Regression Discontinuity Designs Using Covariates*

Sebastian Calonico† Matias D. Cattaneo‡ Max H. Farrell§ Rocío Titiunik¶

March 31, 2016

Abstract

We study identification, estimation, and inference in Regression Discontinuity (RD) designs when additional covariates are included in the estimation. Standard RD estimation and inference is based on nonparametric local polynomial methods using two variables: the outcome variable and the running variable that determines treatment assignment. Applied researchers often include additional "pre-intervention" covariates in their specifications to increase efficiency. However, no results justifying covariate adjustment have been formally derived in the RD literature, leaving applied researchers with little practical guidance and leading to a proliferation of ad-hoc methods that may result in invalid estimation and inference. We examine the properties of a local polynomial estimator that incorporates discrete and continuous covariates in an additive separable, linear-in-parameters way and imposes a common (likely misspecified) covariate effect on both sides of the cutoff. Under intuitive, minimal assumptions, we show that this covariate-adjusted RD estimator remains consistent for the standard RD treatment effect, while also providing point estimation and inference improvements. In contrast, we show that estimating a specification with interactions between treatment status and the covariates leads to an estimator that is inconsistent in general. We present new asymptotic mean squared error expansions, optimal bandwidth choices, optimal point estimators, robust nonparametric inference procedures based on bias-correction techniques, and heteroskedasticity-consistent standard errors. Our results cover sharp, fuzzy, and kink RD designs, and we also discuss extensions to clustered data. Finally, we present two empirical illustrations where we find 5% to 10% reduction in confidence interval length, and an extensive simulation study. All methods are implemented in companion `R` and `Stata` software packages.

Keywords: regression discontinuity, covariate adjustment, causal inference, local polynomial methods, robust inference, bias correction, tuning parameter selection.

*We thank Stephane Bonhomme, David Drukker, Kosuke Imai, Michael Jansson, Lutz Kilian, Pat Kline, Xinwei Ma, Andres Santos, and Gonzalo Vazquez-Bare for thoughtful comments and suggestions. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1357561 and SES-1459931, and Titiunik gratefully acknowledges financial support from the National Science Foundation through grant SES-1357561.

†Department of Economics, University of Miami.

‡Department of Economics and Department of Statistics, University of Michigan.

§Booth School of Business, University of Chicago.

¶Department of Political Science, University of Michigan.

1 Introduction

The Regression Discontinuity (RD) design is widely used in Economics and other social, behavioral, biomedical, and statistical sciences. Within the causal inference framework, this design is considered among the most credible non-experimental strategies because it relies on relatively weak and easy-to-interpret nonparametric identifying assumptions, which permit flexible and robust estimation and inference for (local to the cutoff) treatment effects. The key feature of the design is the existence of a score, index, or running variable for each unit in the sample, which determines treatment assignment via hard-thresholding: all units whose score is above a known cutoff are offered treatment, while all units whose score is below this cutoff are not. Identification, estimation, and inference proceed by comparing the responses of units near the cutoff, taking those below (comparison group) as counterfactuals to those above (treatment or intention-to-treat group). For reviews see Imbens and Lemieux (2008), Lee and Lemieux (2010), Skovron and Titiunik (2016), and references therein.

The most common approach to nonparametric identification in RD designs relies on continuity assumptions. Under this approach, estimation of average treatment effects at the cutoff typically relies on nonparametric local polynomial methods, where the unknown (but assumed-smooth) regression function of the outcome variable given the score is flexibly approximated above and below the cutoff, and then these estimates are used to assess whether there is a discontinuity in levels, derivatives, or ratios thereof, at the cutoff. This discontinuity, if present, is understood as the average response to the treatment, intention-to-treat, treatment effect on the treated, or local average treatment effect, at the cutoff, depending on the specific setting and assumptions under consideration.

A natural estimation strategy is to fit separate local polynomial regressions of the outcome on the score above and below the cutoff, but in practice researchers often augment their models with "pre-intervention" covariates in addition to the running variable. The main motivation behind this practice is to increase the precision of the RD treatment effect estimator. In addition, covariates are sometimes added with the goal of improving the plausibility of the RD design, though this second motivation is much harder to justify because it rests on additional strong assumptions (see Remark 1 below). The practice of including covariates in RD estimation has its roots in the common analogy between RD designs and randomized experiments. Since the RD design is often

(formally or informally) thought of as a randomized experiment near the cutoff (Lee, 2008; Cattaneo et al., 2015), and the inclusion of covariates is often used in the analysis of experiments to increase precision, inclusion of covariates in RD estimation seems natural. However, despite covariate-adjusted RD analysis being widespread in empirical practice, there is no existing justification for using additional covariates for identification, estimation, or inference purposes, employing only continuity/smoothness conditions at the cutoff. This has led to the proliferation of ad-hoc covariate-adjustment practices that, at best, reduce the transparency of the estimation strategy and, at worst, result in generally noncomparable (or even inconsistent) estimators.

We provide a set of results that formalize and justify covariate adjustment in RD designs, and offer valid estimation and inference procedures. Following empirical practice, we augment the standard RD framework in order to codify the inclusion of covariates. In particular, we study local polynomial methods allowing for the inclusion of additional covariates in an additive separable, linear-in-parameters way, which permits continuous, discrete, and mixed regressors and does not require additional smoothing methods (e.g., no need for choosing other bandwidths or kernels). This procedure for *covariate-adjusted RD estimation* covers linear model adjustments, which are popular in applied work, and allows us to characterize not only the conditions under which the inclusion of covariates is appropriate, but also the ways in which adjusting by covariates may lead to inconsistent RD estimators. Thus, our formal results offer concrete guidelines for applied researchers that were previously unavailable.

Under minimal smoothness assumptions, we show that the covariate-adjusted RD estimator that imposes the same adjustment above and below the cutoff is consistent for the standard RD treatment effect if a simple "zero RD treatment effect on covariates" condition holds. For example, in the sharp RD design, the only requirement is that the covariates have equal conditional expectation from above and below at the cutoff, which is often conceived and presented as a falsification or "placebo" test in RD empirical studies (see, e.g., Lee, 2008; Canay and Kamat, 2015, and references therein). This requirement of "balanced" covariates at the cutoff, in the appropriate sense depending on the RD design considered, is the most natural and practically relevant sufficient condition but, more generally, we are able to obtain necessary and sufficient conditions for consistency of the covariate-adjusted RD estimator. We also discuss identification and consistency properties of an alternative covariate-adjusted RD estimator that includes an interaction between treatment and

covariates, and show that this estimator is generally inconsistent for the standard RD parameter of interest. We also characterize the (necessary and) sufficient conditions required for this alternative estimator to be consistent, which are strong and unlikely to hold in empirical settings.

We offer a complete asymptotic analysis for the covariate-adjusted RD estimator, including novel mean squared error (MSE) expansions, several MSE-optimal bandwidths and consistent data-driven implementations thereof, MSE-optimal point estimators, and valid asymptotic inference, covering all empirically relevant RD designs (sharp RD, kink RD, fuzzy RD, and fuzzy kink RD), with both heteroskedastic and clustered data. These results have immediate practical use in any RD analysis and aid in interpreting prior results. In particular, we characterize precisely the source of efficiency gains obtained when using the covariate-adjusted RD estimator (see Remarks 2 and 3). Last but not least, we provide new general purpose `Stata` and `R` packages that implement all our results—see Calonico, Cattaneo, Farrell and Titiunik (2016b) and references therein for more details.

We illustrate our methods with two empirical applications. First, we employ the data of Ludwig and Miller (2007) to re-analyze the effect of Head Start on child mortality, where we find that including nine pre-intervention 1960 census covariates leads to an average reduction of confidence interval length of about 5% to 10% relative to the case without covariate-adjustment. Second, we use the data of Chay, McEwan and Urquiola (2005) on the effect of school improvements on test scores, where we see a 3% to 5% reduction in confidence interval length when six region-indicator covariates are included. Finally, we also discuss the findings from an extensive simulation study investigating the finite-sample properties of our methods.

Our paper contributes to the large and still rapidly expanding methodological literature on RD designs. Instead of giving a (likely incomplete) summary here, we defer to the review articles cited in our opening paragraph for references and the references given throughout the manuscript. In addition, our main results are also connected to the causal inference literature on covariate-adjusted treatment effect estimation in randomized experiments. For a recent review see Imbens and Rubin (2015) and references therein. In the specific context of RD designs, one recent strand of the literature re-interprets the data as being "as good as randomized" within a small window around the cutoff. This so-called "local randomization" RD approach requires conditions stronger than continuity/smoothness of conditional expectations, but allows classical techniques and interpretations from randomized experiments to be brought to bear (Cattaneo et al., 2015, 2016). Although

we maintain the more standard continuity framework throughout this article, our main results show that the large sample implications of a "local randomization" approach to RD designs carry over for the particular case of covariate adjustment—the only requirement is a weak and intuitive restriction on the additional "pre-intervention" covariates distribution included in the estimation.

The plan of presentation is as follows. The bulk of the paper is devoted to an in-depth treatment of the sharp RD design, with a brief section discussing extensions to other popular RD settings. Specifically, Section 2 introduces a framework for covariate-adjustment in sharp RD designs, while Section 3 details important identification and interpretation issues. Section 4 gives a complete analysis of nonparametric inference in sharp RD designs using covariates, including new MSE expansions, MSE-optimal estimators, valid inference based on robust bias-correction techniques, and consistent standard errors. Section 5 discusses extensions to other RD designs. Section 6 presents the results from the empirical illustrations and the simulation study, and Section 7 concludes. The Appendix contains the main formulas concerning the sharp RD design, omitted from the main text to ease the exposition. A supplemental appendix includes: (i) a thorough theoretical treatment of all RD cases and extensions, including proofs of the results herein, (ii) a discussion on implementation and other methodological details, and (iii) the complete set of Monte Carlo results.

2 Sharp RD Designs Using Covariates

The observed data is assumed to be a random sample $(Y_i, T_i, X_i, \mathbf{Z}_i')'$, $i = 1, 2, \ldots, n$, from a large population. The key feature of any RD design is the presence of an observed continuous score or running random variable X_i, which determines treatment assignment for each unit in the sample: all units with X_i greater than a known threshold $\bar{x}$ are assigned to the treatment group, while all units with $X_i < \bar{x}$ are assigned to the control group. In sharp RD designs, treatment compliance is perfect and hence $T_i = \mathbb{1}(X_i \geq \bar{x})$ denotes treatment status. Using the potential outcomes framework, the observed Y_i is given by

$$Y_i = Y_i(0) \cdot (1 - T_i) + Y_i(1) \cdot T_i = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ Y_i(1) & \text{if } T_i = 1, \end{cases}$$

where $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes with and without treatment, respectively, for each unit i in the sample. The parameter of interest is the average treatment effect at the cutoff:

$$\tau = \tau(\bar{x}) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = \bar{x}].$$

Evaluation points of functions are dropped whenever possible throughout the paper. Hahn, Todd and van der Klaauw (2001) gave precise, easy-to-interpret conditions for nonparametric identification of the standard RD treatment effect τ, without additional covariates. The key substantive identifying assumption is that $\mathbb{E}[Y_i(t)|X_i = x]$, $t \in \{0, 1\}$, be continuous at the cutoff $x = \bar{x}$.

