I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11138

Including Covariates in the Regression Discontinuity Design

Markus Frölich
Martin Huber

NOVEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11138

Including Covariates in the Regression Discontinuity Design

Markus Frölich
Center for Evaluation and Development (C4ED), University of Mannheim, IZA and J-PAL

Martin Huber
University of Fribourg

NOVEMBER 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Including Covariates in the Regression Discontinuity Design

This paper proposes a fully nonparametric kernel method to account for observed covariates in regression discontinuity designs (RDD), which may increase precision of treatment effect estimation. It is shown that conditioning on covariates reduces the asymptotic variance and allows estimating the treatment effect at the rate of one-dimensional nonparametric regression, irrespective of the dimension of the continuously distributed elements in the conditioning set. Furthermore, the proposed method may decrease bias and restore identification by controlling for discontinuities in the covariate distribution at the discontinuity threshold, provided that all relevant discontinuously distributed variables are controlled for. To illustrate the estimation approach and its properties, we provide a simulation study and an empirical application to an Austrian labor market reform.

JEL Classification: C13, C14, C21

Keywords: treatment effect, causal effect, complier, LATE, nonparametric regression, endogeneity

Corresponding author:
Markus Frölich
University of Mannheim
68131 Mannheim
Germany
E-mail: froelich@uni-mannheim.de

1 Introduction

The regression discontinuity design (RDD) has received tremendous attention in many fields, e.g. labor markets, political economy, health, education, psychology, criminology, as a credible approach to identifying causal effects without having to resort to fully randomized experiments. Hahn, Todd, and van der Klaauw (2001) formalize the assumptions required to identify causal effects in the RDD and provide nonparametric (local linear) estimators. Porter (2003) complements their work by alternative estimators. Lee and Card (2008) consider the case when the forcing variable is discrete. McCrary (2008) proposes a test for the manipulation of the running variable related to the continuity of its density function. Imbens and Lemieux (2008), van der Klaauw (2008) and Lee and Lemieux (2010) survey the applied and theoretical literature on the RDD. Imbens and Kalyanaraman (2012) discuss optimal bandwidth selection in terms of squared error loss, while Calonico, Cattaneo, and Titiunik (2014) propose methods for robust inference along with optimal bandwidth selection. Dong (2014) presents an alternative to some of the identifying assumptions in Hahn, Todd, and van der Klaauw (2001).

In this paper, the regression discontinuity approach is extended to incorporate covariates in a fully nonparametric way. Our estimator is based on a local nonparametric regression approach, i.e. kernel-based estimation, which allows deriving closed-form expressions for bias and variance.[1] Consider the setup of the RDD: D is a binary treatment indicator, Y is the outcome variable of interest, and Z is the 'forcing variable' with a known threshold z_0 at which the treatment probability $\Pr(D = 1|Z)$ is discontinuous. There are various motivations for accounting for covariates, denoted by X. A first reason is variance reduction, which is well known for the parametric case. But gains in precision can also be achieved in the nonparametric setup, as flexibly including covariates and averaging them out in an appropriate way reduces the asymptotic variance of the estimated treatment effect. We show that under mild regularity conditions, incorporating covariates permits estimating the treatment effect at the rate for *one-*

[1] An alternative approach could use global nonparametric methods such as sieves or polynomials of increasing order. However, such global methods, which are capable of fitting regression curves at many points by means of extrapolation, may perform poorly in the RDD, where a good fit is only needed at the treatment threshold, see Gelman and Imbens (2016). Extrapolation from far-away data points is also inherent in linear regression where one linearly controls for covariates.

dimensional nonparametric regression, i.e. $n^{-\frac{2}{5}}$ (where n is the sample size), irrespective of the dimension of the continuously distributed elements in X. Hence, the curse of dimensionality does not apply due to smoothing over X.

Second, as pointed out in Imbens and Lemieux (2008), covariates may mitigate small sample biases in cases where the number of observations close to the threshold z_0 is small such that one has to include observations in the estimation that are further apart and may potentially differ in X. Controlling for X might eliminate some of the bias that is introduced by observations further away from the threshold, as illustrated in Black, Galdo, and Smith (2007). However, biases related to unobserved characteristics cannot be accounted for.

Third, we also permit for situations where the density $f(X|Z)$ is discontinuous at z_0, which may point to a failure of the RDD assumptions, see Lee (2008), such that the simple RDD estimator is generally inconsistent. Our approach nevertheless identifies a local treatment effect in cases in which X contains all variables that (i) are imbalanced around the threshold and (ii) affect the outcome variable. With this respect, our contribution distinguishes itself from a more recent paper on RDD with covariates by Calonico, Cattaneo, Farrell, and Titiunik (2016), who assume $f(X|Z)$ to be continuous at z_0. Under that stronger identifying condition not needed here, Calonico, Cattaneo, Farrell, and Titiunik (2016) discuss potential precision gains when linearly (rather than nonparametrically as in our method) controlling for X and provide methods for optimal bandwidth selection and robust inference.

One example for $f(X|Z)$ being discontinuous at z_0 is 'classical confounding' where manipulation of Z at the threshold is selective with respect to characteristics that may also affect the outcome, see for instance Urquiola and Verhoogen (2009). If all confounding characteristics are observed in the data, our method yields the treatment effect on compliers at the threshold. See also van der Klaauw (2008) for confounding in the context of dynamic treatment assignment, where observed earlier treatment eligibility or participation (X) jointly affects the (current) forcing variable Z and Y. As a further example, consider the case when Z not only affects D, but also further variables that affect Y. This may occur in spatial RDDs where Z is based on distance to geographical borders. Eugster, Lalive, Steinhauer, and Zweimüller (2017), for instance, use the (mainly French and German) language border within administrative units

of Switzerland to estimate the effects of culture on unemployment. The authors consider a measure of the 'taste for leisure' as one particular indicator of culture. However, in addition to this treatment variable, further community-based covariates that are likely affected by culture also change discontinuously at the border. Controlling for X is therefore necessary as Z would otherwise violate the exclusion restriction with respect to Y at the threshold through its influence on X. Identification of a causal effect is, however, only obtained if X are not 'bad controls' which are affected by unobservables that also influence Y.

The remainder of this paper is organized as follows. Section 2 discusses the identification of the treatment effect in the presence of covariates. Section 3 proposes two estimators and examines their properties and shows that one of them achieves the $n^{-\frac{2}{5}}$ convergence rate. Section 4 provides a simulation study that (among others) illustrates the implications of confounding related to observed covariates at the threshold when applying RDD with and without controlling for X. Section 5 presents an empirical application to Austrian labor market reform previously considered by Lalive (2008) to estimate the effect of age-dependent eligibility to unemployment benefits on unemployment duration. As employees at risk of becoming unemployed might negotiate the exact date of dismissal with their employers, manipulation at the age threshold is a concern. We therefore control for a range of labor market-relevant characteristics that are potential confounders and find our results to differ from RDD without X. Section 6 concludes.

2 RDD with covariates

We define causal effects using the potential-outcome notation in the framework known as the Neyman-Fisher-Rubin causal model.[2] 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 effects $Y_i^1 - Y_i^0$ are permitted to vary across individuals, i.e. no constant treatment effect is assumed. Let Z_i be a variable that influences the treatment variable in a discontinuous way.

In the literature, two distinct designs are examined: the *sharp* design where D_i changes for

[2] See Neyman (1923), Fisher (1935) and Rubin (1978).

everyone at a known threshold z_0, and the *fuzzy* design where D_i changes only for a subset of individuals. In the sharp design (Trochim 1984), participation status is given by a deterministic function of Z, e.g.

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

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. For this reason, the subsequent discussion mostly focusses on the more general fuzzy design.[3] See Hahn, Todd, and van der Klaauw (2001) for more details.

