cemmap

centre for microdata methods and practice

Regression discontinuity design with covariates

Markus Frölich

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP27/07

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Regression discontinuity design with covariates

Markus Frölich

Department of Economics, University of St.Gallen

Last changes: July 22, 2007 (First draft: February 2006)

Abstract

In this paper, the regression discontinuity design (RDD) is generalized to account for differences in observed covariates X in a fully nonparametric way. It is shown that the treatment effect can be estimated at the rate for one-dimensional nonparametric regression irrespective of the dimension of X. It thus extends the analysis of Hahn, Todd and van der Klaauw (2001) and Porter (2003), who examined identification and estimation without covariates, requiring assumptions that may often be too strong in applications. In many applications, individuals to the left and right of the threshold differ in observed characteristics. Houses may be constructed in different ways across school attendance district boundaries. Firms may differ around a threshold that implies certain legal changes, etc. Accounting for these differences in covariates is important to reduce bias. In addition, accounting for covariates may also reduces variance. Finally, estimation of quantile treatment effects (QTE) is also considered.

Keywords: Treatment effect, causal effect, complier, LATE, nonparametric regression, endogeneity
JEL classification: C13, C14, C21

University of St. Gallen, Bodanstrasse 8, SIAW, 9000 St. Gallen, Switzerland; markus.froelich@unisg.ch

1 Introduction

In this paper, the *regression discontinuity* design (RDD) approach is generalized to account for differences in observed *covariates* X in a fully nonparametric way. It is shown that under mild regularity conditions, the treatment effect of interest can be estimated at the rate for *one-dimensional* nonparametric regression irrespective of the dimension of X. It thus extends the analysis of Hahn, Todd, and van der Klaauw (2001) and Porter (2003), who examined identification and estimation without covariates, requiring assumptions that may often be too strong in empirical applications.

The regression discontinuity design is a method frequently used in treatment evaluation, when certain e.g. bureaucratic rules imply a threshold at which many subjects change their treatment status. Consider a law specifying that companies with more than 50 employees have to adhere to certain anti-discrimination legislation whereas smaller firms are exempted. This situation can be considered as a kind of local experiment: Some units, firms or individuals happen to lie on the side of the threshold at which a treatment is administered, whereas others lie on the other side of the threshold. Units close to the threshold but on different sides can be compared to estimate the average treatment effect.

More often than not, however, the units to the left of the threshold differ in their observed characteristics from those to the right of the threshold. Accounting for these differences is important to identify the treatment effect. In the example referred to above, a comparison of firms with 49 employees to those with 51 employees could help to estimate the effects of anti-discrimination legislation on various outcomes. However, firms near the threshold might take the legal effects into account when choosing their employment level. Therefore, firms with 49 employees might thus be quite different in observed characteristics from firms with 51 employees, e.g. with respect to assets, sales, union membership, industry etc. One would therefore like to account for the observed differences between these firms.

Consider a few other examples. Black (1999) examined the impact of school quality on housing prices by comparing houses adjacent to school-attendance district boundaries. School quality varies across the border, which should be reflected in the prices of apartments. How-

ever, if school quality was indeed valued by parents, developers would build different housing structures on the two sides of the boundary: Flats with many bedrooms for families with children on that side of the boundary where the good school is located, and apartments for singles and couples without children on the other side of the border. Black (1999) therefore controls for the number of bedrooms (and other characteristics of the apartments) in a linear model, which could be done fully nonparametrically with the methods developed in this paper.

Such kind of geographic or administrative borders provide opportunities for evaluation in various applications. E.g. individuals living close but on different sides of an administrative border may be living in the same labour market, but in case of becoming unemployed they have to attend different employment offices with potentially rather different types of support or training programmes. These individuals living on the different sides of the border may however also differ in other observed characteristics that one would like to control for.

Angrist and Lavy (1999) exploited a rule that school classes had to be split when *class size* would be larger than 40 otherwise. This policy generates a discontinuity in class size when the enrollment in a grade grows from 40 to 41. But apart from class size there may also be other differences in observed characteristics between the children in a grade with 40 versus 41 children. E.g. rich parents may pull their children out of public school (and send them to private schools) if they realize that their child would be in a class of 40 students, whereas they might not want to do so if class size is only about 20 students.

In these examples,[1] observed covariates are differently distributed across the threshold, which can lead to spurious estimated effects if these differences are not accounted for. The RDD approach without covariates has recently been studied in Hahn, Todd, and van der Klaauw (2001) and Porter (2003). In this paper, I extend the RDD approach to include additional covariates in a fully *nonparametric* way and examine nonparametric identification and estimation of the unconditional treatment effect. It is shown that the rate for univariate nonparametric regression, i.e. $n^{-\frac{2}{5}}$, can be achieved irrespective of the number of variables in X. Hence, the curse of dimensionality does not apply. This is achieved by smoothing over all the covariates X.

[1] Other recent examples include Battistin and Rettore (2002), Lalive (2007) and Puhani and Weber (2007).

Including covariates is often necessary for identification. But even when the estimator would be consistent without controlling for X, *efficiency gains* can be achieved by accounting for covariates. In Section 2, Identification is considered. Section 3 proposes an estimator that achieves $n^{-\frac{2}{5}}$ convergence rate. Section 4 considers estimation of quantile treatment effects (QTE) and other extensions.

2 RDD with covariates

Following the setup of Hahn, Todd, and van der Klaauw (2001), let $D_i \in \{0, 1\}$ be a binary treatment variable, let Y_i^0, Y_i^1 be the individual potential outcomes and $Y_i^1 - Y_i^0$ the individual treatment effect. The potential outcomes as well as the treatment effect are permitted to vary freely across individuals, i.e. no constant treatment effect is assumed. In the examples mentioned, D may represent the applicability of anti-discrimination legislation, school quality, class size etc. Let Z_i be a variable that influences the treatment variable in a discontinuous way, e.g. number of employees, location of house, total school enrollment etc.

In the literature, often two different designs are examined: the *sharp* design where D_i changes for everyone at a known threshold z_0, and the *fuzzy* design where D_i changes only for some individuals. In the sharp design (Trochim 1984), participation status is given by a deterministic function of Z, e.g.

$$D_i = 1(Z_i > z_0). \tag{1}$$

This implies that *all* individuals change programme participation status exactly at z_0. The fuzzy design, on the other hand, permits D to also depend on other factors but assumes that the treatment probability changes discontinuously at z_0:

$$\lim_{\varepsilon \to 0} E\left[D|Z = z_0 + \varepsilon\right] - \lim_{\varepsilon \to 0} E\left[D|Z = z_0 - \varepsilon\right] \neq 0. \tag{2}$$

Note that the fuzzy design includes the sharp design as a special case when the left hand side of (2) is equal to one. Therefore the following discussion focusses on the more general fuzzy design.

The fuzzy design may apply when the treatment decision contains some element of discretion. Case workers may have some discretion about whom they offer a programme, or they

may base their decision also on criteria that are unobserved to the econometrician. It may also often be appropriate in a situation where individuals are offered a treatment or a grant or financial support and decline their participation.[2] (This is further discussed in Section 4.)

