Abstract

I introduce a procedure to nonparametrically estimate local quantile treatment effects in a regression discontinuity (RD) design with a binary treatment. Analogously to Hahn, Todd, and van der Klaauw’s (2001) estimator for average treatment effects using local linear regression, the estimator developed here uses local linear quantile regression to estimate the marginal distributions of potential outcomes to infer the quantile treatment effects for the subgroup of “RD compliers”. I describe the estimation procedure, derive the asymptotic distribution, and provide Monte Carlo results. I apply the procedure to Gormley, et al’s (2005) study of the effects of universal pre-K programs, and find that while evidence for an effect on the upper end of the distribution is weaker, participation in a pre-K program significantly raises the lower end and middle of the distribution of test scores, with the greatest gains in the middle of the distribution.

1 Introduction

The regression discontinuity (RD) design has received increased attention in recent years as a means of quasi-experimentally estimating treatment effects. To cite only a few examples of many recent studies using this design, Jacob and Lefgren (2004) and Matsudaira (2008) estimate the effect of remedial education programs, exploiting assessment test cutoffs in assignment to summer school programs; Black, Smith, Berger,
and Noel (2003) use a feature of the UI “profiling score” to evaluate the effect of the Worker Profiling and Reemployment Services program; Angrist and Lavy (1999) exploit maximum class size rules in Israeli public schools to estimate the effect of class size on educational outcomes; and DiNardo and Lee (2004) use certification elections to estimate the impact of new unions on employers. Studies comparing RD estimates to results based on randomized trials suggest the popularity of the RD design is justifiable.

The studies mentioned above and others using the RD design focus on estimating average treatment effects. In many contexts, however, the effect of a treatment on the distribution of outcomes is of interest. For example, economists often evaluate the social welfare implications of a policy based on the differences in the distribution of outcomes under various alternatives (Atkinson, 1970). Furthermore, a zero average effect may mask significant offsetting effects at different points in the distribution. Examples where distributional effects may be of particular interest include unionization, which is widely believed to compress wages, and progressively oriented social and education programs which may be intended to bring up the lower end of the distribution.

In this paper I introduce a procedure to estimate quantile treatment effects in the fuzzy RD design when selection into treatment is potentially endogenous. As Hahn, Todd, and Klaauw (2001) suggested, the fuzzy RD design leads naturally to instrumental variables (IV) type estimators, and the estimator they develop has an interpretation as a local Wald estimator of a local average treatment effect (LATE). Their insight suggests applying IV quantile treatment effects estimators in order to estimate distributional effects in the RD design.

Two recently developed approaches to IV quantile treatment effects are Chernozhukov and Hansen (2005), and Abadie, Angrist, and Imbens (2002) (see Frandsen (2008) for a comparison of these two estimators). These two approaches rely on distinct sets of identifying assumptions, and the interpretations of the estimands differ. An RD quantile treatment effects estimator in the spirit of Chernozhukov and Hansen (2005) is developed by Guiteras (2008). In some contexts, however, the requirement of rank invariance or rank similarity across treatment status in that model may be less desirable than the LATE assumptions of Abadie, Angrist, and Imbens (2002). I therefore focus on the LATE framework in this paper. A challenge that prevents the trivial application of Abadie, Angrist and Imbens’ (AAI) local quantile treatment effects estimator to the RD design is the fact that the instrument is a deterministic function of the running variable, which must be controlled for. The “non-trivial assignment” condition required by AAI’s estimator therefore fails. One way to deal with this is to use kernel weighting to estimate the effect only at the threshold. In the limit

---

1 See Cook and Wong (2008) for a summary of these studies, which include Black, Galdo, and Smith (2005) and Buddelmeyer and Skoufias (2003).
2 The results in this paper apply equally well to the sharp regression discontinuity design, which I treat as a special case of the fuzzy design.
the running variable plays no role, and therefore can be ignored, allowing a straightforward application of AAI, as in Froelich and Melly (2008). In finite sample, however, ignoring the running variable leads to substantial finite sample bias in this approach. The contribution of my paper is to overcome this difficulty using local linear quantile regression to non-parametrically identify quantile treatment effects at the threshold, in the spirit of Hahn, Todd, and Klaauw (2001).

Another approach to estimating distributional effects in the RD context that makes use of local linear quantile regression is being developed by Baker, Firpo, and Milligan (2005). While their approach overcomes the finite sample sample bias problem inherent in the “local constant” approach of simply applying AAI at the threshold, they rely on a selection-on-observables identifying assumption at the threshold. This rules out cases where selection into treatment is endogenous, even at the threshold of the running variable. The estimator I introduce here allows for endogenous treatment even conditional on being in a neighborhood of the threshold, and thus it has an IV interpretation.

The remainder of this paper is outlined as follows. Section 2 develops the statistical framework and describes the estimation procedure. I derive the asymptotic distribution for the estimator in section 5. I present Monte Carlo simulation results in section 6. In section 7 I apply the procedure to estimate the effect of an Oklahoma universal pre-K program on the distribution of test scores, and section 8 concludes.

2 Statistical framework

In this section I will set up the basic elements of the fuzzy RD framework in the LATE context, define the estimand of interest, establish identification results, and describe the estimation procedure.

Since the motivation for the estimation procedure I develop in this paper is very much in the spirit of Imbens and Angrist’s (1994) LATE framework, I will set up the fuzzy RD framework in terms of potential outcomes. For simplicity, I do not condition on any covariates other than the running variable. The critical elements of the fuzzy RD design, in terms of the LATE notation, are:

\[ Y_0 \equiv \text{potential outcome when untreated} \]

\[ Y_1 \equiv \text{potential outcome when treated} \]

\[ D \equiv \text{indicator for treatment status (possibly endogenous)} \]

\[ Y \equiv Y_0 + (Y_1 - Y_0)D, \text{ observed outcome} \]

\[ \delta \equiv Y_1 - Y_0, \text{ treatment effect (possibly heterogeneous)} \]

\(^3\text{The kernel weighting Froelich and Melly’s (2008) approach implicitly makes this “locally constant” assumption, and suffers from the finite sample bias problems discussed later.}\)
$R \equiv$ scalar running variable

$Z \equiv 1 (R > 0)$, indicator for the running variable exceeding a threshold. I set the threshold equal to zero without loss of generality

$D_0 \equiv$ potential treatment status when $Z = 0$

$D_1 \equiv$ potential treatment status when $Z = 1$.

Some features this setup preserves from the LATE framework are that it allows for heterogeneous treatment effects and endogenous treatment selection, as in a Roy model of selection on gains. Another feature this setup shares with the LATE framework is that we can conceptually classify individuals into one of several mutually exclusive groups, depending on their potential treatment status. I will use the standard nomenclature for these groups, and introduce abbreviations to refer to them:

- **Always takers (AT):** $D_1 = D_0 = 1$
- **Never takers (NT):** $D_1 = D_0 = 0$
- **Compliers (C):** $D_1 > D_0 \iff D_1 = 1, D_0 = 0$
- **Defiers (DE):** $D_1 < D_0 \iff D_1 = 0, D_0 = 1$.

The estimand I consider in this paper is the local quantile treatment effect, or the difference between the marginal distributions of potential outcomes for compliers at a particular quantile near the threshold level of the running variable:

$$\delta_{LQTE} (\tau) \equiv Q_{Y_1|C,R=0} (\tau) - Q_{Y_0|C,R=0} (\tau).$$

An important comment regarding the interpretation of this object is that it reflects the effect of treatment on the distribution, rather than the effect of treatment on any particular individual. Without a rank invariance assumption, as in the Chernozhukov and Hansen (2005) framework, there is no sense in which (1) represents the treatment effect for a particular individual, since an individual with a $Y_0$ of rank $\tau$ need not have a $Y_1$ of rank $\tau$.

## 3 Identification of LQTE

Besides those embodied in the notation given in section 2, I make the following additional assumptions:

**Assumption 1:** $RD \lim_{r \to 0^+} \Pr (D = 1|R = r) > \lim_{r \to 0^-} \Pr (D = 1|R = r)$
Assumption 2: Local Smoothness $F_{Y_d|R}(y|r)$ is continuous in $r$ over an $\varepsilon$-neighborhood of zero, and is strictly increasing in $y_d$ over the same neighborhood, for $d \in \{0, 1\}$. $F_{Y_d|R}(y|r)$ is differentiable in $r$ for $r < 0$ or $r > 0$ in the same neighborhood, and $\Pr(NT|R = r)$ and $\Pr(AT|R = r)$ are continuous in $r$ over that neighborhood.

Assumption 3: Monotonicity $\lim_{r \to 0} \Pr(D_1 \geq D_0|R = r) = 1$

The first assumption is the defining feature of the regression discontinuity design, that the probability of treatment changes discontinuously at the threshold value of the running variable. Without loss of generality I assume the probability of treatment is greater above the threshold. Assumption 2 is a smoothness condition which, intuitively speaking, ensures that after controlling smoothly for the running variable, differences in the distribution of outcomes on either side of the threshold are due to the change in probability of treatment assumed in Assumption 1. Assumption 2 also guarantees quantiles of the potential outcomes are uniquely defined at the threshold. Assumption 3 is the crucial monotonicity assumption that the response of treatment selection to the instrument is monotone. An immediate consequence of this assumption is that the monotonicity condition rules out the existence of defiers—those for whom $D_0 > D_1$—in a neighborhood around the threshold.