Identification is feasible under the continuity of the mean potential outcomes at z_0 and relies on comparing the observed outcomes of those individuals to the left of the threshold with those to the right. In addition to continuity of $E[Y^d|Z = z]$ in z at z_0 for $d = \{0, 1\}$, Hahn, Todd, and van der Klaauw (2001) consider two alternative identifying assumptions:

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

or

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

Assumption (3) is a local selection on observables assumption and identifies the average treatment effect at the threshold: $E[Y^1 - Y^0|Z = z_0]$. Assumption (4) is an instrumental variables assumption that identifies a local average treatment effect (LATE) for a local group of compliers at the threshold:

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

[3] Battistin and Rettore (2008) introduce the mixed sharp fuzzy design as a special case of the fuzzy design.

In the sharp design, everyone is a complier at z_0 and assumption (3) is meaningless (i.e. has no identifying power) such that one needs assumption (4). In the fuzzy design one typically invokes (4), since the conditional independence assumption (3) does not permit treatment selection based on individual gains $Y_i^1 - Y_i^0$. It is worth mentioning that Dong (2014) recently has shown that alternatively to (4), identification of the LATE is obtained by making a continuity assumption of Z in the neighbourhood of z_0.[4]

In the following, we introduce observed covariates X_i and assume that (4) is valid *conditional* on X. As an example, suppose that there exists a liberalized education market in which schools may charge tuition fees, and that by law classes must be split if the number of students surpasses a particular threshold. As argued in Urquiola and Verhoogen (2009) for the case of Chile, schools close to the threshold might adjust tuition fees, thereby causing discontinuities in the admitted students' socioeconomic characteristics such as household income and parents' education. Assume that the latter variables also affect the outcome of interest, e.g. students' educational degree, which implies a violation of HTK2 when assessing the educational effect of class size. However, if household income, parents' education, and all other variables imbalanced at the threshold and affecting the outcome are observed, (4) holds conditional on X_i.[5] By an analogous reasoning as in HTK, and further assumptions made precise below, it follows immediately that the treatment effect on the local compliers *conditional* on X is identified as:

$$\lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 \mid X, D(z_0 + \varepsilon) > D(z_0 - \varepsilon), Z = z_0\right] = \frac{m^+(X, z_0) - m^-(X, z_0)}{d^+(X, z_0) - d^-(X, z_0)}, \tag{5}$$

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.

In this paper, however, we focus on identifying and estimating the *unconditional* effect

$$\lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 \mid D(z_0 + \varepsilon) > D(z_0 - \varepsilon), Z = z_0\right], \tag{6}$$

[4] Continuity of Z implies the smoothness of mean potential outcomes conditional on compliance behavior and of the shares of subgroups defined upon compliance at the threshold, which is sufficient for identification.

[5] Whether it is plausible to assume that all imbalanced covariates affecting the outcome are observed depends on the empirical problem and the richness of data. In in the context of Urquiola and Verhoogen (2009), for instance, ambition might (in addition to parents' education and household income) play a role for selectively (re-)placing students into particular class sizes. One would therefore want to condition on a rich set of socio-economic household characteristics and personality traits, e.g. provided by means of a household survey.

i.e. the effect on all compliers *without* conditioning on X. We identify this effect by first controlling for X and thereafter averaging over X. There are at least three reasons, why estimating the unconditional effect (6) is interesting (or even more interesting than the conditional effect (5)). First, for the purpose of evidence-based policy-making a small number of summary measures can be more easily conveyed to policy makers and the public than a large number of estimated effects at every value of X. Second, unconditional effects can be estimated more precisely than conditional effects. Third, the definition of unconditional effects does not depend on the variables included in X.[6] One can therefore consider different sets of control variables X and still estimate the same object, which is useful for examining robustness of the results to the set of control variables. See also Frölich (2007).

For showing identification of the unconditional effect (6), we first introduce some further notation. Let $\mathcal{N}_\varepsilon$ be a symmetric 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 four subpopulations:

$$\begin{aligned}
\tau_{i,\varepsilon} &= a \quad \text{if} \quad D_i(z) = 1 \quad \forall z \in \mathcal{N}_\varepsilon^- \quad \text{and} \quad D_i(z) = 1 \quad \forall z \in \mathcal{N}_\varepsilon^+ \\
\tau_{i,\varepsilon} &= n \quad \text{if} \quad D_i(z) = 0 \quad \forall z \in \mathcal{N}_\varepsilon^- \quad \text{and} \quad D_i(z) = 0 \quad \forall z \in \mathcal{N}_\varepsilon^+ \\
\tau_{i,\varepsilon} &= c \quad \text{if} \quad D_i(z) = 0 \quad \forall z \in \mathcal{N}_\varepsilon^- \quad \text{and} \quad D_i(z) = 1 \quad \forall z \in \mathcal{N}_\varepsilon^+ \\
\tau_{i,\varepsilon} &= d \quad \text{if} \quad D_i(z) = 1 \quad \forall z \in \mathcal{N}_\varepsilon^- \quad \text{and} \quad D_i(z) = 0 \quad \forall z \in \mathcal{N}_\varepsilon^+.
\end{aligned}$$

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 (if $Z \in \mathcal{N}_\varepsilon$), the second contains those that will *never* be treated (if $Z \in \mathcal{N}_\varepsilon$), and the third and fourth group contains the units that are treated only on one side of z_0.[7] We will assume that the fourth group, i.e. the 'defiers', has measure zero for ε sufficiently small. Note that in the sharp design, everyone is a complier for any $\varepsilon > 0$.

Under the following assumption, we can identify the treatment effect for the local com-

[6] This, of course, is only true if X exclusively contains pre-treatment variables.

[7] In the appendix we also consider a possible fifth group of indefinite units, for which no left-limit of $D_i(z)$ may exist. We assume this group to not exist, i.e. we require that all units have well defined left-limits of $D_i(z)$.

pliers, i.e. for those who switch from $D = 0$ to 1 at z_0.[8] It is assumed throughout that the covariates X are continuously distributed with a Lebesgue density. This assumption is made for convenience to ease exposition, particularly in the derivation of the asymptotic distributions later on. Discrete covariates can (at the expense of more cumbersome notation) easily be included in X, as the derivation of the asymptotic distribution only depends on the number of continuous regressors in X, while discrete variables do not affect the asymptotic properties. In fact, identification does *not* require any continuous X variables. Only Z has to be continuous near z_0, but could have masspoints elsewhere.

Assumption 1: For a symmetric neighbourhood $\mathcal{N}_\varepsilon$ about z_0 and for almost every X

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

Concerning notation, $f^+(x, z_0) = f^+(x|z_0)f(z_0)$ refers to the joint density of X and Z whereas $f^+(x|z_0)$ refers to the conditional density of X.

This assumption requires that in a neighbourhood about z_0, the threshold acts like a local instrumental variable. Assumptions 1 (i) to (iv) are instrumental variable assumptions for a binary instrument, as discussed e.g. in Imbens (2001). The monotonicity assumption 1(ii) rules out defiers at the threshold z_0, while 1(i) requires the existence of compliers. We note that 1(i) and 1(ii) could be relaxed to a local version of the compliers-defiers assumption of de Chaisemartin (2016), which allows for defiers under particular conditions, at the cost of

[8] The conditions in Assumption 1 are very similar, but a little weaker, to a conditional-on-X version of (4).

identifying the effects only for a subset of compliers (the so-called 'comvivors'). Assumptions 1(iii) and 1(iv) represent the exclusion restriction, conditional on X. Assumption 1(v) requires common support because we need to integrate over the support of X in (7).[9] Assumption 1(vi) implies positive density at z_0, such that observations close to z_0 exist.

