Melly, Blaise; Wüthrich, Kaspar

Working Paper

Local quantile treatment effects

Discussion Papers, No. 16-05

Provided in Cooperation with:
Department of Economics, University of Bern

Suggested Citation: Melly, Blaise; Wüthrich, Kaspar (2016) : Local quantile treatment effects, Discussion Papers, No. 16-05, University of Bern, Department of Economics, Bern

This Version is available at:
https://hdl.handle.net/10419/149118

UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Local quantile treatment effects

Blaise Melly
Kaspar Wüthrich

16-05

March 2016

DISCUSSION PAPERS

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

Local quantile treatment effects

Chapter prepared for the *Handbook of Quantile Regression*

Blaise Melly[1] and Kaspar Wüthrich[2]

Acknowledgements: We thank participants at the New Directions in Quantile Regression conference in Cambridge UK and especially the editor Roger Koenker for very helpful comments.

[1]University of Bern, Switzerland, blaise.melly@vwi.unibe.ch
[2]University of Bern, Switzerland, kaspar.wuethrich@vwi.unibe.ch

1

Local quantile treatment effects

CONTENTS

This chapter reviews instrumental variable models of quantile treatment effects. We focus on models that achieve identification through a monotonicity assumption in the treatment choice equation. We discuss the key conditions, the role of control variables as well as the estimands in detail and review the literature on estimation and inference. Then we consider extensions to multiple and continuous instruments, to the regression discontinuity design, and discuss the testability of the assumptions. Finally, we compare this approach to the alternative instrumental variable approach reviewed by Chernozhukov et al. (2016). Two open research problems are highlighted in the conclusion.

Keywords: instrumental variables, local quantile treatment effects, monotonicity, compliers
JEL classification: C21, C26

1.1 Introduction

Policy makers are often interested in the causal effect of a binary treatment D on an outcome variable Y. This effect is frequently summarized by the average treatment effect $E[Y_1] - E[Y_0]$, where Y_1 and Y_0 are the treatment and control potential outcomes that were introduced by Neyman (1923) and popularized by Rubin (1974). This single measure, however, gives only a partial view over the effects of the treatment. The treatment may not affect the mean of the outcome distribution but may increase its variance or change its shape. From a policy perspective, an intervention that helps to raise the lower tail of an income distribution is often more appreciated than an intervention that shifts the median, even if the average treatment effects of both interventions are identical. In another application. In clinical trials, a drug may increase the long-term survival probability without affecting short-term survival.

We naturally have more information about the treatment effect if we know the whole distributions of the potential outcomes, $F_{Y_0}(y)$ and $F_{Y_1}(y)$, or equivalently their inverse, the quantile functions $Q_{Y_0}(\tau)$ and $Q_{Y_1}(\tau)$, where $\tau \in (0,1)$ denotes the quantile index. An intuitive way to summarize the treatment effect consists in reporting the quantile treatment effect (QTE) function

$$\Delta(\tau) = Q_{Y_1}(\tau) - Q_{Y_0}(\tau).$$

QTEs allow for assessing numerous interesting hypotheses. For instance, if the treatment has a pure location shift effect, then QTEs will be constant as a function of the quantile. The location scale shift model implies a monotone QTE function. If the distribution of Y_1 has first-order stochastic dominance over the distribution of Y_0, then the QTE function will be above the zero line. Thus, the QTE function represents an intuitive way to report the effect of a treatment on the marginal distribution of the outcome variable.

It would be even more informative to know the whole joint distribution of Y_1 and Y_0. Heckman et al. (1997) discuss several of the additional evaluation questions that could be analyzed if we knew this joint distribution. For instance, we could calculate the distribution of the individual treatment effects, $F_{Y_1-Y_0}(y)$, and in particular recover the proportion of units who benefit from the treatment. Alas, this joint distribution is not identified even in the ideal situation where the treatment has been randomized. Without further assumptions we can only bound the joint distribution based on the classical inequalities by Hoeffding and Fréchet and the resulting identified sets are almost always very wide.

Rank preservation is a common assumption that allows for recovering the joint distribution from the marginals. If we assume that each individual maintains its rank in the distribution of the outcome regardless of his treatment status, then the τ QTE is the treatment effect for individuals at the τ quantile of the potential outcome distributions. Doksum (1974) and Lehmann (1975)

were the first to suggest this measure of the treatment effect and to discuss its properties. The idea is that each subject possesses an underlying "proneness" or ability – depending on the application proneness to die early, to learn fast, to grow taller – which does not change with the treatment. In some applications, rank preservation is natural because it seems unlikely that the treatment makes weak subjects robust and strong subjects weak. On the opposite, if we consider an application where the outcome is the wage and the treatment is the sector of employment, it is implausible that the best professors of philosophy are also the best bricklayers. The methods reviewed in this chapter do not rely on the rank invariance assumption. We therefore interpret QTE as the difference between the same quantile of two distributions. However, the rank preservation assumption can always be added on the top to enrich the interpretation of the results.

In a randomized control trial with perfect compliance, the sample quantiles in the treated and control groups are consistent for $Q_{Y_1}(\tau)$ and $Q_{Y_0}(\tau)$ and, consequently, their difference is consistent for $\Delta(\tau)$. In practice, $\Delta(\tau)$ can be estimated using a simple quantile regression of Y on D and a constant. However, it is often impossible to impose perfect compliance: some subjects may be assigned randomly to the treatment but they may refuse to take it, other may be assigned to the control group but may find a way to get the treatment anyway. As a result, the observed treatment variable is self-selected and its endogeneity renders standard quantile regression inconsistent just as it is the case for least squares methods. On the other hand, since the assignment has been randomized, it is still possible to identify the causal effect of the assignment on the outcome distribution. In the context of clinical trial, this is called the intention-to-treat (ITT) effect. While the ITT effect might be an interesting parameter, especially when the only possible intervention is to assign the treatment, the ITT effect does not provide an estimate of the treatment effect.

Instrumental variable methods provide a powerful tool to address this problem. There are several approaches to instrumental variable identification and estimation of QTE. In this chapter, we focus on models that achieve identification through a monotonicity assumption in the treatment choice equation and do not rely on rank preservation or other similar assumptions. This is in sharp contrast to the instrumental variable approach reviewed by Chernozhukov et al. (2016). We compare both approaches in Section 1.5. We also limit our survey to models for binary endogenous variables because almost no results have been obtained for multi-valued or continuous treatments within the framework that we consider.

In the simplest set-up, both the treatment and the assignment (instrument) are binary. With imperfect compliance, there are some individuals who do not react to a change in the assignment; either they always refuse the treatment or they always find a way to get the treatment. Since we allow the treatment effect to be arbitrarily heterogeneous, we cannot identify the treatment effects for these units. There is no information in the data for them about one of

both potential outcomes. This implies that it is also impossible to identify the treatment effect for the whole population (QTE) or, if some units are always treated independently of the assignment, for the treated sub-population (QTE on the treated). We are only able to identify effects for the individuals that responds to a change in the value of the instrument. These individuals are called compliers because they comply with the assignment.