If the conditional mean of Y^0 is continuous at z_0, a treatment effect can be identified. Identification essentially relies on comparing the outcomes of those individuals to the left of the threshold with those to the right of the threshold. Hahn, Todd, and van der Klaauw (2001) consider two alternative identifying assumptions (in addition to continuity of $E\left[Y^d|Z=z\right]$ in z at z_0 for $d=\{0,1\}$):

$$\text{HTK1:} \qquad Y_i^1 - Y_i^0 \perp\!\!\!\perp D_i | Z_i \qquad \text{for } Z_i \text{ near } z_0 \tag{3}$$

or

$$\text{HTK2:} \qquad \left\{Y_i^1 - Y_i^0, D_i(z)\right\} \perp\!\!\!\perp Z_i \qquad \text{near } z_0 \qquad \text{and there exists } \varepsilon > 0$$
$$\text{such that } D_i(z_0+e) \geq D_i(z_0-e) \text{ for all } 0 < e < \varepsilon. \tag{4}$$

The former assumption (3) is some kind of selection on observables assumption and identifies $E[Y^1 - Y^0|Z=z_0]$. The second assumption (4) is some kind of instrumental variables assumption and identifies the treatment effect only for a group of local compliers

$$\lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 | D(z_0+\varepsilon) > D(z_0-\varepsilon), Z=z_0\right]$$

and corresponds to some kind of local LATE effect. As discussed in Section 4, in the frequent situation of a mixed sharp-fuzzy RDD design, it corresponds to the average treatment effect on the treated (ATET) $E\left[Y^1 - Y^0|D=1, Z=z_0\right]$. This is e.g. the case with one-sided non-compliance. Whichever of these two assumptions is invoked, the estimator is the same.

Both assumptions above are in many applications too strong. The conditional independence assumption (3) does not permit any kind of deliberate treatment selection which incorporates

[2] For example, van der Klaauw (2002) analyses the effect of financial aid offers to college applicants on their probability of subsequent enrollment. College applicants are ranked according to their test score achievements into a small number of categories. The amount of financial aid offered depends largely on this classification. Yet, the financial aid officer also takes other characteristics into account, which are not observed by the econometrician. Hence the treatment assignment is not a deterministic function of the test score Z, but the conditional expectation $E[D|Z]$ displays jumps because of the test-score rule.

the individual gains $Y_i^1 - Y_i^0$. But even the local IV assumption (4) can be too strong without conditioning on any covariates. It requires that the individuals to the left and right of the threshold have the same unobserved gains and also that there is no deliberate selection into $Z_i < z_0$ versus $Z_i > z_0$. When the individuals left and right of the threshold differ in their observed characteristics, one would be doubtful of the assumptions (3) or (4). In the following, I will first examine identification and estimation under a weaker version of the local IV condition (4) in the fuzzy design. A discussion of a weaker version of (3) is postponed to Section 4.

We start with an informal discussion to provide intuition for what follows. As discussed by examples in the introduction, the IV assumption may become more credible[3] if we control for a number of observed covariates X that may be related to Y, D and/or Z:

$$\left\{Y_i^1 - Y_i^0, D_i(z)\right\} \perp\!\!\!\perp Z_i | X_i \qquad \text{for } Z_i \text{ near } z_0. \tag{5}$$

We also maintain the monotonicity assumption:

$$D_i(z_0 + e) \geq D_i(z_0 - e) \text{ for all } 0 < e < \varepsilon \text{ and some } \varepsilon > 0. \tag{6}$$

By an analogous reasoning as in HTK, and some more assumptions made precise below, the treatment effect on the local compliers conditional on X will be:

$$\lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 \mid X, D(z_0 + \varepsilon) > D(z_0 - \varepsilon), Z = z_0\right] = \frac{m^+(X, z_0) - m^-(X, z_0)}{d^+(X, z_0) - d^-(X, z_0)}, \tag{7}$$

where $m^+(X, z) = \lim_{\varepsilon \to 0} E\left[Y | X, Z = z + \varepsilon\right]$ and $m^-(X, z) = \lim_{\varepsilon \to 0} E\left[Y | X, Z = z - \varepsilon\right]$ and $d^+(X, z)$ and $d^-(X, z)$ defined analogously with D replacing Y.

Estimating the conditional treatment effect for every value of X by (7), although sometimes informative, has two disadvantages, particularly if the number of covariates in X is very large: First, precision of the estimate decreases with the dimensionality of X, which is known as the *curse of dimensionality*. Second, policy makers and other users of evaluation studies often prefer to see one number and not a multidimensional estimate. We may therefore be interested in the *unconditional* treatment effect, in particular in estimating the average treatment effect

[3] In the following, it is assumed that the local *conditional* IV assumption is valid, but even it were not exactly true it is nevertheless rather likely that accounting for observed differences between units to the left and to the right of the threshold would help to *reduce* bias, even if not eliminating it completely.

in the *largest* subpopulation for which it is identified. More precisely, we may be interested in the treatment effect on all compliers:

$$\lim_{\varepsilon \to 0} E\left[Y_i^1 - Y_i^0 \mid D_i(z_0+\varepsilon) > D_i(z_0-\varepsilon), Z = z_0\right],$$

i.e. without conditioning on X. Under the assumptions (5) and (6), this is the largest subpopulation, since only the treatment status of the local compliers is affected by variation in Z. In a one-sided non-compliance design, this is the ATET, see Section 4.

From inspecting the right-hand side of (7) one might imagine to estimate the unconditional effect by integrating out the distribution of X and plugging in nonparametric estimators in the resulting expression:

$$\int \left(\frac{m^+(X, z_0) - m^-(X, z_0)}{d^+(X, z_0) - d^-(X, z_0)} \right) dF_X. \tag{8}$$

This approach, however, has two disadvantages. First, when X is high dimensional, the denominator in (8) may often be very close to zero, leading to a very high variance of (8) in small samples. Second, it does not correspond to a well-defined treatment effect for a specific population. The following theorem, however, shows that a nicer expression can be obtained for the treatment effect on the *local compliers*, which is in the form of a ratio of two integrals.

For stating the result, it is helpful to introduce more precise notation first. Let $\mathcal{N}_\varepsilon$ be an ε neighbourhood about z_0 and partition $\mathcal{N}_\varepsilon$ into $\mathcal{N}_\varepsilon^+ = \{z : z \geq z_0, z \in \mathcal{N}_\varepsilon\}$ and $\mathcal{N}_\varepsilon^- = \{z : z < z_0, z \in \mathcal{N}_\varepsilon\}$. According to their reaction to the instrument z over $\mathcal{N}_\varepsilon$ we can partition the population into five subpopulations:

$$\begin{aligned}
\tau_{i,\varepsilon} &= a \quad \text{if} \quad \min_{z \in \mathcal{N}_\varepsilon} D_i(z) = \max_{z \in \mathcal{N}_\varepsilon} D_i(z) = 1 \\
\tau_{i,\varepsilon} &= n \quad \text{if} \quad \min_{z \in \mathcal{N}_\varepsilon} D_i(z) = \max_{z \in \mathcal{N}_\varepsilon} D_i(z) = 0 \\
\tau_{i,\varepsilon} &= c \quad \text{if} \quad \min_{z \in \mathcal{N}_\varepsilon} D_i(z) < \max_{z \in \mathcal{N}_\varepsilon} D_i(z) \quad \text{and} \quad D_i(z) \text{ monotone over } \mathcal{N}_\varepsilon \\
\tau_{i,\varepsilon} &= d \quad \text{if} \quad \min_{z \in \mathcal{N}_\varepsilon} D_i(z) > \max_{z \in \mathcal{N}_\varepsilon} D_i(z) \quad \text{and} \quad D_i(z) \text{ monotone over } \mathcal{N}_\varepsilon \\
\tau_{i,\varepsilon} &= i \quad \text{if} \quad \min_{z \in \mathcal{N}_\varepsilon} D_i(z) \neq \max_{z \in \mathcal{N}_\varepsilon} D_i(z) \quad \text{and} \quad D_i(z) \text{ non-monotone.}
\end{aligned}$$

These subpopulations are a straightforward extension of the LATE concept of Imbens and Angrist (1994). The first group contains those units that will *always* be treated (for $z \in \mathcal{N}_\varepsilon$),

the second contains those that will *never* be treated, the third and fourth group (the compliers and defiers) contain those units that react (weakly) monotonously over $\mathcal{N}_\varepsilon$ while the fifth group (labelled indefinite) contains all units that react non-monotonously, e.g. they may first switch from $D = 0$ to 1 and then back for increasing values of z.