We also assume the existence of the limit density functions $f^+(x|z_0)$ and $f^-(x|z_0)$ at the threshold z_0. So far, we do not assume anything about their continuity with respect to z. In other words, the conditional density could be discontinuous, i.e. $f^+(x|z_0) \neq f^-(x|z_0)$, in which case controlling for X is important for identification and thus consistent estimation, or it could be continuous, i.e. $f^+(x|z_0) = f^-(x|z_0)$, in which case identification does not hinge on controlling for observed covariates. The latter may, however, reduce the variance of the point estimator, as discussed below.[10]

Assumption (1vii) requires the conditional expectation functions to be bounded 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.[11]

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

$$\gamma = \lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 \mid Z \in \mathcal{N}_\varepsilon, \tau_\varepsilon = c\right] = \frac{\int \left(m^+(x,z_0) - m^-(x,z_0)\right) \cdot \frac{f^+(x|z_0)+f^-(x|z_0)}{2} dx}{\int \left(d^+(x,z_0) - d^-(x,z_0)\right) \cdot \frac{f^+(x|z_0)+f^-(x|z_0)}{2} dx}. \quad (7)$$

Proof: See the appendix.

Under Assumption 1, the treatment effect for the local compliers is identified as a *ratio* of two integrals, as shown in Theorem 1. The numerator in (7) is the intention-to-treat

[9] If this assumption is not satisfied, one can redefine (7) by restricting it to the common support.

[10] Note that Assumption 1 is somewhat stronger than needed for identification. Assumptions (1i) to (1iv) could be replaced with other assumptions that identify the local treatment effect conditional on X. For instance, if local compliers and local defiers had the same treatment effect, one could drop the monotonicity assumption. In addition, the existence of a density function for X is not needed.

[11] This assumption is certainly stronger than needed and could be replaced with some other smoothness conditions on $E[Y^d|X,Z]$ in a neighbourhood of z_0.

(ITT) effect of Z on Y, weighted by the conditional density of X, at z_0. (In the limit, the density of X conditional on Z being within a symmetric neighbourhood around z_0 is given by $\frac{f^+(x|z_0)+f^-(x|z_0)}{2}$.) The denominator in (7) gives the effect of Z on D, i.e. the fraction of compliers, at z_0. Thus, the ratio of integrals gives the ITT effect multiplied with the inverse of the number of compliers, corresponding to the LATE at z_0.

The ratio of integrals expression in (7) is obtained by applying iterated expectations to

$$E\left[Y^1 - Y^0 \mid Z \in \mathcal{N}_\varepsilon, \tau_\varepsilon = c\right]$$

to obtain

$$= \int E\left[Y^1 - Y^0 \mid X = x, Z \in \mathcal{N}_\varepsilon, \tau_\varepsilon = c\right] \cdot f_{X|Z\in\mathcal{N}_\varepsilon,\tau_\varepsilon=c}(x)\, dx. \tag{8}$$

Clearly, the density $f(X|Z \in \mathcal{N}_\varepsilon, \tau_\varepsilon = c)$ among the local compliers is not identified since the type τ_ε is unobservable. However, by applying Bayes' theorem to $f(X|Z \in \mathcal{N}_\varepsilon, \tau_\varepsilon = c)$ and replacing the first term in (8) with (5) (before taking limits), several terms cancel out and we obtain after various calculations the expression (7), which relies on observed variables only. See the supplementary appendix for detailed derivations. We thereby have identified the average effect. Similarly, we could identify Quantile Treatment Effects by combining the previous derivations with the reasoning in Frölich and Melly (2013) and Frandsen, Frölich, and Melly (2012).

So far, we have identified the treatment effect for the compliers in the fuzzy design. Without restrictions on treatment effect heterogeneity, it is impossible to identify the effects for always- and never-participants since they would never change treatment status in a neighbourhood of z_0. However, in the *sharp design*, everyone is a complier at z_0, i.e. $d^+(x, z_0) - d^-(x, z_0) = 1$, and the expression (7) simplifies to

$$\lim_{\varepsilon\to 0} E\left[Y^1 - Y^0 \mid Z \in \mathcal{N}_\varepsilon\right] = \int \left(m^+(x, z_0) - m^-(x, z_0)\right) \cdot \frac{f^+(x|z_0) + f^-(x|z_0)}{2} dx. \tag{9}$$

The estimand (9) in the sharp design is identical to the numerator of (7). The following discussion focusses on the estimation of (7), where the numerator and denominator of (7) are analyzed separately. Therefore, the asymptotic distribution of (9) in the sharp design is

immediately obtained by using the results for the numerator of (7) only. We also note that the estimands (7) and (9) bear some resemblance to the partial means estimator of Newey (1994). Both the numerator and denominator of (7) have a partial means form, in that averages over the covariates X are taken, at the left and the right limit at z_0.

Instead of generalizing assumption (4) to permit for further covariates X, we could alternatively start from the conditional independence assumption (3). To conserve space, we, however, do not analyze this in much detail since most applied work either uses a sharp design (where (3) is meaningless) or otherwise refers to (4). Consider an extension of (3) by including covariates X:

$$Y_i^1 - Y_i^0 \perp\!\!\!\perp D_i | X_i, Z_i \qquad \text{for } Z_i \text{ near } z_0. \tag{10}$$

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

$$E\left[Y^1 - Y^0 | Z = z_0\right] = \int \frac{m^+(x, z_0) - m^-(x, z_0)}{d^+(x, z_0) - d^-(x, z_0)} \cdot \frac{f^+(x|z_0) + f^-(x|z_0)}{2} dx. \tag{11}$$

This expression differs from (7) and (9) in that it is an integral of a ratio and not a ratio of integrals. The results derived in Section 3 therefore do not apply to (11). In addition, expression (11) may be difficult to estimate in small samples as the denominator can be close to zero for some values of x.[12]

Instead of using (10), one might be willing to strengthen the latter assumption to

$$Y_i^1, Y_i^0 \perp\!\!\!\perp D_i | X_i, Z_i \qquad \text{for } Z_i \text{ near } z_0. \tag{12}$$

This permits identifying the treatment effect as

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

[12] This problem is of much less concern for estimators of (7) and (9) as those are 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.

where $E[Y|D,X,Z=z_0]$ can be estimated by a combination of the left and right hand side limits. This approach does not exclusively rely on comparing observations across the threshold but also uses variation within either side of the threshold. The estimand has a similar structure as (7) and (9) and the estimation properties derived later could easily be extended to this case.

3 Estimation

A straightforward estimator of (7) is

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

where $\hat{m}$ and $\hat{d}$ are nonparametric estimators and $K_h(u)$ a kernel function.[13]

For practical convenience, we will mostly work with product kernel functions below. Product kernel functions also have the advantage that one can easily incorporate discrete X in the spirit of Racine and Li (2004). Define κ and $\bar{\kappa}$ as *univariate* kernel functions, where κ is a second-order kernel (assumed to be *symmetric* and *integrating to one*) and $\bar{\kappa}$ is a kernel of order $\lambda \geq 2$. The following kernel constants for κ 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 $\ddot{\mu}_l = \int_{0}^{\infty} u^l \kappa^2(u) du$.[14] The kernel constants for $\bar{\kappa}$ are defined as $\eta_l = \int_{-\infty}^{\infty} u^l \bar{\kappa}(u) du$ and $\dot{\eta}_l = \int_{-\infty}^{\infty} u^l \bar{\kappa}^2(u) du$.[15]

We will consider two different choices for $K_h(u)$ in (13). The conventional choice would be to use a positive (i.e. second order) and symmetric kernel

$$K_h(u) = \frac{1}{h}\kappa(u). \tag{14}$$

However, as shown below, the use of this 'naive' kernel function (14) leads at best to a convergence rate of $n^{-\frac{1}{3}}$ of (13).

[13] For the sharp design (9) the estimator simplifies to $\frac{\sum\left(\hat{m}^+(X_i,z_0) - \hat{m}^-(X_i,z_0)\ \right) \cdot K_h\left(\frac{Z_i - z_0}{h}\right)}{\sum K_h\left(\frac{Z_i - z_0}{h}\right)}$.

[14] 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 $\mu_0 = 1$, $\mu_1 = \mu_3 = \mu_5 = 0$, $\mu_2 = 0.2$, $\mu_4 = 6/70$, $\bar{\mu}_0 = 0.5$, $\bar{\mu}_1 = 3/16$, $\bar{\mu}_2 = 0.1$, $\bar{\mu}_3 = 1/16$, $\bar{\mu}_4 = 3/70$.