Our main estimand is, therefore, the QTE for the compliers, which we call the local quantile treatment effect (LQTE) because it corresponds to the QTE for a subpopulation. This terminology is in analogy to the local average treatment effect (LATE) of Imbens and Angrist (1994) and is *not* related to nonparametric methods that are local with respect to the covariates. Whether or not the subpopulation of compliers is of interest depends heavily on the empirical context; see, for instance, the controversial discussion in Imbens (2010), Deaton (2010), and Heckman and Urzúa (2010). While other populations would naturally be of interest, we would like to emphasize again that there is no information in the data about their treatment effects. The LQTE is the QTE for the the largest population for which the effect is identified. As will be obvious in Section 1.5, the approaches that recover treatment effects for larger populations achieve identification by means of extrapolation from the compliers.

The methods surveyed in this chapter have already been applied numerous times in very different setups. For instance, Abadie et al. (2002) estimate the effect of JTPA training programs on the earnings distribution of previously unemployed individuals. Eren and Ozbeklik (2014) similarly study the effects of the Job Corps programs. They both identify the causal effects of the programs by exploiting a randomized experiment with imperfect compliance. The methods are, however, not limited to the analysis of experimental data. Many variables of interest cannot be randomized in economics and other social sciences. For this reason, researchers use 'natural experiments' to identify treatment effects. Ananat and Michaels (2008) note that the probability of a divorce is higher when the first-born child is a female. Using this source of variation, they find that a divorce has little effect on women's mean houshold income but it increases womens odds of having very high or very low income. Cawley and Meyerhoefer (2012) estimate the impact of obesity on the distribution of medical costs, instrumenting the respondents weight with the weight of a biological relative. Frölich and Melly (2013) use twin birth as an instrument for having several children and estimate its effect on the houshold income.

The basic framework was already developed in the 90s; Angrist and Pischke (2008), Imbens et al. (2014) and Imbens and Rubin (2015) provide interesting surveys of this approach to instrumental variables. In Section 1.2, after briefly summarizing the basics, we focus on the particularities of the quantile estimands and their implications for the identification of the effects in setups without and with covariates. Section 1.3 briefly reviews the literature on estimation and inference. In Section 1.4 we consider two extensions within the same framework – nonbinary instruments and the regression discontinuity

design – as well as the testability of the identifying assumptions. Section 1.5 compares the model in this chapter to the instrumental variable quantile regression model introduced by Chernozhukov and Hansen (2005) and reviewed in Chernozhukov et al. (2016). Finally, Section 1.6 briefly summarizes the findings and highlights two important open problems in this literature.

1.2 Framework, estimands and identification

1.2.1 Without covariates

We develop our presentation in the context of a randomized trial with noncompliance. The units of observation are assigned to a treatment but this assignment cannot be perfectly enforced. We consider the simplest case where both the assignment Z as well as the treatment D are binary. In many applications it is reasonable to assume that there is no interference between units. This assumption, first introduced by Cox (1958), is called the Stable Unit of Treatment Assignment (SUTVA) by Rubin (1980).

Assumption 1 *SUTVA: For any unit the value of the treatment when exposed to the assignment z and of the outcome when exposed to the assignment z and the treatment d is the same regardless of the treatments and assignments that other units receive.*

SUTVA excludes any kind of interaction between units. A classical violation of this assumption is a setting where the treatment is a vaccine that immunizes the unit against a contagious disease. The effect of this vaccine will necessarily be a function of the number of persons who have already been vaccinated. Peer effects and general equilibrium effects are also excluded; e.g. SUTVA is violated if the wage effect of getting a college degree depends on the proportion of the labor force having the same degree. Thus, as all partial equilibrium approaches, it is well suited only for small-scale interventions.

We use the potential outcomes notation. D_z denotes the potential treatment status when the assignment is set exogenously to $z \in \{0, 1\}$. The observed treatment is related to the potential treatments as $D = D_1 Z + D_0(1 - Z)$. In the case of perfect compliance, $D_z = z$ such that $D = Z$ and the treatment itself has been randomized. With imperfect compliance this equality breaks down and we have to take into account the fact that the treatment has been self-selected and is, therefore, endogenous. Similarly, we define four potential outcomes Y_{zd} for each combination of z and $d \in \{0, 1\}$. Note that without Assumption (1) the potential outcomes of each unit would depend on the assignments and treatments of all units in the population. It is impossible to identify any treatment effect without imposing some restriction on the dependence between units.

With a randomized assignment it is easy to identify the average causal effect of the assignment on the outcome, $E\left[Y_{1D_1} - Y_{0D_0}\right]$, as well as its effect on the distribution and quantile functions of the outcome. In the context of clinical trial, this is called the intention-to-treat (ITT) effect. While the ITT effect might be an interesting estimate, especially when the only possible intervention is to assign the treatment, the ITT effect does not take into account the information about the treatment. If compliance is low, it will underestimate the absolute value of the effect of the treatment. In the following, we will take the extent of non-compliance into account in order to estimate the effect of the treatment.

With D and Z being binary, we can partition the population into four types $\mathcal{T}$ defined by D_1 and D_0:

D_1	D_0	Type
1	1	Always-takers ($\mathcal{T} = a$)
1	0	Compliers ($\mathcal{T} = c$)
0	1	Defiers ($\mathcal{T} = d$)
0	0	Never-takers ($\mathcal{T} = n$)

The never-takers and the always-takers do not react to the assignment and do not contribute to the ITT effect. We have no source of random variation for these types and we will not be able to identify any treatment effects for them. Both the compliers and the defiers react to a change in the assignment but they do so in opposite directions. Therefore, the ITT effect is a weighted average of the individual treatment effects with positive weights for the compliers, negative weights for the defiers and zero weights for both other types (e.g., Angrist et al., 1996). Given that we observe each unit only in one state of the world, we are unable to determine the type of each unit to separate compliers from defiers. There are at least two ways to solve this mixture problem: either we restrict the outcome to be the same for the compliers and the defiers (homogeneity assumption for the treatment effect) or we assume that there are no defiers (homogeneity assumption for the assignment effect). The approach to instrumental variables that we review in this chapter follows the second way. De Chaisemartin (2014) shows that these two types of assumptions can be combined and would lead to a similar interpretation of the same estimators.

We impose the following assumptions:

Assumption 2

1. *Independent instrument:* $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0) \perp\!\!\!\perp Z$
2. *Exclusion restriction:* $P(Y_{1d} = Y_{0d}) = 1$ *for* $d \in \{0, 1\}$
3. *Relevant instrument:* $\Pr(\mathcal{T} = c) > 0$ *and* $\Pr(Z = 1) \in (0, 1)$
4. *Monotonicity:* $\Pr(\mathcal{T} = d) = 0$

Assumption 2.1 is an unconfounded instrument restriction. It is mechanically satisfied if the assignment has been randomized. Full independence is required for the identification of the whole LQTE process but a slighter weaker

local quantile independence condition is enough to identify a single LQTE, see Assumption 4 in Section 1.4.3. The exclusion restriction (Assumption 2.2) requires that the assignment Z must have no direct effect on the potential outcomes to be a valid instrumental variable. It is important to note that this exclusion restriction is unrelated to – and, in particular, not implied by – the randomization of the instrument (Assumption 2.1). In medical studies it is well-known that there is an psychological effect of being assigned to a treatment. For this reason, even the control group receives a placebo treatment and the exclusion restriction is then likely to be satisfied. There are cases where it is more difficult to find a placebo treatment. For instance, if unemployed individuals receive an invitation to attend a CV-writing course, they may decide not to attend the course but start reading books to improve their application documents. This would violate the exclusion restriction. Under Assumptions 1 and 2.2 we can define potential outcomes in terms of D alone: $Y_0 = Y_{00} = Y_{10}$ and $Y_1 = Y_{01} = Y_{11}$. Assumption 2.3 requires that at least some individuals react to changes in the value of the instrument. The strength of the instrument can be measured by $\Pr(\mathcal{T} = c)$. These first three assumptions are common to all instrumental variable models. Assumption 2.4, which is often referred to as monotonicity, is specific to the approach we review in this chapter. It requires that D_z weakly increases with z for all individuals. Note that it is mechanically satisfied if there is perfect one-sided compliance, which is quite common. If those assigned to receive the control treatment can be denied access to the active treatment, then there are no defiers and no always-takers. In addition to automatically satisfying the monotonicity assumption this also implies that the treated are all compliers such that the LQTE is equal to the QTET.