Under the assumptions given below, we can identify the treatment effect for the local compliers, i.e. those that switch from $D = 0$ to 1 at z_0. When the group of always-treated has measure zero, as in the one-sided non-compliance case, this also corresponds to ATET, as discussed in Section 4.

Theorem 1 (Identification of complier treatment effect) *Under the Assumption 1 given below, the local average treatment effect γ for the subpopulation of local compliers is nonparametrically identified as:*

$$\gamma = \lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 \mid Z \in \mathcal{N}_\varepsilon, \tau_\varepsilon = c\right] = \frac{\int \left(m^+(x, z_0) - m^-(x, z_0)\right) \cdot \left(f^+(x|z_0) + f^-(x|z_0)\right) dx}{\int \left(d^+(x, z_0) - d^-(x, z_0)\right) \cdot \left(f^+(x|z_0) + f^-(x|z_0)\right) dx}. \tag{9}$$

(Proof in appendix.)

A straightforward estimator of (9) is

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} \left(\hat{m}^+(X_i, z_0) - \hat{m}^-(X_i, z_0)\right) K_h\left(\frac{Z_i - z_0}{h}\right)}{\sum_{i=1}^{n} \left(\hat{d}^+(X_i, z_0) - \hat{d}^-(X_i, z_0)\right) K_h\left(\frac{Z_i - z_0}{h}\right)}, \tag{10}$$

where $\hat{m}$ and $\hat{d}$ are nonparametric estimators and $K_h(u) = \frac{1}{h}\kappa(u)$ is a positive, symmetric kernel function with h converging to zero with growing sample size. In addition to its well defined causal meaning, the estimator (10) is likely to behave more stable in finite samples than an estimator of (8) because the averaging over the distribution of X is conducted first before the ratio is taken.

In the following the assumptions for identification are discussed. They are presented somewhat differently from (5) and (6), on the one hand to relax these assumptions a little bit and state them more rigorously, but also to provide a more intuitive exposition, which may help to judge their plausibility for a given application. It is assumed throughout that the covariates

X are continuously distributed with a Lebesgue density. This is an assumption made for convenience to ease the exposition, particularly in the derivation of the asymptotic distributions. Discrete covariates can easily be included in X and identification does *not* require any continuous X variables. The derivation of the asymptotic distribution only depends on the number of continuous regressors in X. Discrete random variables do not affect the asymptotic properties and could easily be included at the expense of a more cumbersome notation. Only Z has to be continuous near z_0, but could have masspoints elsewhere.

Assumption 1:

i) Existence of compliers	$\lim_{\varepsilon \to 0} \Pr(\tau_\varepsilon = c \vert Z = z_0) > 0$
ii) Monotonicity	$\lim_{\varepsilon \to 0} \Pr\left(\tau_\varepsilon = t \vert Z \in \mathcal{N}_\varepsilon\right) = 0 \qquad \text{for } t \in \{d, i\}$
iii) Independent IV	$\lim_{\varepsilon \to 0} \Pr\left(\tau_\varepsilon = t \vert X, Z \in \mathcal{N}_\varepsilon^+\right) - \Pr\left(\tau_\varepsilon = t \vert X, Z \in \mathcal{N}_\varepsilon^-\right) = 0 \qquad \text{for } t \in \{a, n, c\}$
iv) IV Exclusion	$\lim_{\varepsilon \to 0} E\left[Y^1 \vert X, Z \in \mathcal{N}_\varepsilon^+, \tau_\varepsilon = t\right] - E\left[Y^1 \vert X, Z \in \mathcal{N}_\varepsilon^-, \tau_\varepsilon = t\right] = 0 \qquad \text{for } t \in \{a, c\}$
	$\lim_{\varepsilon \to 0} E\left[Y^0 \vert X, Z \in \mathcal{N}_\varepsilon^+, \tau_\varepsilon = t\right] - E\left[Y^0 \vert X, Z \in \mathcal{N}_\varepsilon^-, \tau_\varepsilon = t\right] = 0 \qquad \text{for } t \in \{n, c\}$
v) Common support	$\lim_{\varepsilon \to 0} Supp(X \vert Z \in \mathcal{N}_\varepsilon^+) = \lim_{\varepsilon \to 0} Supp(X \vert Z \in \mathcal{N}_\varepsilon^-)$
vi) Density at threshold	$F_Z(z)$ is differentiable at z_0 and $f_Z(z_0) > 0$
	$\lim_{\varepsilon \to 0} F_{X \vert Z \in \mathcal{N}_\varepsilon^+}(x)$ and $\lim_{\varepsilon \to 0} F_{X \vert Z \in \mathcal{N}_\varepsilon^-}(x)$ exist and are differentiable in x with pdf $f^+(x \vert z_0)$ and $f^-(x \vert z_0)$, respectively.
vii) Bounded moments	$E[Y^d \vert X, Z]$ are bounded away from $\pm$ infinity *a.s.* over $\mathcal{N}_\varepsilon \qquad \text{for } d \in \{0, 1\}$

(Regarding notation: $f^+(x, z_0) = f^+(x \vert z_0) f(z_0)$ refers to the joint distribution of X and Z whereas $f^+(x \vert z_0)$ refers to the conditional distribution of X.)

Assumption (1ii) requires that, in a very small neighbourhood of z_0, the instrument has a weakly monotonous impact on $D_i(z)$: Increasing z does never decrease $D_i(z)$ a.s. Assumption (1i) requires that $E[D \vert Z]$ is in fact discontinuous at z_0, i.e. assumes that some units change their treatment status exactly at z_0. Assumptions (1iii) and (1iv) essentially correspond to (5). Assumption (1v) ensures that the integral in (9) is well defined. If it is not satisfied, one can re-define (9) by restricting it to the common support. Assumption (1vi) requires that there is positive density at z_0, such that observations close to z_0 exist. The assumption (1vii) requires the conditional expectation functions to be bounded at some value from above and below in

a neighbourhood of z_0. It is invoked to permit interchanging the operations of integration and taking limits via the Dominated Convergence Theorem. (It is certainly stronger than needed and could be replaced with some kind of smoothness conditions on $E[Y^d|X,Z]$ in a neighbourhood of z_0.)

As argued before, the IV restrictions (1iii) and (1iv) will often be plausible only if X contains several covariates, depending on the process that generated the observed Z. The other substantial assumption is the monotonicity condition (1ii), whereas the remaining assumptions are mostly testable.