The new feature studied in this paper is the presence of additional "pre-intervention", "predetermined", or "exogenous" covariates, collected in the random vector $\mathbf{Z}_i \in \mathbb{R}^d$, which could be continuous, discrete, or mixed. Without loss of generality, we assume

$$\mathbf{Z}_i = \mathbf{Z}_i(0) \cdot (1 - T_i) + \mathbf{Z}_i(1) \cdot T_i = \begin{cases} \mathbf{Z}_i(0) & \text{if } T_i = 0 \\ \mathbf{Z}_i(1) & \text{if } T_i = 1, \end{cases}$$

where $\mathbf{Z}_i(1)$ and $\mathbf{Z}_i(0)$ denote the (potential) covariates on either side of the threshold. In practice, it is natural to assume that some features of the marginal distributions of $\mathbf{Z}_i(1)$ and $\mathbf{Z}_i(0)$ are equal at the cutoff $\bar{x}$ or, more extremely, that $\mathbf{Z}_i(1) = \mathbf{Z}_i(0)$, which would match the definition of a "pre-treatment" covariate in the context of randomized controlled trials.

A large portion of the literature on estimation and inference in RD designs focuses on nonparametric local polynomial estimators. In practice, researchers first choose a neighborhood around the cutoff, usually via a bandwidth choice, and then conduct local polynomial inference—that is, they rely on linear regression fits using only units whose scores lay within that pre-selected neighborhood. The standard RD treatment effect estimator is then given by

$$\hat{\tau}(h) = \mathbf{e}_0' \hat{\boldsymbol{\beta}}_{Y+,p}(h) - \mathbf{e}_0' \hat{\boldsymbol{\beta}}_{Y-,p}(h),$$

where $\hat{\boldsymbol{\beta}}_{Y-,p}(h)$ and $\hat{\boldsymbol{\beta}}_{Y+,p}(h)$ correspond to the weighted least squares coefficients

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_{Y-,p}(h) \\ \hat{\boldsymbol{\beta}}_{Y+,p}(h) \end{bmatrix} = \underset{\boldsymbol{\beta}_-, \boldsymbol{\beta}_+}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{r}_{-,p}(X_i - \bar{x})'\boldsymbol{\beta}_- - \mathbf{r}_{+,p}(X_i - \bar{x})'\boldsymbol{\beta}_+)^2 K_h(X_i - \bar{x}), \qquad (1)$$

with $\boldsymbol{\beta}_-, \boldsymbol{\beta}_+ \in \mathbb{R}^{p+1}$, $\mathbf{r}_{-,p}(x) = \mathbb{1}(u < 0)(1, x, \cdots, x^p)'$, $\mathbf{r}_{+,p}(x) = \mathbb{1}(u \geq 0)(1, x, \cdots, x^p)'$, $\mathbf{e}_0$ the $(p+1)$-vector with a one in the first position and zeros in the rest, and $K_h(u) = K(u/h)/h$ for a kernel function $K(\cdot)$ and a positive bandwidth sequence h. The kernel and bandwidth serve to localize the regression fit near the cutoff. We assume the following standard regularity conditions for the kernel.

Assumption 1 (Kernel)**.** *$k(\cdot) : [0,1] \mapsto \mathbb{R}$ is bounded and nonnegative, zero outside its support, and positive and continuous on $(0,1)$. Set $K(u) = \mathbb{1}(u < 0)k(-u) + \mathbb{1}(u \geq 0)k(u)$.*

The most popular choices of kernel are (i) the uniform kernel, giving equal weighting to observations $X_i \in [\bar{x} - h, \bar{x} + h]$, and (ii) the triangular kernel that assigns linear down-weighting to the same observations. The preferred choice of polynomial order is $p = 1$, which gives the standard local-linear RD point estimator. The estimators $\hat{\boldsymbol{\beta}}_{Y-,p}(h)$ and $\hat{\boldsymbol{\beta}}_{Y+,p}(h)$ are, of course, numerically equivalent to the coefficients that would be obtained from two separate weighted regressions, using only observations on one side of the cutoff (with the same kernel and bandwidth). We set the problem as a single joint least-squares linear regression fit to ease the upcoming comparisons with the covariate-adjusted RD estimator.

While the standard estimator $\hat{\tau}(h)$ is popular in empirical work, and readily justified by local smoothness assumptions, it is extremely common for empirical researchers to augment their specification with the additional covariates $\mathbf{Z}_i$. One way of introducing covariates leads to the following estimator, which we call the *covariate-adjusted RD estimator*:

$$\tilde{\tau}(h) = \mathbf{e}_0'\tilde{\boldsymbol{\beta}}_{Y+,p}(h) - \mathbf{e}_0'\tilde{\boldsymbol{\beta}}_{Y-,p}(h),$$

where $\tilde{\boldsymbol{\beta}}_{Y+,p}(h)$ and $\tilde{\boldsymbol{\beta}}_{Y-,p}(h)$ are defined through

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}}_{Y-,p}(h) \\ \tilde{\boldsymbol{\beta}}_{Y+,p}(h) \\ \tilde{\boldsymbol{\gamma}}_{Y,p}(h) \end{bmatrix} = \underset{\boldsymbol{\beta}_-, \boldsymbol{\beta}_+, \boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{r}_{-,p}(X_i - \bar{x})'\boldsymbol{\beta}_- - \mathbf{r}_{+,p}(X_i - \bar{x})'\boldsymbol{\beta}_+ - \mathbf{Z}_i'\boldsymbol{\gamma})^2 K_h(X_i - \bar{x}), \quad (2)$$

where $\boldsymbol{\beta}_-, \boldsymbol{\beta}_+ \in \mathbb{R}^{p+1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^d$. Throughout the paper and supplemental appendix we employ the following notational convention whenever possible: a quantity with a tilde ($\tilde{\theta}$, say) is estimated with additional covariates, while a quantity with a hat ($\hat{\theta}$, say) is not; cf. (1) vs. (2).

The estimator $\tilde{\tau}(h)$ broadly captures the common empirical practice of first choosing a neighborhood around the cutoff, and then conducting local "flexible" linear least-squares estimation and inference with covariates. But our approach formalizes two restrictions in the way that the additional covariates $\mathbf{Z}_i$ enter the least-squares fit locally to the cutoff: (i) additive separability between the basis expansion of the running variable and the additional covariates, and (ii) a linear-in-parameters specification for these covariates. We avoid full nonparametric estimation over $(X_i, \mathbf{Z}_i')' \in \mathbb{R}^{1+d}$, which would introduce d additional bandwidths and kernels, quickly leading to a curse of dimensionality and hence rendering empirical application infeasible. Further, in practice, $\mathbf{Z}_i$ could include power expansions, interactions, and other "flexible" transformations of the original set of covariates. This approach to RD covariate adjustment allows for any type of additional regressors, including fixed effects or other discrete variables.

The typical motivation for using the covariate-adjusted RD estimator $\tilde{\tau}(h)$ is to improve precision in estimating the RD treatment effect, τ, which arguably stems from least squares analysis of randomized controlled trials. We build on this intuition and make precise the conditions required for consistency of the covariate-adjusted RD estimator $\tilde{\tau}(h)$ for τ. We also show that much more stringent conditions are required if treatment-covariate interactions are included in the estimation model.

Remark 1 (Covariates and Identification)**.** A common misconception among some researchers relying on continuity-based nonparametric identification results for RD designs is that including additional (pre-intervention) covariates can enhance the plausibility of the design. Specifically, it is sometimes claimed that even if $\mathbb{E}[Y_i(t)|X_i = x]$, $t \in \{0, 1\}$, is not continuous at the cutoff, adding covariates could restore nonparametric identification of the RD average treatment effect at the cutoff. However, within the continuity-based RD framework, if $\mathbb{E}[Y_i(t)|X_i = x, \mathbf{Z}_i(t)]$, $t \in \{0, 1\}$, is indeed continuous, then $\mathbb{E}[Y_i(t)|X_i = x] = \mathbb{E}[\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]|X_i = x]$ will be continuous under most reasonable assumptions. For example, suppose that $\mathbf{Z}_i(0)$ and $\mathbf{Z}_i(1)$ are binary (e.g., gender), then $\mathbb{E}[Y_i(t)|X_i = x] = \mathbb{E}[Y_i(t)|X_i = x, \mathbf{Z}_i(t) = 0]\mathbb{P}[\mathbf{Z}_i(t) = 0|X_i = x] + \mathbb{E}[Y_i(t)|X_i = x, \mathbf{Z}_i(t) = 1]\mathbb{P}[\mathbf{Z}_i(t) = 1|X_i = x]$ is a linear combination of assumed-continuous functions and hence must be continuous as well. Thus, covariate adjustment does not solve identification problems when $\mathbb{E}[Y_i(t)|X_i = x]$ is discontinuous. On the other hand, adding covariates for identification purposes can be rationalized and be useful within a local randomization framework for RD designs (e.g.,

Angrist and Rokkanen, 2015; Keele et al., 2015). Below we discuss in more detail the distinctions between these frameworks (see also Cattaneo et al., 2016). □

2.1 Notation and Regularity Conditions

To make precise the difference in population parameters recovered with and without additional covariates, and to analyze the asymptotic properties of the covariate-adjusted RD estimators, it is useful to establish notation and state all the regularity conditions employed. This is done simultaneously in the assumption below, which is the only assumption imposed on the data generating process.

Assumption 2 (Sharp RD Designs)**.** *For $\varrho \geq p+2$ and all $x \in [x_l, x_u]$, where $x_l, x_u \in \mathbb{R}$ such that $x_l < \bar{x} < x_u$:*

(a) The Lebesgue density of X_i, denoted $f(x)$, is continuous and bounded away from zero.

(b) $\mu_{Y-}(x) := \mathbb{E}[Y_i(0)|X_i = x]$, $\mu_{Y+}(x) := \mathbb{E}[Y_i(1)|X_i = x]$, $\boldsymbol{\mu}_{Z-}(x) := \mathbb{E}[\mathbf{Z}_i(0)|X_i = x]$, $\boldsymbol{\mu}_{Z+}(x) := \mathbb{E}[\mathbf{Z}_i(1)|X_i = x]$, $\mathbb{E}[\mathbf{Z}_i(0)Y_i(0)|X_i = x]$, and $\mathbb{E}[\mathbf{Z}_i(1)Y_i(1)|X_i = x]$ are ϱ times continuously differentiable.

(c) $\mathbb{V}[\mathbf{S}_i(t)|X_i = x]$, with $\mathbf{S}_i(t) := (Y_i(t), \mathbf{Z}_i(t)')'$, $t \in \{0,1\}$, are continuously differentiable and invertible.

(d) $\mathbb{E}[|\mathbf{S}_i(t)|^4|X_i = x]$, $t \in \{0,1\}$, are continuous, where $|\cdot|$ denotes the Euclidean norm.

Assumption 2 imposes standard continuity/smoothness assumptions common to all nonparametric analyses of RD designs, properly enlarged to allow for the inclusion of additional covariates. Indeed, if one simply ignores all statements involving these covariates, the conditions are exactly those found in the RD literature.