Imbens and Angrist (1994) show that, under Assumptions 1 and 2, the Wald (1940) estimator is consistent for the average treatment effect for the compliers, usually referred to as the LATE:

$$E\left(Y_1 - Y_0 | \mathcal{T} = c\right) = \frac{E\left[Y|Z=1\right] - E\left[Y|Z=0\right]}{E\left[D|Z=1\right] - E\left[D|Z=0\right]}.$$

The result for the average of the dependent variable naturally also apply to transformations of the dependent variable such as the distribution function ($F_Y\left(y\right) = E\left[1\left(Y \leq y\right)\right]$). This directly provides results for the distributional treatment effect for the compliers – $F_{Y_1}\left(y|\mathcal{T} = c\right) - F_{Y_0}\left(y|\mathcal{T} = c\right)$. Unfortunately, this approach does not yield quantile treatment effects for the compliers, $Q_{Y_1}\left(\tau|\mathcal{T} = c\right) - Q_{Y_0}\left(\tau|\mathcal{T} = c\right)$, because $F_{Y_1}\left(y|\mathcal{T} = c\right)$ and $F_{Y_0}\left(y|\mathcal{T} = c\right)$ must be obtained separately such that one can invert them to obtain quantile functions. Obtaining the effect on the distribution is enough to test some hypotheses such as the absence of any effect or first order stochastic dominance, but it is not enough to recover important information about the treatment effects. The quantiles and LQTEs have an intuitive and straightforward interpretation. Their unit of measurement is the same as the unit of the outcome itself. This allows, for example, to test the hypothesis that the treatment effect

exerts a pure location shift or a location-scale shift of the distribution of the outcome. With distributional treatment effects we may find that the effect is positive over a part of the support of Y and negative over another part, but we cannot identify how relevant these two parts are without knowing separately the distributions of Y_1 and Y_0. For instance, it may be that the treatment effect is positive over 99% of the quantiles and negative only over 1%; or it may be the opposite.

Imbens and Rubin (1997) show that the distributions of Y_1 and Y_0 are identified for the compliers by working directly with the densities. Abadie (2002) gives more convenient representations for the cumulative distribution functions (cdf):

$$F_{Y_1}(y|\mathcal{T}=c) = \frac{E[1(Y \leq y)D|Z=1] - E[1(Y \leq y)D|Z=0]}{E[D|Z=1] - E[D|Z=0]} \tag{1.1}$$

The reason is that $E[1(Y \leq y)D|Z=1] = F_{Y_1}(y|\mathcal{T}=a)\Pr(\mathcal{T}=a) + F_{Y_1}(y|\mathcal{T}=c)\Pr(\mathcal{T}=c)$ and $E[1(Y \leq y)D|Z=0] = F_{Y_1}(y|\mathcal{T}=a)\Pr(\mathcal{T}=a)$, such that the numerator simplifies to $F_{Y_1}(y|\mathcal{T}=c)\Pr(\mathcal{T}=c)$. Similarly, the denominator simplifies to $\Pr(\mathcal{T}=c)$. A similar reasoning can be used to identify the distribution of the control potential outcome:

$$F_{Y_0}(y|\mathcal{T}=c) = \frac{E[1(Y \leq y)(1-D)|Z=1] - E[1(Y \leq y)(1-D)|Z=0]}{E[1-D|Z=1] - E[1-D|Z=0]}. \tag{1.2}$$

These two distributions can then be inverted to obtain the quantile functions:

$$Q_{Y_1}(\tau|\mathcal{T}=c) = \inf\{y : F_{Y_1}(y|\mathcal{T}=c) \geq \tau\}$$

and

$$Q_{Y_0}(\tau|\mathcal{T}=c) = \inf\{y : F_{Y_0}(y|\mathcal{T}=c) \geq \tau\}.$$

Consequently, the LQTEs are identified as

$$\Delta(\tau|\mathcal{T}=c) = Q_{Y_1}(\tau|\mathcal{T}=c) - Q_{Y_0}(\tau|\mathcal{T}=c).$$

1.2.2 In the presence of covariates: conditional LQTE

In almost all applications we also observe a vector of covariates X. We may want to include them in the estimation for two reasons. First, the validity of the independence assumption 2.1 may be plausible only after conditioning on covariates. This is the case for stratified randomized experiment when the assignment probabilities are different across the strata. This is also the case in observational studies when the instrument has not been randomized. For instance, Frölich and Melly (2013) use twin birth as an instrument for having

several children. Since it is well known that the probability of a twin birth is a function of the race of the parents and increases with the age of the mother, they control for these characteristics to satisfy the exclusion restriction. Even if this assumption is valid unconditionally, e.g. because the instrument has been unconditionally randomized, we will see in Section 1.3 that we may wish to include covariates in the estimation for efficiency reasons. Therefore, we consider the case where the instrument is valid after conditioning on the covariates, which also covers randomized instruments as a special case,

Assumption 3 *For all* $x \in supp(X)$

1. *Independent instrument:* $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0) \perp\!\!\!\perp Z|X = x$
2. *Exclusion restriction:* $P(Y_{1d} = Y_{0d}|X = x) = 1$ *for* $d \in \{0, 1\}$
3. *Relevant instrument:* $\Pr(\mathcal{T} = c|X = x) > 0$ *and* $\Pr(Z = 1|X = x) \in (0, 1)$
4. *Monotonicity:* $\Pr(\mathcal{T} = d|X = x) = 0$

Assumption 3 is simply the conditional version of Assumption 2. Note that Assumption 3.3 requires the support of X to be identical in the $Z = 0$ and the $Z = 1$ subpopulations. If the support condition is not met initially, we need to define the parameters relative to the common support.

The conditional distributions for the compliers are identified using the same approach as above, e.g. for Y_1,

$$F_{Y_1}(y|X = x, \mathcal{T} = c) = \frac{E[1(Y \leq y) D|X = x, Z = 1] - E[1(Y \leq y) D|X = x, Z = 0]}{E[D|X = x, Z = 1] - E[D|X = x, Z = 0]}. \quad (1.3)$$

The nonparametric estimation of these conditional distributions will naturally suffer from the curse of dimensionality. For this reason, Abadie et al. (2002) impose a linear restriction for the compliers

$$Q_{Y_0}(\tau|X = x, \mathcal{T} = c) = X'\beta(\tau)$$

and

$$Q_{Y_1}(\tau|X = x, \mathcal{T} = c) = X'\beta(\tau) + \Delta(\tau|\mathcal{T} = c).$$

The particular form of conditional quantile functions does not matter for what follows. In particular, we can easily introduce interactions between the treatment and the covariates, which allows for heterogeneous treatment effects with respect to the observables.