These assumptions are quite similar to Hahn, Todd, and Klaauw’s (2001) conditions for identifying the local average treatment effect in an RD setting. Assumption 1 here is precisely their RD condition, and Assumption 3 is equivalent to their monotonicity condition. One difference is that because I am identifying distributional effects, I require smoothness of the conditional distribution function, while Hahn, Todd, and Klaauw assume smoothness of the conditional expectation function, because they are identifying local average treatment effects.

The assumptions I make are analogous to those required for Abadie, Angrist, and Imbens’s (2002) local quantile treatment effects estimator, or Imbens and Angrist’s (1994) LATE assumptions. Instead of independence between an instrument and potential outcomes and potential treatment status, I make continuity assumptions on the distribution of potential outcomes and potential treatment status. The LATE first stage assumption is replaced by the analogous RD assumption that the probability of treatment jumps discretely as the running variable hits the threshold value. This is essentially a local first stage requirement. Assumption 3, local monotonicity, is directly analogous to the monotonicity assumption in the LATE framework. The most striking difference between my assumptions and Abadie, Angrist, and Imbens’s (2002) assumptions is the absence here of the “Non-trivial assignment” condition which they require. Indeed, the principal

---

4There are several settings in which monotonicity holds automatically, including when then RD “fuzziness” is one-sided, with either no treatment below the threshold, or 100 percent treatment above the threshold. Another setting in which monotonicity holds automatically is in latent index models of selection, as discussed below.
challenge of applying Abadie, Angrist, and Imbens’s (2002) quantile treatment effects estimator in an RD setting is that the non-trivial assignment condition fails here, since conditional on the running variable, the “instrument”, $Z = 1 (R > 0)$, is deterministically either zero or one.

While the notation and assumptions I listed above constitute the essential elements of the framework, it may help to fix ideas and link the setup to economic models to describe a selection model incorporating the elements above, which I do in the following example. I emphasize that I write down the selection model for pedagogical reasons only; while many of the restrictions of this selection model are equivalent to the assumptions I made above, any further restrictions that may be implied by the selection model are not required for the estimation procedure I will describe in section 4.

Example 1 Latent index model

A model incorporating the assumptions of my setup has outcome $Y_i$ for individual $i$ depending on treatment and a smooth function of the running variable:

$$Y_i = g(R_i) + \delta_i D_i + \eta_i.$$ 

I incorporate the RD assumption and monotonicity into this example by using Vytlacil’s (2002) insight that in this context the monotonicity assumption can always be rationalized by a latent index model for treatment selection. I have treatment endogenously determined via a latent index model:

$$D_i = 1 \{ \delta_i + \gamma Z_i \geq \varepsilon_i^D \},$$

where $\gamma > 0$. In this model of selection, $\delta_i$ can be thought of as the gross benefit of treatment in terms of the effect on $Y_i$, $\varepsilon_i^D$ represents an unobserved individual-specific cost of treatment which is not reflected in $Y_i$, and $Z_i = 1 (R_i > 0)$ is a cost shifter that reduces the private cost of treatment by $\gamma$ when the running variable exceeds the threshold. An example would be a scholarship offer made when an individual scores above some threshold on an achievement test. Individuals select into treatment when the benefit exceeds the private cost, net of the effect of the cost shifter, $Z_i$. Always takers (AT) in this case are individuals for whom the benefit exceeds the private cost, even without the cost shifter:

$$AT = \{ i : \delta_i \geq \varepsilon_i^D \}. $$
Never takers (NT) are those for whom even with the cost shifter, the private cost exceeds the benefit:

\[ NT = \{ i : \delta_i + \gamma < \varepsilon_i^D \} , \]

and compliers are those for whom the private cost exceeds the benefit without the cost shifter, but the benefit exceeds the cost with the cost shifter:

\[ C = \{ i : \delta_i < \varepsilon_i^D \leq \delta_i + \gamma \} . \]

The compliers are those for whom the instrument (in this case an indicator for being above the threshold) directly determines treatment status, and therefore it is the treatment effect on the distribution of potential outcomes for this group that is identified.

Given our assumptions, at the threshold we can adapt Imbens and Rubin’s (1997) and Abadie’s (2002) method of identifying counterfactual distributions for compliers\(^5\). Intuitively, we can nonparametrically identify the marginal distribution of \( Y_0 \) at the threshold for never takers (they are just the untreated ones with \( Z = 1 \)) and for a mixture of never takers and compliers (untreated ones with \( Z = 0 \)), and thus we can back out the marginal distribution of \( Y_0 \) for compliers. We can also identify the marginal distribution at the threshold of \( Y_1 \) for always takers (treated with \( Z = 0 \)) and for a mixture of always takers and compliers (treated with \( Z = 1 \)), and thus we can back out the threshold marginal distribution of \( Y_1 \) for compliers. The local quantile treatment effect is then simply the difference between the inferred marginal distributions of the potential outcomes for compliers at a particular quantile.

Because I will use the result several times, I provide the following lemma, which makes the above intuition precise, and generalizes it to expectations of any function of potential outcomes, not just distribution functions. Again, I emphasize that this lemma (and indeed all the identification results in this section) are adaptations of Imbens and Rubin’s (1997) and Abadie’s (2002) results, but applied to the regression discontinuity context.

**Lemma 2** *Expectations of Functions of Potential Outcomes for Compliers*

Let \( h(\cdot) \) be any measurable function on the real line such that \( \mathbb{E}|h(Y)| < \infty \) and \( \mathbb{E}[h(Y_d)|R=r] \) is continuous and differentiable in \( r \) over an \( \varepsilon \)-neighborhood of zero for \( d \in \{0,1\} \). Then under Assumptions

\(^5\)The intuition and steps I give in this section to derive the marginal distributions of the potential outcomes for compliers are algebraically equivalent to the argument in Abadie (2002), except I condition everywhere on \( R = 0 \) by taking the appropriate limits. I express things more explicitly in terms of distributions for various subgroups, along the lines of Imbens and Rubin (1997), to make the intuition for the mechanics of my estimation procedure clear.
\( \text{1-3, } E [h(Y_d) | C, R = 0], \quad d \in \{0, 1\} \) is identified from the joint distribution of \((Y, D, R)\) as

\[
\begin{align*}
E [h(Y_1) | C, R = 0] &= \lim_{r \to 0^+} E [h(Y) | D = 1, R = r] - (1 - \Pr (C | AT \cup C, R = 0)) \lim_{r \to 0^-} E [h(Y) | D = 1, R = r] \\
E [h(Y_0) | C, R = 0] &= \lim_{r \to 0^-} E [h(Y) | D = 0, R = r] - (1 - \Pr (C | NT \cup C, R = 0)) \lim_{r \to 0^+} E [h(Y) | D = 0, R = r]
\end{align*}
\]

where \( \Pr (C | NT \cup C, R = 0) \) and \( \Pr (C | AT \cup C, R = 0) \) are given by

\[
\begin{align*}
\Pr (C | NT \cup C, R = 0) &= \frac{\lim_{r \to 0^+} E [D | R = r] - \lim_{r \to 0^-} E [D | R = r]}{1 - \lim_{r \to 0^-} E [D | R = r]} \\
\Pr (C | AT \cup C, R = 0) &= \frac{\lim_{r \to 0^+} E [D | R = r] - \lim_{r \to 0^-} E [D | R = r]}{\lim_{r \to 0^+} E [D | R = r]}. \quad (2)
\end{align*}
\]

**Proof.** The proof follows the intuition given in the paragraph before the theorem. The expectation at the threshold of \( h(Y_0) \) for never takers is given by:

\[
E [h(Y_0) | NT, R = 0] = \lim_{r \to 0^+} E [h(Y_0) | NT, R = r]
\]

\[
= \lim_{r \to 0^+} E [h(Y_0) | D = 0, R = r]
\]

\[
= \lim_{r \to 0^+} E [h(Y) | D = 0, R = r].
\]

The first equality follows from the smoothness hypothesis of the lemma. The second equality follows from monotonicity, Assumption 3, which implies that conditional on \( R = r > 0 \), the set of never takers (NT) is identical to the set of untreated. The final equality follows from the fact that conditional on \( D = 0 \), we have \( Y = Y_0 \). Next, I show in a similar manner that the expectation of \( h(Y_0) \) at the threshold for a mixture of never takers and compliers is identified:

\[
E [h(Y_0) | NT \cup C, R = 0] = \lim_{r \to 0^-} E [h(Y_0) | NT \cup C, R = r]
\]

\[
= \lim_{r \to 0^-} E [h(Y_0) | D = 0, R = r]
\]

\[
= \lim_{r \to 0^-} E [h(Y) | D = 0, R = r].
\]