The assumptions are placed only on features such as the mean and variance of the conditional distributions given the running variable X_i alone. Importantly, Assumption 2 does not restrict in any way the "long" conditional expectation $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]$, $t \in \{0,1\}$, which implies that our methods allow for discrete, continuous, and mixed additional covariates, and do not require any semiparametric or parametric modeling of this regression function. That is, as we discuss in more detail below, we allow for complete misspecification of $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]$, $t \in \{0,1\}$, for fixed n, and hence give a "best linear approximation" interpretation to the regression coefficients obtained in (2).

Assumption 2 is not intended to be minimal, but rather parsimonious and easily applicable to nonparametric estimation and inference. Finally, all limits are taken as $n \to \infty$, unless otherwise noted.

3 Identification and Estimation in Sharp RD Designs Using Covariates

We now present the first main result of the paper, which connects and gives an interpretation to the implicit estimand associated with the covariate-adjusted RD estimator.

Lemma 1 (Sharp RD with Covariates)**.** *Let Assumptions 1 and 2 hold. If* $nh \to \infty$ *and* $h \to 0$*, then*

$$\tilde{\tau}(h) \to_{\mathbb{P}} \tau - \left[\boldsymbol{\mu}_{Z+} - \boldsymbol{\mu}_{Z-}\right]' \boldsymbol{\gamma}_Y,$$

where

$$\boldsymbol{\gamma}_Y = (\boldsymbol{\sigma}^2_{Z-} + \boldsymbol{\sigma}^2_{Z+})^{-1} \mathbb{E}\left[\left(\mathbf{Z}_i(0) - \boldsymbol{\mu}_{Z-}(X_i)\right) Y_i(0) + \left(\mathbf{Z}_i(1) - \boldsymbol{\mu}_{Z+}(X_i)\right) Y_i(1) \middle| X_i = \bar{x} \right],$$

where $\boldsymbol{\sigma}^2_{Z-} := \mathbb{V}[\mathbf{Z}_i(0)|X_i = x]$ *and* $\boldsymbol{\sigma}^2_{Z+} := \mathbb{V}[\mathbf{Z}_i(1)|X_i = x]$.

It is well known in the RD literature that, under the conditions of Lemma 1, $\hat{\tau}(h) \to_{\mathbb{P}} \tau$. The conclusion of this lemma gives a precise description of the probability limit of the covariate-adjusted sharp RD estimator, when implemented according to (2). A similar result is discussed for all other popular RD designs in Section 5.

Lemma 1 shows that this covariate-adjusted RD estimation approach is consistent for the standard RD treatment effect at the cutoff, $\tau = \mu_{Y+} - \mu_{Y-}$, plus an additional term that depends on the RD treatment effect at the cutoff for the additional covariates, $\boldsymbol{\tau}_Z := \boldsymbol{\mu}_{Z+} - \boldsymbol{\mu}_{Z-}$. It follows that, given the smoothness conditions imposed in Assumption 2, a sufficient condition for $\tilde{\tau}(h) \to_{\mathbb{P}} \tau$ is that $\boldsymbol{\mu}_{Z+} = \boldsymbol{\mu}_{Z-}$. This is weaker than assuming that the marginal distributions of $\mathbf{Z}_i(0)$ and $\mathbf{Z}_i(1)$ are equal at the cutoff. In other words, $\boldsymbol{\mu}_{Z+} = \boldsymbol{\mu}_{Z-}$ is implied by, but does not require that, $\mathbb{P}[Z_{ki}(0) \leq z|X_i = \bar{x}] = \mathbb{P}[Z_{ki}(1) \leq z|X_i = \bar{x}]$ for all z and $k = 1, 2, \cdots, d$, where $\mathbf{Z}_i(t) = [Z_{1i}(t), Z_{2i}(t), \cdots, Z_{di}(t)]'$ with $t \in \{0, 1\}$.

The typical motivation for including covariates in RD analyses is to gain precision in estimating

the RD treatment effect of interest, τ, which has its roots in the analysis of randomized experiments. Even if implicitly, researchers employing covariates in RD designs assume some form of "local randomization", where units are thought to be assigned to treatment or control at random near the cutoff: $\{Y_i(0), Y_i(1), \mathbf{Z}_i(0), \mathbf{Z}_i(1)\} \perp\!\!\!\perp T_i \mid X_i \in [\bar{x}-h, \bar{x}+h]$. Such local randomization is discussed intuitively by Lee (2008) and Lee and Lemieux (2010); more recently, Cattaneo et al. (2015, 2016) discuss the stronger conditions, beyond continuity, required for the interpretation and valid analysis of RD designs as local randomized experiments. (See also de la Cuesta and Imai, 2016, for a recent discussion of the distinction between continuity and local randomization.) From this perspective, "pre-intervention" covariates would satisfy $\mathbf{Z}_i(0) = \mathbf{Z}_i(1)$ conditional on $X_i \in [\bar{x}-h, \bar{x}+h]$, that is, their distributions would be equal among control and treatment units near the cutoff.

In contrast, in this paper we do not assume a local randomization condition of any form, but rather focus on continuity-based methods and hence (superpopulation) nonparametric identification. In this setting, Lemma 1 shows that only continuity of the conditional expectations of the additional covariates at the cutoff is the key condition required for consistency of the covariate-adjusted RD estimator. In other words, whenever additional covariates satisfying $\boldsymbol{\mu}_{Z+} = \boldsymbol{\mu}_{Z-}$ are included as in (2), the estimator $\tilde{\tau}(h)$ will remain consistent for the standard RD estimand τ.

3.1 Identification and Estimation with Treatment Interaction

In addition to efficiency gains, another common motivation for examining covariates in randomized experiments is to discover treatment effect heterogeneity, and covariate-adjusted linear regression is a frequently used method for doing so. A potentially interesting extension of our work, and in particular Lemma 1, would be to further augment the covariate-adjusted RD estimator $\tilde{\tau}(h)$ implemented as in (2) with interactions between $\mathbf{r}_p(X_i - \bar{x})$ and $\mathbf{Z}_i$. This alternative estimation method might be useful to assess treatment effect heterogeneity at the cutoff, as well as to provide a more "flexible" approximation of the unknown conditional expectations in finite samples. While such general approach is beyond the scope of this paper, we do discuss a special case of this idea to illustrate the potential pitfalls of allowing for interactions in the local polynomial fits.

To be specific, we will examine what we call the *treatment-interacted covariate-adjusted RD estimator*. This estimator, sometimes used in empirical work, is given by

$$\check{\eta}(h) = \mathbf{e}_0' \check{\boldsymbol{\beta}}_{Y+,p}(h) - \mathbf{e}_0' \check{\boldsymbol{\beta}}_{Y-,p}(h),$$

where now $\check{\boldsymbol{\theta}}_p(h) = [\check{\boldsymbol{\beta}}_{Y-,p}(h)', \check{\boldsymbol{\beta}}_{Y+,p}(h)', \check{\boldsymbol{\gamma}}_{Y-,p}(h)', \check{\boldsymbol{\gamma}}_{Y+,p}(h)']'$ is computed by

$$\min_{\boldsymbol{\beta}_-,\boldsymbol{\beta}_+,\boldsymbol{\gamma}_-,\boldsymbol{\gamma}_+} \sum_{i=1}^{n} (Y_i - \mathbf{r}_{-,p}(X_i - \bar{x})'\boldsymbol{\beta}_- - \mathbf{r}_{+,p}(X_i - \bar{x})'\boldsymbol{\beta}_+ - \mathbf{Z}'_{-,i}\boldsymbol{\gamma}_- - \mathbf{Z}'_{+,i}\boldsymbol{\gamma}_+)^2 K_h(X_i - \bar{x}),$$

where $\mathbf{Z}_{-,i} = \mathbb{1}(X_i < \bar{x})\mathbf{Z}_i$ and $\mathbf{Z}_{+,i} = \mathbb{1}(X_i \geq \bar{x})\mathbf{Z}_i$, and $\boldsymbol{\beta}_-, \boldsymbol{\beta}_+ \in \mathbb{R}^{p+1}$ and $\boldsymbol{\gamma}_-, \boldsymbol{\gamma}_+ \in \mathbb{R}^d$. In words, this alternative estimator fits a weighted least squares regression with full interactions between treatment assignment and both the polynomial basis expansion of X_i and the additional covariates $\mathbf{Z}_i$. Thus, $\check{\boldsymbol{\theta}}_p(h)$ is numerically equivalent to fitting two separate weighted linear regressions on each side of the cutoff, leading to $[\check{\boldsymbol{\beta}}_{Y-,p}(h)', \check{\boldsymbol{\gamma}}_{Y-,p}(h)']$ and $[\check{\boldsymbol{\beta}}_{Y+,p}(h)', \check{\boldsymbol{\gamma}}_{Y+,p}(h)']$. We present the estimation approach in a fully interacted version only for notation simplicity, so that $\boldsymbol{\gamma}_-$ and $\boldsymbol{\gamma}_+$ have a symmetric interpretation.

As shown in the next lemma, including the treatment-covariate interaction has important implications for interpretation. The difference follows from the fact that including this interaction allows $\boldsymbol{\gamma}_- \neq \boldsymbol{\gamma}_+$ in the estimation, whereas the covariate-adjusted RD estimator of (2) forces equality.

Lemma 2 (Sharp RD with Covariates and Treatment Interaction)**.** *Let Assumptions 1 and 2 hold. If $nh \to \infty$ and $h \to 0$, then*

$$\check{\eta}(h) \to_{\mathbb{P}} \eta := \tau - \left[\boldsymbol{\mu}'_{Z+}\boldsymbol{\gamma}_{Y+} - \boldsymbol{\mu}'_{Z-}\boldsymbol{\gamma}_{Y-}\right],$$

where

$$\boldsymbol{\gamma}_{Y-} = (\boldsymbol{\sigma}^2_{Z-})^{-1}\mathbb{E}\left[(\mathbf{Z}_i(0) - \boldsymbol{\mu}_{Z-}(X_i))\, Y_i(0) \middle| X_i = \bar{x}\right],$$

$$\boldsymbol{\gamma}_{Y+} = (\boldsymbol{\sigma}^2_{Z+})^{-1}\mathbb{E}\left[(\mathbf{Z}_i(1) - \boldsymbol{\mu}_{Z+}(X_i))\, Y_i(1) \middle| X_i = \bar{x}\right].$$

The conclusion of this lemma defines a new RD parameter, η, recovered when additional covariates interacted with treatment assignment are included linearly in the local polynomial estimation. A similar result is established for all other RD designs in the supplement. This result gives a precise and general interpretation to the probability limit of the interacted covariate-adjusted RD estimator: $\check{\eta}(h)$ is consistent for the standard RD average treatment effect at the cutoff, $\tau = \mu_{Y+} - \mu_{Y-}$, plus an additional term which can be interpreted as the difference of the best linear approximations at the cutoff of the unknown conditional expectations $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(0)]$, $t \in \{0, 1\}$, based on the additional covariates included in the RD estimation. Alternatively, the "bias" due to the inclu-

sion of additional covariates interacted with treatment assignment, $\boldsymbol{\mu}_{Z+}'\boldsymbol{\gamma}_{Y+} - \boldsymbol{\mu}_{Z-}'\boldsymbol{\gamma}_{Y-}$, can be interpreted as the difference of the best linear predictions of $Y_i(t)$ on $\mathbf{Z}_i(t)$, $t \in \{0,1\}$, at the cutoff.