Equation (1.3) shows that $\Delta(\tau|\mathcal{T} = c)$ and $\beta(\tau)$ are identified but does not suggest a convenient parametric estimator. Abadie et al. (2002), applying a result in Abadie (2003), give the following weighted quantile regression

representation

$$(\beta(\tau), \Delta(\tau|\mathcal{T}=c)) = \arg\min_{\tilde{\beta},\tilde{\Delta}} E\left[W^{AAI} \cdot \rho_\tau\left(Y - X\tilde{\beta} - D\tilde{\Delta}\right)\right] \tag{1.4}$$

$$W^{AAI} = 1 - \frac{D(1-Z)}{1-E(Z|X)} - \frac{(1-D)Z}{E(Z|X)}. \tag{1.5}$$

The proof follows from the fact that

$$\begin{aligned}
E\left[W^{AAI}|Y,X,\mathcal{T}=a\right] &= 1 - \frac{1-E\left[Z|Y,X,D_0=D_1=1\right]}{1-E(Z|X)} = 0,\\
E\left[W^{AAI}|Y,X,\mathcal{T}=n\right] &= 1 - \frac{E\left[Z|Y,X,D_0=D_1=0\right]}{E(Z|X)} = 0,\\
E\left[W^{AAI}|Y,X,\mathcal{T}=c\right] &= 1.
\end{aligned}$$

In other words, on average, the weights W^{AAI} find the conditional compliers and annihilate the always- and never-takers. This suggests a weighted quantile regression estimator. However, the sample analog of (1.4) is not globally convex because the weights are positive and negative. We discuss the solution suggested by Abadie et al. (2002) below in Section 1.3.

1.2.3 In the presence of covariates: unconditional LQTE

In Section 1.2.2 we have discussed *conditional* quantile treatment effects, i.e. the quantile is defined within the population with the same X. For instance, Abadie et al. (2002) are interested in the effect of a training program on earnings. Assuming rank invariance, the conditional LQTE for $\tau = 0.1$ is the effect for the individuals who are at the 0.1 quantile of the earnings distribution given their age, education level, race, etc. This includes white workers in their peak earnings years with a college degree who earn well above the median of the whole population. This is not necessarily the estimand that the policy maker is interested in: she may be interested in the effects for unconditionally poor households. Similarly, in other applications, she may be interested in unconditionally low-birthweight, or unconditionally low achieving students. This distinction does not really play a role for averages because the unconditional mean is simply the average of the conditional means, but it matters for quantiles because the law of iterated expectation does not apply. While we may be interested in unconditional LQTE, we may need to include control variables in order to satisfy Assumption 3, in particular the (conditional) independence assumption, or we may wish to add other covariates for efficiency reasons. In other words, the set of covariates that we wish to include depends on both the estimand that we are interested in and on the exclusion restriction that we have to satisfy. There may be a tension between these objectives.

It is possible to separate the definition of the estimand from the assumptions needed for a causal interpretation. A first representation of the uncon-

ditional cdf of the potential outcomes for the compliers is given by

$$\begin{aligned}
F_{Y_1}(y|\mathcal{T}=c) &= \int F_{Y_1}(y|X=x,\mathcal{T}=c)dF_X(x|\mathcal{T}=c) \\
&= \int F_{Y_1}(y|X=x,\mathcal{T}=c)\frac{\Pr(\mathcal{T}=c|X=x)}{\Pr(\mathcal{T}=c)}dF_X(x) \\
&= \frac{\int \left(E\left[1\left(Y\leq y\right)D|X=x,Z=1\right]-E\left[1\left(Y\leq y\right)D|X=x,Z=0\right]\right)dF_X(x)}{\int \left(E\left[D|X=x,Z=1\right]-E\left[D|X=x,Z=0\right]\right)dF_X(x)}
\end{aligned} \tag{1.6}$$

where the first equality holds by the law of total expectation, the second by Bayes' law and the third by the representation of the conditional distribution in (1.3) and the fact that $\Pr(\mathcal{T}=c|X=x) = E\left[D|X=x,Z=1\right] - E\left[D|X=x,Z=0\right]$ as seen above. A similar result applies to $F_{Y_0}(y|\mathcal{T}=c)$.

Parts (b) and (c) of Theorem 3.1 in Abadie (2003) applied to the cdf provide a weighted representation of $F_{Y_1}(y|\mathcal{T}=c)$ and $F_{Y_0}(y|\mathcal{T}=c)$ (see Frölich and Melly, 2013):

$$\begin{aligned}
F_{Y_1}(y|\mathcal{T}=c) &= \frac{E\left[1\left(Y<y\right)DW^{FM}\right]}{E\left[DW^{FM}\right]} \\
F_{Y_0}(y|\mathcal{T}=c) &= \frac{E\left[1\left(Y<y\right)(1-D)W^{FM}\right]}{E\left[DW^{FM}\right]},
\end{aligned} \tag{1.7}$$

where

$$W^{FM} = \frac{Z-\Pr(Z=1|X)}{\Pr(Z=1|X)\left(1-\Pr(Z=1|X)\right)}(2D-1).$$

This result can be obtained from (1.6) via iterated expectations arguments. These weights are different from the W^{AAI} weights defined in Section 2.3. They find the conditional compliers in the following sense

$$\begin{aligned}
E\left[W^{FM}|X,Y,\mathcal{T}=a\right] &= E\left[W^{FM}|X,Y,\mathcal{T}=n\right] = 0 \\
E\left[W^{FM}|X,Y,\mathcal{T}=c\right] &= 2
\end{aligned}$$

but, in addition, they also balance covariates between treated compliers and non-treated compliers in the following sense

$$\begin{aligned}
E\left[W^{FM}|X,Y,\mathcal{T}=c,D=1\right] &= \frac{1}{\Pr(Z=1|X)} \\
&= \frac{1}{\Pr(Z=1|X,\mathcal{T}=c)} \\
&= \frac{1}{\Pr(D=1|X,\mathcal{T}=c)} \\
E\left[W^{FM}|X,Y,\mathcal{T}=c,D=0\right] &= \frac{1}{1-\Pr(Z=1|X)} \\
&= \frac{1}{1-\Pr(D=1|X,\mathcal{T}=c)}
\end{aligned}$$

This means that the weights W^{FM} not only find the conditional compliers but also reweight them such that the treated and control compliers have the same covariates distributions. They share the first property with the weights W^{AAI} of Abadie et al. (2002) and the second property with the inverse probability weights used when the treatment is exogenous conditional on the covariates. W^{FM} indeed nests inverse probability weights as a strict special case when the treatment is used as its own instrument, which is justified under exogeneity. Firpo (2007) discusses the estimation of unconditional QTEs following this approach.