What happens if the monotonicity assumption is not valid? If there are defiers (but no individuals of the indefinite type), the right hand side of (9) nevertheless still identifies the treatment effect γ if the average treatment effect is the same for compliers and defiers.[4] (Proof see appendix.) Hence, γ is still identified and the same estimators, discussed below, can be used in this case.

3 Statistical properties

In this section, the statistical properties of two different estimators of (9) are discussed. It is shown that the most obvious estimator (10) achieves at best a convergence rate of $n^{-\frac{1}{3}}$. An alternative estimator, however, achieves a convergence rate of $n^{-\frac{2}{5}}$, i.e. the rate of univariate nonparametric regression. This is achieved through smoothing with implicit double boundary correction.

All three estimators proceed in two steps and require nonparametric first step estimates of m^+, m^-, d^+ and d^-. These can be estimated nonparametrically by considering only observations to the right or the left of z_0, respectively. Since this corresponds to estimation at a boundary point, local linear regression is suggested, which is known to display better boundary behaviour than conventional Nadaraya-Watson kernel regression. $m^+(x,z_0)$ is estimated by

[4] And assuming that Assumptions (1iii) and (1iv) also hold for the defiers.

local linear regression as the value of a that solves

$$\underset{a,b,c}{\arg\min} \sum_{j=1}^{n} \left(Y_j - a - b\left(Z_j - z_0\right) - c'\left(X_j - x\right)\right)^2 \cdot K_j I_j^+ \tag{11}$$

where $I_j^+ = 1(Z_j > z_0)$ and a product kernel is used

$$K_j = K_j(x, z_0) = \kappa\left(\frac{Z_j - z_0}{h_z}\right) \cdot \prod_{l=1}^{L} \bar{\kappa}\left(\frac{X_{jl} - x_l}{h_x}\right). \tag{12}$$

where κ and $\bar{\kappa}$ are univariate kernel functions, where κ is a second-order kernel and $\bar{\kappa}$ is a kernel of order $\lambda \geq 2$. The kernel κ is assumed to be *symmetric* and *integrating to one*. The following kernel constants will be used later: $\mu_l = \int_{-\infty}^{\infty} u^l \kappa(u) du$ and $\bar{\mu}_l = \int_{0}^{\infty} u^l \kappa(u) du$ and $\tilde{\mu} = \frac{\bar{\mu}_2}{2} - \bar{\mu}_1^2$. (With symmetric kernel $\bar{\mu}_0 = \frac{1}{2}$.) Furthermore define $\ddot{\mu}_l = \int_{0}^{\infty} u^l \kappa^2(u) du$. The kernel function $\bar{\kappa}$ is a univariate kernel of order λ, with kernel constants of this kernel be denoted as $\eta_l = \int u^l \bar{\kappa}(u) du$ and $\dot{\eta}_l = \int_{-\infty}^{\infty} u^l \bar{\kappa}^2(u) du$. The kernel function being of order λ means that $\eta_0 = 1$ and $\eta_l = 0$ for $0 < l < \lambda$ and $\eta_\lambda \neq 0$.[5]

A result derived later will require higher-order kernels if the number of continuous regressors is larger than 3. For applications with at most 3 continuous regressors, a second-order kernel will suffice such that $\bar{\kappa} = \kappa$ can be chosen.

Notice that three different bandwidths h_z, h_x, h are used. h is the bandwidth in the matching estimator to compare observations to the left and right of the threshold, whereas h_z and h_x determine the local smoothing area for the local linear regression, which uses observations only to the right or only to the left of the threshold. We will need some smoothness assumptions as well as conditions on the bandwidth values.

Assumption 2:

i) IID sampling: The data $\{(Y_i, D_i, Z_i, X_i)\}$ are iid from $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^L$

ii) Smoothness:

- $m^+(x, z)$, $m^-(x, z)$, $d^+(x, z)$, $d^-(x, z)$ are λ times continuously differentiable with respect

[5] For the Epanechnikov kernel with support $[-1, 1]$, i.e. $K(u) = \frac{3}{4}\left(1 - u^2\right) 1\left(|u| < 1\right)$ the kernel constants are $\mu_0 = 1$, $\mu_1 = \mu_3 = \mu_5 = 0$, $\mu_2 = 0.2$, $\mu_4 = 6/70$, $\bar{\mu}_0 = 0.5$, $\bar{\mu}_1 = 3/16$, $\bar{\mu}_2 = 0.1$, $\bar{\mu}_3 = 1/16$, $\bar{\mu}_4 = 3/70$.

to x at z_0 with λ-th derivative Hölder continuous in an interval around z_0,

- $f^+(x,z)$ and $f^-(x,z)$ are $\lambda-1$ times continuously differentiable with respect to x at z_0 with $(\lambda-1)$-th derivative Hölder continuous in an interval around z_0,
- $m^+(x,z)$, $d^+(x,z)$ and $f^+(x,z)$ have two continuous right derivatives with respect to z at z_0 with second derivative Hölder continuous in an interval around z_0,
- $m^-(x,z)$, $d^-(x,z)$ and $f^-(x,z)$ have two continuous left derivatives with respect to z at z_0 with second derivative Hölder continuous in an interval around z_0,

iii) the univariate Kernel functions κ and $\bar{\kappa}$ in (12) are bounded, Lipschitz and zero outside a bounded set; κ is a second-order kernel and $\bar{\kappa}$ is a kernel of order λ,

iv) Bandwidths: The bandwidths satisfy h, h_z, $h_x \to 0$ and $nh \to \infty$ and $nh_z \to \infty$ and $nh_z h_x^L \to \infty$.

v) Conditional variances: The left and right limits of the conditional variances $\lim_{\varepsilon \to 0} E\left[\left(Y - m^+(X,Z)\right)^2 | X, Z = z + \varepsilon\right]$ and $\lim_{\varepsilon \to 0} E\left[\left(Y - m^-(X,Z)\right)^2 | X, Z = z - \varepsilon\right]$ exist at z_0.

With these preliminaries we consider two estimators in turn. The estimator $\hat{\gamma}$ (10) will be considered last as it has the worst statistical properties. The first estimator $\hat{\gamma}_{RDD}$ considered below is a modification of (10) where some type of boundary kernel is used in the second smoothing step. Thereby a faster convergence rate can be achieved. The asymptotic distribution is derived for this estimator and it is shown that the asymptotic variance becomes smaller the more covariates X are included. For the $\hat{\gamma}$ estimator it is then shown that its convergence rate is lower than that of $\hat{\gamma}_{RDD}$.[6] All estimators are straightforward to implement with any statistical software package.