It follows that, when a treatment-covariate interaction is included in the estimation, a necessary and sufficient condition for the resulting covariate-adjusted RD estimator to be consistent for the standard RD treatment effect is that $\boldsymbol{\mu}_{Z+}'\boldsymbol{\gamma}_{Y+} = \boldsymbol{\mu}_{Z-}'\boldsymbol{\gamma}_{Y-}$. This condition, however, is harder to justify in practice than the condition required for the model without the interaction. In particular, the previously sufficient condition $\boldsymbol{\mu}_{Z+} = \boldsymbol{\mu}_{Z-}$ ("covariate balance") is now no longer sufficient because one needs also to assume that $\boldsymbol{\gamma}_{Y+} = \boldsymbol{\gamma}_{Y-}$. The latter additional assumption can be regarded as an "homogeneous partial effect of covariates on potential outcomes".

The interacted covariate-adjusted estimand η can also be interpreted as a "partial effect" RD estimator. To make this clear, consider the following example. Suppose that $\mathbb{E}[Y_i(0)|X_i, \mathbf{Z}_i(0)] = \xi_-(X_i) + \mathbf{Z}_i(0)'\boldsymbol{\delta}_{Y-}$ and $\mathbb{E}[Y_i(1)|X_i, \mathbf{Z}_i(1)] = \xi_+(X_i) + \mathbf{Z}_i(1)'\boldsymbol{\delta}_{Y+}$ near the cutoff. Then $\boldsymbol{\gamma}_{Y-} = \boldsymbol{\delta}_{Y-}$ and $\boldsymbol{\gamma}_{Y+} = \boldsymbol{\delta}_{Y+}$ and $\check{\eta}(h) \to_{\mathbb{P}} \eta = \xi_+(\bar{x}) - \xi_-(\bar{x}) \neq \tau$, and hence the interacted covariate-adjusted RD estimator $\check{\eta}(h)$ is consistent for a partial effect at the cutoff. In this example, the additional condition required for $\check{\eta}(h) \to_{\mathbb{P}} \tau$ becomes $\boldsymbol{\delta}_{Y-} = \boldsymbol{\delta}_{Y+}$, which in turn is implied by $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)] = \mathbb{E}[Y_i(t)|X_i]$ near the cutoff, though the latter is not required.

3.2 Practical Implications

Lemmas 1 and 2 not only give general, precise, and intuitive characterizations of the probability limits of two popular covariate-adjusted RD estimators, but also have interesting implications for the analysis and interpretation of RD designs using covariates. Most notably, the lemmas above show the conditions under which a covariate-adjusted RD estimator is consistent for the standard (causal) RD treatment effect of interest, τ, and, by implication, establish when estimators with and without covariate adjustment can be compared to each other.

Since in most applications τ is the parameter of interest, comparing $\hat{\tau}(h)$ vs. $\tilde{\tau}(h)$ requires the assumption $\boldsymbol{\mu}_{Z+} = \boldsymbol{\mu}_{Z-}$ (Lemma 1), while comparing $\hat{\tau}(h)$ vs. $\check{\eta}(h)$ requires $\boldsymbol{\mu}_{Z+}'\boldsymbol{\gamma}_{Y+} = \boldsymbol{\mu}_{Z-}'\boldsymbol{\gamma}_{Y-}$ (Lemma 2). Adjusting for covariates in RD settings seems most useful when the estimand of interest remains unchanged, in which case comparing precision becomes meaningful (as in Remarks 2 and 3). In applications, there is no *a priori* reason to blindly compare different estimators ($\hat{\tau}(h)$, $\tilde{\tau}(h)$, $\check{\eta}(h)$) without imposing (and, possibly, testing for) the underlying sufficient assumptions required

to retain the same target RD treatment effect of interest.

4 Inference in Sharp RD Designs using Covariates

Estimation and inference in RD designs using local polynomial methods without covariates (i.e. using only Y_i and X_i) has been studied in great detail in recent years—see, among others, Hahn et al. (2001), Porter (2003), Imbens and Kalyanaraman (2012), Calonico et al. (2014), Gelman and Imbens (2014), Armstrong and Kolesar (2015), Kamat (2015), Calonico et al. (2016a), and references therein. These papers give asymptotic MSE expansions, MSE-optimal point estimators, data-driven bandwidth selection methods, asymptotically valid inference procedures based on bias-correction and non-standard distributional approximations, and even valid Edgeworth expansions to inform empirical practice.

We study the asymptotic properties of the covariate-adjusted RD estimator, $\tilde{\tau}(h)$, building on this literature. We assume that $\boldsymbol{\mu}_{Z+} = \boldsymbol{\mu}_{Z-}$ in order to maintain the same standard RD treatment effect of interest (Lemma 1). We present new MSE expansions, several data-driven optimal bandwidth selectors, valid distributional approximations based on bias-correction techniques, and consistent standard errors for $\tilde{\tau}(h)$. Analogous results for other RD designs are briefly discussed in Section 5. A full treatment of all cases, including several other extensions, is given in the supplemental appendix. All our methods are implemented in companion general purpose `R` and `Stata` software packages (Calonico et al., 2016b).

To characterize the asymptotic properties of the covariate-adjusted RD estimator $\tilde{\tau}(h)$, we rely on the following representation (valid for each n):

$$\tilde{\tau}(h) = \hat{\tau}(h) - \hat{\boldsymbol{\tau}}_Z(h)'\tilde{\boldsymbol{\gamma}}_{Y,p}(h),$$

where $\hat{\tau}(h)$ and $\tilde{\boldsymbol{\gamma}}_{Y,p}(h)$ were defined above (see (1) and (2)), and $\hat{\boldsymbol{\tau}}_Z(h)$ is a d-dimensional vector containing the standard RD treatment effect estimator for each covariate. In other words, each element of $\hat{\boldsymbol{\tau}}_Z(h)$ is constructed using the corresponding covariate as outcome variable in (1). In the appendix and supplemental appendix we give exact details. Using this partial-out representation,

it follows that

$$\tilde{\tau}(h) - \tau = \mathbf{s}(h)' \begin{bmatrix} \hat{\tau}(h) - \tau \\ \hat{\boldsymbol{\tau}}_Z(h) \end{bmatrix} = \mathbf{s}' \begin{bmatrix} \hat{\tau}(h) - \tau \\ \hat{\boldsymbol{\tau}}_Z(h) \end{bmatrix} \{1 + o_{\mathbb{P}}(1)\}$$

where $\mathbf{s}(h) = (1, \tilde{\boldsymbol{\gamma}}_{Y,p}(h)')'$ and $\mathbf{s} = (1, \boldsymbol{\gamma}_Y')'$, and because $\mathbf{s}(h) \to_{\mathbb{P}} \mathbf{s}$ using the results underlying Lemma 1 (and, later, we will also use $\boldsymbol{\tau}_Z = \mathbf{0}$).

Therefore, the asymptotic analysis proceeds by studying the (joint) large-sample properties of the vector $\hat{\boldsymbol{\tau}}_S(h) := (\hat{\tau}(h), \hat{\boldsymbol{\tau}}_Z(h)')'$ and then taking the linear combination $\mathbf{s}(h)$ or $\mathbf{s}$, as appropriate. Note that $\hat{\boldsymbol{\tau}}_S(h) \to_{\mathbb{P}} \boldsymbol{\tau}_S := (\tau, \boldsymbol{\tau}_Z')'$ under the conditions in Lemma 1. In fact, most of the results presented in this paper do not require the assumption $\boldsymbol{\tau}_Z = \mathbf{0}$, though without this assumption the parameter of interest changes, undermining the practical usefulness of the results. Finally, we re-emphasize that our results do not impose any restrictions on the distribution of $Y_i(t)|X_i, \mathbf{Z}_i(t)$, $t \in \{0, 1\}$, and impose instead minimal restrictions on the distributions of $Y_i(t), \mathbf{Z}_i(t)|X_i$, $t \in \{0, 1\}$ (see Assumption 2). For example, we do not place any parametric or semiparametric assumption on $\mathbb{E}[Y_i(0)|X_i, \mathbf{Z}_i(0)]$ or $\mathbb{E}[Y_i(1)|X_i, \mathbf{Z}_i(1)]$.

4.1 MSE Expansion and Optimal Point Estimators

We first establish a valid asymptotic MSE expansion for the covariate-adjusted RD estimator. This expansion will aid in developing optimal bandwidth choices and MSE-optimal point estimators. Furthermore, the bias expressions will be instrumental to develop valid inference procedures based on robust bias-correction techniques. Let $\mathbf{X} = [X_1, X_2, \cdots, X_n]'$, and define

$$\mathsf{MSE}[\tilde{\tau}(h)] = \mathbb{E}[(\mathbf{s}'\hat{\boldsymbol{\tau}}_S(h) - \mathbf{s}'\boldsymbol{\tau}_S)^2|\mathbf{X}] = (\mathsf{Bias}[\tilde{\tau}(h)])^2 + \mathsf{Var}[\tilde{\tau}(h)],$$

where $\mathsf{Bias}[\tilde{\tau}(h)] := \mathbb{E}[\mathbf{s}'\hat{\boldsymbol{\tau}}_S(h) - \mathbf{s}'\boldsymbol{\tau}_S|\mathbf{X}]$ and $\mathsf{Var}[\tilde{\tau}(h)] := \mathbb{V}[\mathbf{s}'\hat{\boldsymbol{\tau}}_S(h)|\mathbf{X}]$.

Theorem 1 (MSE Expansion)**.** *Let Assumptions 1 and 2 hold. If $nh \to \infty$ and $h \to 0$, then*

$$\mathsf{MSE}[\tilde{\tau}(h)] = h^{2(1+p)}\mathcal{B}_{\tilde{\tau}}(h)^2 \{1 + o_{\mathbb{P}}(1)\} + \frac{1}{nh}\mathcal{V}_{\tilde{\tau}}(h),$$

where the precise expressions for all bias and variance terms are given in the appendix.

The bias and variance expressions in Theorem 1 are different from those available in the literature (Imbens and Kalyanaraman, 2012; Calonico et al., 2014) due to the presence of the additional

covariates $\mathbf{Z}_i$. As a consequence, MSE-optimal bandwidth selection and MSE-optimal point estimators in RD designs using covariates are different from their counterparts without covariates. Bias-correction techniques and standard errors constructions will also be different, as discussed below.

The leading bias and variance formulas in Theorem 1 are derived in pre-asymptotic form. For the bias, the random term $\mathcal{B}_{\tilde{\tau}}(h)$ gives a pre-asymptotic stochastic approximation to the conditional bias of the linearized estimator (hence the presence of the $o_{\mathbb{P}}(1)$ term), whereas the variance term $\mathcal{V}_{\tilde{\tau}}(h)$ is simply obtained by a conditional on $\mathbf{X}$ calculation for the linearized estimator. Calonico et al. (2016a) prove, using valid Edgeworth expansions, that employing pre-asymptotic approximations when conducting asymptotic inference in nonparametrics can lead to superior performance. Furthermore, fewer unknown features of the data generating process must be characterized and estimated.