Frölich and Melly (2013) suggest to consider unconditional quantile treatment effects

$$\Delta\left(\tau|\mathcal{T}=c\right)=Q_{Y_1}\left(\tau|\mathcal{T}=c\right)-Q_{Y_0}\left(\tau|\mathcal{T}=c\right)$$

while keeping Assumption 3 (i.e. we need covariates for the identification). Both unconditional quantile functions can be obtained by inverting the unconditional cdfs obtained in (1.6) or (1.7). Instead, it is also possible to directly obtain the LQTE by a weighted quantile regression

$$\left(Q_{Y_0}\left(\tau|\mathcal{T}=c\right),\Delta\left(\tau|\mathcal{T}=c\right)\right)=\underset{\tilde{Q}_{Y_0},\tilde{\Delta}}{\arg\min}E\left[W^{FM}\cdot\rho_\tau(Y-\tilde{Q}_{Y_0}-\tilde{\Delta}D)\right]. \tag{1.8}$$

1.3 Estimation and inference

In the absence of covariates, the sample analogs of (1.1) and (1.2) provide natural estimators of the cdfs of the potential outcomes. The estimated cdfs will necessarily be non-monotone but they can be monotonized for instance using the re-arrangement method of Chernozhukov et al. (2010). This allows to invert the cdfs and to obtain the quantile functions and subsequently the LQTEs.

In the presence of covariates, Abadie et al. (2002) suggest to estimate conditional LQTE based on the weighted quantile regression representation (1.4). Estimation based on (1.4) may be a difficult task because the weights (1.5) are positive and negative, implying that the objective function has many local minima. Therefore, Abadie et al. (2002) suggest to replace the weights (1.5) with their projection on (Y,D,X), which can be shown to be always positive. Their estimation strategy consists in three steps: (i) estimation of the instrument propensity score $E\left(Z|X\right)$ using nonparametric power series, (ii) estimation of the positive weights using nonparametric power series of the estimated weights on (Y,D,X), and (iii) weighted quantile regression using the estimated positive weights. The resulting estimator is $\sqrt{n}$ consistent and asymptotically normal. However, Hong and Nekipelov (2010) show that the

estimator suggested in Abadie et al. (2002) does not attain the semiparametric efficiency bound. The alternative and efficient estimator that they suggest uses density weighting for the compliers, which is similar to the efficient weighted least squares estimator in the presence of heteroscedasticity. This approach is nevertheless not popular because (i) it requires estimating the conditional density of Y for the compliers, and (ii) the weighted estimator is more difficult to interpret if the conditional model is misspecified.

For unconditional effects, analog estimators based on all three representations (1.6), (1.7), and (1.8) have been suggested. While parametric restrictions (e.g. linearity) are necessary to achieve the $\sqrt{n}$ consistency of the conditional LQTE estimators, unconditional LQTE can be estimated at the $\sqrt{n}$ rate without any parametric restrictions because the unconditional distributions are averages of conditional distributions.

Belloni et al. (2013) provide estimators based on (1.6) for environments with many control variables, either because many variables are available in the raw dataset or because we want to include interactions and other transformations of the control variables. They suggest Lasso-type methods that automatically select the relevant ones. Assuming that reduced form relationships are approximately sparse, they show that valid inference can be performed after data-driven selection of control variables. Moreover, they derive the limiting laws of the estimators of the whole quantile treatment effect process. This allows to construct confidence band for the LQTE function over a continuum of quantile, to test functional hypotheses and for dominance relations between the potential outcomes. Hsu et al. (2015) suggest a weighted cdf estimator based on (1.7) and derive the asymptotic distribution for the whole LQTE process. Finally, Frölich and Melly (2013) analyze a weighted quantile regression estimator based on (1.8). Their estimator is $\sqrt{n}$-consistent, asymptotically normal, and achieves the the semiparametric efficiency bound. In addition, Frölich and Melly (2013) show that adding covariates increases the precision of the estimator of the unconditional effect if these additional covariates (i) do not affect the instrument propensity score and (ii) affect the outcome. These conditions are often satisfied, for instance, when the instrument has been randomized. On the other hand, including covariates that affect the instrument propensity score and do not affect the outcome will increase the variance of the estimator. Note that such a comparison of variances does not make sense for the conditional LQTE because the definition of the estimand is a function of the covariates included.

Note that all the asymptotic results mentioned in this section rely on the continuity of the dependent variable. This assumption is not an identifying condition. On the contrary, the LQTE framework accommodates discrete outcomes and outcomes with mass points very naturally. It is also possible to show that the analog estimators of the cdfs of the potential outcomes are asymptotically normally distributed even for discrete outcomes. On the other hand, continuity of the dependent variable is a condition for obtaining well-behaved asymptotic distributions for the quantiles and LQTE estimators.

1.4 Extensions: regression discontinuity design, nonbinary instruments and testing

1.4.1 Regression discontinuity design

A fuzzy regression discontinuity design (RDD) exploits a discontinuity in the probability of treatment when a running variable R exceeds a threshold r_0. This section also covers the sharp RDD as a special case where the probability jumps from 0 to 1 at the threshold. If the distribution of the potential outcomes is continuous in R, then the discontinuity becomes a valid instrumental variable for the treatment. In this context, compliers are the units that switch their treatment status at the discontinuity and monotonicity means that all units that switch treatment at the discontinuity do it in the same direction. This design fits in the framework outlined in Section 1.2 with the exception that the identification is local at $R = r_0$, which requires nonparametric estimation.

The local version of (1.1) is given by

$$F_{Y_1}\left(y|\mathcal{T}=c, R=r_0\right) = \lim_{\varepsilon\to 0}\frac{E\left[1\left(Y\leq y\right)D|r_0<R<r_0+\varepsilon\right]-E\left[1\left(Y\leq y\right)D|r_0-\varepsilon<R<r_0\right]}{E\left[D|r_0<R<r_0+\varepsilon\right]-E\left[D|r_0-\varepsilon<R<r_0\right]} \tag{1.9}$$

and similarly for the control potential outcome. Note that the distributions are identified only for the local compliers, i.e. the compliers whose running variable is arbitrarily close to r_0. A natural estimand to consider is the QTE for the local compliers:

$$\Delta(\tau|\mathcal{T}=c, R=r_0) \equiv Q_{Y_1}(\tau|\mathcal{T}=c, R=r_0) - Q_{Y_0}(\tau|\mathcal{T}=c, R=r_0).$$

The representation of $F_{Y_1}\left(y|\mathcal{T}=c, R=r_0\right)$ in (1.9) is a function of four conditional means at boundary points (from the left and the right of r_0). Frandsen et al. (2012) use local linear techniques to estimate these means, which is not subject to bias from ignoring the running variable and automatically corrects for boundary effects. The cdfs are subsequently inverted to obtain an estimator of $\Delta(\tau|\mathcal{T}=c, R=r_0)$. They prove uniform consistency and asymptotic Gaussianity of the estimators for the whole QTE process for the local compliers. Of course, this nonparametric estimator only converges at the one-dimensional nonparametric rate.

1.4.2 Multi-valued and continuous instruments

In the previous two sections we focused on the case with a binary instrument. If the instrument is multi-valued (or there are several instruments), then it is obviously possible to identify a LQTE with respect to any pair

of distinct values of Z, satisfying Assumption 2. Instead of estimating many pairwise effects, one may prefer to estimate the LQTE for the largest complying sub-population. This is simply obtained by considering the value of Z that minimizes the treatment probability and the value that maximizes the treatment probability (given a monotonicity assumption defined with respect to $\Pr(D=1|Z)$).