3.1 Boundary RDD kernel estimator

As will be seen later, the estimator (10) suffers from a low convergence rate. As an alternative, we could use a kernel function which implicitly adapts to the boundary. We define the RDD

[6] In an earlier version of the paper, also a two-step local linear estimator was considered which also has lower convergence rate than $\hat{\gamma}_{RDD}$.

estimator as

$$\hat{\gamma}_{RDD} = \frac{\sum_{i=1}^{n} \left(\hat{m}^{+}(X_i, z_0) - \hat{m}^{-}(X_i, z_0)\right) K_h^* \left(\frac{Z_i - z_0}{h}\right)}{\sum_{i=1}^{n} \left(\hat{d}^{+}(X_i, z_0) - \hat{d}^{-}(X_i, z_0)\right) K_h^* \left(\frac{Z_i - z_0}{h}\right)}, \tag{13}$$

where the kernel function is

$$K_h^*(u) = (\bar{\mu}_2 - \bar{\mu}_1 u) \cdot K_h(u). \tag{14}$$

By using this kernel function, the estimator $\hat{\gamma}_{RDD}$ achieves the convergence rate of a one dimensional nonparametric regression estimator, irrespective of the dimension of X. Loosely speaking, it achieves thus the fastest convergence rate possible and is not affected by a curse of dimensionality. This is achieved by smoothing over all other regressors and by an implicit boundary adaptation.

In addition, the bias and variance terms due to estimating m^+, m^-, d^+, d^- and due to estimating the density functions $\frac{f^-(x|z_0)+f^+(x|z_0)}{2}$ by the empirical distribution functions converge at the same rate.

For an optimal convergence result further below, we need to be specific about the choice of the bandwidth values.

Assumption 3:

The bandwidths satisfy the following conditions:

$$\begin{aligned} \lim_{n\to\infty} \sqrt{nh^5} &= r < \infty \\ \lim_{n\to\infty} \frac{h_z}{h} &= r_z \qquad \text{with } 0 < r_z < \infty \\ \lim_{n\to\infty} \frac{h_x^{\lambda/2}}{h} &= r_x < \infty. \end{aligned}$$

This assumption ensures that bias and standard deviation of the estimator converge at rate $n^{-\frac{2}{5}}$ to zero, i.e. bias and variance converge to zero at the rate of a univariate nonparametric regression.

Note that the last condition of Assumption 3 provides an upper bound on h_x, whereas Assumption (2iv) provides a lower bound on h_x. Suppose that h_x depends on the sample size

in the following way:

$$h_x \propto n^{\zeta},$$

then the bandwidth conditions of Assumption 2 and 3 together require that

$$-\frac{4}{5L} < \zeta \leq -\frac{2}{5\lambda}. \tag{15}$$

This implies that h_x converges at a slower rate to zero than h and h_z when $L \geq 4$, i.e. when X contains 4 or more continuous regressors.

A necessary condition for Assumption 2 and 3 to hold jointly thus is that $-\frac{4}{5L} < -\frac{2}{5\lambda}$ or equivalently $\lambda > \frac{L}{2}$. As further discussed below, this requires higher-order kernels if X contains 4 or more continuous regressors, whereas conventional kernels are sufficient otherwise.

Assumption 3 is sufficient for bias and variance to converge at the univariate nonparametric rate, which is summarized in the following theorem:

Theorem 2 (Asymptotic distribution of $\hat{\gamma}_{RDD}$) *a)Under Assumptions 1 and 2, the bias and variance terms of $\hat{\gamma}_{RDD}$ are of order*

$$\begin{aligned} Bias(\hat{\gamma}_{RDD}) &= O(h^2 + h_z^2 + h_x^{\lambda}) \\ Var(\hat{\gamma}_{RDD}) &= O\left(\frac{1}{nh} + \frac{1}{nh_z}\right) \end{aligned}$$

b) Under Assumptions 1, 2 and 3 the estimator is asymptotically normally distributed and converges at the univariate nonparametric rate

$$\sqrt{nh}\left(\hat{\gamma}_{RDD} - \gamma\right) \rightarrow N\left(\mathcal{B}_{RDD}, \mathcal{V}_{RDD}\right).$$

where $\mathcal{B}_{RDD} =$

$$\begin{aligned} &\frac{r}{\Gamma}\frac{\bar{\mu}_2^2 - \bar{\mu}_1\bar{\mu}_3}{4\tilde{\mu}f(z_0)}\int \left(m^+(x,z_0) - m^-(x,z_0) - \gamma\left(d^+(x,z_0) - d^-(x,z_0)\right)\right)\left(\frac{\partial^2 f^+}{\partial z^2}(x,z_0) + \frac{\partial^2 f^-}{\partial z^2}(x,z_0)\right)dx \\ &+\frac{rr_z^2}{\Gamma}\frac{\bar{\mu}_2^2 - \bar{\mu}_1\bar{\mu}_3}{2\tilde{\mu}}\int\left(\frac{\partial^2 m^+(x,z_0)}{\partial z^2} - \frac{\partial^2 m^-(x,z_0)}{\partial z^2} - \gamma\frac{\partial^2 d^+(x,z_0)}{\partial z^2} + \gamma\frac{\partial^2 d^-(x,z_0)}{\partial z^2}\right)\frac{f^-(x,z_0) + f^+(x,z_0)}{2f(z_0)}dx \\ &+\frac{rr_x^2\eta_\lambda}{\Gamma}\int\sum_{l=1}^{L}\left\{\frac{\partial^\lambda m^+(x,z_0)}{\lambda!\cdot\partial x_l^\lambda} + \sum_{s=1}^{\lambda-1}\frac{\partial^s m^+(x,z_0)}{\partial x_l^s}\omega_s^+ - \frac{\partial^\lambda m^-(x,z_0)}{\lambda!\cdot\partial x_l^\lambda} - \sum_{s=1}^{\lambda-1}\frac{\partial^s m^-(x,z_0)}{\partial x_l^s}\omega_s^-\right\}\frac{f^-(x,z_0) + f^+(x,z_0)}{2f(z_0)}dx \\ &-\frac{\gamma rr_x^2\eta_\lambda}{\Gamma}\int\sum_{l=1}^{L}\left\{\frac{\partial^\lambda d^+(x,z_0)}{\lambda!\cdot\partial x_l^\lambda} + \sum_{s=1}^{\lambda-1}\frac{\partial^s d^+(x,z_0)}{\partial x_l^s}\omega_s^+ - \frac{\partial^\lambda d^-(x,z_0)}{\lambda!\cdot\partial x_l^\lambda} - \sum_{s=1}^{\lambda-1}\frac{\partial^s d^-(x,z_0)}{\partial x_l^s}\omega_s^-\right\}\frac{f^-(x,z_0) + f^+(x,z_0)}{2f(z_0)}dx \end{aligned}$$

where $\Gamma = \int \left(d^+(x,z_0) - d^-(x,z_0)\right) \cdot \frac{f^-(x|z_0)+f^+(x|z_0)}{2} dx$

and $\omega_s^+ = \left\{ \frac{\partial^{\lambda-s} f^+(X_i,z_0)}{s!(\lambda-s)!\cdot \partial x_l^{\lambda-s}} - \frac{\partial^{\lambda-1} f^+(x_0,z_0)}{\partial x_1^{\lambda-1}} \cdot \left(\frac{\partial^{\lambda-2} f^+(x_0,z_0)}{\partial x_l^{\lambda-2}} \right)^{-1} \frac{(\lambda-2)!}{(\lambda-1)!s!(\lambda-1-s)!} \frac{\partial^{\lambda-1-s} f^+(X_i,z_0)}{\partial x_l^{\lambda-1-s}} \right\} / f^+(X_i,z_0)$