The first equality again uses the smoothness hypothesis of the lemma. The second equality follows from monotonicity, Assumption 3, which implies that conditional on \( R = r < 0 \), the set \( NT \cup C \) is identical to the set of untreated. The final equality again follows from the fact that conditional on \( D = 0 \), we have \( Y = Y_0 \).
Furthermore, the fraction of compliers in the set $NT \cup C$ can be computed by Bayes’ Rule:

$$
\Pr (C|NT \cup C, R = 0) = \frac{\Pr (C|R = 0)}{\Pr (NT \cup C|R = 0)}
= \frac{1 - \Pr (NT|R = 0) - \Pr (AT|R = 0)}{\Pr (NT \cup C|R = 0)}
= \frac{1 - \lim_{r \to 0^+} \Pr (NT|R = r) - \lim_{r \to 0^-} \Pr (AT|R = r)}{\lim_{r \to 0^+} \Pr (D = 0|R = r)}
= \frac{1 - \lim_{r \to 0^+} \Pr (D = 0|R = r) - \lim_{r \to 0^-} \Pr (D = 1|R = r)}{\lim_{r \to 0^-} \Pr (D = 0|R = r)}
= \frac{\lim_{r \to 0^+} \Pr (D = 1|R = r) - \lim_{r \to 0^-} \Pr (D = 1|R = r)}{1 - \lim_{r \to 0^-} E[D|R = r]}. \quad (3)
$$

which is just a rescaling of the Hahn, Todd, and van der Klaauw “first stage”. The first equality is Bayes’ Rule. The second equality follows from monotonicity, Assumption 3. The third follows from continuity of $\Pr (NT|R = r)$ at $R = 0$ (Assumption 2). The denominator in the right hand side of the third equality follows from the fact that conditional on $R = r < 0$, the set $NT \cup C$ is identical to the set of untreated. The fourth equality also follows from the definitions of $AT$ and $NT$. The fifth equality simply uses

$$
1 - \lim_{r \to 0^+} \Pr (D = 0|R = r) = \lim_{r \to 0^+} [1 - \Pr (D = 0|R = r)]
= \lim_{r \to 0^+} \Pr (D = 1|R = r),
$$

and the final equality rewrites probabilities as expectations. Knowing $\Pr (C|NT \cup C, R = 0)$, we can back out the expectation of $h(Y_0)$ for compliers, using the identity:

$$
E [h(Y_0)|NT \cup C, R = 0] \quad (y) \quad = \quad (1 - \Pr (C|NT \cup C, R = 0)) E [h(Y_0)|NT, R = 0] \quad + \quad \Pr (C|NT \cup C, R = 0) E [h(Y_0)|C, R = 0].
$$

Rearranging this expression gives the identification result for $E [h(Y_0)|C, R = 0]$:

$$
E [h(Y_0)|C, R = 0] = \lim_{r \to 0^-} E [h(Y)|D = 0, R = r] - (1 - \Pr (C|NT \cup C, R = 0)) \lim_{r \to 0^+} E [h(Y)|D = 0, R = r] \quad \frac{\Pr (C|NT \cup C, R = 0)}{Pr (C|NT \cup C, R = 0)}.
$$
where \( \Pr (C|NT \cup C, R = 0) \) is given by (3), which by the RD condition, Assumption 2, is strictly positive.

Using exactly symmetrical arguments, I can write down the expectation of \( h(Y) \) for compliers:

\[
E[h(Y_1)|C, R = 0] = \lim_{r \to 0^+} \frac{\lim_{r \to 0^-} E[h(Y)|D = 1, R = r] - (1 - \Pr (C|AT \cup C, R = 0)) \lim_{r \to 0^-} E[h(Y)|D = 1, R = r]}{\Pr (C|AT \cup C, R = 0)}
\]

(4)

where \( \Pr (C|AT \cup C, R = 0) = \lim_{r \to 0^+} \frac{\lim_{r \to 0^-} E[D|R=r]}{\Pr (C|NT \cup C, R = 0)} \). This establishes the result in the lemma.

Identification of the local quantile treatment effect follows from a special case of this lemma, where \( h(Y_d) = 1(Y_d \leq y) \), as the following theorem shows:

**Theorem 3** LQTE Identification.

Under Assumptions 1-3, the local quantile treatment effect, \([1]\), is identified from the joint distribution of \((Y, D, R)\) as

\[
\delta_{LQTE} (\tau) = F^{-1}_{Y_1|C,R=0} (\tau) - F^{-1}_{Y_0|C,R=0} (\tau),
\]

where \( F_{Y_1|C,R=0} (y) \) and \( F_{Y_0|C,R=0} (y) \) are given by

\[
F_{Y_1|C,R=0} (y) = \lim_{r \to 0^+} \frac{\lim_{r \to 0^-} E[Y|D=1, R=r] - (1 - \Pr (C|AT \cup C, R = 0)) \lim_{r \to 0^-} E[Y|D=1, R=r]}{\Pr (C|AT \cup C, R = 0)}
\]

(5)

\[
F_{Y_0|C,R=0} (y) = \lim_{r \to 0^+} \frac{\lim_{r \to 0^-} E[Y|D=0, R=r] - (1 - \Pr (C|NT \cup C, R = 0)) \lim_{r \to 0^-} E[Y|D=0, R=r]}{\Pr (C|NT \cup C, R = 0)}
\]

(6)

and \( \Pr (C|NT \cup C, R = 0) \) and \( \Pr (C|AT \cup C, R = 0) \) are given by (3) and (3).

**Proof.** The proof follows from lemma 2. Let \( h(Y_d) = 1(Y_d \leq y) \). Then by the smoothness condition, Assumption 2, \( E[h(Y_d)] = F_{Y_d} (y) \) satisfies the smoothness hypothesis of the lemma, and by Assumptions 1-3, the remaining hypotheses of lemma 2 are satisfied, establishing results 5 and 6. By Assumption 2, \( F_{Y_1|C,R=0}^{-1}(\tau) \) and \( F_{Y_0|C,R=0}^{-1}(\tau) \) exist, which establishes the result of the theorem.

**4 Estimation procedure**

The local quantile treatment effect, \([1]\), may be consistently estimated in a number of ways. I will briefly discuss two approaches, and spend greater time developing the second approach, which has much better finite sample properties than the first. Both approaches involve kernel weighting in the running variable with a bandwidth that shrinks as the sample size grows to narrow in on the threshold.
4.1 Naive Approach: Abadie-weighted QR at the threshold

The first approach is what I call a naive application of Abadie, Angrist, and Imbens’s (2002) local quantile effects estimator. This naive estimator combines the kernel weights which narrow in on the threshold with Abadie’s (2000) “complier finding” weights. Although the non-trivial assignment condition for this estimator to work fails in the RD setup, as the sample size grows, in the limit we are conditioning on $R = 0$, and so there is no longer any need to include $R$ as a regressor, and thus the technique is applicable. The estimator simply applies Abadie, Angrist, and Imbens’s (2002) result that the local quantile treatment effect, $\delta_{LQTE}$, satisfies the following:

$$
(Q_{Y_0}|C (\tau), \delta_{LQTE} (\tau)) = \arg\min_{(a,d)} E \left[ \rho_{\tau} (Y - a - dD) | D_1 > D_0, R = 0 \right]
$$

where $\rho_{\tau} (u) = u (\tau - 1 (u < 0))$ is the usual check function, and $\kappa$ is Abadie’s (2002) “complier finding” weight at the threshold:

$$
\kappa = 1 - \frac{D (1 - Z)}{1 - \Pr (Z = 1 | R = 0)} - \frac{(1 - D) Z}{\Pr (Z = 1 | R = 0)}.
$$

Note that if we take $\Pr (Z = 1 | R = 0)$ to be equal to one half, then $\kappa = 1$ when $D = Z$ and $\kappa = -1$ otherwise. This suggests using the following naive estimator to consistently estimate the local quantile treatment effect:

$$
\hat{\delta}_{Naive} (\tau) = \arg\min_{d} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau} (Y_i - a - dD_i) \cdot \hat{\kappa}_i \cdot K \left( \frac{R_i}{h_n} \right),
$$

where $K (u)$ is a kernel weighting function and $\hat{\kappa}_i$ is an estimate of $\kappa$ at $R = 0$ (possibly also estimated via a kernel weighting function), and $h_n$ is a bandwidth. The estimator is consistent as long as $h_n$ shrinks to zero with the sample size, since in the limit we have conditioned on $R = 0$ and Abadie, Angrist, and Imbens’s (2002) result applies. However, in finite sample, there will be variation in $R$ over the window defined by the kernel weights. Since we do not include $R$ in the quantile regression in (7), this leads to a sort of omitted variables bias, since the instrument, $Z$, is not independent of the potential outcomes when we do not condition on $R$. This bias can be quite large, as the Monte Carlo results in section 6 suggest.

There may be ways to reduce the finite sample bias of the kernel-weighted Abadie quantile regression estimator. One approach that seems at once to reduce some of the finite sample bias and make the estimator computable via linear programming techniques is to apply Abadie, Angrist, and Imbens’s (2002) strategy of substituting $\kappa$ for $\kappa_v = E [\kappa | Y, D, R = 0]$. A non-parametric estimator for $\kappa_v$ that works well in
simulations to reduce finite sample bias uses local linear regression on either side of the threshold to estimate $\kappa_V$:

$$
\begin{align*}
\tilde{\kappa}_v &= 1 - \frac{D \left(1 - \tilde{V}\right)}{1 - \tilde{E} \left[Z \cdot K \left(\frac{R}{h}\right)\right]} - \frac{(1 - D) \tilde{V}}{\tilde{E} \left[Z \cdot K \left(\frac{R}{h}\right)\right]}, \\
\tilde{V} &= \tilde{E} \left[Z \cdot K \left(\frac{R}{h}\right) | \tilde{Y}, \tilde{D}\right],
\end{align*}
$$

where $\tilde{E}$ denotes non-parametric estimation (series regression, for example), and $\tilde{Y}$ and $\tilde{D}$ are residuals from local linear regressions of $Y$ and $D$, respectively, on $R$. The local linear regressions naturally are done separately on either side of the threshold for each variable. While this approach seems to partially reduce the finite sample bias of the naive application of Abadie-weighted regression in simulations, there is still substantial bias. More importantly, I do not have theoretical results on the properties, including consistency, of this modified estimator.