When the instrument is continuous, it is possible to identify a continuum of treatment effects. Heckman and Vytlacil (2005) developed this approach for average treatment effects and called the resulting parameters marginal treatment effects. They show that the LATE can be represented as a weighted average of marginal treatment effects. Inversely, the marginal treatment effects can be considered as the limit form of the LATE parameter. Carneiro and Lee (2009) extend these ideas to the estimation of the quantile analogs, the marginal quantile treatment effects (MQTE). To simplify the notation we do not incorporate covariates in this section. Suppose that individuals choose their treatment status according to the following equation:

$$D = 1\{\Pr(D=1|Z) \geq \eta\}, \tag{1.10}$$

where $\eta|Z \sim U(0,1)$ is a scalar error term. Z is a continuous instrument that is independent of Y_0 and Y_1. For binary Z Vytlacil (2002) shows that (1.10) is equivalent to the monotonicity Assumption 2.4. Hence, this model can be seen as a generalization of the LQTE framework to general instruments. In this model, the main estimands of interest are the MQTE

$$\Delta(\tau|p) \equiv Q_{Y_1}(\tau|\eta=p) - Q_{Y_0}(\tau|\eta=p).$$

Carneiro and Lee (2009) show that $\Delta(\tau|p)$ is identified if

$$p \in supp\left(\Pr\left(D=1|Z,D=1\right)\right) \cap supp\left(\Pr(D=1|Z,D=0)\right).$$

They note that

$$\begin{aligned} F_Y\left(y|\Pr\left(D=1|Z\right)=p, D=1\right)\cdot p &= F_{Y_1}\left(y|\Pr\left(D=1|Z\right)=p, D=1\right)\cdot p \\ &= F_{Y_1}\left(y|\eta\leq p\right)\cdot p \\ &= \int_0^p F_{Y_1}\left(y|\eta=h\right)dh. \end{aligned}$$

By taking the derivative with respect to p on both side, this implies that

$$\begin{aligned} F_{Y_1}\left(y|\eta=p\right) =& F_Y\left(y|\Pr\left(D=1|Z\right)=p, D=1\right) \\ &+ \frac{\partial F_Y\left(y|\Pr\left(D=1|Z\right)=p, D=1\right)}{\partial p}. \end{aligned}$$

A similar result applies to the cdf of Y_0. Both cdfs can be inverted to obtain the quantile functions.

Yu (2014) suggests semiparametric estimators of the marginal quantile

treatment effects based on this representation. He allows for the presence of control variables and considers both conditional and unconditional MQTE. Finally, he derives the corresponding weak limits and shows the validity of the bootstrap for inference.

1.4.3 Testing instrument validity

Assumption 2 imposes testable implications on the joint distribution of (Y, D, Z). Under this set of asumptions, $F_Y(y|D = 1, Z = 1)$ is a mixture of $F_{Y_1}(y|\mathcal{T} = c)$ and $F_{Y_1}(y|\mathcal{T} = a)$ with identified mixing probabilities. One of the mixing distributions is also separately identified because $F_{Y_1}(y|\mathcal{T} = a) = F_Y(y|D = 1, Z = 0)$. Imbens and Rubin (1997) note that a violation of Assumption 2 may lead the resulting density function for the compliers, $f_{Y_1}(y|\mathcal{T} = c)$, to be negative. An equivalent result holds for the distribution of the control outcome.

More formally, Assumption 2 implies that, for every Borel set B in $supp(Y)$,

$$P(Y \in B, D = 1|Z = 1) - P(Y \in B, D = 1|Z = 0) = P(Y_1 \in B, D_1 > D_0),$$
$$P(Y \in B, D = 0|Z = 0) - P(Y \in B, D = 0|Z = 1) = P(Y_0 \in B, D_1 > D_0).$$

Because the right sides are nonnegative by the definition of probabilities, we obtain the following testable restriction (Balke and Pearl (1997), Heckman and Vytlacil (2005)):

$$P(Y \in B, D = 1|Z = 1) - P(Y \in B, D = 1|Z = 0) \geq 0, \tag{1.11}$$
$$P(Y \in B, D = 0|Z = 0) - P(Y \in B, D = 0|Z = 1) \geq 0. \tag{1.12}$$

Kitagawa (2015) shows that this testable restriction possesses two important features. First, it is optimal in the sense that any other feature of the observable data distribution cannot contribute to further screen out violations of Assumption 2 further. Second, validity of Assumption 2 is a refutable but nonverifiable hypothesis. In particular, it is possible to contruct a joint probability law (Y_1, Y_0, D_1, D_0, Z) that satisfies (1.11) and (1.12) but violates Assumption 2. Consequently, accepting the null hypothesis never allows us to confirm Assumption 2 no matter how large the sample is.

To implement this testing idea, Kitagawa (2015) proposes a variance-weighted Kolmogorov-Smirnov test statistic based on the empirical distribution and a bootstrap algorithm to obtain critial values. He also provides an extension of the test to settings with conditioning covariates. Mourifié and Wan (2014) show that an alternative formulation of (1.11) and (1.12) fits into the intersection bounds framework of Chernozhukov et al. (2013), which provides an alternative test of the model.

This test exploits the independence Assumption 2.1. Huber and Mellace (2015) note that this independence assumption is too strong if one is only interested in the average effect. Therefore, they propose an alternative test for

a slightly weaker mean independence assumption. The fact that the observed distribution $F_Y(y|D=1,Z=1)$ is a mixture of the distributions of Y_1 for the always-takers and for the compliers allows bounding $E[Y_1|\mathcal{T}=a]$ using a result in Horowitz and Manski (1995): in one extreme scenario the always-takers are all at the bottom of the distribution and in the other extreme scenario they are all at the top of the distribution. The test suggested by Huber and Mellace (2015) consists in checking that $E[Y_1|\mathcal{T}=a]$, which is point identified by $E[Y|D=1,Z=0]$, lies within the bounds obtained from the mixture. Similarly, $E[Y_1|\mathcal{T}=n]$ must lie within its bounds.

The same line of reasoning applies to quantile effects. Full independence is needed for the identification of the whole LQTE process. If one is interested in a single LQTE, then Assumption 2.1 may be replaced by

Assumption 4 $(D_1, D_0) \perp\!\!\!\perp Z$ *and*

$$F_{Y_d}\left(Q_{Y_d}\left(\tau|\mathcal{T}=c\right)|\mathcal{T}=t, Z=0\right) = F_{Y_d}\left(Q_{Y_d}\left(\tau|\mathcal{T}=c\right)|\mathcal{T}=t, Z=1\right),$$

for $d \in \{0,1\}$ *and* $t \in \{c,a,n\}$.

