similar arguments, Carneiro and Lee (2009) extended these ideas to the identification and estimation of the quantile analogs, the marginal quantile treatment effect. For more details, see Heckman and Vytlacil (2001), Heckman and Vytlacil (2005), Carneiro and Lee (2009) and the references therein.

In some applications, researchers may confront the problem of the possible existence of unobserved confounders and proper IV is not available. In these settings, difference-in-differences (DID) models, which aim to attain identification by restricting the way in which unobserved confounders affect the outcome of interest over time with finite time periods, can be used to identify the treatment effect parameters of interest. Recent studies on QTEs under this framework include, but not limited to, the papers by Fan and Yu (2012), Callaway and Li (2019), and Callaway, Li and Oka (2018) and among others. Because of space limitations, we refer the interested reader to the original references for further details.

§5 Conclusion

This paper provides a selective survey on some recent methodological advancements in the evaluation of quantile treatment effect under the framework of selection on observables or unobservables. First, the review goes to the classical framework that the treatment variable is exogenous conditional on some observable variables. Under this framework, it is shown that the parameters of interest, such as quantile treatment effect or quantile treatment effect on the treated, can be identified directly from the joint distribution of the observed data. However, when the treatment variable is endogenous or self-selected even though conditioning on some observable variables, to identify meaningful treatment parameters, instrumental variable is needed and it provides a powerful tool to address this problem. Therefore, we then proceed by reviewing instrumental variable methods to estimate quantile treatment effect. In addition to the traditional exclusion and relevance conditions for the instrument, the models considered impose that the treatment either weakly increases or weakly decreases with the instrument for all individuals in the population. Under these assumptions, the entire distributions of the control and treated outcomes are identified for the individuals which react to the instrument without imposing any restrictions on treatment effect heterogeneity. Finally, we have also summarized some recent extensions to the data-rich environments, to the regression discontinuity design and discussed some other approaches to identify quantile treatment effect, some of which are still in the international research frontier.

References

[1] A Abadie. *Bootstrap tests for distributional treatment effects in instrumental variable models*, J Am Stat Assoc, 2002, 97(457):284-292.

[2] A Abadie. *Semiparametric instrumental variable estimation of treatment response models*, J Econometrics, 2003, 113(2):231-263.

[3] A Abadie, J D Angrist, G W Imbens. *Instrumental variable estimates of the effect of subsidized training on the quantile of trainee earnings*, Econometrica, 2002, 70(1): 91-117.

[4] A Abadie, G W Imbens. *Large sample properties of matching estimators for average treatment effects*, Econometrica, 2006, 74(1):235-267.

[5] A Abadie, G W Imbens. *Matching on the estimated propensity score*, Econometrica, 2016, 84(2):781-807.

[6] J Abrevaya, Y C Hsu, R P Lieli. *Estimating conditional average treatment effects*, J Bus Econ Stat, 2015, 33(4):485-505.

[7] J D Angrist, G W Imbens, D B Rubin. *Identification of causal effects using instrumental variables*, J Am Stat Assoc, 1996, 91(434):444-455.

[8] J D Angrist, J S Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2008.

[9] J D Angrist, J S Pischke. *Mastering 'Metrics: The Path from Cause to Effect*, Princeton University Press, 2014.

[10] S Athey, G W Imbens. *The state of applied econometrics: Causality and policy evaluation*, J Econ Persp, 2017, 31(2):3-32.

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

[12] Z W Cai. *Regression quantiles for time series*, Econometric Theory, 2002, 18(1):169-192.

[13] Z W Cai, Y Fang, M Lin, S F Tang. *Inferences for partially conditional quantile treatment effect models*, Working Paper, Department of Economics, University of Kansas, 2019a.

[14] Z W Cai, Y Fang, M Lin, S F Tang. *Partially conditional quantile treatment effect models with endogenous treatment*, Working Paper, Department of Economics, University of Kansas, 2019b.