Other possibilities to correct for the finite sample bias in this approach include jackknife or analytic bias corrections, in the spirit of Hahn and Newey (2004), who show how to reduce bias in nonlinear panel models. I hope to explore this avenue of bias reduction in future work. However, with the finite-sample problems of Abadie-type approaches still extant, I will now turn to another approach to estimating (1) which has much better finite sample properties.

### 4.2 Better Approach: Local Linear Quantile Regression

The second approach solves the finite sample bias problems inherent in the “local constant” approximation implicit in omitting the running variable from the quantile regression in (7). Instead, the second approach uses local linear techniques such as Yu and Jones (1998) to estimate conditional quantile functions or distributions at the threshold. As Yu and Jones (1998) discuss, there are two natural approaches to local linear estimation of quantile functions. One is to estimate quantile functions directly, while another is to estimate distribution functions and then invert. Just as in their setting, in the RD quantile treatment effects setting I consider here, while the first approach seems more direct, there are reasons to prefer the latter method of inverting distribution functions. In particular, as I will show in section 6, the two methods perform nearly identically in terms of finite sample bias and variance, but the method of inverting distribution functions is much faster to compute. Nevertheless I will develop both approaches, and discuss the relative merits, as well as compare the two estimators in Monte Carlo simulations in section 6.

---

6 In Monte Carlo simulations, the method of inverting distribution functions took half the time of estimating the quantile function directly.
4.2.1 Estimating Quantile Functions Directly

Possibly the most natural approach to using local linear methods to estimate the local quantile treatment effect, \( (1) \), is to directly estimate the quantiles functions on the right hand side of \( (1) \), and take the difference to be an estimator for the local quantile treatment effect. The following theorem uses Koenker and Bassett’s (1978) insight that quantiles satisfy a minimization problem to suggest a consistent local linear estimator for the local quantile effect.

**Theorem 4** Consistent LQTE Estimation via Quantile Functions.

Augment Assumption 2 to also include continuity of 
\[
\mathbb{E} \left[ Y_d \mid R = r \right] \text{ in a neighborhood of } r = 0.
\]
Then under Assumptions 1-3 a consistent estimator for the local quantile treatment effect, \( (1) \), is

\[
\hat{\delta}_{LQTE} (\tau) = \hat{Q}_{Y_1 \mid C, R = 0} (\tau) - \hat{Q}_{Y_0 \mid C, R = 0} (\tau),
\]

where

\[
\hat{Q}_{Y_1 \mid C, R = 0} (\tau) = \arg \min_q \mathbb{E} [\rho_\tau (Y_1 - q) \mid C, R = 0],
\]

\[
\hat{Q}_{Y_0 \mid C, R = 0} (\tau) = \arg \min_q \mathbb{E} [\rho_\tau (Y_0 - q) \mid C, R = 0],
\]

and \( \mathbb{E} [\rho_\tau (Y_1 - q) \mid C, R = 0] \) and \( \mathbb{E} [\rho_\tau (Y_0 - q) \mid C, R = 0] \) are consistent, local linear estimates of

\[
\mathbb{E} [\rho_\tau (Y_1 - q) \mid D = 1, R = r] = \lim_{r \to 0^-} \frac{\mathbb{E} [\rho_\tau (Y_1 - q) \mid D = 1, R = r]}{\Pr (C \mid AT \cup C, R = 0)}
\]

\[
\mathbb{E} [\rho_\tau (Y_0 - q) \mid D = 0, R = r] = \lim_{r \to 0^-} \frac{\mathbb{E} [\rho_\tau (Y_0 - q) \mid D = 0, R = r]}{\Pr (C \mid NT \cup C, R = 0)}
\]

where \( \Pr (C \mid NT \cup C, R = 0) \) and \( \Pr (C \mid AT \cup C, R = 0) \) are given by \( (3) \) and \( (3) \).

**Proof.** Let \( h (Y_d) = \rho_\tau (Y_d - q), \ d \in \{0, 1\} \), where \( \rho_\tau (u) = u (\tau - 1 \ (u < 0)) \) is the check function. Note that \( F_{Y_d \mid R} (y, r) \) and \( E [Y_d \mid R = r] \) continuous in \( r \) on a neighborhood around zero implies that \( E [\rho_\tau (Y_d - q) \mid R = r] \) is also continuous on that same neighborhood\(^7\). Then the hypotheses of lemma 2 are satisfied, establishing results \( (11) \) and \( (12) \). Then the conclusion of the theorem follows from Koenker and Bassett (1978).

\(^7\)Integrate \( \int \rho_\tau (Y_d - q) f_{Y_d \mid R} (y, r) \) by parts to see this result.
4.2.2 Inverting Estimated Distribution Functions

An alternative consistent estimator for the local quantile treatment effect is the (horizontal) difference between local linear estimates of the conditional distribution functions \( \hat{F}_{Y_1|C,R=0}(\tau) \) and \( \hat{F}_{Y_0|C,R=0}(\tau) \) at a particular quantile, as the following theorem establishes.

**Theorem 5** Consistent LQTE Estimation via Inverting Distribution Functions.

Under Assumptions 1-3 a consistent estimator for the local quantile treatment effect, \( \hat{\delta}_{LQTE} \), is

\[
\hat{\delta}_{LQTE}(\tau) = \hat{F}_{Y_1|C,R=0}^{-1}(\tau) - \hat{F}_{Y_0|C,R=0}^{-1}(\tau),
\]

(13)

where

\[
\hat{F}_{Y_1|C,R=0}^{-1}(\tau) \in \left\{ a : \hat{F}_{Y_1|C,R=0}(a) = \tau \right\},
\]

(14)

\[
\hat{F}_{Y_0|C,R=0}^{-1}(\tau) \in \left\{ b : \hat{F}_{Y_0|C,R=0}(b) = \tau \right\},
\]

(15)

and \( \hat{F}_{Y_1|C,R=0}(y), \hat{F}_{Y_0|C,R=0}(y) \) are local linear, consistent estimates of \( \hat{F}_{Y_1|C,R=0}(a) \) and \( \hat{F}_{Y_0|C,R=0}(b) \).

**Proof.** The result follows from Slutsky’s theorem and the definition of a quantile function. ■

At this point I emphasize again that expressions \( \hat{F}_{Y_1|C,R=0}(a) \) and \( \hat{F}_{Y_0|C,R=0}(b) \) are simply algebraic re-arrangements of Imbens and Rubin’s (1997) and Abadie’s (2002) results for the marginal distribution of potential outcomes.
for compliers\(^8\), but here I condition on \( R = 0 \) by taking the appropriate limits. The key observation is that local linear techniques can be used to avoid the substantial finite sample bias that a trivial (albeit kernel weighted) application of Abadie’s results would suffer from. Since using local linear techniques to estimate (6) and (6) is the main insight, I will discuss possible ways that might be done. The estimator I propose is a function of first step estimates of the component quantities in (6) and (6)\(^7\) consisting of the following four conditional distributions:

\[
\lim_{r \to 0^-} F_{Y|D=0,R=r}(y), \\
\lim_{r \to 0^+} F_{Y|D=0,R=r}(y), \\
\lim_{r \to 0^-} F_{Y|D=1,R=r}(y), \\
\lim_{r \to 0^+} F_{Y|D=1,R=r}(y)
\]

\(^8\) To see this, start with Abadie’s (2002) expression for the cdf of \( Y_1 \) for compliers, and re-arrange using Bayes’ rule to arrive at an unconditional (on \( R = 0 \)) version of my expression (5):

\[
F_{Y_1}(y) = \frac{E[1(Y \leq y)D|Z = 1] - E[1(Y \leq y)D|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} \\
= \frac{E[1(Y \leq y)|D = 1, Z = 1]Pr(D = 1|Z = 1) - E[1(Y \leq y)|D = 1, Z = 0]Pr(D = 1|Z = 0)}{Pr(C)} \\
= \frac{F_{Y|D=1,Z=1}(y)Pr(D = 1|Z = 1) - F_{Y|D=1,Z=0}(y)Pr(D = 1|Z = 0)}{Pr(C)} \\
= \frac{F_{Y|D=1,Z=1}(y)Pr(D = 1|Z = 1) - F_{Y|D=1}(y)Pr(1|Z = 1)}{Pr(C)} \\
= \frac{F_{Y|D=1,Z=1}(y) - F_{Y|D=1}(y)Pr(1|Z = 1, D = 1)}{Pr(1|Z = 1, D = 1)} \\
= \frac{F_{Y|D=1,Z=1}(y) - F_{Y|D=1}(y)(1 - Pr(C|Z = 1, D = 1))}{Pr(C|D = 1, Z = 1)} \\
= \frac{F_{Y|D=1,Z=1}(y) - F_{Y|D=1}(y)(1 - Pr(C|Z = 1, D = 1))}{Pr(C|Z = 1, D = 1)} \\
= \frac{F_{Y|D=1,Z=1}(y) - F_{Y|D=1}(y)(1 - Pr(C|Z = 1, D = 1))}{Pr(C|Z = 1, D = 1)}.
\]