Compared with Assumption 2.1, full independence between the instrument and the potential outcomes has been replaced by a local quantile independence restriction. Chesher (2003), for instance, emphasizes that this local condition is weaker than the global one. The proofs of the results in Section 2 clearly go through with Assumption 4 replacing Assumption 2.1. Under this alternative assumption, Kitagawa's first testable restriction (1.11) is valid only for two sets B, $(-\infty, Q_{Y_1}(\tau|\mathcal{T}=c)]$ and $(Q_{Y_1}(\tau|\mathcal{T}=c), \infty)$ and the second restriction (1.12) only for two other sets, $(-\infty, Q_{Y_0}(\tau|\mathcal{T}=c)]$ and $(Q_{Y_0}(\tau|\mathcal{T}=c), \infty)$. Interestingly, these four restrictions correspond to the moment inequalities of Huber and Mellace (2015) but applied to $1(Y_1 \leq Q_{Y_1}(\tau|\mathcal{T}=c))$ and $1(Y_0 \leq Q_{Y_0}(\tau|\mathcal{T}=c))$ instead of Y_1 and Y_0.

To summarize, Assumption 4 is sufficient for the consistency of the LQTE estimator at a single quantile, it is still testable but it is more difficult to reject. Note, however, that the goal of an instrument validity test is not necessarily to test the weakest set of assumptions under which the estimator is consistent but may be to assess the plausibility of the model that motivates the estimator. For instance, if an economic model delivers an exclusion restriction or the instrument has been randomized, then it makes sense to test the model that includes the full independence assumption. As discussed above, even asymptotically it is difficult to reject an invalid instrument; it therefore judicious to test all the implications of the model to increase the power of the test.

1.5 Comparison to the Instrumental Variable Quantile Regression Model

In this section, we compare the framework reviewed in this chapter with another popular approach to instrumental variable estimation that accommodates binary treatments and binary instruments, the instrumental variable quantile regression (IVQR) model introduced by Chernozhukov and Hansen (2005) and reviewed in Chernozhukov et al. (2016).

It is instructive to depart from the potential outcomes notation and to consider the following general two-equation structural model,

$$Y = q(D, \varepsilon), \tag{1.13}$$
$$D = h(Z, \eta), \tag{1.14}$$

where $q(\cdot)$ and $h(\cdot)$ are general nonseparable functions and ε and η are unobservable components that can be scalars or vectors. We refer to equation (1.13) as the *outcome equation* and to equation (1.14) as the *selection equation*. Note that this system of structural equations is recursive or triangular. To simplify the exposition, we omit covariates throughout this section and refer the interested reader to the original references for further details.

The equation-based model can be related to potential outcomes as

$$Y_d = q(d, \varepsilon_d),$$
$$D_z = h(z, \eta_z).$$

This formulation plays a key role in understanding different approaches to instrumental variables estimation. Vytlacil (2002) shows that Assumption 3 is equivalent to the following latent index selection model

$$D_z = 1\{v(z) \geq \eta\}, \tag{1.15}$$

where $v(\cdot)$ is a nontrivial function of Z and $(Y_1, Y_0, \eta) \perp\!\!\!\perp Z$. Thus, Assumption 2 restricts the unobserved heterogeneity in the selection equation to be scalar, i.e., $\eta_1 = \eta_0 = \eta$, while leaving the heterogeneity in the outcome equation unrestricted. As will be discussed below, restrictions on the dimensionality of the unobservables play a key role in nonseparable IV models.

A natural alternative is to consider a model that restricts the unobservables in the outcome equation, while leaving the selection equation unconstrained. This leads to the IVQR model. By the Skorohod representation of random variables, potential outcomes can be related to their structural quantile functions as $Y_d = q(d, \varepsilon_d)$, where $q(\cdot)$ is the structural quantile function of Y_d and $\varepsilon_d \sim U(0, 1)$. ε_d can be interpreted as a rank variables because it determine individual ranks in the distribution of Y_d.

The key assumption underlying the IVQR model is the rank preservation

assumption. Formally, rank preservation requires that conditional on Z, $\varepsilon_1 = \varepsilon_0$. Given the interpretation of ε_d as a rank variable, this assumption thus restricts the ranks to be invariant accross potential outcome distributions, whence the name of this assumption. Chernozhukov and Hansen (2005) show that rank preservation can be weakened to rank similarity. Formally, rank similarity requires that conditional on Z and the disturbance in the selection equation, ε_1 and ε_0 are identically distributed. Rank similarity thus allows for random slippages from an individuals rank level ε. Chernozhukov and Hansen (2005) show that under the aforementioned assumptions and a full rank condition on the Jacobian of the moment condition (1.16), $q(D, \tau)$ is identified for the whole population from the following conditional moment restriction, for all $\tau \in (0, 1)$

$$P(Y \leq q(D, \tau)|Z) = P(Y < q(D, \tau)|Z) = \tau \tag{1.16}$$

It is interesting to compare the identification strategy of the IVQR model to the instrumental variable framework reviewed in this chapter. Restricting the dimensionality of unobservables in the outcome equation yields point identification of the effect for the whole population, while restricting the dimensionality in the selection equation only point identifies QTEs for the compliers. Intuitively, the reason is that the one-to-one mapping from the outcome to the unobserved error term with continuous outcome variables is lost when there are mass-points.

On the surface, the IVQR model does not seem to be connected to the approach reviewed in this chapter – the two estimands differ (i.e., the QTE and the LQTE respectively) and the underlying assumptions are non-nested, non-contradictory, and concern different aspects of the models (i.e., the outcome equation respectively the selection equation). For these reasons, Chernozhukov and Hansen (2013) describe both model as complements and, for example, Chernozhukov and Hansen (2004) use comparisons of both models as specification checks for the underlying assumptions.

Wüthrich (2016) shows that there is actually a close connection between the estimands of both models. Under Assumptions 1 and 3, the model model captures LQTE at transformed quantile levels:

$$\Delta^{IVQR}(\tau) = \Delta(\tau'|\mathcal{T} = c) \tag{1.17}$$

where

$$\tau' = F_{Y_0}\left(q(0, \tau)|\mathcal{T} = c\right) = F_{Y_1}\left(q(1, \tau)|\mathcal{T} = c\right)$$

This result has interesting implications for the connection between both models: (i) if the LQTE estimands are constant across quantiles, the estimates of both models converge to the same true effect, (ii) if the LQTEs are positive (or negative) at all quantiles, then the sign of the quantile estimands will be the same in both models, and (iii) monotonicity of the LQTE function (which is

implied, for example, by a location scale shift model for the compliers) implies monotonicity of the IVQR estimands.

It is important to note that (1.17) does not rely on the rank preservation assumption and thus also provides a characterization of the IVQR estimands absent the rank preservation assumption. To this end, (1.17) implies that the IVQR estimands are quite robust: they preserve sign and monotonicity of $\Delta(\tau|\mathcal{T} = c)$ whenever these properties are invariant across quantiles. Furthermore, the results show that the estimates based on the IVQR model are not arbitrary under misspecification but correspond to well-defined (functions of) causal effects for the compliers.

The results in Wüthrich (2016) confirm that with unrestricted treatment effect heterogeneity all the information about the treatment effects has to come from the compliers. Moreover, they show how the IVQR model extrapolates from the compliers to the whole population. This motivates the use of the IVQR as an approach to extrapolation in the LQTE framework.