[15] Z W Cai, Y Fang, M Lin, S F Tang. *Testing unconfoundedness assumption using auxiliary variables*, Working Paper, Department of Economics, University of Kansas, 2019c.

[16] Z W Cai, X Wang. *Nonparametric estimation of conditional VaR and expected shortfall*, J Econometrics, 2008, 147(1):120-130.

[17] B Callaway, T Li. *Quantile treatment effects in difference in differences models with panel data*, Quantitative Economics, 2019, 10(4): 1579-1618.

[18] B Callaway, T Li, T Oka. *Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods*, J Econometrics, 2018, 206(2):395-413.

[19] P Carneiro, S Lee. *Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality*, J Econometrics, 2009, 149(2):191-208.

[20] V Chernozhukov, I Fernández-Val, A Galichon. *Quantile and probability curves without crossing*, Econometrica, 2010, 78(3):1093-1125.

[21] V Chernozhukov, C Hansen. *An IV model of quantile treatment effects*, Econometrica, 2005, 73(1):245-261.

[22] V Chernozhukov, C Hansen. *Instrumental quantile regression inference for structural and treatment effect models*, J Econometrics, 2006, 132(2):491-525.

[23] D R Cox. *Planning of Experiments*, Wiley, New York, 1958.

[24] R K Crump, V J Hotz, G W Imbens, O A Mitnik. *Nonparametric tests for treatment effect heterogeneity*, Rev Econ Stud, 2008, 90(3):389-405.

[25] R H Dehejia, S Wahba. *Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs*, J Am Stat Assoc, 1999, 94(448):1053-1062.

[26] K Doksum. *Empirical probability plots and statistical inference for nonlinear models in the two-sample case*, Ann Stat, 1974, 2(2):267-277.

[27] S G Donald, Y C Hsu. *Estimation and inference for distribution functions and quantile functions in treatment effect models*, J Econometrics, 2014, 178(3):383-397.

[28] Y Q Fan, Z F Yu. *Partial identification of distributional and quantile treatment effects in difference-in-differences models*, Econom Lett, 2012, 115(3):511-515.

[29] S Firpo. *Efficient semiparametric estimation of quantile treatment effects*, Econometrica, 2007, 75(1):259-276.

[30] B R Frandsen, M Frölich, B Melly. *Quantile treatment effects in the regression discontinuity design*, J Econometrics, 2012, 168(2):382-395.

[31] M Frölich. *Propensity score matching without conditional independence assumption–with an application to the gender wage gap in the United Kingdom*, Econom J, 2007, 10(2):359-407.

[32] M Frölich, B Melly. *Unconditional quantile treatment effects under endogeneity*, J Bus Econ Stat, 2013, 31(3):346-357.

[33] J Hahn. *On the role of the propensity score in efficient semiparametric estimation of average treatment effects*, Econometrica, 1998, 66(2):315-331.

[34] P Hall, Q Yao. *Approximating conditional distributions using dimension reduction*, Ann Stat, 2005, 13(3): 1404-1421.

[35] J J Heckman, H Ichimura, J A Smith, P E Todd. *Characterizing selection bias using experimental data*, Econometrica, 1998, 66(5):1017-1098.

[36] J J Heckman, H Ichimura, P E Todd. *Matching as an econometric evaluation estimator: evidence from evaluating a job training programe*, Rev Econ Stud, 1997, 64(4): 605-654.

[37] J J Heckman, H Ichimura, P E Todd. *Matching as an econometric estimator evaluation*, Rev Econ Stud, 1998, 65(2): 261-294.

[38] J J Heckman, J R Robb. *Alternative methods for evaluating the impact of interventions: An overview*, J Econometrics, 1985, 30(1-2):239-267.

[39] J J Heckman, J Smith, N Clements. *Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts*, Rev Econ Stud, 1997, 64(4):487-535.

[40] J J Heckman, E Vytlacil. *Local instrumental variables*, In: Hsiao, C, Morimune, K, Powell, J (Eds), Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya, Cambridge University Press, Cambridge, 2001.