An alternative approach is to recognize that (9) and (10) and be rewritten as ‘local Wald’ ratios:

\[
F_{Y_1|C,R=0}(y) = \lim_{r \to 0^+} \frac{E[1(Y \leq y)D|R = r] - \lim_{r \to 0^-} E[1(Y \leq y)D|R = r]}{\lim_{r \to 0^+} E[D|R = r] - \lim_{r \to 0^-} E[D|R = r]}, \\
F_{Y_0|C,R=0}(y) = \lim_{r \to 0^+} \frac{E[1(Y \leq y)(1-D)|R = r] - \lim_{r \to 0^-} E[1(Y \leq y)(1-D)|R = r]}{\lim_{r \to 0^+} E[1-D|R = r] - \lim_{r \to 0^-} E[1-D|R = r]},
\]

and estimate these quantities in one step via local linear two-stage least squares, as suggested by Imbens and Lemieux (2008). However, in practice it is often optimal to use different bandwidths to estimate the various pieces. The approach I propose estimates the pieces separately, allowing optimal choice of bandwidth for each piece separately.
and the following two conditional expectations:

\[
\lim_{r \to 0^+} E[D|R = r] \tag{16a}
\]
\[
\lim_{r \to 0^-} E[D|R = r]. \tag{16b}
\]

Since each of these quantities must be estimated at a boundary, local linear approaches are most suitable (Fan, 1992). For the conditional distributions, possibly the most straightforward local linear estimator of 
\[
\lim_{r \to 0^+} F_{Y|D=0,R=r}(y), \tag{17}
\]
that satisfies:

\[
\left( \hat{F}_{Y|D=0,R=0^-}(y), \hat{b}(y) \right) = \arg \min_{a,b} \sum_{i:D=0,Z=0} [1(Y_i \leq y) - a - bR_i]^2 K \left( \frac{R_i}{h} \right), \tag{18}
\]

However, to optimize the tradeoff between bias and variance, and to ensure the estimated function is continuous Yu and Jones (1998) propose to introduce smoothing “in the y-direction”, as well, so that the local linear estimator of 
\[
\lim_{r \to 0^-} F_{Y|D=0,R=r}(y), \tag{17}
\]
would be \( \hat{F}_{Y|D=0,R=0^-}(y) \), where:

\[
\left( \hat{F}_{Y|D=0,R=0^-}(y), \hat{b}_y(y) \right) = \arg \min_{a,b} \sum_{i:D=0,Z=0} \left[ \Omega \left( \frac{y - Y_i}{h_2} \right) - a - bR_i \right]^2 K \left( \frac{R_i}{h} \right), \tag{18}
\]

where \( \Omega(\cdot) \) is the distribution function associated with associated with a kernel density function, \( W \). This is essentially a conditional distribution version of Fan, Yao, and Tong’s (1996) “double-kernel” conditional density estimator. This double-smoothed estimator has the following closed form:

\[
\hat{F}_{Y|D=0,R=0^-}(y) = \frac{1}{\sum_{j:D=0,Z=0} w_j(h_1)} \sum_{j:D=0,Z=0} w_j(h_1) \Omega \left( \frac{y - Y_j}{h_2} \right), \tag{18}
\]

where weighting function associated with local linear fitting is given by:

\[
w_j(h_1) = K \left( \frac{R_j}{h_1} \right) [S_{n,2} - R_jS_{n,1}],
\]

with

\[
S_{n,l} = \sum_{i:D=0,Z=0} K \left( \frac{R_i}{h_1} \right) R_i, \quad l = 1, 2.
\]

The bandwidths used for smoothing in the x-direction (\( K \)) and y-direction (\( W \)) are, respectively, \( h_1 \) and \( h_2 \).

Yu and Jones (1998) give operational rules of thumb for choosing these bandwidths.

One drawback to the local linear distribution estimators such as (17) and (18), however, is that in finite sample the estimate need not be bounded between zero and one, and may be non-monotone. An approach
that solves these difficulties, but preserves the attractive bias properties of local linear estimators is an
adjusted Nadaraya-Watson estimator proposed by Hall, Wolff, and Yao (1999). This estimator is given by:

\[ F_{ANW}^{Y|D=0,R=0-}(y) = \frac{\sum_{i:D=0,Z=0} 1(Y_i \leq y) p_i K\left(\frac{R_i}{h}\right)}{\sum_{i:D=0,Z=0} p_i K\left(\frac{R_i}{h}\right)}, \]

where the weights \( \{p_i\} \) are chosen to satisfy \( p_i \geq 0 \) \( \forall i \), \( \sum_{i:D=0,Z=0} p_i = 1 \), and

\[ \sum_{i:D=0,Z=0} p_i R_i K\left(\frac{R_i}{h}\right) = 0. \]

Hall, Wolff, and Yao (1999) show how to pick the weights \( \{p_i\} \).

The conditional expectations in (16) are precisely the same quantities as those in the denominator of
Hahn, Todd, and Klaauw’s (2001) local Wald estimator, and can be estimated in the local linear fashion
they suggest.

With these methods for estimating the necessary components of the local quantile treatment effect, (13),
in hand, I now provide asymptotic distribution theory and present Monte Carlo simulations to demonstrate
the performance of the estimation procedure, and compare the various methods of local linear distribution
estimation.

5 Asymptotic Distribution Theory

In this section I derive the limiting distribution for the local linear quantile treatment effects estimator,
(13), obtained via inverting estimated distribution functions. First I will define some notation that will be
useful for the theorem, as well as additional regularity assumptions. Define \( m_d (r) = F_{Y|D=d,R=r}(y) \) for
\( d \in \{0, 1\} \) and \( p(r) = E[D|R=r] \). Define the limits \( m_d^+(r) = \lim_{e \to 0^+} m_d (r + e) \), \( m_d^-(r) = \lim_{e \to 0^+} m_d (r - e) \),
\( p^+(r) = \lim_{e \to 0^+} p(r + e) \), \( p^-(r) = \lim_{e \to 0^+} p(r - e) \). Additionally, define

\[ \sigma_d^2 (0) = \lim_{r \to 0^+} \text{Var} (Y|D=d,R=r) , \quad \sigma_d^2 (0) = \lim_{r \to 0^+} \text{Var} (Y|D=d,R=r) \\
\eta_{L_1,L_2} (0) = \lim_{r \to 0^+} \text{Cov} (L_1,L_2|R=r) , \quad \eta_{L_1,L_2} (0) = \lim_{r \to 0^+} \text{Cov} (L_1,L_2|R=r) , \]

for random variables \( L_1 \) and \( L_2 \). Finally, define the following constant:

\[ \omega^+ = \frac{\int_0^\infty \left( (\int_0^\infty s^2 K(s) \, ds) - (\int_0^\infty s K(s) \, ds) \cdot u \right)^2 K(u)^2 \, du}{\int_R(0) f_{R}(0) \gamma_{h_1} \cdot \left[ (\int_0^\infty u^2 K(u) \, du) (\int_0^\infty K(u) \, du) - (\int_0^\infty u K(u) \, du)^2 \right]^2} , \]

(20)
with \( \omega^- \) similarly defined, but now with the integral in the limits of integration over \((-\infty, 0)\).

In addition to assumptions 1-3, I make the following assumptions in the derivation of the limiting distribution:

**A1** For \( r > 0 \) and \( d \in \{0, 1\} \), \( m_d(r) \) and \( p(r) \) are equicontinuous and twice continuously differentiable.

There exists some \( M > 0 \) such that \( |m_d^+(r)|, |m_d^-(r)|, |m_d''(r)| \), and \( |p^+(r)|, |p'(r)|, |p''(r)| \) are uniformly bounded on \((0, M]\). Similarly, \( |m_d^+(0)|, |m_d^-(0)|, |m_d''(0)| \), and \( |p^+(0)|, |p'(0)|, |p''(0)| \) are uniformly bounded on \([-M, 0]\).

**A2** The limits \( m_d^+(0), m_d^-(0), m_d''(0), m_d''(0), p^+(0), p'(0), p''(0) \) are well-defined and finite, for \( d \in \{0, 1\} \).

**A3** The density of \( R, f_R(r) \) is continuous and bounded near 0. It is also bounded away from zero near 0. \( R \) has bounded support.

**A4** \( K(\cdot), W(\cdot) \) are Borel measurable, bounded, continuous, symmetric, nonnegative-valued with compact support.

**A5** \( \sigma^2(r) = \text{Var}(Y|R = r) \) is uniformly bounded near 0. Similarly, \( \eta(r) = \text{Cov}(Y, D|R = r) \) is uniformly bounded near 0. Furthermore, the limits \( \sigma^2(0), \sigma^2(0), \eta^+(0), \text{ and } \eta^+(0) \) are well-defined and finite.