[15] The kernel function $\bar{\kappa}$ being of order λ means that $\eta_0 = 1$ and $\eta_l = 0$ for $0 < l < \lambda$ and $\eta_\lambda \neq 0$.

As an alternative, we consider a *boundary* kernel

$$K_h(u) = (\bar{\mu}_2 - \bar{\mu}_1 |u|) \cdot \frac{1}{h}\kappa(u) \tag{15}$$

in (13), and we will see that this leads to a convergence rate of $n^{-\frac{2}{5}}$ of (13), i.e. the rate of univariate nonparametric regression. This is achieved through smoothing with implicit double boundary correction.[16]

In the following, we will refer to estimator (13) with kernel function (14) as $\hat{\gamma}_{naive}$. Estimator (13) with kernel function (15) is denoted as $\hat{\gamma}_{RDD}$. Because of the asymptotic properties derived below we recommend the use of $\hat{\gamma}_{RDD}$.

In either case, estimation proceeds in two steps and requires nonparametric *first step* estimates of m^+, m^-, d^+ and d^-.[17] 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 local linear regression as the value of a that solves

$$\underset{a,b,c}{\arg\min} \sum_{j=1}^{n} \left(Y_j - a - b(Z_j - z_0) - c'(X_j - x)\right)^2 \cdot K_j I_j^+ \tag{16}$$

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

$$K_j = K_j(x, z_0) = \kappa\left(\frac{Z_j - z_0}{h_z}\right) \cdot \prod_{l=1}^{L} \bar{\kappa}\left(\frac{X_{jl} - x_l}{h_x}\right), \tag{17}$$

where L is the dimension of X, and κ and $\bar{\kappa}$ are univariate kernel functions with κ a second-order kernel and $\bar{\kappa}$ a kernel of order $\lambda \geq 2$.

A result derived later will require higher-order kernels (i.e. $\lambda > 2$) 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. Note that three different bandwidths

[16] See e.g. Jones (1993) or Jones and Foster (1996) for similar boundary kernels, or Gasser and Müller (1979), Gasser, Müller, and Mammitzsch (1985), Müller (1991) or Tenreiro (2013) for a more general discussion on various forms of boundary kernels or boundary corrections including the derivation of optimal boundary kernels for density estimation, estimation of distribution functions or estimation of nonparametric curves etc.

[17] In the sharp design (9), d^+ and d^- are not estimated but set to 1 and 0, respectively.

h_z, h_x, h are used. h is the bandwidth in the matching estimator (13) 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 in (16), which uses observations only to the right or only to the left of the threshold. We need some smoothness assumptions as well as conditions on the bandwidth values.[18]

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 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 (17) are symmetric, bounded, Lipschitz, integrate to one and are 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 $\lim_{\varepsilon \to 0} E\left[(Y - m^+(X,Z))^2 | X, Z = z + \varepsilon\right]$ and $\lim_{\varepsilon \to 0} E\left[(Y - m^-(X,Z))^2 | X, Z = z - \varepsilon\right]$ exist at z_0.

[18] Note that the above setup includes global linear regression for the special case where all bandwidth values are set to infinity. In this case, the estimator (16) corresponds to a linear regression using only data points to the right; and analogously on the left hand side. While a bandwidth value of infinity minimizes variance it could lead to a large bias if the true regression curve is non-linear. The estimator analyzed below seeks to minimize mean squared error, i.e. the sum of the squared bias and variance.

3.1 Properties of $\hat{\gamma}_{naive}$

With these preliminaries we consider the properties of $\hat{\gamma}_{naive}$ and $\hat{\gamma}_{RDD}$. The estimator $\hat{\gamma}_{naive}$ is, in essence, 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 one for estimating (7), it has worse statistical properties than $\hat{\gamma}_{RDD}$ in the sense that it achieves a lower rate of convergence. This is due to the missing boundary correction in the second step.

Proposition 2 (Asymptotic properties of $\hat{\gamma}_{naive}$) *Under Assumptions 1, 2 and 3, the bias and variance terms of $\hat{\gamma}_{naive}$, which is the estimator (13) with kernel function (14), are of order*

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

For the sharp design (9), the same results apply. 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 for $\hat{\gamma}_{naive}$ by appropriate bandwidth choices is $n^{-\frac{1}{3}}$.[19] It is straightforward to show asymptotic normality for this estimator, but the (first order) approximation may not be very useful in practice 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. Hence, such an approximation might not be very accurate in small samples. A more useful approximation can be obtained by retaining also the lower order terms. However, it seems more promising to use $\hat{\gamma}_{RDD}$ instead.

[19] In the special case where the density is continuous, i.e. $f^-(x|z_0) = f^+(x|z_0)$, the bias term with respect to the bandwidth h is $O(h^2)$ such that a convergence rate of $n^{-\frac{2}{5}}$ is possible. In this paper, we focus on the estimator proposed in the next section, though, because it can obtain $n^{-\frac{2}{5}}$ rate irrespective of whether the density is continuous or not.

3.2 Properties of $\hat{\gamma}_{RDD}$

The estimator $\hat{\gamma}_{RDD}$ is based on (13), but uses the boundary kernel (15) in the second smoothing step, instead of (14). It thereby attains the convergence rate of a one dimensional nonparametric regression estimator, irrespective of the dimension of X. It thus obtains the fastest convergence rate possible and is not affected by a curse of dimensionality. This is achieved by smoothing over all regressors X and by an implicit boundary adaptation with respect to Z. (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.)

We derive the asymptotic distribution of this estimator and show that the asymptotic variance becomes smaller the more covariates X are included. For the 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 the bias and standard deviation of the estimator converge at rate $n^{-\frac{2}{5}}$ to zero, i.e. 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

$$-\frac{4}{5L} < \zeta \leq -\frac{2}{5\lambda}. \tag{18}$$

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. Therefore, a necessary condition for Assumptions

2 and 3 to hold jointly 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 the bias and variance to converge at the univariate nonparametric rate, which is summarized in the following theorem.

Theorem 3 (Asymptotic distribution of $\hat{\gamma}_{RDD}$) *a)Under Assumptions 1 and 2, the bias and variance terms of $\hat{\gamma}_{RDD}$, which is the estimator (13) with kernel function (15), 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

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

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

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

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

For the sharp design (9), the same results are obtained but the formulae are simpler. d^+ and d^- are not estimated but set to 1 and 0, respectively. This implies that $\Gamma = 1$ and the terms σ_D^{2+}, σ_D^{2-}, σ_{YD}^{2+}, σ_{YD}^{2-} and all derivatives of $d^+(x, z_0)$ and $d^-(x, z_0)$ are zero.

Note that Assumption 3 is stronger than needed for the results of Theorem 3. For obtaining $n^{-\frac{2}{5}}$ convergence weaker rate conditions would suffice. In other words, it would not be needed that the ratios of the bandwidths converge to a well defined limit point. Assumption 3 permits obtaining concise and explicit expressions for bias and variance, though. We also see that undersmoothing is permitted: For a choice of $r = 0$ in Assumption 3, the limit bias term is zero, i.e. $\mathcal{B}_{RDD} = 0$. Such undersmoothing is convenient, e.g. for developing test statistics.[20]

Part (18) 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 (16). 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 (17) 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

[20] We thank a referee for pointing this out.