and ω_s^- *defined analogously*

and $\mathcal{V}_{RDD} =$

$$\frac{\bar{\mu}_2^2 \ddot{\mu}_0 - 2\bar{\mu}_2 \bar{\mu}_1 \ddot{\mu}_1 + \bar{\mu}_1^2 \ddot{\mu}_2}{\Gamma^2 4 \tilde{\mu}^2 f^2(z_0)} \times \Bigg(\quad \frac{1}{r_z} \int \left(f^+(x,z_0) + f^-(x,z_0)\right)^2$$
$$\times \left(\frac{\sigma_Y^{2+}(x,z_0) - 2\gamma\sigma_{YD}^{2+}(X,z_0) + \gamma^2 \sigma_D^{2+}(x,z_0)}{f^+(x,z_0)} + \frac{\sigma_Y^{2-}(x,z_0) - 2\gamma\sigma_{YD}^{2-}(X,z_0) + \gamma^2 \sigma_D^{2-}(x,z_0)}{f^-(x,z_0)} \right) dx$$
$$+ \int \left\{ m^+(x,z_0) - \gamma d^+(x,z_0) - m^-(x,z_0) + \gamma d^-(x,z_0) \right\}^2 \cdot \left(f^+(x,z_0) + f^-(x,z_0)\right) dx \quad \Bigg),$$

where $\sigma_Y^{2+}(X,z) = \lim_{\varepsilon \to 0} E\left[\left(Y - m^+(X,Z)\right)^2 | X, Z = z + \varepsilon \right]$
and $\sigma_{YD}^{2+}(X,z) = \lim_{\varepsilon \to 0} E\left[\left(Y - m^+(X,Z)\right)\left(D - d^+(X,Z)\right) | X, Z = z + \varepsilon \right]$ *and* $\sigma_D^{2+}(X,z) = \lim_{\varepsilon \to 0} E\left[\left(D - d^+(X,Z)\right)^2 | X, Z = z + \varepsilon \right]$ *and analogously for* $\sigma_Y^{2+}(X,z)$, $\sigma_{YD}^{2+}(X,z)$ *and* $\sigma_D^{2+}(X,z)$.

The part (15) of Assumption 3 requires that $\lambda > \frac{L}{2}$ to control the bias due to smoothing in the X dimension. If X contains at most 3 continuous regressors, a second order kernel $\lambda = 2$ can be used. Otherwise, higher order kernels are required to achieve a $n^{-\frac{2}{5}}$ convergence rate. Instead of using higher order kernels, one could alternatively use local higher order polynomial regression instead of local linear regression (11). However, when the number of regressors in X is large, this could be inconvenient to implement in practice since a large number of interaction and higher order terms would be required, which could give rise to problems of local multicollinearity in small samples and/or for small bandwidth values. On the other hand, higher order kernels are very convenient to implement when a product kernel (12) is used. Higher order kernels are only necessary for smoothing in the X dimension but not for smoothing along Z.

When a second order kernel is used and X contains at most 3 continuous regressors, the

bias term $\mathcal{B}_{RDD}$ simplifies to

$$\frac{r}{\Gamma}\frac{\bar{\mu}_2^2-\bar{\mu}_1\bar{\mu}_3}{4\tilde{\mu}f(z_0)}\int\left(m^+(x,z_0)-m^-(x,z_0)-\gamma\left(d^+(x,z_0)-d^-(x,z_0)\right)\right)\left(\frac{\partial^2 f^+}{\partial z^2}(x,z_0)+\frac{\partial^2 f^-}{\partial z^2}(x,z_0)\right)dx$$
$$+\frac{rr_z^2}{\Gamma}\frac{\bar{\mu}_2^2-\bar{\mu}_1\bar{\mu}_3}{2\tilde{\mu}}\int\left(\frac{\partial^2 m^+(x,z_0)}{\partial z^2}-\frac{\partial^2 m^-(x,z_0)}{\partial z^2}-\gamma\frac{\partial^2 d^+(x,z_0)}{\partial z^2}+\gamma\frac{\partial^2 d^-(x,z_0)}{\partial z^2}\right)\cdot\frac{f^-(x,z_0)+f^+(x,z_0)}{2f(z_0)}dx$$
$$+\frac{rr_x^2\mu_2}{2\Gamma}\int\sum_{l=1}^{L}\left\{\frac{\partial^2 m^+(x,z_0)}{\partial x_l^2}-\frac{\partial^2 m^-(x,z_0)}{\partial x_l^2}-\gamma\frac{\partial^2 d^+(x,z_0)}{2\cdot\partial x_l^2}+\gamma\frac{\partial^2 d^-(x,z_0)}{2\cdot\partial x_l^2}\right\}\cdot\frac{f^-(x,z_0)+f^+(x,z_0)}{2f(z_0)}dx.$$

3.2 Variance reduction through the use of control variables

In most of the discussion so far, the role of the X variables was to make the identifying assumptions more plausible. However, the X variables may also contribute to reducing the variance of the estimator, which is shown in the following proposition.

Suppose that the characteristics are identically distributed on both sides of the threshold such that γ is identified without controlling for any X. In this case one could estimate γ consistently by (13) with X being the empty set. This estimator is denoted $\hat{\gamma}_{no\,X}$ henceforth. Alternatively, one could use a set of control variables X in (13), which we denote as $\hat{\gamma}_{RDD}$ as before. Suppose that both estimators are consistent for γ.[7] As shown below, $\hat{\gamma}_{no\,X}$ generally has a *larger* asymptotic variance than $\hat{\gamma}_{RDD}$, i.e. than the estimator that controls for X. On the other hand, an ordering of squared bias seems not to be possible under general conditions. However, one can always choose a bandwidth sequence such that r is very small in Assumption 3, which would imply that the bias is negligible for both estimators. Hence, there are precision gains by controlling for X even if the RDD estimator would be consistent without any covariates.

For stating Proposition 3 in a concise way, some new notation is required. Let $w^+(X,z)=\lim_{\varepsilon\to 0}E\left[Y-\gamma D|X,Z=z+\varepsilon\right]$ be the right limit of the difference between Y and γD, and $w^+(z)=\lim_{\varepsilon\to 0}E\left[Y-\gamma D|Z=z+\varepsilon\right]$ be the corresponding expression without conditioning on X. Define further the variance of $w^+(X,z_0)$ as $V^+=\int\left\{w^+(x,z_0)-w^+(z_0)\right\}^2 f(x|z_0)dx$. Define $w^-(X,z)$, $w^-(z)$ and V^- analogously as the left limits. Proposition 3 shows that there are efficiency gains if $V^+\neq 0$ and/or $V^-\neq 0$.

[7] Hence, X should not include e.g. variables that are on the causal pathway or causally affected by D.