**A6** \( \lim_{r \to 0} \mathbb{E} \left[ |Y - m_d(r)|^3 \right] I = d, R = r \) is well-defined and finite.

**A7** The bandwidth sequences satisfy \( h_1 = \gamma_{h_1} \cdot n^{-b} \) and \( h_2 = \gamma_{h_2} h_1 \) for some \( \gamma_{h_1} \) and \( \gamma_{h_2} \), with \( \frac{1}{5} < b < 1 \).

Asymptotic normality of the quantile treatment effect estimator, (13), follows from the weak convergence as a process of conditional distribution function estimators, such as (17) by the functional delta method (van der Vaart, 1998). The following lemma establishes the weak convergence as a process of the local linear conditional distribution function estimator, (17).

**Lemma 6** Under assumptions 1-3 and A1-A7 the sequence \( \sqrt{n h_n} \left( \hat{F}_{Y|D=0, R=0} - F_{Y|D=0, R=0} - (y) \right) : y \in \mathbb{R} \) is asymptotically tight in \( C^\infty(\mathbb{R}) \) and converges in distribution to a Gaussian process.

**Proof.** The local linear conditional distribution function estimator, (17), can be written as a function of
sample averages:

\[
\hat{F}_{Y|D=0,R=0^-}(y) = \frac{A_{n,2}B_{n,0}(y) - A_{n,1}B_{n,1}(y)}{A_{n,2}A_{n,0} - A_{n,1}^2},
\]

\[
A_{n,l} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n} K \left( \frac{R_i}{h_n} \right)^l,
\]

\[
B_{n,l}(y) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_n} 1(Y_j \leq y) K \left( \frac{R_j}{h_n} \right)^l.
\]

For simplicity, in this proof I take \( n \) to be the sample size drawn from the distribution of \((Y, R)\) conditional on \( D = 0, R \leq 0 \), and therefore I omit explicit conditioning on \( D = 0, R \leq 0 \) in sums. I write the bandwidth as \( h_n \) to emphasize dependence on the sample size. To show the weak convergence of \( \hat{F}_{Y|D=0,R=0^-}(y) \) I establish that each of the terms \( A_{n,0}, A_{n,1}, A_{n,2}, B_{n,0}(y), B_{n,1}(y) \) converge weakly as processes, and apply a functional delta method. I start by establishing the convergence as a process of

\[
\sqrt{n h_n} (B_{n,l}(y) - E[B_{n,l}(y)]), \quad l = 0, 1,
\]

since the \( A_{n,1} \) terms are trivial functions of \( y \). Define a vector of random variables, \( X_i \), and indexing set \( T \):

\[
X_i = \begin{pmatrix} Y_i \\ R_i \end{pmatrix}, \quad T = \mathbb{R}.
\]

Define the set of functions \( F_n = \{f_{n,t} : t \in T\} \), with:

\[
f_{n,t}(X_i) = 1(Y_i \leq t) \frac{1}{\sqrt{h_n}} K \left( \frac{R_i}{h_n} \right)^l, \quad l = 0, 1.
\]

Then the process (22) can be written:

\[
n^{-\frac{1}{2}} \sum_{i=1}^{n} (f_{n,t}(X_i) - Pf_{n,t}) : t \in T,
\]

which corresponds to van der Vaart and Wellner’s (1996) setup for Theorem 2.11.22 for convergence of processes indexed by classes of functions changing with \( n \). Letting \( P \) and \( P^* \) denote measure and outer measure, respectively, and \( \rho(s,t) \) a pseudonorm on \( \mathbb{R} \), the conditions needed for convergence are the following:

1. There exist envelope functions \( F_n : |f_{n,t}(x)| \leq F_n(x) \quad \forall x, f, n \) which satisfy

19
(a) \[ P^* F_n^2 = O(1), \]

and

(b) \[ P^* F_n^2 \{ F_n > \eta \sqrt{n} \} \to 0, \quad \text{for every } \eta > 0 \]

2. \( \mathcal{F}_{n, \delta} = \{ f_{n,s} - f_{n,t} : \rho(s,t) < \delta \} \) and \( \mathcal{F}_{n, \delta}^2 \) are \( P \)-measurable for every \( \delta > 0 \)

3. \( f_{n,t} \) satisfy:
\[
\sup_{\rho(s,t) < \delta_n} P (f_{n,s} - f_{n,t})^2 \to 0, \quad \text{for every } \delta_n \downarrow 0,
\]

4. The uniform entropy condition on page 220 of van der Vaart and Wellner holds.

Start with the first condition (envelope functions). Define a set of envelope functions to be:
\[
F_n = \frac{1}{\sqrt{h_n}} K \left( \frac{R_j}{h_n} \right) \left( \frac{R_j}{h_n} \right)^l, \quad l = 0, 1.
\]

Clearly these are envelope functions for class \( \mathcal{F}_n \). Under the measurability assumption, condition 1a can be written:
\[
PF_n^2 = \int \left( \frac{1}{\sqrt{h_n}} K \left( \frac{R_j}{h_n} \right) \left( \frac{R_j}{h_n} \right)^l \right)^2 dF_{R|R \leq 0, D=0}(r), \quad l = 0, 1
\]
\[
= \int \left( K(u) (u)^l \right)^2 f_R(h_n u) du, \quad l = 0, 1,
\]

making the change of variables \( u = \frac{r}{h_n} \). Condition 1a then holds under our boundedness assumptions on \( R \) and \( K(\cdot) \). Condition 1b holds trivially for \( l = 0 \) for bounded \( K(\cdot) \). For \( l = 1, 1b \) is essentially the Lindberg-Feller condition, and holds if, for example, \( R \) is bounded. Condition 2 is implied by our assumption that \( K(\cdot) \) is measurable.
The quantity in condition 3 can be written:

$$\sup_{\rho(s,t) < \delta_n} P \left( f_{n,s} - f_{n,t} \right)^2$$

$$= \sup_{\rho(s,t) < \delta_n} P \left( (Y_i \leq s) - 1(Y_i \leq t) \right) \cdot \frac{1}{\sqrt{h_n}} K \left( \frac{R_i}{h_n} \right) \left( \frac{R_i}{h_n} \right)^2$$

$$= \int_{r \leq 0} \left\{ \sup_{\rho(s,t) < \delta_n} \int_y (1(y \leq s) - 1(y \leq t))^2 dF_{Y|R=r,R\leq 0,D=0} (y) \right\}$$

$$\times \left( \frac{1}{\sqrt{h_n}} K \left( \frac{r}{h_n} \right) \left( \frac{r}{h_n} \right)^2 \right) dF_{R|R\leq 0,D=0} (r) .$$

In view of condition 1a holding, condition 3 holds if we have:

$$\sup_{\rho(s,t) < \delta_n} \int_y (1(y \leq s) - 1(y \leq t))^2 dF_{Y|R=r,R\leq 0,D=0} (y)$$

$$= \sup_{\rho(s,t) < \delta_n} F_{Y|R=r,R\leq 0,D=0} (s) - 2F_{Y|R=r,R\leq 0,D=0} (s \land t) + F_{Y|R=r,R\leq 0,D=0} (t)$$

$$= \sup_{\rho(s,t) < \delta_n} F_{Y|R=r,R\leq 0,D=0} (s \lor t) - F_{Y|R=r,R\leq 0,D=0} (s \lor t) \to 0 ,$$

for every $\delta_n \downarrow 0$. This holds under our equicontinuity assumption on the conditional CDF of $Y$.

Finally, by example 2.11.24 on page 221 of van der Vaart and Wellner, condition 4 is satisfied since $F_n$ is VC class with a VC index of 2. To see this, note that every one-point set is shattered, but a two-point set:

$$\{ x_1, x_2 \} = \left\{ \begin{pmatrix} y_1 \\ r_1 \end{pmatrix}, \begin{pmatrix} y_2 \\ r_2 \end{pmatrix} \right\} ,$$

with, say, $y_1 < y_2$ is not shattered because the function cannot pick out $\{ x_2 \}$. This establishes that the $B_{n,t}(y)$ terms converge. A similar argument applies to the $A_{n,t}$ terms. By the Cramér-Wold device the terms converge jointly. Noting that as a map from $\mathbb{R}^5$ to $\mathbb{R}$ the usual differentiability of the function in (21) implies Hadamard differentiability, the conclusion follows by the functional delta method. ■

Having established that the local linear conditional distribution function estimators converge as processes, we can apply a functional delta method to derive the limiting distribution of the quantile treatment effect estimator in the following theorem.