The main constants in Theorem 1 have a familiar form: the bias and variance are, respectively, $\mathcal{B}_{\tilde{\tau}}(h) = \mathcal{B}_{\tilde{\tau}+}(h) - \mathcal{B}_{\tilde{\tau}-}(h)$ and $\mathcal{V}_{\tilde{\tau}}(h) = \mathcal{V}_{\tilde{\tau}-}(h) + \mathcal{V}_{\tilde{\tau}+}(h)$, where each component stems from estimating the unknown regression function on one side of the cutoff. Here, the bias is entirely due to estimating the unknown functions $\mu_{Y-}(\cdot)$ and $\boldsymbol{\mu}_{Z-}(\cdot)$ for the control group and $\mu_{Y+}(\cdot)$ and $\boldsymbol{\mu}_{Z+}(\cdot)$ for the treatment group. When the additional covariates are not included, these constants reduce exactly to those already available in the literature. In the appendix, we also give the limiting version of the bias and variance constants; that is, we characterize the fixed, real scalars $\mathcal{B}_{\tilde{\tau}}$ and $\mathcal{V}_{\tilde{\tau}}$ that satisfy $(\mathcal{B}_{\tilde{\tau}}(h), \mathcal{V}_{\tilde{\tau}}(h)) \to_{\mathbb{P}} (\mathcal{B}_{\tilde{\tau}}, \mathcal{V}_{\tilde{\tau}})$.

Assuming that $\mathcal{B}_{\tilde{\tau}} \neq 0$, the MSE-optimal bandwidth choice for the covariate-adjusted RD estimator $\tilde{\tau}(h)$ is:

$$\mathfrak{h}_{\tilde{\tau}} = \left[\frac{1}{2(1+p)} \frac{\mathcal{V}_{\tilde{\tau}}/n}{\mathcal{B}_{\tilde{\tau}}^2} \right]^{\frac{1}{3+2p}}.$$

This choice can be used to construct a consistent and MSE-optimal covariate-adjusted sharp RD point estimator: $\tilde{\tau}(\mathfrak{h}_{\tilde{\tau}}) \to_{\mathbb{P}} \tau$, provided that $\boldsymbol{\tau}_Z = \mathbf{0}$. Note that $d = \dim(\mathbf{Z}_i)$ does not impact the rate of decay because we do not employ any nonparametric smoothing methods on the additional covariates.

We address the issue of data-driven implementations of the new optimal bandwidth choices further below, after discussing valid large sample inference.

Remark 2 (Asymptotic Efficiency)**.** In addition to finite-sample efficiency considerations, which

are well known from the literature on linear least-squares, we can give a precise characterization of the asymptotic efficiency gains from introducing additional covariates in the RD estimation. Using the explicit formulas given in the appendix, it is easy to show that the asymptotic variance of the covariate-adjusted estimator $\tilde{\tau}$ (denoted by $\mathcal{V}_{\tilde{\tau}}$) is equal to the asymptotic variance of the standard RD estimator $\hat{\tau}$ (denoted by $\mathcal{V}_{\tau}$) plus a linear combination (based on $\boldsymbol{\gamma}_Y$) of terms involving the $\mathbb{C}ov[Y_i(t), \mathbf{Z}_i(t)|X_i = \bar{x}]$ and $\mathbb{V}[\mathbf{Z}_i(t)|X_i = \bar{x}]$. Therefore, $\tilde{\tau}$ can be asymptotically more efficient than $\hat{\tau}$ when the term $2\mathbb{C}ov[Y_i(t), \mathbf{Z}_i(t)|X_i = \bar{x}]'\boldsymbol{\gamma}_Y$ is negative and larger in absolute value than $\boldsymbol{\gamma}_Y'\mathbb{V}[\mathbf{Z}_i(t)|X_i = \bar{x}]\boldsymbol{\gamma}_Y$. □

Remark 3 (MSE-optimal Point Estimation)**.** The results above also show that $\tilde{\tau}(\mathfrak{h}_{\tilde{\tau}})$ can be a better point estimator in a MSE sense that its MSE-optimal counterpart without covariates, $\hat{\tau}(\mathfrak{h}_{\hat{\tau}})$, where $\mathfrak{h}_{\hat{\tau}}$ denotes the MSE-optimal bandwidth choice for the standard RD estimator $\hat{\tau}$ (Imbens and Kalyanaraman, 2012; Calonico et al., 2014). Using the explicit formulas given in the appendix, it is easy to give conditions such that $\mathsf{MSE}[\tilde{\tau}(\mathfrak{h}_{\tilde{\tau}})] < \mathsf{MSE}[\hat{\tau}(\mathfrak{h}_{\hat{\tau}})]$ (both have the same rate of decay), although this is not the main goal of our paper. We still recommend that $\hat{\tau}(\mathfrak{h}_{\hat{\tau}})$ be the benchmark RD point estimator, and thus that researchers incorporate covariates to increase precision relative to it, rather than to replace it. □

Remark 4 (Other Optimal Bandwidth Choices)**.** In the supplemental appendix we discuss other MSE-optimal bandwidth selectors based on the results underlying Theorem 1, which are specifically tailored to one-sided and two-sided estimation problems in RD designs. Specifically, we present: (i) separate MSE optimizations on either side of the cutoff, (ii) the MSE for the sum rather than the difference of the one-sided estimators, and (iii) several regularized versions of the plug-in bandwidth selectors. In all cases, the decay rate of these bandwidths matches the MSE-optimal choice, but the exact leading constants differ, implying that any of these could be used to construct sharp RD point estimators with an MSE-optimal convergence rate. These choices may be more stable in finite samples or more robust to situations where the smoothing bias may be small. □

4.2 Asymptotic Distribution and Valid Inference

To develop valid asymptotic distributional approximations and inference procedures we employ robust nonparametric bias-correction. It is by now well understood that inference based on large-sample distribution theory using MSE-optimal bandwidths will suffer from a first order bias, leading

to invalid hypothesis testing and confidence intervals because of misspecification errors near the cutoff. This local smoothing bias involves the bias term in Theorem 1, $\mathcal{B}_{\tilde{\tau}}(h)$, which can be estimated and removed. Following recent ideas and results in Calonico et al. (2014, 2016a), we propose robust bias-corrected distributional approximations for the covariate-adjusted RD estimator, and discuss the associated inference procedures based on such approximations.

The bias terms of Theorem 1 are known up to a higher-order derivative of the unknown regression functions, $\mu_{Y-}(\cdot)$, $\boldsymbol{\mu}_{Z-}(\cdot)$, $\mu_{Y+}(\cdot)$, and $\boldsymbol{\mu}_{Z+}(\cdot)$, all capturing the misspecification error introduced by the local polynomial approximation. These objects can be estimated nonparametrically—the complete details are available in the appendix (we replace $\mathbf{s}$ by $\mathbf{s}(h)$ for implementation). At present, let us simply take as given the bias estimator $\tilde{\mathcal{B}}_{\tilde{\tau}}(b)$ based on local polynomial techniques, which depends on a preliminary bandwidth $b \to 0$, possibly different from h, and on a polynomial order $q > p$. Then, the bias-corrected covariate-adjusted sharp RD estimator is

$$\tilde{\tau}^{\mathsf{bc}}(h,b) = \tilde{\tau}(h) - h^{1+p}\tilde{\mathcal{B}}_{\tilde{\tau}}(b) \tag{3}$$

A particularly empirically useful choice is $b = h$, which is both allowed for by our asymptotic theory and has some optimal properties (Calonico et al., 2016a). This bias correction approach is standard in the literature (e.g., Fan and Gijbels, 1996, Section 4.4), and captures nicely "flexible" regression adjustments to account for misspecification in finite samples (Calonico et al., 2014, Remark 7).

The key idea behind the robust bias-corrected distributional approximation is to employ an estimator of the variability of $\tilde{\tau}^{\mathsf{bc}}(h,b)$ for Studentization purposes, rather than an estimator of the variability of $\tilde{\tau}(h)$ only. Thus, the final missing ingredient before we can state our asymptotic Gaussianity result is characterizing the (conditional) variance of the bias-corrected covariate-adjusted RD estimator. Its fixed-n variability is easily characterized due to its approximate (conditional) linearity, and is given by

$$\mathcal{V}^{\mathsf{bc}}_{\tilde{\tau}}(h,b) = [\mathbf{s}' \otimes \mathbf{e}_0'\mathbf{P}^{\mathsf{bc}}_{-,p}(h,b)]\boldsymbol{\Sigma}_{S-}[\mathbf{s}' \otimes \mathbf{e}_0'\mathbf{P}^{\mathsf{bc}}_{-,p}(h,b)]' + [\mathbf{s}' \otimes \mathbf{e}_0'\mathbf{P}^{\mathsf{bc}}_{+,p}(h,b)]\boldsymbol{\Sigma}_{S+}[\mathbf{s}' \otimes \mathbf{e}_0'\mathbf{P}^{\mathsf{bc}}_{+,p}(h,b)]',$$

where the $(1+p) \times n$ matrices $\mathbf{P}^{\mathsf{bc}}_{-,p}(h,b)$ and $\mathbf{P}^{\mathsf{bc}}_{+,p}(h,b)$ can be computed directly from the data, and the $n(1+d) \times n(1+d)$ matrices of variances and covariances $\boldsymbol{\Sigma}_{S-}$ and $\boldsymbol{\Sigma}_{S+}$ are unknown. Specifically, $\boldsymbol{\Sigma}_{S-} = \mathbb{V}[\mathbf{S}(0)|\mathbf{X}]$ and $\boldsymbol{\Sigma}_{S+} = \mathbb{V}[\mathbf{S}(1)|\mathbf{X}]$, where $\mathbf{S}(0) = (\mathbf{Y}(0)', \mathrm{vec}(\mathbf{Z}(0))')'$ and $\mathbf{S}(1) =$

$(\mathbf{Y}(1)', \text{vec}(\mathbf{Z}(1))')'$, with $\mathbf{Y}(t) = [Y_1(t), Y_2(t), \cdots, Y_n(t)]'$ and $\mathbf{Z}(t) = [\mathbf{Z}_1(t), \mathbf{Z}_2(t), \cdots, \mathbf{Z}_n(t)]'$ for $t \in \{0, 1\}$. The appendix collects tedious details and specific formulas, including the exact form of $\mathbf{P}^{\mathsf{bc}}_{-,p}(h, b)$ and $\mathbf{P}^{\mathsf{bc}}_{+,p}(h, b)$.

The (infeasible) variance formula $\mathcal{V}^{\mathsf{bc}}_{\bar{\tau}}(h, b)$ differs from that presented in Theorem 1, $\mathcal{V}_{\bar{\tau}}(h)$, because it also accounts for the leading additional variability injected by the bias estimation, $h^{1+p}\tilde{\mathcal{B}}_{\bar{\tau}}(b)$. By virtue of the variance formula being computed both conditionally and pre-asymptotic, up to the linear combination term $\mathbf{s}$, it involves only one unknown feature, $\mathbf{\Sigma}_{S-}$ and $\mathbf{\Sigma}_{S+}$, which must be estimated, thereby simplifying implementation considerably.