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

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

It remains to be discussed how the bandwidth values h, h_z and h_x should be chosen in practice. It is beyond the scope of this paper to develop a data driven bandwidth selector, and we therefore limit ourselves to a procedure that is rate optimal, i.e. satisfies Assumptions 2 and 3 as n increases to infinity. The first part of Assumption 3 suggests to choose h proportional to $n^{-\frac{1}{5}}$, which corresponds to the rate for univariate nonparametric regression. A simple procedure is to choose h via (least squares) cross-validation with respect to a nonparametric regression of Y on Z (outside of a neighbourhood around z_0), which is known to provide a bandwidth that converges at the desired rate.[21]

With an estimate for h, we can choose $h_z = h$ which is permitted by Assumptions 2 and 3. If X contains at most three continuous regressors, we can also choose $h_x = h$. On the other hand, if $L \geq 4$, then h_x should converge at a slower rate than h and h_z. Assumptions 2 and 3 give us some leeway in the exact choice of h_x. If we would like to make the bias small (for reasons discussed in the next section), we would choose the lower bound of (18) to set $h_x = c_1 \cdot n^{-\frac{4}{5L}+\delta}$ for a small positive δ and some positive constant c_1. This contrasts with the choice for h which is given as $h = c_2 \cdot n^{-\frac{1}{5}}$. We do not know the optimal c_1 and c_2, but since we only aim for a rate optimal choice, we can set $c_1 = c_2$ to obtain $h_x = c_1 \cdot n^{-\frac{4}{5L}+\delta} = c_1 n^{-\frac{4}{5L}+\delta} \cdot n^{\frac{1}{5}} n^{-\frac{1}{5}}$ such that

$$h_x = n^{\frac{1-4/L+5\delta}{5}} \cdot h.$$

We can thus use the bandwidth h obtained via cross-validation and multiply it with $n^{\frac{1-4/L+5\delta}{5}}$

[21] At the same time it is known that the bandwidth obtained by cross-validation converges only very slowly to the true optimal bandwidth. Nevertheless, many applied researchers proceed by using the bandwidth obtained from cross-validation and then examine the sensitivity of the final estimation results to changes in the bandwidth values by re-estimating with various multiples and/or fractions of the original bandwidth values.

for some small δ to obtain the (larger) bandwidth value for h_x. Having estimated $\hat{\gamma}_{RDD}$ with these bandwidths, one would usually examine the robustness of the results to the bandwidths values.

3.3 Variance reduction through the use of control variables

In most of the discussion so far it was permitted that $f(x|z)$ is discontinuous at z_0 such that controlling for X allows reducing bias. In the case where $f(x|z)$ is continuous, controlling for X is still helpful: It can reduce the variance of the estimator, which is shown in the following theorem. Suppose that the covariates are identically distributed on both sides of the threshold (i.e. $f(x|z)$ is continuous) such that γ is identified with and without controlling for any X. In this case one could use $\hat{\gamma}_{RDD}$ with X being the empty set. This estimator is henceforth denoted as $\hat{\gamma}_{no\,X}$. Alternatively, one could use a set of control variables X in the estimator, which we denote as $\hat{\gamma}_{RDD}$ as before. Suppose that both estimators are consistent for γ. As shown below, $\hat{\gamma}_{no\,X}$ generally has a *larger* asymptotic variance than $\hat{\gamma}_{RDD}$.[22] On the other hand, an ordering of squared biases seems impossible under general conditions. However, by Assumption 3 we can set $r = 0$, i.e. choose a bandwidth sequence such that the ratio of the squared bias to variance converges to zero. Such undersmoothing implies that the asymptotic bias $\mathcal{B}_{RDD}$ is zero and the mean-squared-error is thus identical to $\mathcal{V}_{RDD}$. With such undersmoothing, we only need to analyze the asymptotic variance. As outlined below, there are precision gains by controlling for X even if the RDD estimator would be consistent without covariates.

For stating Theorem 4 in a concise way, some further 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.[23] Define the variance of $w^+(X, z_0)$ as $V^+ = \int \left\{w^+(x, z_0) - w^+(z_0)\right\}^2 f(x|z_0)dx$.

[22] We would like to point out that the result in Theorem 4 only refers to the variance. While we find that covariates reduce variance, we do not have a corresponding result for the bias. Hence, in certain situations, asymptotic bias could possibly increase and we, therefore, cannot rule out that the inclusion of covariates X in certain cases could even increase MSE if in such situations an increase in squared bias is larger than the decrease of variance due to the inclusion of X.

[23] This also contains the sharp design (9) as a special case, where $w^+(X,z) = \lim_{\varepsilon \to 0} E\left[Y - \gamma | X, Z = z + \varepsilon\right]$ and $w^-(X,z) = \lim_{\varepsilon \to 0} E\left[Y | X, Z = z - \varepsilon\right]$.

Define $w^-(X,z)$, $w^-(z)$ and V^- analogously as the left limits. Theorem 4 shows that there is a reduction in variance if $V^+ \neq 0$ and/or $V^- \neq 0$.

To gain some intuition, note that V^+ is the variance of the conditional expectation of Y given X plus the variance of the conditional expectation of γD given X minus the covariance between these two terms. Hence, V^+ is usually 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.[24] Define further the covariance C as $\int (w^+(x,z_0) - w^+(z_0))\,(w^-(x,z_0) - w^-(z_0))\,f(x|z_0)dx$. For the case where V^+ and V^- are both non-zero, we define the correlation coefficient $R = \frac{C}{\sqrt{V^+V^-}}$. Now, we can state the result in terms of the variances and the correlation coefficient, which also depends on the bandwidth sequences. The variance of $\hat{\gamma}_{RDD}$ is a function of 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 $h = h_z$.[25] This implies $r_z = 1$ in Assumption 3. Using this notation, the difference in the asymptotic variances can be written as

$$\mathcal{V}_{RDD} - \mathcal{V}_{no\,X} = \left\{ \frac{r_z - 2}{2} V^+ + \frac{r_z - 2}{2} V^- - r_z C \right\} \left(\frac{\bar{\mu}_2^2 \ddot{\mu}_0 - 2\bar{\mu}_2 \bar{\mu}_1 \ddot{\mu}_1 + \bar{\mu}_1^2 \ddot{\mu}_2}{\Gamma^2 \tilde{\mu}^2 f(z_0) r_z} \right)$$

or, if V^+ and V^- are both non-zero, as $= \left\{ \frac{r_z-2}{2} V^+ + \frac{r_z-2}{2} V^- - r_z R \sqrt{V^+V^-} \right\} \left(\frac{\bar{\mu}_2^2 \ddot{\mu}_0 - 2\bar{\mu}_2 \bar{\mu}_1 \ddot{\mu}_1 + \bar{\mu}_1^2 \ddot{\mu}_2}{\Gamma^2 \tilde{\mu}^2 f(z_0) r_z} \right)$, as derived in the appendix. This implies the following:

Theorem 4 *Let $\hat{\gamma}_{RDD}$ be the estimator (13) with kernel function (15) 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

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

(b) Under any of the following conditions

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

[24] This discussion excludes the unreasonable case where it predicts both but not $Y - \gamma D$.

[25] 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.

- *if* $V^+ = 0$ *and* $V^- \neq 0$ *or vice versa and* $r_z < 2$

- *or if* $V^+ \neq 0$ *and* $V^- \neq 0$ *and* $R \geq 0$ *and* $r_z < 2$

- *or if* $V^+ \neq 0$ *and* $V^- \neq 0$ *and* $-1 < R < 0$ *and* $r_z < 2\frac{1+R}{1-R^2}$.

- *or if* $V^+ \neq 0$ *and* $V^- \neq 0$ *and* $R = -1$ *and* $r_z < 1$.

Hence, if, in case (a) of Theorem 4, where 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* 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.[26] 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, however, one would rather expect a positive correlation.[27][28]

4 Simulations

This section presents a simulation study in order to investigate the finite sample performance of the suggested method in the context of the sharp and fuzzy RDD. Starting with the former,

[26] In the sharp design (9), X cannot have predictive power for D (conditional on Z), hence predictive power for Y is needed.