**Theorem 7** \textit{LQTE Asymptotic Distribution.}

Let $c$ be the vector \( \begin{pmatrix} 1 & -1 \end{pmatrix} \). Then under assumptions 1-4 and A1-A7 the local linear quantile treatment effects estimator, $\{ \hat{L}_{13} \}$ is asymptotically normally distributed with a limiting distribution given by the
following:
\[ n^{\frac{1-b}{b}} \left( \delta_{\text{LQTE}} (\tau) - \delta_{\text{LQTE}} (\tau) \right) \xrightarrow{d} N \left[ 0, \epsilon J_{F_C} J_{P} \Sigma_{\bar{P}} J_{P} J'_{F_C} \epsilon \right], \]

where the matrices \( J_{F_C}, J_{P}, \) and \( \Sigma_{\bar{P}} \) are defined in the following proof, and \( \frac{1}{2} < b < 1. \)

**Proof.** The estimator (13) is a (Hadamard) differentiable function of several intermediate estimators. Let the vector of component quantities in (5) and (6) be
\[
P = \begin{pmatrix}
\lim_{r \to 0^+} F_{Y|D=1, R=r} (y) \\
\lim_{r \to 0^-} F_{Y|D=1, R=r} (y) \\
\lim_{r \to 0^+} F_{Y|D=0, R=r} (y) \\
\lim_{r \to 0^-} F_{Y|D=0, R=r} (y) \\
\lim_{r \to 0^+} E [D|R=r] \\
\lim_{r \to 0^-} E [D|R=r]
\end{pmatrix},
\]
with \( \hat{P} \) as the corresponding vector of estimators. Each of the estimators in \( \hat{P} \) is local linear estimator of the conditional expectation of some variable, \( L \), approaching \( R = 0 \) from the right or the left. The \( j \)-th element of \( \hat{P} \) can be written:
\[
\hat{P}_j = \lim_{r \to 0^+} E [L_j|R=r]
\]
if it estimates a right limit, or
\[
\hat{P}_j = \lim_{r \to 0^-} E [L_j|R=r]
\]
if it estimates a left limit, where \( L_j \) is the left hand side variable for estimator \( \hat{P}_j \) (e.g., \( 1 (Y_i \leq y) | D_i = 1 \)).

Given assumptions A1-A7, we can apply Lemmas 1-7 from Hahn, Todd, and der Klaauw (1999) to establish the joint convergence in distribution of \( \hat{P} \):
\[ n^{\frac{1-b}{b}} \left[ \hat{P} - P \right] \xrightarrow{d} N \left[ 0, \Sigma_{\bar{P}} \right], \]

where the \((j, k)\)-th element of \( \Sigma_{\bar{P}} \) is given by
\[ \Sigma_{\bar{P}_{j,k}} = \omega^+ \eta_{L_j, L_k}^+ (0) \]
if \( \hat{P}_j \) and \( \hat{P}_k \) both estimate right limits, and similarly if they both estimate left limits. Since the first four estimators in \( \hat{P} \) use separate observations, and they all involve conditioning in \( D = 1 \) or \( D = 0 \), we have
\[ \eta_{L_j, L_k} = 0 \text{ for } j \neq k, \text{ so } \Sigma_{P} \text{ is a diagonal matrix. The } j\text{-th diagonal element of } \Sigma_{P} \text{ is the limiting variance:} \]

\[ \Sigma_{P_{j,j}} = \omega^+ \sigma_{L_j}^2 (0) \] (24a)

or

\[ \Sigma_{P_{j,j}} = \omega^- \sigma_{L_j}^2 (0), \] (24b)

depending on whether \( \hat{P}_j \) estimates a right or a left limit, and \( \omega^+ \) and \( \omega^- \) are defined by (20). Choosing the bandwidth according to Assumption A7 in this section undersmooths, in the sense of Horowitz (2001), causing the bias squared to converge to zero at a faster rate than the variance, correctly centering the asymptotic distribution.

Next I turn to the joint limiting distribution of the local linear conditional distribution function estimators, \((\hat{F}_{Y_1|C, R=0}(y) \quad \hat{F}_{Y_0|C, R=0}(y))\). This vector of estimators is a differentiable function of \( \hat{P} \), so by the multivariate delta method and Lemma 6 we have that \((\hat{F}_{Y_1|C, R=0}(y) \quad \hat{F}_{Y_0|C, R=0}(y))\) converges as a process:

\[ n^{\frac{1}{2}} \left[ \begin{array}{c} \hat{F}_{Y_1|C, R=0}(y) \\ \hat{F}_{Y_0|C, R=0}(y) \end{array} \right] \overset{d}{\longrightarrow} N \left( \mathbf{0}, J_P \Sigma_P J'_P \right), \]

where \( J_P \) is the Jacobian of the map from \( \hat{P} \) to \((\hat{F}_{Y_1|C, R=0}(y) \quad \hat{F}_{Y_0|C, R=0}(y))\) in (5,6) evaluated at the truth:

\[
J_P = \begin{pmatrix}
\frac{1}{\Pr(C|AT \cup C, R=0)} & 0 & 0 \\
-\frac{\Pr(C|AT \cup C, R=0)}{\Pr(C|AT \cup C, R=0)} & 1 & -1 \\
0 & 0 & \frac{1}{\Pr(C|NT \cup C, R=0)} \\
\frac{\lim_{r \rightarrow 0^+} E[D|R=r]}{\lim_{r \rightarrow 0^+} E[D|R=r]^2} & \frac{1}{\lim_{r \rightarrow 0^+} E[D|R=r]} & -\frac{D_{P1}}{\lim_{r \rightarrow 0^+} E[D|R=r]} \\
-\frac{\lim_{r \rightarrow 0^+} D_{P1}}{\lim_{r \rightarrow 0^+} E[D|R=r]} & \frac{D_{P1}}{\lim_{r \rightarrow 0^+} E[D|R=r]} & 1
\end{pmatrix},
\]

where

\[
D_{P1} = \lim_{r \rightarrow 0^+} \frac{F_{Y|D=1, R=r}(y) - F_{Y_1|C, R=0}(y)}{\Pr(C|AT \cup C, R=0)}
\]
Finally I apply the functional delta method to derive the limiting distribution of the quantile treatment effect estimator, \( \{13\} \). In terms of vectors, \( \{13\} \) can be written:

\[
\delta_{LQTE}(\tau) = c' \begin{bmatrix}
\hat{F}_{Y|C,R=0}^{-1}(\tau) \\
\hat{F}_{Y_0|C,R=0}^{-1}(\tau)
\end{bmatrix},
\]

which leads to the conclusion:

\[
n^{1-b} \left( \delta_{LQTE}(\tau) - \delta_{LQTE}(\tau) \right) \overset{d}{\to} N \left[ 0, c' J_{FC} \Sigma_{PP} J_{P} J_{FC} c \right],
\]

where \( J_{FC} \) is the Jacobian of the inverse in \( \{25\} \) evaluated at the true quantile:

\[
J_{FC} = \begin{pmatrix}
\frac{1}{f_{Y_1|C,R=0}(Q_{Y_1|C,R=0}(\tau))} & 0 \\
0 & \frac{1}{f_{Y_0|C,R=0}(Q_{Y_0|C,R=0}(\tau))}
\end{pmatrix}.
\]

This completes the derivation of the result.

Estimates of \( f_{Y_1|C,R=0}(Q_{Y_1|C,R=0}(\tau)) \) and \( f_{Y_0|C,R=0}(Q_{Y_0|C,R=0}(\tau)) \) in \( J_{FC} \) can be obtained via formulae \( \{5\} \) and \( \{6\} \), but substituting in the corresponding density functions for the distribution functions there. The densities in those formulae can be estimated using Fan, Yao, and Tong’s (1996) local linear conditional density estimator.

The preceding theorem was for the simplest local linear distribution without smoothing in the y-direction. If we add smoothing in the y direction, then using Yu and Jones’s (1998) Lemmas 1 and 2, a typical diagonal element of the variance covariance matrix of \( \hat{P} \) is:

\[
Var \left( \hat{P}_j | R \right) = \frac{R(K)}{nh_1 f_R(0)} \left( F(y|0) (1 - F(y|0)) - f_y(y|0) \alpha(W) h_2 \right) + o_p \left( \frac{h_2^2}{n h_1} \right),
\]

where \( R(K) = \int K(u)^2 du \) and \( \alpha(W) = \int \Omega(t) (1 - \Omega(t)) dt \). If we continue to set the x-direction bandwidth as \( h_1 = \gamma_{h_1} n^{-b} \) and \( h_2 = \gamma_{h_2} h_1^2 \), as suggested by Yu and Jones (1998) then the limiting variance becomes:

\[
n^{1-b} \text{Var} \left( \hat{P}_j | R \right) = \frac{R(K)}{\gamma_{h_1} f_R(0)} \left( F(y|0) (1 - F(y|0)) \right) + O_p \left( \frac{1}{n^{2b}} \right),
\]

and the \( f_y(y|0) \alpha(W) h_2 \) term drops out in the limiting variance, and the limiting variance derived in the
proof continues to hold. The bias-squared term in the double smoothed estimator is:

\[
B^2 = \left\{ \frac{1}{2} F^{00}(y|0) \mu_2(K) h_1^2 + \frac{1}{2} F^{02}(y|0) \mu_2(W) h_2^2 \right\}^2,
\]

where \( F^{ab}(y|z) = \frac{\partial^2 F(y|z)}{\partial z^a \partial y^b} \). Plugging in the rule for bandwidth, we get:

\[
n^{1-b}B^2 = O_p \left( \frac{1}{n^{b/2}} \right).
\]

Since \( b > \frac{1}{5} \), the limiting bias is zero in the double-smoothed case as well, and thus under our conditions the limiting distribution for the double-smoothed estimator is the same as for the simpler estimator.