To gain some intuition: V^+ is the variance of the conditional expectation of Y given X plus the variance of the conditional expectation of D given X minus the covariance between these two terms. Hence, V^+ is nonzero if X is a predictor of Y and/or of D. On the other hand, V^+ and V^- are zero only if X *neither* predicts Y *nor* D.[8]

Define further the covariance C as $\int \left(w^+(x,z_0) - w^+(z_0)\right)\left(w^-(x,z_0) - w^-(z_0)\right) f(x|z_0)dx$ and the correlation coefficient $R = \frac{C}{\sqrt{V^+V^-}}$. Now, we can state the result in terms of the variances and the correlation coefficient. The results also depend on the bandwidth sequences. The variance of $\hat{\gamma}_{RDD}$ depends on the smoothing in the Z dimension via h and h_z. The $\hat{\gamma}_{no\,X}$ estimator only depends on h_z since there is no smoothing in the second step. A natural choice would thus be $h = h_z$.[9]

Proposition 3 *Let $\hat{\gamma}_{RDD}$ be the estimator (13) using the set of regressors X and let $\hat{\gamma}_{no\,X}$ be the estimator with X being the empty set. Denote the asymptotic variance of $\hat{\gamma}_{no\,X}$ by $\mathcal{V}_{no\,X}$ and assume that both estimators consistently estimate γ and satisfy Assumptions 2 and 3. Assume further that the distribution of X is continuous at z_0, i.e. $f^+(X,z_0) = f^-(X,z_0)$ a.s..*
(a) If $V^+ = V^- = 0$ then

$$\mathcal{V}_{RDD} - \mathcal{V}_{no\,X} = 0.$$

(b) Under any of the following conditions

$$\mathcal{V}_{RDD} - \mathcal{V}_{no\,X} < 0,$$

- if $V^+ = 0$ and $V^- \neq 0$ or vice versa and $r_z < 2$
- or if $V^+ \neq 0 \neq V^-$ and $R \geq 0$ and $r_z < 2$
- or if $V^+ \neq 0 \neq V^-$ and $-1 < R < 0$ and $r_z < 2\frac{1+R}{1-R^2}$.
- or if $V^+ \neq 0 \neq V^-$ and $R = -1$ and $r_z < 1$.
(Proof see appendix).

Hence, if, in case (a) of Proposition 3, X has no predictive power neither for Y nor for D, the asymptotic variances are the same. On the other hand, if X has predictive power *either*

[8] Excluding the unreasonable case where it predicts both but not $Y - \gamma D$.

[9] The variance of $\hat{\gamma}_{RDD}$ can be reduced even further relative to $\hat{\gamma}_{no\,X}$ by choosing $h_z < h$, but this would be more of a technical trick than a substantive result.

for Y *or* for D and one uses the same bandwidths for both estimators ($h_z = h$), the RDD estimator with covariates has a strictly smaller variance. This holds in all cases except for the very implausible scenario where $w^+(X, z_0)$ and $w^-(X, z_0)$ are negatively correlated with a correlation coefficient of -1. In most economic applications one would rather expect a clearly positive correlation.

Proposition 3 can easily be extended to show that the RDD estimator with a larger regressor set $\mathbf{X}$, i.e. where $X \subset \mathbf{X}$, has smaller asymptotic variance than the RDD estimator with X. (The proof is analogous and is omitted.) Hence, one can combine including some X for eliminating bias with adding further covariates to reduce variance. The more variables are included in $\mathbf{X}$ the smaller the variance will be.[10]

3.3 Naive RDD estimator

Finally, we examine the properties of the straightforward estimator (10):

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} \left(\hat{m}^+(X_i, z_0) - \hat{m}^-(X_i, z_0)\right) K_h\left(\frac{Z_i - z_0}{h}\right)}{\sum_{i=1}^{n} \left(\hat{d}^+(X_i, z_0) - \hat{d}^-(X_i, z_0)\right) K_h\left(\frac{Z_i - z_0}{h}\right)},$$

i.e. which uses the conventional Nadaraya-Watson type weighting by $K_h(u)$. In essence, it is a combination between local linear regression in the first step and Nadaraya-Watson regression in the second step. Although this estimator appears to be the most obvious and straightforward way to estimate (9) it has worse statistical properties than the previous estimator in the sense that it achieves only a lower rate of convergence. The intuition for this is the missing boundary correction in the second step, which is implicitly included in the previous estimator.

Proposition 4 (Aymptotic properties of $\hat{\gamma}$) *Under Assumptions 1 and 2, the bias and variance terms of $\hat{\gamma}$ are of order*

$$\begin{aligned} Bias(\hat{\gamma}) &= O(h + h_z^2 + h_x^\lambda) \\ Var(\hat{\gamma}) &= O\left(\frac{1}{nh} + \frac{1}{nh_z}\right). \end{aligned}$$

[10] Proposition 3 is derived under the assumption that $\dim(\mathbf{X})$ does not grow with sample size. If $\dim(\mathbf{X})$ is very large in a particular application, the result of Proposition 3 may not be appropriate anymore. This will be examined in further research.

(The exact expressions for bias and variance are given in the appendix).

From this result it can be seen that the fastest rate of convergence possible by appropriate bandwidth choice is $n^{-\frac{1}{3}}$. It is straightforward to show asymptotic normality for this estimator, but the (first order) approximation may not be very useful as it would be dominated by the bias and variance terms $O(h)$ and $O(\frac{1}{nh})$. The terms corresponding to the estimation error of $\hat{m}^+(x,z_0), \hat{m}^-(x,z_0), \hat{d}^+(x,z_0), \hat{d}^-(x,z_0)$ would be of lower order and thus ignored in the first-order approximation. The bias and variance approximation thus obtained would be the same as in a situation where $m^+(x,z_0), m^-(x,z_0), d^+(x,z_0), d^-(x,z_0)$ were known and not estimated. A more useful approximation can be obtained by retaining also the lower order terms. Overall, however, it seems to be more useful to use the previously proposed estimator $\hat{\gamma}_{RDD}$ instead.

4 Extensions and quantile treatment effects

4.1 Conditional independence and sharp design

The previous sections showed how to incorporate differences in covariates X in a regression discontinuity design (RDD) in a fully nonparametric way. Identification and estimation was examined for the *fuzzy design* under a local IV condition (5), which permits *unobserved heterogeneity*.

Alternatively, one could consider weakening the conditional independence assumption (3) to

$$Y_i^1 - Y_i^0 \perp\!\!\!\perp D_i | X_i, Z_i \qquad \text{for } Z_i \text{ near } z_0. \tag{16}$$

Analogously to the derivations in Hahn, Todd, and van der Klaauw (2001) it follows then that

$$E\left[Y^1 - Y^0 | X, Z = z_0\right] = \frac{m^+(X,z_0) - m^-(X,z_0)}{d^+(X,z_0) - d^-(X,z_0)}.$$

Similarly to the derivations for Theorem (1), one can show that the unconditional treatment effect for the population near the threshold would then be

$$E\left[Y^1 - Y^0 | Z = z_0\right] = \int \frac{m^+(x,z_0) - m^-(x,z_0)}{d^+(x,z_0) - d^-(x,z_0)} \cdot \frac{f^+(x|z_0) + f^-(x|z_0)}{2} dx.$$

This expression may be difficult to estimate in small samples as the denominator can be very small for some values of x.[11] In this case it may be helpful to strengthen the CIA assumption (16) to

$$Y_i^1, Y_i^0 \perp\!\!\!\perp D_i | X_i, Z_i \qquad \text{for } Z_i \text{ near } z_0. \tag{17}$$

This permits to identify the treatment effect as

$$\begin{aligned} &E\left[Y^1 - Y^0 | Z = z_0\right] \\ &= \int \left(E\left[Y|D=1, X=x, Z=z_0\right] - E\left[Y|D=0, X=x, Z=z_0\right]\right) \cdot \frac{f^+(x|z_0) + f^-(x|z_0)}{2} dx, \end{aligned}$$

where $E\left[Y|D, X, Z=z_0\right]$ can be estimated by a combination of the left hand side and the right hand side limit. This approach does no longer have to rely only on comparing observations across the threshold but also uses variation within either side of the threshold.