To operationalize the variance formula we replace unknown quantities by plug-in estimators thereof. The estimators need to account for the specific data structure at hand, such as heteroskedasticity and/or clustering. In particular, we discuss two type of plug-in variance estimators, one based on a nearest neighbor (NN) approach and the other based on a plug-in residuals (PR) approach, covering both conditional heteroskedasticity and clustered data. We defer the notationally cumbersome details to the supplemental appendix, and instead we provide here only a brief summary of the main ideas and results. The unknown matrices $\mathbf{\Sigma}_{S-}$ and $\mathbf{\Sigma}_{S+}$ contain, under conditional heteroskedasticity, diagonal submatrices with representative elements, respectively,

$$\sigma_{YZ_k-,i} = \mathbb{C}ov[Y_i(0), Z_{ki}(0)|X_i], \qquad \sigma_{YZ_k+,i} = \mathbb{C}ov[Y_i(1), Z_{ki}(1)|X_i]$$

for $k = 1, 2, \cdots, d$. The feasible variance estimators are then constructed by replacing these unknown objects with unbiased estimators thereof, as follows.

- **NN Variance Estimation**. Employing ideas in Muller and Stadtmuller (1987) and Abadie and Imbens (2006, 2008), we replace $\sigma_{YZ_k-,i}$ and $\sigma_{YZ_k+,i}$ by, respectively,

$$\hat{\sigma}_{YZ_k-,i}(J) = \mathbb{1}(X_i < \bar{x})\frac{J}{J+1}\left(Y_i - \frac{1}{J}\sum_{j=1}^{J} Y_{\ell_{-,j}(i)}\right)\left(Z_{ki} - \frac{1}{J}\sum_{j=1}^{J} Z_{k\ell_{-,j}(i)}\right),$$

$$\hat{\sigma}_{YZ_k+,i}(J) = \mathbb{1}(X_i \geq \bar{x})\frac{J}{J+1}\left(Y_i - \frac{1}{J}\sum_{j=1}^{J} Y_{\ell_{+,j}(i)}\right)\left(Z_{ki} - \frac{1}{J}\sum_{j=1}^{J} Z_{k\ell_{+,j}(i)}\right),$$

for $k = 1, 2, \cdots, d$, and where $\ell_{-,j}(i)$ is the index of the j-th closest unit to unit i among $\{X_i : X_i < \bar{x}\}$ and $\ell_{+,j}(i)$ is the index of the j-th closest unit to unit i among $\{X_i : X_i \geq \bar{x}\}$,

and J denotes a (fixed) the number of neighbors chosen. Replacing the non-zero entries of $\mathbf{\Sigma}_{S-}$ and $\mathbf{\Sigma}_{S+}$ in this fashion (which depend on the sampling structured assumed), and $\mathbf{s}$ by $\mathbf{s}(h)$, we obtain the NN variance estimator of $\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h,b)$, denoted by $\check{\mathcal{V}}_{\tilde{\tau}}^{\mathsf{bc}}(h,b)$.

- **PR Variance Estimation**. This method applies ideas from least-squares methods; see Long and Ervin (2000), MacKinnon (2012), and Cameron and Miller (2015) for review on variance estimation in this context. We replace $\sigma_{YZ_k-,i}$ and $\sigma_{YZ_k+,i}$ by, respectively,

$$\hat{\sigma}_{YZ_k-,q,i}(h) = \mathbb{1}(X_i < \bar{x})\omega_{-,i}\left(Y_i - \mathbf{r}_q(X_i-\bar{x})'\hat{\boldsymbol{\beta}}_{Y-,q}(h)\right)\left(Z_{ki} - \mathbf{r}_q(X_i-\bar{x})'\hat{\boldsymbol{\beta}}_{Z_k-,q}(h)\right),$$

$$\hat{\sigma}_{YZ_k+,q,i}(h) = \mathbb{1}(X_i \geq \bar{x})\omega_{+,i}\left(Y_i - \mathbf{r}_q(X_i-\bar{x})'\hat{\boldsymbol{\beta}}_{Y+,q}(h)\right)\left(Z_{ki} - \mathbf{r}_q(X_i-\bar{x})'\hat{\boldsymbol{\beta}}_{Z_k+,q}(h)\right),$$

for $k = 1, 2, \cdots, d$, and where $\hat{\boldsymbol{\beta}}_{V-,q}(h)$ and $\hat{\boldsymbol{\beta}}_{V+,q}(h)$ denote the q-th order local polynomial fits using as outcome variable $V \in \{Y, Z_1, Z_2, \cdots, Z_d\}$, with bandwidth h, as described in (1), and $\omega_{-,i}$ and $\omega_{+,i}$ denote finite-sample adjustments used to construct the HC_k variance estimators. See the supplement for details. Replacing the non-zero entries of $\mathbf{\Sigma}_{S-}$ and $\mathbf{\Sigma}_{S+}$ in this fashion (which depend on the sampling structured assumed), and $\mathbf{s}$ by $\mathbf{s}(h)$, we obtain the PR variance estimator of $\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h,b)$, denoted by $\hat{\mathcal{V}}_{\tilde{\tau}}^{\mathsf{bc}}(h,b)$.

Putting together all the pieces, we obtain the following distributional approximation result.

Theorem 2 (Asymptotic Normality)**.** *Let the conditions of Theorem 1 hold, and* $\boldsymbol{\tau}_Z = 0$. *If* $\sqrt{nh}h^{1+p}\min\{h, b^{q-p}\} \to 0$ *and* $\overline{\lim}\,(h/b) < \infty$, *then*

$$\tilde{T}_{\tilde{\tau}} = \frac{\tilde{\tau}^{\mathsf{bc}}(h,b) - \tau}{\sqrt{\frac{1}{nh}\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h,b)}} \to_d \mathcal{N}(0,1).$$

Furthermore, $\check{\mathcal{V}}_{\tilde{\tau}}^{\mathsf{bc}}(h,b)/\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h,b) \to_{\mathbb{P}} 1$ *and* $\hat{\mathcal{V}}_{\tilde{\tau}}^{\mathsf{bc}}(h,b)/\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h,b) \to_{\mathbb{P}} 1$.

Theorem 2 provides valid inference in sharp RD designs using covariates. To our knowledge, this is the first such result available in the literature for the covariate-adjusted RD estimation. Extensions of this result to all other popular RD designs are discussed in Section 5 and the supplemental appendix. Once bandwidths are chosen, asymptotically valid inference procedures are easily constructed. For example, an approximately 95% robust bias-corrected covariate-adjusted confidence interval for the RD treatment effect τ, using $h = b$ and NN variance estimation, is given

by

$$\left[\check{\tau}^{\mathsf{bc}}(h,h) - \frac{1.96}{\sqrt{nh}} \cdot \sqrt{\check{\mathcal{V}}^{\mathsf{bc}}_{\check{\tau}}(h,h)}\ ,\ \check{\tau}^{\mathsf{bc}}(h,h) + \frac{1.96}{\sqrt{nh}} \cdot \sqrt{\check{\mathcal{V}}^{\mathsf{bc}}_{\check{\tau}}(h,h)}\right].$$

We explore the performance in finite samples of our proposed methods in Section 6.

Remark 5 (Clustered Data)**.** Theorem 2 can also be established under clustered sampling. All derivations and results remain valid, but the variance formulas will depend on the particular form of clustering. In this case, asymptotics are conducted under the standard assumptions: (i) each unit i belongs to exactly one of G clusters, and (ii) $G \to \infty$ and $Gh \to \infty$. See Cameron and Miller (2015) for a review of cluster-robust inference, and Bartalotti and Brummet (2016) for a discussion in the context of MSE-optimal bandwidth selection for sharp RD designs. This extension is conceptually straightforward but notationally cumbersome, and is deferred to the supplement. Our companion software in `R` and `Stata` also includes optional cluster-robust (i) bandwidth selection, (ii) MSE-optimal point estimation, and (iii) robust bias-corrected inference. □

4.3 Data-driven Bandwidth Selection

We now discuss bandwidth selection briefly, leaving full details to the supplement (where we also discuss several alternative bandwidth selectors as in Remark 4). Here we focus exclusively on two main, distinct approaches: (i) the MSE-optimal choice derived previously ($\mathfrak{h}_\eta$), which can be used to construct MSE-optimal RD point estimators, and (ii) a novel bandwidth selection approach constructed to obtain the fastest decay of the coverage error rate (CER) of robust bias-corrected confidence intervals (denoted $\mathfrak{h}_{\mathtt{CER},\eta}$), motivated by the valid Edgeworth expansions for RD inference.

One of the strengths of Theorem 2 is that the distributional approximation is valid under a large set of tuning parameter choices (strictly more than would be possible without bias correction), which in particular includes the MSE-optimal choice (which is not valid for standard procedures). Assuming the bias is not zero, Theorem 1 can be used in the familiar way to construct feasible MSE-optimal bandwidth choices. For bandwidths $b \to 0$ and $v \to 0$, these will be given by

$$\tilde{\mathfrak{h}}_{\tilde{\tau}} = \left[\frac{1}{2(1+p)} \frac{\tilde{\mathcal{V}}_{\tilde{\tau}}(v)/n}{\tilde{\mathcal{B}}_{\tilde{\tau}}(b)^2}\right]^{\frac{1}{3+2p}}$$

where the exact form of the bias estimator, $\tilde{\mathcal{B}}_{\tilde{\tau}}(b)$, and variance estimator, $\tilde{\mathcal{V}}_{\tilde{\tau}}(v)$, are given in the supplemental appendix. Heuristically, these estimators are formed as plug-in versions of the pre-

asymptotic formulas obtained in Theorem 1, following the previous discussion leading to Theorem 2. In the supplemental appendix, we also show that these feasible versions of the optimal bandwidths are consistent for their infeasible analogues; i.e., $\tilde{\mathfrak{h}}_{\tilde{\tau}}/\mathfrak{h}_{\tilde{\tau}} \to_{\mathbb{P}} 1$.

A particularly attractive alternative to MSE-optimal bandwidth selection is to develop coverage error optimal bandwidth choices. Following the results in Calonico et al. (2016a), we also recommend the following plug-in bandwidth selector

$$\tilde{\mathfrak{h}}_{\mathtt{CER},\tilde{\tau}} = n^{-\frac{p}{(3+p)(3+2p)}} \cdot \tilde{\mathfrak{h}}_{\tilde{\tau}}.$$

This bandwidth choice minimizes the coverage error rate for confidence intervals based on Theorem 2, and can be preferred for inference purposes.

5 Other RD designs

We extend our main results to cover other popular RD designs, including fuzzy, kink, and fuzzy kink RD. Here we give a short overview of the main ideas, deferring all details to the supplemental appendix. There are two wrinkles to the standard sharp RD design discussed so far that must be accounted for: ratios of estimands/estimators for fuzzy designs and derivatives in estimands/estimators for kink designs.