[27] $\hat{\gamma}_{RDD}$ has a smaller variance than $\hat{\gamma}_{no\,X}$ as it exploits the available information more effectively. Consider, for simplicity, the sharp design. $\hat{\gamma}_{no\,X}$ estimates the conditional mean of Y left and right of the threshold. In terms of iterated expectations, the left limit of the mean of Y at the threshold could be estimated as the left limit of the mean of Y conditional on X averaged out with respect to the distribution of X, using only data points to the left of the threshold. In contrast, $\hat{\gamma}_{RDD}$ estimates the left limit of the mean of Y conditional on X, but then takes averages with respect to the distribution of X in the neigbourhood of z_0. In the case where the distribution of X is continuous at z_0, i.e. $f^+(X, z_0) = f^-(X, z_0)$, the estimator $\hat{\gamma}_{RDD}$ uses the data points X_i in the left *and* in the right neigbourhood of z_0 in order to estimate $f(X, z_0)$, whereas $\hat{\gamma}_{no\,X}$ uses only the data on one side of the threshold. This implies that $\hat{\gamma}_{RDD}$ uses more information in the estimation of the empirical distribution function $F(X, z_0)$, which leads to the variance reductions in Theorem 4.

[28] Theorem 4 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 specific covariates for eliminating bias with adding further covariates to reduce variance. The more variables are included in $\mathbf{X}$ the smaller the variance will be.

we consider the following data generating process (DGP):

$$
\begin{aligned}
&Z, U, V, W \sim \mathcal{N}(0,1) \text{ independently of each other,} \qquad (19)\\
&D = I\{Z > 0\}, \quad X_1 = \alpha D + 0.5U, \quad X_2 = \alpha D + 0.5V,\\
&Y = D + 0.5Z - 0.25DZ + 0.25Z^2 + \beta(X_1 + X_2) + \frac{\beta}{2}(X_1^2 + X_2^2) + W.
\end{aligned}
$$

Both the running variable Z and the unobservables U, V, W, which affect the covariates X_1, X_2 and the outcome Y, respectively, are standard normally distributed. The parameter α reflects the strength of the association between the distributions of X_1, X_2 and the treatment state D. β determines the impact of X_1, X_2 and their higher order terms on Y. In the simulations, we consider various combinations of α and β. First, we set $\alpha = 0$ and $\beta = 0.4$ such that the covariates affect the outcome, but are balanced around the threshold. In this case, controlling for $X = (X_1, X_2)$ is not necessary for the consistency of RDD, but might reduce the variance. Second, we set $\alpha = 0.2$ and $\beta = 0.4$, implying that the distribution of X differs across treatment states at the threshold and that X affects Y.

We run 1000 simulations and consider sample sizes of $n = 1000$ and 4000 to analyse RDD estimation based on the boundary kernel $\hat{\gamma}_{RDD}$, see (15). Least squares cross-validation (CV) is used to select the bandwidths for the estimation of $m^+(x, z)$ and $m^-(x, z)$ (using local linear regression) as well as $K_h(u)$ required in (13),[29] based on the 'np' package for the statistical software 'R' by Hayfield and Racine (2008). In addition, we also make use of undersmoothing and oversmoothing by taking half or twice the CV bandwidth, respectively (CV/2, 2CV).[30]

We compare our method to conventional RDD estimation without covariates as implemented in the 'rdd' package for 'R' by Dimmery (2016), which is based on a local linear regression of Y on Z. We consider several bandwidths choices, namely the values picked by the CV procedure for $\hat{\gamma}_{RDD}$; the method of Imbens and Kalyanaraman (2012) (IK) for optimal bandwidth selection in RDD; the robust inference approach of Calonico, Cattaneo,

[29] For $m^+(X, Z)$ and $m^-(X, Z)$, CV only uses treated and non-treated observations, respectively.

[30] We also considered a local cross-validation procedure that only used observations with values of the running variable not smaller than its median among observations below the threshold and not larger than its median among observations above the threshold, see Ludwig and Miller (2007). For 'CV' and '2CV', results were similar to those reported in Tables 1 and 2. Results available upon request.

Table 1: Simulations - sharp RDD

	$\hat{\gamma}_{RDD}$			RDD without X				$\hat{\gamma}_{RDD}$			RDD without X			
bandwidth	CV	CV/2	2CV	CV	IK	CCT	LM	CV	CV/2	2CV	CV	IK	CCT	LM
$\alpha = 0, \beta = 0.4$	n=1000							n=4000						
bias	0.00	-0.00	0.00	-0.00	0.01	0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.00	-0.01	-0.00
sdev	0.15	0.27	0.13	0.43	0.20	0.22	0.15	0.09	0.12	0.09	0.27	0.10	0.10	0.08
rmse	0.15	0.27	0.13	0.43	0.20	0.22	0.15	0.09	0.12	0.09	0.27	0.10	0.10	0.08
$\alpha = 0.2, \beta = 0.4$	n=1000							n=4000						
bias	-0.00	-0.00	-0.01	0.18	0.18	0.18	0.17	-0.00	-0.00	-0.01	0.16	0.17	0.17	0.17
sdev	0.17	0.27	0.14	0.45	0.20	0.22	0.15	0.09	0.13	0.09	0.30	0.10	0.10	0.08
rmse	0.17	0.27	0.14	0.48	0.27	0.28	0.23	0.09	0.13	0.09	0.34	0.20	0.20	0.19

Note: 'CV', 'CV/2', '2CV' stands for bandwidth selection based on least squares cross-validation, as well as twice and half that value. 'IK' is the optimal Imbens-Kalyanaraman (2012) bandwidth. 'CCT' is the robust inference approach of Calonico, Cattaneo, and Titiunik (2014) (CCT). 'LM' is the local cross-validation approach of Ludwig and Miller (2007) based on the median values of the running variable above and below the threshold. 'bias', 'sdev', and 'rmse' report the bias, standard deviation, and root mean squared error of the respective method.

and Titiunik (2014) (CCT) as implemented as default option in the 'rdrobust' package for 'R' by Calonico, Cattaneo, and Titiunik (2015); and the local cross-validation approach of Ludwig and Miller (2007) (LM) based on the median values of the running variable above and below the threshold. In all estimations, the Epanechnikov kernel is used.

Table 1 reports the bias, standard deviation, and root mean squared error (RMSE) of the estimators for various choices of α, β in the sharp RDD. When setting $\alpha = 0, \beta = 0.4$, all procedures are unbiased as expected. Under either sample size, $\hat{\gamma}_{RDD}$ outperforms RDD without X in terms of precision when using the same CV bandwidth for both estimators. Furthermore, $\hat{\gamma}_{RDD}$ with CV is in most cases also more precise than RDD without X based on the IK, CCT, and LM bandwidths.[31] As expected, a smaller bandwidth (CV/2) increases

[31] Under $n = 1000, \alpha = 0, \beta = 0.4$, the means (standard deviations) of the CV, IK, CCT, and LM bandwidths for Z are 0.16 (0.06), 0.84 (0.29), 0.66 (0.11), 1.58 (0.51), respectively. The means and standard deviations are

the standard deviation of $\hat{\gamma}_{RDD}$, while a larger bandwidth (2CV) slightly decreases it. For $n = 4000$, however, the differences in precision are quite moderate for various bandwidth choices.

When setting $\alpha = 0.2$ and $\beta = 0.4$, the biases of $\hat{\gamma}_{RDD}$ are again close to zero, while this is no longer the case for RDD without X. For $n = 1000$, $\hat{\gamma}_{RDD}$ with CV and 2CV dominates any RDD without X in terms of bias, standard deviation, and root mean squared error (RMSE), while $\hat{\gamma}_{RDD}$ with CV/2 is less precise. Under $n = 4000$, all three versions of $\hat{\gamma}_{RDD}$ have a considerably smaller RMSE than any RDD without X.