For the *sharp* design, *all* these different parameters are identified as

$$\int \left(m^+(x, z_0) - m^-(x, z_0)\right) \cdot \frac{f^+(x|z_0) + f^-(x|z_0)}{2} dx. \tag{18}$$

This follows because $d^+(x, z_0) - d^-(x, z_0) = 1$ in the sharp design since everyone is a complier. The estimators discussed in the previous section can be used to estimate (18) by straightforward modifications and all the previously obtained results apply analogously.

4.2 Combination of sharp and fuzzy design

An interesting situation occurs when the RDD is sharp on the one side but fuzzy on the other. An important case is when *eligibility* depends strictly on observed characteristic but participation in *treatment* is voluntary. For example, eligibility to certain treatments may be means tested (e.g. food stamps programmes) with a strict eligibility threshold z_0, but take-up of the treatment may be less than 100 percent. As another example, eligibility to certain labor market programmes may depend on the duration of unemployment or on the age of individuals. E.g. the "New Deal for Young People" in the UK *offers* job-search assistance (and

[11] This problem is of much less concern for the estimators of the previous section as those were based on a ratio of two integrals and not on an integral of a ratio. For those estimators the problem of very small denominators for some values of X averages out.

other programmes) to all individuals aged between eighteen and twenty-four who have been claiming unemployment insurance for six months.

Accordingly, the population consists of three subgroups (near the threshold): ineligibles, eligible non-participants and participants. This setup thereby rules out the existence of defiers such that the monotonicity condition (Assumption 1ii) is automatically fulfilled close to z_0. In addition, the average *treatment effect on the treated* (ATET) is now identified[12] and is given by

$$\lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 \mid D = 1, Z \in \mathcal{N}_\varepsilon\right] = \frac{\int \left(m^+(x, z_0) - m^-(x, z_0)\right) \left(f^+(x|z_0) + f^-(x|z_0)\right) dx}{\int d^+(x, z_0) \cdot \left(f^+(x|z_0) + f^-(x|z_0)\right) dx}. \quad (19)$$

where it has been supposed that only individuals *above* a threshold z_0 are eligible. The previously examined estimators apply with obvious modifications.

4.3 Quantile treatment effects

The previous discussion only referred to the estimation of the *average* treatment effect. In many situations one might be interested in distributional aspects as well, e.g. educational inequality of a particular schooling intervention or wage inequality effects of a labour market intervention. Quantile treatment effects (QTE) have received considerable attention in recent years, see e.g. Abadie, Angrist, and Imbens (2002). The following theorem shows identification of the distribution of the potential outcomes under a local IV condition, which permits unobserved heterogeneity. (The adjustments necessary for the sharp or mixed sharp-fuzzy design or when using the conditional independence assumption (16) or (17) are straightforward and are omitted here.) By strengthening Assumption 1 a little, the potential outcome distributions are identified:

Theorem 5 (Distribution of potential outcomes) *Under Assumption 1, where in Assumption (1iv) the symbols* Y^1 *and* Y^0 *are replaced by* $1\left(Y^1 \leq u\right)$ *and* $1\left(Y^0 \leq u\right)$,

[12] This is because there are no always-participants in this setup. Hence, the treated are the compliers.

respectively, the potential outcome distributions for the local compliers are identified as

$$\lim_{\varepsilon\to 0} F_{Y^1|Z\in\mathcal{N}_\varepsilon,\tau_\varepsilon=c}(u) = \frac{\int(\ddot{F}^1(u,D,x,z_0)-\ddot{F}^0(u,D,x,z_0))\,(f^+(x|z_0)+f^-(x|z_0))\,dx}{\int(d^+(x,z_0)-d^-(x,z_0))\,(f^+(x|z_0)+f^-(x|z_0))\,dx} \quad (20)$$

$$\lim_{\varepsilon\to 0} F_{Y^0|Z\in\mathcal{N}_\varepsilon,\tau_\varepsilon=c}(u) = \frac{\int(\ddot{F}^1(u,D-1,x,z_0)-\ddot{F}^0(u,D-1,x,z_0))\,(f^+(x|z_0)+f^-(x|z_0))\,dx}{\int(d^+(x,z_0)-d^-(x,z_0))\,(f^+(x|z_0)+f^-(x|z_0))\,dx}$$

where $\ddot{F}^1(u,d,x,z)=\lim_{\varepsilon\to 0}E\left[1\left(Y\le u\right)\cdot d|X=x,Z=z+\varepsilon\right]$

and $\ddot{F}^0(u,d,x,z)=\lim_{\varepsilon\to 0}E\left[1\left(Y\le u\right)\cdot d|X,Z=z-\varepsilon\right]$. *(Proof see appendix.)*

The quantile treatment effect for quantile τ is now obtained by inversion

$$QTE^\tau=\lim_{\varepsilon\to 0}F^{-1}_{Y^1|Z\in\mathcal{N}_\varepsilon,\tau_\varepsilon=c}(\tau)-\lim_{\varepsilon\to 0}F^{-1}_{Y^0|Z\in\mathcal{N}_\varepsilon,\tau_\varepsilon=c}(\tau). \quad (21)$$

The previous methods could thus be used for the estimation of quantile treatment effects by straightforward modifications. The asymptotic variance formula will be somewhat more complex for QTE^τ because of the correlation between the two terms in (21). Alternative estimators based on direct estimation of the quantiles could be developed along the lines of Frölich and Melly (2006). This is left for future research.

5 Conclusions

In this paper, the *regression discontinuity* design (RDD) has been generalized to account for differences in observed *covariates* X. Incorporating covariates X will often be important to eliminate (or at least reduce) bias. In addition, accounting for covariates also reduces variance. It has been shown that the curse of dimensionality does not apply and that the average treatment effect (on the local compliers) can be estimated at rate $n^{-\frac{2}{5}}$ irrespective of the dimension of X. For achieving this rate, a boundary RDD estimator and a 2SLL estimator have been suggested. (A naive kernel estimator would only achieve a lower convergence rate.)

If X contains at most 3 continuous regressors, conventional second order kernels can be used. If X contains more continuous regressors, higher order kernels are required, which can conveniently be implemented in the form of product kernels.

In a mixed sharp-fuzzy design, e.g. when eligibility rules are strict but treatment is voluntary, the treatment effect on the treated (ATET) is identified. Finally, estimation of QTE has been considered, which can be achieved at the same rate.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Angrist, J., and V. Lavy (1999): "Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–575.

Battistin, E., and E. Rettore (2002): "Testing for programme effects in a regression discontinuity design with imperfect compliance," *Journal of Royal Statistical Society Series A*, 165, 39–57.

Black, S. (1999): "Do 'Better' Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, 114, 577–599.

Frölich, M., and B. Melly (2006): "Nonparametric quantile treatment effects under endogeneity," *mimeo*.

Hahn, J., P. Todd, and W. van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Lalive, R. (2007): "How do extended benefits affect unemployment duration? A regression discontinuity approach," *forthcoming in Journal of Econometrics*, x, x–x.

Porter, J. (2003): "Estimation in the regression discontinuity model," mimeo.

Puhani, P., and A. Weber (2007): "Does the early bird catch the worm? Instrumental variable estimates of early educational effects of age of school entry in Germany," *Empirical Economics*, 32, 359–386.

Trochim, W. (1984): *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Sage Publications, Beverly Hills.

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