5.1 Fuzzy RD Designs

The distinctive feature of fuzzy RD designs is that treatment compliance is imperfect. This implies that $T_i = T_i(0) \cdot \mathbb{1}(X_i < \bar{x}) + T_i(1) \cdot \mathbb{1}(X_i \geq \bar{x})$, that is, the treatment status T_i of each unit $i = 1, 2, \cdots, n$ is no longer a deterministic function of the running variable X_i, but $\mathbb{P}[T_i = 1 | X_i = x]$ still changes discontinuously at the RD threshold level $\bar{x}$. Here, $T_i(0)$ and $T_i(1)$ denote the two potential treatment status for each unit i when, respectively, $X_i < \bar{x}$ (not offered treatment) and $X_i \geq \bar{x}$ (offered treatment).

To analyze the case of fuzzy RD designs, we first recycle notation for potential outcomes and covariates as follows:

$$Y_i(t) := Y_i(0) \cdot (1 - T_i(t)) + Y_i(1) \cdot T_i(t)$$

$$\mathbf{Z}_i(t) := \mathbf{Z}_i(0) \cdot (1 - T_i(t)) + \mathbf{Z}_i(1) \cdot T_i(t)$$

for $t = 0, 1$. That is, in this setting, potential outcomes and covariates are interpreted as their "reduced form" (or intention-to-treat) counterparts. Giving causal interpretation to covariate-adjusted instrumental variable type estimators is delicate; see e.g. Abadie (2003) for more discussion. Nonetheless, the above re-definitions enable us to use the same notation, assumptions, and results, already given for the sharp RD design, taking the population target estimands as simply the probability limits of the RD estimators.

The following assumption complements Assumption 2, now concerning the (potential) treatment variables.

Assumption 3 (Fuzzy RD Designs)**.** *For $\varrho \geq p+2$ and all $x \in [x_l, x_u]$, where $x_l, x_u \in \mathbb{R}$ such that $x_l < \bar{x} < x_u$:*

(a) $\mu_{T-}(x) := \mathbb{E}[T_i(0)|X_i = x]$, $\mu_{T+}(x) := \mathbb{E}[T_i(1)|X_i = x]$, $\mathbb{E}[\mathbf{Z}_i(0)T_i(0)|X_i = x]$, and $\mathbb{E}[\mathbf{Z}_i(1)T_i(1)|X_i = x]$ are ϱ times continuously differentiable.

(b) $\mathbb{V}[\mathbf{F}_i(t)|X_i = x]$, with $\mathbf{F}_i(t) := (Y_i(t), T_i(t), \mathbf{Z}_i(t)')'$, $t \in \{0, 1\}$, are continuously differentiable and invertible.

(c) $\mathbb{E}[|T_i(t)|^4|X_i = x]$, $t \in \{0, 1\}$, are continuous.

(d) $\mu_{T-}(x) \neq \mu_{T+}(x)$.

The standard fuzzy RD estimand is

$$\varsigma = \frac{\tau_Y}{\tau_T}, \qquad \tau_Y = \mu_{Y+} - \mu_{Y-}, \qquad \tau_T = \mu_{T+} - \mu_{T-},$$

where recall that we continue to omit the evaluation point $x = \bar{x}$, and we have redefined the potential outcomes and additional covariates to incorporate imperfect treatment compliance. Furthermore, now τ has a subindex highlighting the outcome variable being considered (Y or T), and hence $\tau = \tau_Y$ by definition. See Hahn, Todd and van der Klaauw (2001) and Imbens and Lemieux (2008) for further discussion on identification and interpretation of this estimand.

The standard estimator of ς, without covariate adjustment, is

$$\hat{\varsigma}(h) = \frac{\hat{\tau}_Y(h)}{\hat{\tau}_T(h)}, \qquad \hat{\tau}_V(h) = \mathbf{e}_0'\hat{\boldsymbol{\beta}}_{V+,p}(h) - \mathbf{e}_0'\hat{\boldsymbol{\beta}}_{V-,p}(h),$$

with $V \in \{Y, T\}$, according to (1). Similarly, the covariate-adjusted fuzzy RD estimator is

$$\tilde{\varsigma}(h) = \frac{\tilde{\tau}_Y(h)}{\tilde{\tau}_T(h)}, \qquad \tilde{\tau}_V(h) = \mathbf{e}_0' \tilde{\boldsymbol{\beta}}_{V+,p}(h) - \mathbf{e}_0' \hat{\boldsymbol{\beta}}_{V-,p}(h),$$

with $V \in \{Y, T\}$, according to (2). Our notation makes clear that the fuzzy RD estimators, with or without additional covariates, are simply the ratio of two sharp RD estimators, with or without covariates.

The properties of the standard fuzzy RD estimator $\hat{\varsigma}(h)$ were studied in great detail before, while the covariate-adjusted fuzzy RD estimator $\tilde{\varsigma}(h)$ has not been studied in the literature before. With these preliminaries, we can give the analogue of Lemma 1 for fuzzy RD designs using covariates.

Lemma 3 (Fuzzy RD with Covariates)**.** *Let Assumptions 1, 2, and 3 hold. If* $nh \to \infty$ *and* $h \to 0$*, then*

$$\tilde{\varsigma}(h) \to_{\mathbb{P}} \frac{\tau_Y - [\boldsymbol{\mu}_{Z+} - \boldsymbol{\mu}_{Z-}]' \boldsymbol{\gamma}_Y}{\tau_T - [\boldsymbol{\mu}_{Z+} - \boldsymbol{\mu}_{Z-}]' \boldsymbol{\gamma}_T},$$

where $\boldsymbol{\gamma}_V = (\boldsymbol{\sigma}^2_{Z-} + \boldsymbol{\sigma}^2_{Z+})^{-1} \mathbb{E}[(\mathbf{Z}_i(0) - \boldsymbol{\mu}_{Z-}(X_i))V_i(0) + (\mathbf{Z}_i(1) - \boldsymbol{\mu}_{Z+}(X_i))V_i(1) | X_i = \bar{x}]$ *with* $V \in \{Y, T\}$.

Under the same conditions, when no additional covariates are included, it is well known that $\hat{\varsigma}(h) \to_{\mathbb{P}} \varsigma$. Thus, this lemma clearly shows that both probability limits will coincide under the same sufficient condition as in the sharp RD design: $\boldsymbol{\mu}_{Z-} = \boldsymbol{\mu}_{Z+}$. Therefore, at least asymptotically, a (causal) interpretation for the probability limit of the covariate-adjusted fuzzy RD estimator can be deduced from the corresponding (causal) interpretation for the probability limit of the standard fuzzy RD estimator, whenever the condition $\boldsymbol{\mu}_{Z-} = \boldsymbol{\mu}_{Z+}$ holds.

Since the fuzzy RD estimators are constructed as a ratio of two sharp RD estimators, their asymptotic properties can be characterized by studying the asymptotic properties of the corresponding sharp RD estimators, which have already been analyzed in previous sections. Specifically, the asymptotic properties of covariate-adjusted fuzzy RD estimator $\tilde{\zeta}(h)$ can be characterized by employing the following linear approximation:

$$\tilde{\varsigma}(h) - \varsigma = \mathbf{f}_{\tilde{\varsigma}}'(\tilde{\boldsymbol{\tau}}(h) - \boldsymbol{\tau}) + \epsilon_{\tilde{\varsigma}},$$

with

$$\mathbf{f}_{\tilde{\varsigma}} = \begin{bmatrix} \frac{1}{\tau_T} \\ -\frac{\tau_Y}{\tau_T^2} \end{bmatrix}, \qquad \tilde{\boldsymbol{\tau}}(h) = \begin{bmatrix} \tilde{\tau}_Y(h) \\ \tilde{\tau}_T(h) \end{bmatrix}, \qquad \boldsymbol{\tau} = \begin{bmatrix} \tau_Y \\ \tau_T \end{bmatrix},$$

and where the term $\epsilon_{\tilde{\varsigma}}$ is a quadratic (high-order) error. Therefore, it is sufficient to study the asymptotic properties of the bivariate vector $\tilde{\boldsymbol{\tau}}(h)$ of covariate-adjusted sharp RD estimators, provided that $\epsilon_{\tilde{\varsigma}}$ is asymptotically negligible relative to the linear approximation, which is proven in the supplement. As before, while not necessary for most of our results, we continue to assume that $\boldsymbol{\mu}_{Z-} = \boldsymbol{\mu}_{Z+}$ so the standard RD estimand is recovered by the covariate-adjusted fuzzy RD estimator.

Employing the linear approximation and parallel results as those discussed above for the sharp RD design (now also using T_i as outcome variable as appropriate), it is conceptually straightforward to conduct inference in fuzzy RD designs with covariates. All the same results outlined in the previous section are established for this case: in the supplemental appendix we present MSE expansions, MSE-optimal bandwidth, MSE-optimal point estimators, consistent bandwidth selectors, robust bias-corrected distribution theory and consistent standard errors under either heteroskedasticity or clustering, for the covariate-robust fuzzy RD estimator $\tilde{\varsigma}(h)$. We do not attempt to present these results here because they are notationally cumbersome, with little new conceptual insight. Nevertheless, all these results are implemented in the general purpose software packages for `R` and `Stata` described in Calonico et al. (2016b).

5.2 Kink RD Designs

Our final extension concerns the so-called kink RD designs. See Card, Lee, Pei and Weber (2015) for a discussion on identification and Calonico et al. (2014) for a discussion on estimation and inference, both covering sharp and fuzzy settings without additional covariates. Dong and Lewbel (2015) also study derivative estimation in RD designs, without additional covariates. We briefly outline identification and consistency results when additional covariates are included in kink RD estimation (i.e., derivative estimation at the cutoff), but relegate all other inference results to the supplemental appendix.

To describe the estimands of interest in this context, let $g^{(s)}(x) = \partial^s g(x)/\partial x^s$ for any sufficiently

smooth function $g(\cdot)$. The standard sharp kink RD parameter is (proportional to)

$$\tau_{Y,1} = \mu_{Y+}^{(1)} - \mu_{Y-}^{(1)},$$

while the fuzzy kink RD parameter is

$$\varsigma_1 = \frac{\tau_{Y,1}}{\tau_{T,1}}$$

where $\tau_{T,1} = \mu_{T+}^{(1)} - \mu_{T-}^{(1)}$. In the absence of additional covariates in the RD estimation, these RD treatment effects are estimated by using the local polynomial plug-in estimators:

$$\hat{\tau}_{Y,1}(h) = \mathbf{e}_1'\hat{\boldsymbol{\beta}}_{Y+,p}(h) - \mathbf{e}_1'\hat{\boldsymbol{\beta}}_{Y-,p}(h) \qquad \text{and} \qquad \hat{\varsigma}_1(h) = \frac{\hat{\tau}_{Y,1}(h)}{\hat{\tau}_{T,1}(h)},$$

where $\mathbf{e}_1$ denote the conformable 2nd unit vector (i.e., $\mathbf{e}_1 = (0, 1, 0, 0, \cdots, 0)'$). Therefore, the covariate-adjusted kink RD estimators in sharp and fuzzy settings are

$$\tilde{\tau}_{Y,1}(h) = \mathbf{e}_1'\tilde{\boldsymbol{\beta}}_{Y+,p}(h) - \mathbf{e}_1'\tilde{\boldsymbol{\beta}}_{Y-,p}(h)$$

and

$$\tilde{\varsigma}_1(h) = \frac{\tilde{\tau}_{Y,1}(h)}{\tilde{\tau}_{T,1}(h)}, \qquad \tilde{\tau}_{V,1}(h) = \mathbf{e}_1'\tilde{\boldsymbol{\beta}}_{V+,p}(h) - \mathbf{e}_1'\tilde{\boldsymbol{\beta}}_{V-,p}(h), \qquad V \in \{Y, T\},$$

respectively. The following lemma gives our main identification and consistency results.