Table 2: Simulations - fuzzy RDD

	$\hat{\gamma}_{RDD}$			RDD without X				$\hat{\gamma}_{RDD}$			RDD without X			
bandwidth	CV	CV/2	2CV	CV	IK	CCT	LM	CV	CV/2	2CV	CV	IK	CCT	LM
$\alpha = 0, \beta = 0.4$	n=1000							n=4000						
bias	-0.01	0.00	-0.02	-0.05	-0.02	-0.01	-0.01	0.01	-0.00	0.01	-0.01	0.00	-0.01	-0.01
sdev	0.27	0.42	0.22	0.76	0.34	0.34	0.24	0.16	0.18	0.14	0.34	0.16	0.16	0.12
rmse	0.27	0.42	0.22	0.76	0.34	0.34	0.24	0.16	0.18	0.14	0.34	0.16	0.16	0.12
$\alpha = 0.2, \beta = 0.4$	n=1000							n=4000						
bias	-0.01	-0.00	-0.03	0.25	0.27	0.27	0.27	0.01	0.01	0.00	0.25	0.28	0.27	0.27
sdev	0.28	0.52	0.23	0.67	0.33	0.34	0.23	0.15	0.20	0.15	0.39	0.16	0.16	0.12
rmse	0.28	0.52	0.23	0.72	0.43	0.43	0.36	0.15	0.20	0.15	0.46	0.32	0.31	0.30

Note: 'CV', 'CV/2', '2CV' stands for bandwidth selection based on least squares cross-validation, as well as twice and half that value. 'IK' is the optimal Imbens-Kalyanaraman (2012) bandwidth. 'CCT' is the robust inference approach of Calonico, Cattaneo, and Titiunik (2014) (CCT). 'LM' is the local cross-validation approach of Ludwig and Miller (2007) based on the median values of the running variable above and below the threshold. 'bias', 'sdev', and 'rmse' report the bias, standard deviation, and root mean squared error of the respective method.

Secondly, we consider the case of a fuzzy RDD. We modify the DGP by replacing $D = I\{Z > 0\}$ in (19) with $D = I\{-1 + 2I\{Z > 0\} + 0.5U + Q > 0\}$, with $Q \sim \mathcal{N}(0, 1)$ independently of any other variable. D is now endogenous even at the threshold due to U entering both

very similar under $n = 1000, \alpha = 0.2, \beta = 0.4$.

the treatment and outcome equation. The bandwidths used for the estimation of $d^+(x,z)$ and $d^-(x,z)$ required for the fuzzy RDD method are selected in an analogous way as for $m^+(x,z)$ and $m^-(x,z)$. We also consider fuzzy RDD estimation without covariates based on Dimmery (2016) with CV, IK, CCT, and LM bandwidth choices, repectively.[32] The results are reported in Table 2 and show a qualitatively similar pattern as for the sharp RDD. However, standard errors are generally larger as estimation is based on the compliers only, which by the definition of the DGP make up for about 65% of the population.

5 Application

As an empirical illustration of our method we use data from Lalive (2008), who studies a labor market program introduced in June 1988 that extended the maximum duration of unemployment benefits from 30 to 209 weeks for job seekers aged 50 or older in certain regions of Austria under particular conditions. This suggests the use of a sharp RDD for assessing the program's effect on labor market outcomes such as unemployment duration. The treatment is defined based on the age threshold of 50. As acknowledged by Lalive (2008), however, a concern is that employees and companies could manipulate age at entry into unemployment, for example, by postponing a layoff in a way that the age requirement is just satisfied. This is a common concern in many applications. If such manipulations are selective with respect to employee characteristics that also affect labor market outcomes, conventional RDD without covariates fails to identify the effect of the program due to confounding related to an imbalance of the characteristics around the threshold. In contrast, our method remains consistent if all labor market relevant characteristics are plausibly observed in the data. As a word of caution, however, we would like to point out that this cannot be taken for granted in our application. For instance, unobserved individual characteristics like motivation, (dis-)utility from work, and self-confidence might predict both manipulation and labor market success. To consistently estimate the program effect by our method, it is required that these factors do not

[32] Under $n = 1000, \alpha = 0, \beta = 0.4$, the means (standard deviations) of the CV, IK, CCT, and LM bandwidths for Z are 0.23 (0.07), 0.84 (0.29), 0.66 (0.11), 1.73 (0.59), respectively. The means and standard deviations are very similar under $n = 1000, \alpha = 0.2, \beta = 0.4$.

entail confounding conditional on the socio-economic and employment-related characteristics available in the data (see the discussion below).

Table 3: Covariate sample means and balance tests at the threshold

		IK		IK/2	
	sample mean	difference	p-value	difference	p-value
married (binary)	0.75	0.16	0.00	0.16	0.01
single (binary)	0.09	-0.05	0.05	-0.05	0.13
education: medium (binary)	0.22	0.02	0.51	-0.00	0.99
education: high (binary)	0.08	0.04	0.03	0.04	0.14
foreign (binary)	0.02	0.01	0.37	0.01	0.59
replacement rate	0.44	-0.01	0.01	-0.01	0.03
log wage in last job	6.15	0.12	0.00	0.18	0.00
actual to potential work experience	0.89	0.02	0.06	0.00	0.77
white collar worker (binary)	0.32	0.16	0.00	0.15	0.00
industry: agriculture (binary)	0.02	-0.01	0.65	0.02	0.20
industry: utilities (binary)	0.00	0.00	0.32	0.00	0.32
industry: food (binary)	0.05	-0.02	0.31	-0.03	0.44
industry: textiles (binary)	0.12	0.02	0.54	-0.03	0.38
industry: wood (binary)	0.03	0.00	0.82	0.02	0.20
industry: machines (binary)	0.08	0.04	0.05	0.06	0.06
industry: other manufactoring (binary)	0.11	0.03	0.31	0.04	0.33
industry: construction (binary)	0.03	0.03	0.03	0.04	0.02
industry: tourism (binary)	0.32	-0.03	0.46	-0.02	0.73
industry: traffic (binary)	0.02	-0.03	0.07	-0.02	0.37
industry: services (binary)	0.17	-0.05	0.14	-0.03	0.50

Note: 'IK', 'IK/2' denote the optimal Imbens-Kalyanaraman (2012) bandwidth and half that value in an RDD estimation when using each of the covariates as outcome. P-values are based on analytic standard errors and account for clustering of age (measured in months).

Our analysis makes use of the Austrian social security database, which includes information on job seekers (age, employment, unemployment and earnings history) and the employers (region and industry), and the Austrian unemployment register, which contains information on the place of residence and socio-economic characteristics. The universe of inflows into

unemployment between 1986 and 1995 is covered, and the inflow sample can be followed up until the end of 1998. We refer to Lalive (2008) for a description of sample adjustments made to the data set. Specifically, we consider the female subsample in the age bracket 46 to 53 years living in a region where the program had been introduced, consisting of 5659 observations. The outcome variable Y is unemployment duration, measured as weeks registered at the unemployment office. The running variable Z is distance to the age threshold of 50, measured in months divided by 12. Table 3 reports sample means and balancing tests at the threshold for potentially labor market relevant characteristics, which serve as X. The tests are based on running RDD estimations with the elements in X as outcome variables using the 'rdd' package, which performs local linear regression around the threshold. Estimates, standard errors, and p-values are reported for the IK bandwidth and half of it. Indeed, several covariates are imbalanced around the threshold, which concerns among others marital status, wage in the last job, and being a white collar worker.[33] The results therefore suggest that observations slightly above the age threshold have somewhat more favorable labor market relevant characteristics than those slightly below.

Our RDD estimator derived from equation (7) controls for differences in X by giving appropriate weights to each of these characteristics, according to their distribution about the thresh-old. Consider, for example, the variable marital status, which is significantly different in Table 3. On average, 75% of the observations in the sample are married, but the (conditional) probability of being married is discontinuous at the threshold: The nonparametric estimates of the probability from the left and right are 63.7% and 79.9%, respectively. In a symmetric neighbourhood about the threshold, the probability of being married is thus 71.8%. Our method proceeds by estimating the outcome unemployment duration for married women left and right of the threshold and multiplying with a weight of 0.718. An analogous approach applies to unmarried women using a weight of 0.282.

[33] To control the family-wise error rate of multiple testing in Table 3, one may apply the (conservative) Bonferroni correction: divide the nominal level of significance by the number of tested covariates (in our case 20) and reject an individual null hypothesis of covariate balance if the corresponding p-value is even lower. For log wage in last job and white collar worker, the null hypothesis is rejected under either bandwidth at the nominal 5% level of significance.