1.6 Conclusion and open problems

In this chapter we have reviewed instrumental variable methods to estimate QTEs. 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 units in the population. This monotonicity assumption is sometimes satisfied by construction (e.g. one-sided perfect compliance) but there are naturally also cases where it is a strong assumption, see for instance the examples discussed by de Chaisemartin (2014). If it can be made, the whole distribution of the control and treated outcomes are identified for the units which react to the instrument without imposing any restrictions on treatment effect heterogeneity. Estimation and inference methods have been developed for the standard set-up with a binary instrument, a binary treatment and a continuous outcome. Stata and R codes are available for implementing most of these methods. For instance, Frölich and Melly (2010) provide a Stata package for estimating conditional and unconditional LQTEs based the approaches by Abadie et al. (2002) and Frölich and Melly (2013) respectively. We have also summarized extensions to multi-valued and continuous instruments, which are now well-understood.

From our point of view, the most pressing open research questions are inference methods for discrete outcomes and identification of the effects of nonbinary treatments. Since the monotonicity assumption does not restrict the outcomes at all, identification follows for discrete outcomes in exactly the same way as it does for continuous outcomes. The LQTE framework therefore accommodates discrete outcomes and outcomes with mass points very naturally. This is in sharp contrast to the instrumental variable quantile regression

model (Chernozhukov and Hansen, 2005), where continuity is essential for point identification. It is also possible to show that the analog estimators of the cdfs of the potential outcomes are asymptotically normally distributed even for discrete outcomes. However, the existing literature still assumes continuity to provide inference tools based on the asymptotic Gaussianity of the quantile estimators. Inference procedures that accommodate discrete outcomes would be useful and should deserve closer attention by future research.

On the other hand, the LQTE framework does not easily extend to nonbinary treatments. When an independent instrument is nonbinary, any binary transformation of this instrument will also satisfy the independence assumption and the results for binary instrument can be used. But when the endogenous treatment is nonbinary, then the instrument does not necessarily satisfy the exclusion restriction for any binary transformation of the treatment. In addition, the number of types of compliers increases exponentially in the number of points in the support of the treatment. For average effects, Angrist and Imbens (1995) show that a weighted average of local effects is identified. This is unfortunately not the case for quantile effects. This raises new challenges that have not yet been overcome.

Instead of restricting heterogeneity in the treatment choice equation, an alternative literature reviewed by Chernozhukov et al. (2016) restricts heterogeneity in the outcome equation by imposing a stochastic rank preservation condition. This assumption allows extrapolating the treatment effects from the compliers (or, more generally, from the population for which the effects are identified) to the whole population. This implies that there is a close connection between these models in the sense that the QTEs identified by one model corresponds to the treatment effect identified by the other model at another quantile.

This relationship holds for the standard setup. Since the IVQR model imposes assumptions on the outcome equation but not on the selection equation, which is the opposite of the LQTE model, it will accommodate well the opposite types of generalizations. For instance, it applies without modification to multi-valued and continuous treatments if the instrument is rich enough to identify all the parameters. Torgovitsky (2015) and D'Haultfoeuille and Février (2015) show that combining the rank invariance and the monotonicity assumption allows to suppress the large support condition for the instrument. On the contrary, point identification breaks down when the outcome is not continuous, see Chesher (2010). Intuitively, the one-to-one mapping from the outcome to the unobserved error term is lost when there are mass-points.

Bibliography

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., Angrist, J., Imbens, G. W., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70, 91–117.

Ananat, E. O., Michaels, G., 2008. The effect of marital breakup on the income distribution of women with children. Journal of Human Resources 43 (3), 611–629.

Angrist, J., Imbens, G. W., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of American Statistical Association 90, 431–442.

Angrist, J. D., Imbens, G. W., Rubin, D. B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91 (434), 444–455.

Angrist, J. D., Pischke, J.-S., 2008. Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Balke, A., Pearl, J., 1997. Bounds on treatment effects from studies with imperfect compliance. Journal of the American Statistical Association 92 (439), 1171–1176.

Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2013. Program evaluation with high-dimensional data. arXiv preprint arXiv:1311.2645.

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

Cawley, J., Meyerhoefer, C., 2012. The medical care costs of obesity: an instrumental variables approach. Journal of health economics 31 (1), 219–230.

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

Chernozhukov, V., Fernandez-Val, I., Hansen, C., 2016. Handbook chapter.

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

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

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

Chernozhukov, V., Lee, S., Rosen, A., 2013. Intersection bounds: Estimation and inference. Econometrica 81, 667–737.

Chesher, A., 2003. Identification in nonseparable models. Econometrica 71, 1405–1441.

Chesher, A., 2010. Instrumental variable models for discrete outcomes. Econometrica 78 (2), 575–601.

Cox, D. R., 1958. Planning of experiments.

de Chaisemartin, C., 2014. Tolerating defiance? local average treatment effects without monotonicity. Warwick Economics Research Paper Series 1020.

Deaton, A., 2010. Instruments, randomization, and learning about development. Journal of Economic Literature 48 (2), pp. 424–55.

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

Doksum, K., 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. The Annals of Statistics, 267–277.

Eren, O., Ozbeklik, S., 2014. Who benefits from job corps? a distributional analysis of an active labor market program. Journal of Applied Econometrics 29 (4), 586–611.

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

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

Frölich, M., Melly, B., 2010. Estimation of quantile treatment effects with stata. Stata Journal 10 (3), 423.

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

Heckman, J. J., Smith, J., Clements, N., 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., Urzúa, S., 2010. Comparing IV with structural models: What simple IV can and cannot identify. Journal of Econometrics 156 (1), 27 – 37, structural Models of Optimization Behavior in Labor, Aging, and Health.

Heckman, J. J., Vytlacil, E., May 2005. Structural equations, treatment effects, and econometric policy evaluation 1. Econometrica 73, 669–738.

Hong, H., Nekipelov, D., 2010. Semiparametric efficiency in nonlinear late models. Quantitative Economics 1 (2), 279–304.

Horowitz, J., Manski, C., 1995. Identification and robustness with contaminated and corrupted data. Econometrica 63, 281–302.

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

Huber, M., Mellace, G., 2015. Testing instrument validity for LATE identification based on inequality moment constraints. Review of Economics and Statistics 97 (2), 398–411.

Imbens, G. W., 2010. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). Journal of Economic Literature 48 (2), 399–423.

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

Imbens, G. W., Rubin, D., 1997. Estimating outcome distributions for compliers in instrumental variables models. Review of Economic Studies 64, 555–574.

Imbens, G. W., Rubin, D. B., 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Imbens, G. W., et al., 2014. Instrumental variables: An econometricians perspective. Statistical Science 29 (3), 323–358.

Kitagawa, T., 2015. A test for instrument validity. Econometrica 83 (5), 2043–2063.

Lehmann, E. L., 1975. Nonparametrics: statistical methods based on ranks. San Francisco: Holden-Day.

Mourifié, I., Wan, Y., 2014. Testing LATE assumptions. Available at SSRN.

Neyman, J., 1923. On the application of probability theory to agricultural experiments. essay on principles. Statistical Science Reprint, 5, 463–480.

Rubin, D. B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Rubin, D. B., 1980. Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American Statistical Association 75 (371), 591–593.

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

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

Wald, A., 1940. The fitting of straight lines if both variables are subject to error. The Annals of Mathematical Statistics 11 (3), 284–300.

Wüthrich, K., 2016. A comparison of two quantile models with endogeneity, mimeo, Universitaet Bern, Departement Volkswirtschaft.

Yu, P., 2014. Marginal quantile treatment effect, mimeo, Department of Economics, University of Auckland.