6 Monte Carlo results

In this section I will present results from Monte Carlo simulations of the local linear quantile regression approach I outlined in the previous section. First I will briefly describe the underlying model I use in the simulations. The primitive of the model is the joint distribution of \((Y_0, Y_1, D, R)\), which I specify as follows:

\[
R \sim N(0, \sigma^2_R) \\
Y_0 = R + \varepsilon_0, \\
Y_1 = Y_0 - \varepsilon_1, \\
D = 1(Y_1 - Y_0 + \gamma I(R > 0) \geq \varepsilon_D),
\]

where the disturbance terms are jointly normal and independent:

\[
\begin{pmatrix}
\varepsilon_0 \\
\varepsilon_1 \\
\varepsilon_D
\end{pmatrix} \sim N(0, \begin{pmatrix}
\sigma^2_0 & 0 \\
0 & \sigma^2_1 \\
0 & \sigma^2_D
\end{pmatrix}).
\]

This model exhibits the key features of the RD design with heterogeneous treatment effects. Note that the average treatment effect (ATE) is zero. In this model, the complier group consists of those individuals for whom the following holds:

\[
C = \{Y_1 - Y_0 < \varepsilon_D \leq Y_1 - Y_0 + \gamma\} \\
= \{0 < \varepsilon_D + \varepsilon_1 \leq \gamma\}.
\]
The local average treatment effect (LATE) is therefore:

\[
LATE = E[Y_1 - Y_0 | C]
\]

\[
= -E[\varepsilon_1 | 0 < \varepsilon_D + \varepsilon_1 \leq \gamma]
\]

\[
= -\sigma_1 E\left[\frac{\phi\left(\frac{-\varepsilon_D}{\sigma}\right) - \phi\left(\frac{\gamma - \varepsilon_D}{\sigma}\right)}{\Phi\left(\frac{2 - \varepsilon_D}{\sigma}\right) - \Phi\left(\frac{-\varepsilon_D}{\sigma}\right)}\right].
\]

The true local quantile treatment effect is:

\[
\delta_{LQTE}(\tau) = Q_{Y_1|C}(\tau) - Q_{Y_0|C}(\tau).
\]

I perform 1,000 simulations with a sample size of 100000, using the parameter values \(\gamma = 0.5, \sigma_R = \sigma_0 = \sigma_1 = \sigma_D = 1\). Figure 1 shows the results from the simulations. While the confidence intervals are quite wide, the figure shows that the bias is very small, despite the fact that the estimator consists of nonlinear functions of estimated quantities.

As a comparison, I also ran simulations using the same model to assess the performance of the approach using Abadie-weighted quantile regression, kernel-weighted at the threshold. As I mentioned in the discussion above, such an approach relies on a “local constant” approximation, rather than a local linear approximation. This leads to considerable bias in finite sample, as Figure 2 shows. The confidence intervals for the Abadie approach are narrower, but this combined with the substantial bias could actually be a drawback, since the truth is nearly outside of the confidence interval, especially for more central quantiles, where ordinarily we would expect to have the most confidence in our estimates.

7 Application: Effects of Universal Pre-K

In this section I will apply the RD quantile treatment effects procedure to an example from the literature which will both illustrate how the procedure might be applied to real-life questions, as well as point out some challenges faced by nonparametric estimation of distributional effects.

Policies designed to improve educational performance are one setting in which distributional effects may be important to policy makers. One such policy that specifically targets the lower end of the distribution is the introduction of universal pre-K programs. Gormley, Gayer, Phillips, and Dawson (2005) use a regression discontinuity design to analyze an Oklahoma universal pre-K program, and find significant positive effects on average test scores measuring cognitive development along a variety of dimensions. By conditioning on various socio-economic status indicators, they find indirect suggestive evidence that the program also has
Figure 1: Monte Carlo LQTE point estimates and confidence intervals
Figure 2: Monte Carlo Abadie LQTE point estimates and confidence intervals
positive effects on the lower end of the distribution. The quantile treatment effects estimator developed in this paper allows direct investigation of the effect of the policy on the lower end of the distribution.

Oklahoma introduced a universal pre-K program for four-year-olds in 1998, and by 2002-2003 (the period I analyze) 91 percent of the state’s school districts were participating, including Tulsa Public Schools (TPS), the largest district in the state, and the district from which my sample is drawn. A child’s participation in the pre-K program is voluntary (on the part of the parents), but is subject to a birthday cut-off eligibility rule. Children who had turned four years old by September 1, 2002 were eligible for the program, while younger children were not. Figure 3 shows the discontinuity in probability of treatment that the eligibility rule induced. Because the participation among children who missed the cutoff is essentially nil, local treatment effects in this setting correspond to the effect of treatment on the treated.

At the start of the 2003-2004 school year, all incoming kindergartners and TPS pre-K participants were given the Woodcock-Johnson Achievement Test, a nationally normed test that has been widely used in studies of early education. Treated students are those who participated in a TPS pre-K program the previous year.

Figure 3: The figure plots the probability of attending TPS pre-K in 2002-2003 as a function of birthdate relative to cutoff.
Because the sample consists of entering kindergartners and students enrolled in TPS pre-K for 2003-2004, and the number of students who participated in pre-K despite being younger than the cutoff age is extremely small\textsuperscript{10}, the estimand has a “treatment on the treated” interpretation.

The estimated local quantile treatment effects of TPS pre-K programs on scores on the three subtests of the Woodcock-Johnson tests are plotted in figure\textsuperscript{11} Panel A shows a relatively precisely estimated two to four point effect on the lower end of the distribution of the Letter-Word Identification score, with the effects in the middle of the distribution somewhat larger than at the lowest end. Effects on the upper end of the distribution are less precisely estimated, but the point estimates decline at the upper end, and the effect ceases to be significantly different than zero. We cannot rule out, however, that the effect on the upper end of the distribution are as large (or larger) than those at lower points in the distribution. Effects are on the distribution of the Spelling score are plotted in panel B, and are similar to the results for the Letter-Word Identification scores. Panel C shows the effects of pre-K participation on the Applied Problems score. The point estimates are largest and most precisely estimated for the bottom end of the distribution, and similarly to the other two subtests, point estimates of the effects for the top of the distribution are smaller and less precise.

The estimation results imply that universal pre-K in Oklahoma succeeded in significantly raising the lower end of the distribution of test scores, especially for the Applied Problems subtest. These results are consistent with the Gormley, Gayer, Phillips, and Dawson’s (2005) findings that point estimates of average effects were larger for children from potentially disadvantaged socio-economic groups\textsuperscript{11}. These results are subject to the caveat that they measure the net effect of participating in a TPS pre-K program versus alternatives parents might have chosen in absence of the program. The alternatives may have been different for children at different points in the distribution\textsuperscript{12}, and thus we cannot draw conclusions about the gross impact of universal pre-K programs on the distribution of outcomes. An additional caveat is that these results reflect the short-term effect. It’s possible that children who did not participate may catch up over time, although evidence from the Perry Preschool Study (Schweinhart, Barnes, Weikart, Barnett, and Epstein, 1993; Anderson, 2008) suggests there may be significant long term impacts of pre-K programs.

This application illustrated the ability of the estimation procedure to evaluate distributional policy effects, but it also highlighted some challenges involved with nonparametric estimation in general, and especially with nonparametric estimators of distributional effects. Meaningful inference requires large samples. The

\textsuperscript{10}Out of 1,510 four-year-olds (as of September 1, 2002) in the sample, only 2 children were treated, despite being too young according to the cutoff rule.

\textsuperscript{11}Gormley, Gayer, Phillips, and Dawson (2005) find that point estimates of effects for children receiving reduced-price lunch, Hispanic children, and Native American children are larger than for the sample as a whole, although they cannot reject equality.

\textsuperscript{12}For example, the alternative for a child at the upper end of the distribution may have been a private pre-K program, while the alternative for a child at the lower end may have been a child-care program of lower quality.
Figure 4: The figure plots point estimates and confidence intervals for the effect of TPS pre-K participation on the distribution of scores on three WJ subtests.
imprecision of the estimates for the upper end of the distribution reflects the modest sample size compounded by the skewness of the distribution of test scores, with much lower densities above the median.

8 Conclusion

In this paper I have introduced a new approach to estimating local quantile treatment effects in an RD design, showing consistency and asymptotic normality. The estimator is the horizontal difference between the marginal distributions of the potential outcomes for compliers, which are estimated via local linear quantile regression techniques. In contrast to other possible approaches to estimating distribution effects in an RD context, the procedure I have developed here relies only on the LATE assumptions, and avoids the significant finite sample bias that other “local constant” approaches suffer from, including the Abadie quantile regression kernel-weighted at the threshold. Monte Carlo simulations confirm that the bias of the approach I suggest is minimal compared to other approaches. A less desirable trait of the procedure I introduced is that it involves inverted distribution functions estimated in a first step, but in practice the inversion works very well.

An application of the procedure to estimating the distributional effects of an Oklahoma universal pre-K program shows that the lower end of the distribution is significantly raised, while estimates at the top of the distribution are less precise. Other possibilities for applying the methodology are numerous, and include the study of remedial education programs by Jacob and Lefgren (2004) and Matsudaira (2008), the study of the UI Worker Profiling and Reemployment Services program by Black, Smith, Berger, and Noel (2003), and the effect of unions on wages by DiNardo and Lee (2004). I leave the application of the RD quantile treatment effects estimation to these questions and others to future research.

References