Hence, a weighted average with respect to the fraction of married women in a symmetric neighbourhood about the threshold is taken. This removes the discontinuity in marital status: The 63.7% married women to the left are up-weighted with 0.718/0.637, while the 79.9% married women to the right are down-weighted with 0.718/0.799. Accordingly, the 36.3% unmarried women to the left are down-weighted with 0.282/0.363, while those 20.1% to the right are up-weighted with 0.282/0.201. In contrast, RDD estimation not controlling for X compares the unemployment duration left and right of the threshold without weighting, thereby ignoring that there are for instance fewer married women to the left than to the right of the threshold.

Table 4 presents the results for $\hat{\gamma}_{RDD}$ when using cross-validation for the bandwidth selection of h_x, h_z in the first step estimation of m^+ and m^-. Different from the simulations in Section 4, however, the covariates now contain both continuous and discrete elements. We therefore apply the method of Racine and Li (2004), which allows for both continuous and discrete regressors by means of product kernels and is implemented in the 'np' package of Hayfield and Racine (2008). We use the Epanechnikov, Wang and van Ryzin (1981), and Aitchison and Aitken (1976) kernel functions for continuous, ordered discrete, and unordered discrete covariates, respectively. We consider several choices for bandwidth h in the Epanechnikov-based boundary kernel function for the running variable in (13): 0.1, 0.2,..., 0.5. We also compare the results to RDD regression without covariates based on the 'rdd' package with the same bandwidth choice h. The standard errors of any method are based on nonparametrically bootstrapping the respective estimates 999 times, i.e. randomly resampling the original data with replacement and applying the estimators to the bootstrap samples. The $\hat{\gamma}_{RDD}$ estimates point to a substantial increase in unemployment duration by about 110 weeks.

The results are highly significant, as the standard errors of roughly 4 weeks are quite moderate. When using RDD without X, both the effect of about 140 weeks and the standard error of about 10 weeks are substantially higher. For each bandwidth value considered, the estimates are statistically significantly different between the methods (at the 5% level based on bootstrapping the differences in the estimates 999 times). This indicates that there might be some confounding due to observed covariates. Also the effects reported in Table 3 columns (3)

Table 4: Effect estimates

	$\hat{\gamma}_{RDD}$					RDD without X				
Bandwidth h	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
treatment effect	115.31	112.74	110.76	109.71	108.64	134.25	143.67	141.41	137.99	132.55
standard error	4.23	4.09	4.14	4.03	4.41	9.72	12.49	9.90	8.45	8.03
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: The bandwidths h_x, h_z for the first step estimates of m^+ and m^- entering $\hat{\gamma}_{RDD}$ (see Section 3) are picked by least squares cross-validation. For bandwidth h on the running variable Z in $\hat{\gamma}_{RDD}$ and RDD without X, several values are considered as indicated in the table. Standard errors are based on bootstrapping the estimate 999 times. Sample size is 5659 observations. X includes the variables given in Table 3: marital status, education, migration status, replacement rate, log wage in last job, actual to potential work experience, white collar worker, and industry.

and (4) of Lalive (2008) when omitting X and either using a global RDD model with a higher order polynomial for the running variable or a local linear model with a very small bandwidth are somewhat higher than $\hat{\gamma}_{RDD}$ (122 to 126 weeks). In contrast, the effect of 103 weeks presented in column (6) of Table 3 in Lalive (2008) is based on linearly controlling for covariates. Our somewhat higher (and at the 5% level statistically significantly different) estimates (when bootstrapping the differences) are likely due to using a more flexible specification with respect to the association of Y and X.

6 Conclusion

In this paper, the *regression discontinuity* design (RDD) has been generalized to incorporate *covariates* X in a fully nonparametric way. Including covariates can reduce the variance and eliminate biases if X is discontinuously distributed at the threshold. 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 has been suggested. We investigated the finite sample

properties of our estimator in simulations and applied it to estimate the effect of age-dependent unemployment benefits on unemployment duration in Austrian labor market reform, where manipulation at the threshold is a potential concern.

References

Aitchison, J., and C. Aitken (1976): "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

Battistin, E., and E. Rettore (2008): "Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs," *Journal of Econometrics*, 142, 715–730.

Black, D., J. Galdo, and J. Smith (2007): "Evaluating the regression discontinuity design using experimental data," *mimeo*, University of Michigan, USA.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2016): "Regression Discontinuity Designs Using Covariates," *working paper, University of Michigan.*

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

——— (2015): "rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs," *R Journal*, 7, 38–51.

de Chaisemartin, C. (2016): "Tolerating defiance? Identification of treatment effects without monotonicity," *working paper, University of Warwick.*

Dimmery, D. (2016): "Package 'rdd'," *Manual for the statistical software 'R'.*

Dong, Y. (2014): "An Alternative Assumption to Identify LATE in Regression Discontinuity Designs," *Unpublished Manuscript, University of California Irvine.*

Eugster, B., R. Lalive, A. Steinhauer, and J. Zweimüller (2017): "Culture, Work Attitudes and Job Search: Evidence from the Swiss Language Border," *forthcoming in Journal of the European Economic Association.*

Fisher, R. (1935): *Design of Experiments.* Oliver and Boyd, Edinburgh.

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

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

Frölich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

Gasser, T., and H. Müller (1979): "Kernel estimation of regression functions," in *Smoothing techniques for curve estimation, Lecture Notes in Mathematics 757*, ed. by T. Gasser, and M. Rosenblatt, pp. 23–68. Springer, Berlin.

Gasser, T., H. Müller, and V. Mammitzsch (1985): "Kernels for nonparametric curve estimation," *Journal of the Royal Statistical Society Series B*, 47, 238–252.

Gelman, A., and G. Imbens (2016): "Why high-order polynomials should not be used in regression discontinuity designs," *working paper, Stanford University*.

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.

Hayfield, T., and J. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Imbens, G. (2001): "Some remarks on instrumental variables," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 17–42. Physica/Springer, Heidelberg.

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

Imbens, G., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 79, 933–959.

Imbens, G. W., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

Jones, M. (1993): "Simple boundary correction for kernel density estimation," *Statistics and Computing*, 3, 135–146.

Jones, M., and P. Foster (1996): "A simple nonnegative boundary correction method for kernel density estimation," *Statistica Sinica*, 6, 1005–1013.

Lalive, R. (2008): "How do extended benefits affect unemployment duration? A regression discontinuity approach," *Journal of Econometrics*, 142, 785 – 806.

Lee, D. (2008): "Randomized experiments from non-random selection in U.S. House elections," *Journal of Econometrics*, 142, 675–697.

Lee, D., and D. Card (2008): "Regression discontinuity inference with specification error," *Journal of Econometrics*, 142, 655–674.

Lee, D., and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

Ludwig, J., and D. L. Miller (2007): "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *The Quarterly Journal of Economics*, 122, 159–208.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142, 698–714.

Müller, H. (1991): "Smooth optimum kernel estimators near endpoints," *Biometrika*, 78, 521–530.

Newey, W. (1994): "Kernel estimation of partial means and a general variance estimator," *Econometric Theory*, 10, 233–253.

Neyman, J. (1923): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles.," *Statistical Science*, Reprint, 5, 463–480.

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

Racine, J., and Q. Li (2004): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130.

Rubin, D. (1978): "Bayesian inference for causal effects: the role of randomization," *Annals of Statistics*, pp. 34–58.

Tenreiro, C. (2013): "Boundary kernels for distribution function estimation," *REVSTAT Ű Statistical Journal*, 11, 169–190.

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

Urquiola, M., and E. Verhoogen (2009): "Class-Size Caps, Sorting, and the Regression-Discontinuity Design," *The American Economic Review*, 99, 179–215.

van der Klaauw, W. (2008): "Breaking the link between poverty and low student achievement: An evaluation of Title I," *Journal of Econometrics*, 142, 731–756.

Wang, M., and J. van Ryzin (1981): "A class of smooth estimators for discrete distributions," *Biometrika*, 68, 301–309.