[41] J J Heckman, E Vytlacil. *Structural equations, treatment effects, and econometric policy evaluation*, Econometrica, 2005,73(3):669-738.

[42] K Hirano, G W Imbens, G Ridder. *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica, 2003, 71(4):1161-1189.

[43] Y C Hsu, T C Lai, R P Lieli. *Estimation and inference for distribution functions and quantile functions in endogenous treatment effect models*, IEAS Working Paper, 2015.

[44] G W Imbens, J D Angrist. *Identification and estimation of local average treatment effects*, Econometrica, 1994, 62(2):467-475.

[45] G W Imbens, T Lemieux. *Regression discontinuity designs: a guide to practice*, J E-conometrics, 2008, 142(2):615-635.

[46] G W Imbens, D B Rubin. *Estimating outcome distributions for compliers in instrumental variables models*, Rev Econ Stat, 1997, 64(4):555-574.

[47] G W Imbens, D B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, 2015.

[48] G W Imbens, J M Wooldridge. *Recent developments in the econometrics of program evaluation*, J Econ Lit, 2009, 47(1):5-86.

[49] R Koenker, G Bassett. *Regression quantiles*, Econometrica, 1978, 46(1):33-50.

[50] J Li, J Lv. *High-dimensional varying index coefficient quantile regression model*, Working Paper, Department of Statistics and Applied Probability, National University of Singapore, 2019.

[51] M J Lee. *Matching, Regression Discontinuity, Difference in Differences, and Beyond*, Oxford University Press, 2016.

[52] S Lee, R Okui, Y J Whang. *Doubly robust uniform confidence band for the conditional average treatment effect function*, J Appl Economet, 2017, 32(7):1207-1225.

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

[54] Z Q Liu, Z W Cai, Y Fang, M Lin. *Statistical analysis and evaluation of macroeconomic policies: A selective review*, Appl Math J Chinese Univ, 2020, 35(1):57-83.

[55] B Melly. *Estimation of counterfactual distributions using quantile regression*, Discussion Paper, Universität St Gallen, 2006.

[56] B Melly, K Wüthrich. *Local quantile treatment effects*, Handbook of quantile regression, Chapman and Hall/CRC, 2017.

[57] P R Rosenbaum, D B Rubin. *The central role of the propensity score in observational studies for causal effects*, Biometrika, 1983, 70(1):41-55.

[58] P R Rosenbaum, D B Rubin. *Reducing bias in observational studies using sub-classfication on the propensity score*, J Am Stat Assoc, 1985, 79(387):516-524.

[59] D B Rubin. *Estimating causal effects of treatments in randomized and nonrandomized studies*, J Educ Psychol, 1974, 66(5):688-701.

[60] D B Rubin. *Randomization analysis of experimental data: The Fisher randomization test comment*, J Am Stat Assoc, 1980, 75(371):591-593.

[61] D B Rubin. *Formal mode of statistical inference for causal effects*, J Statist Plann Inference, 1990, 25(3):279-292.

[62] S Shen, X H Zhang. *Distributional tests for regression discontinuity: Theory and empirical examples*, Rev Econ Stat, 2016, 98(4):685-700.

[63] S F Tang, Z W Cai, Y Fang, M Lin. *A new quantile treatment effect model to study smoking effect on birth weight during mother's pregnancy*, Working Paper, Department of Economics, University of Kansas, 2019.

[64] G Vazquez-Bare. *Identification and estimation of spillover effects in randomized experiments*, arXiv preprint arXiv:1711.02745, 2019a.

[65] G Vazquez-Bare. *Causal Spillover effects using instrumental variables*, Working Paper, Department of Economics, University of California, Santa Barbara, 2019b.

[66] K Wüthrich. *A comparison of two quantile models with endogeneity*, J Bus Econ Stat, 2019, 34(3):1-28.

Department of Statistics, School of Economics, Xiamen University, Xiamen 361005, China.
E-mail: tangshengfang103@163.com