Lemma 4 (Kink RD with Covariates)**.** *Let Assumptions 1, 2, and 3 hold. If* $nh \to \infty$ *and* $h \to 0$*, then*

$$\tilde{\tau}_{Y,1}(h) \to_{\mathbb{P}} \tau_{Y,1} - [\boldsymbol{\mu}_{Z+}^{(1)} - \boldsymbol{\mu}_{Z-}^{(1)}]'\boldsymbol{\gamma}_Y$$

and

$$\tilde{\varsigma}_1(h) \to_{\mathbb{P}} \frac{\tau_{Y,1} - [\boldsymbol{\mu}_{Z+}^{(1)} - \boldsymbol{\mu}_{Z-}^{(1)}]'\boldsymbol{\gamma}_Y}{\tau_{T,1} - [\boldsymbol{\mu}_{Z+}^{(1)} - \boldsymbol{\mu}_{Z-}^{(1)}]'\boldsymbol{\gamma}_T},$$

where $\boldsymbol{\gamma}_Y$ *and* $\boldsymbol{\gamma}_T$ *are defined in Lemma 3, and recall that* $\boldsymbol{\mu}_{Z-}^{(1)} = \boldsymbol{\mu}_{Z-}^{(1)}(\bar{x})$ *and* $\boldsymbol{\mu}_{Z+}^{(1)} = \boldsymbol{\mu}_{Z+}^{(1)}(\bar{x})$ *with* $\boldsymbol{\mu}_{Z-}^{(1)}(x) = \partial\boldsymbol{\mu}_{Z-}(x)/\partial x$ *and* $\boldsymbol{\mu}_{Z+}^{(1)}(x) = \partial\boldsymbol{\mu}_{Z+}(x)/\partial x$*.*

As before, in this setting it is well known that $\hat{\tau}_{Y,1}(h) \to_{\mathbb{P}} \tau_{Y,1}$ (sharp kink RD) and $\hat{\varsigma}_1(h) \to_{\mathbb{P}} \varsigma_1$ (fuzzy kink RD), formalizing once again that the estimand when covariates are included is in general different from the standard kink RD estimand without covariates. In this case, a sufficient condition

for the estimands with and without covariates to agree is $\boldsymbol{\mu}_{Z+}^{(1)} = \boldsymbol{\mu}_{Z-}^{(1)}$ for both sharp and fuzzy kink RD designs.

While the above results are in qualitative agreement with the sharp and fuzzy RD cases, and therefore most conclusions transfer directly to kink RD designs, there is one interesting difference concerning the sufficient conditions guaranteeing that both estimands coincide: a sufficient condition now requires $\boldsymbol{\mu}_{Z+}^{(1)} = \boldsymbol{\mu}_{Z-}^{(1)}$. This requirement is not related to the typical falsification test conducted in empirical work, that is, $\boldsymbol{\mu}_{Z+} = \boldsymbol{\mu}_{Z-}$, but rather a different feature of the conditional distributions of the additional covariates given the score—the first derivative of the regression function at the cutoff. Therefore, this finding suggests a new falsification test for empirical work in kink RD designs: testing for a zero sharp *kink* RD treatment effect on "pre-intervention" covariates. For example, this can be done using standard sharp kink RD treatment effect results, using each covariate as outcome variable.

As before, inference results follow the same logic already discussed. Complete details are given in the supplement and fully implemented in the `R` and `Stata` software described by Calonico et al. (2016b).

6 Numerical Results

We now illustrate our methods empirically and present an extensive simulation study conducted to assess the finite sample properties of the covariate-adjusted RD estimator and the associated large sample inference procedures developed in this paper. To conserve space, we only discuss the main findings and relegate details to the supplemental appendix.

6.1 Empirical illustrations

We first illustrate our methods in two empirical applications. First, we re-analyze the effect of Head Start assistance on child mortality in the U.S., which was first studied by Ludwig and Miller (2007). In this application, the unit of observation is the U.S. county, the treatment is receiving technical assistance to apply for Head Start funds, and the running variable is the county-level poverty index constructed in 1965. The RD design arises because the treatment was given only to counties whose poverty index was $\bar{x} = 59.1984$ or above, a cutoff that was chosen to ensure that the 300th poorest counties received the treatment. The outcome of interest is the child mortality rate

(for children of age five to nine) due to causes affected by Head Start's health services component.

Next, we revisit the effect of school improvements on student language test scores in Chile, first studied by Chay, McEwan and Urquiola (2005). The unit of observation is the school, the treatment is receiving the school improvement program P-900, which was assigned in 1990 based on an index (the running variable) constructed from previous school-level test scores. The outcome we study is school-level language score gain between 1988 and 1992.

We compare the standard RD estimator to the covariate-adjusted RD estimator employing heteroskedasticity-robust nearest neighbor variance estimation for both applications. In the Head Start application, the additional regressors $\mathbf{Z}_i$ are nine county-level covariates from the pre-intervention 1960 U.S. Census: total population, percentage of black and urban population, and levels and percentages of population in three age groups (children age 3 to 5, children age 14 to 17, and adults older than 25). In the education application, the additional covariates are seven binary variables indicating the school's region group (Chile's 13 administrative regions were divided into seven groups, with schools in each region group facing a different cutoff value).

Table 1 presents the main results. In each application's panel in that table, the first row reports the RD local-linear ($p = 1$) point estimate using the corresponding MSE-optimal bandwidth h (depending on the column). The next three rows report 95% robust bias-corrected confidence intervals, the percentage length change of the covariate-adjusted confidence interval relative to the unadjusted one, and the p-value associated with a hypothesis of zero RD treatment effect. These three rows appear twice, first when h for the RD point estimator and b for the bias estimator are chosen separately (in this case, $\rho = h/b$ is unrestricted), and then when $b = h$ (in this case, $\rho = 1$). Finally, the last two rows in each application's panel report, respectively, the two bandwidths and the number of observations to the left and to the right of the cutoff with $X_i \in [\bar{x} - h, \bar{x} + h]$.

The columns in Table 1 correspond to different RD approaches. The first two columns employ the MSE-optimal bandwidth without covariates—the first column reports inference results without covariates while the second column presents covariate-adjusted inference. Thus, the second column is intended to mimic a common practice among practitioners, who sometimes estimate the MSE-optimal bandwidth without covariates and then include covariates in the estimation and inference using the same observations (i.e., keeping the bandwidth choice fixed). It follows that, in the second column, the h and b bandwidths used are the MSE-optimal bandwidths without covariates, and

therefore the point estimator in this second column is no longer MSE-optimal, since the optimal bandwidths are not used; however, the confidence intervals and p-values are still valid because the optimal bandwidths with and without covariates have the same rate of decay. The third column reports covariate-adjusted RD estimation using the asymptotic approximations derived in the previous sections; in this column, both bandwidths are chosen according to the MSE-optimal formulas with covariates. In all cases, we use triangular kernel weights and nearest neighbor residual estimates. Employing other kernels or variance estimators give very similar empirical results.

Our empirical findings are quite consistent across applications: employing covariate-adjusted RD inference leads to precision improvements while the point estimators remain stable. In the Head Start application, the point estimator ranges from -2.41 to -2.51, an effect that is statistically different from zero at 5% significance level in all cases. As should be expected when the additional covariates are truly pre-determined, including covariates does not substantially alter the point estimates. (We also implemented "placebo tests" on the additional covariates and found, as expected, no statistical evidence of RD treatment effects.) Including additional covariates in this application can lead to sizable efficiency gains: for example, when both h and b are estimated ($\rho = h/b$ unrestricted), adding covariates within the MSE-optimal h without covariates ($h = 6.81$) results in a 8.25% reduction in the length of the 95% confidence interval (column 2), as this confidence interval shrinks from $(-5.49, -0.10)$ to $(-5.37, -0.45)$. The length of the confidence interval is even shorter when both bandwidths are chosen optimally using covariates (column 3), one of the novel results in this paper, with a length reduction of roughly 10%.

In the case of the education data, we also find that including additional covariates does not affect the point estimators, while providing some efficiency improvements. In this case, the point estimates range from 3.45 to 3.49, and the confidence interval length shrinks approximately 3% to 5% depending on the case considered.

Our empirical results suggest that including pre-intervention covariates can be empirically useful in real RD applications, thereby illustrating the usefulness of the new methods developed in this paper.

6.2 Simulation Evidence

We now illustrate our methods using simulated data. We consider four data generating processes constructed using the data of Lee (2008): all parameters are obtained from the real data unless explicitly noted otherwise. This model has been used extensively before, see Imbens and Kalyanaraman (2012) and Calonico et al. (2014, 2016a), among many others. The additional covariate included is previous democratic vote share, and the four models are distinguished by the importance of this covariate: (i) in Model 1, the covariate is irrelevant; (ii) in Model 2 it enters the conditional expectation of the potential outcomes $\mathbb{E}[Y_i(t)|X_i = x, \mathbf{Z}_i(t)]$, $t \in \{0, 1\}$, according to the real data; (iii) Model 3 takes Model 2 but sets the residual correlation between the outcome and covariate to zero; (iv) Model 4 takes Model 2 but doubles the residual correlation between the outcome and covariate equations. Note that Models 3 and 4 do not imply $\mathbb{C}ov[Y_i(t), \mathbf{Z}_i(t)|X_i = x] = 0$, $t \in \{0, 1\}$. The constructions allowed $\mathbb{E}[Y_i(t)|X_i = x, \mathbf{Z}_i(t)]$ to have different coefficients on each side of the cutoff, while the conditional expectation of the potential covariates $\mathbb{E}[\mathbf{Z}_i(t)|X_i = x]$, $t \in \{0, 1\}$, were constructed assuming they are continuous at the cutoff (but still with different coefficients on either side otherwise). Therefore, our covariate-adjusted RD estimator will be "misspecified" when viewed as a local weighted least-square fit. The exact details of our Monte Carlo experiment are given in the supplemental appendix to conserve space.

We use a sample of size $n = 1,000$ and consider $5,000$ replications. We compare the standard RD estimator ($\hat{\tau}$) and the covariate-adjusted RD estimator ($\tilde{\tau}$), with both infeasible and data-driven MSE-optimal and CER-optimal bandwidth choices. To analyze the performance of our inference procedures, we report average bias of the point estimators and average coverage rate and interval length of nominal 95% confidence intervals. In addition, we also explore the performance of our data-driven bandwidth selectors by reporting some of their main statistical features, such as mean, median, standard deviation, across the $5,000$ replications. We report only one table that presents estimates using triangular kernel and nearest neighbor (NN) heteroskedasticity-robust variance estimators; complete details and results are in the supplement.

The numerical results are given in Tables 2 and 3. All findings are highly consistent with our large sample theory. Table 2 shows that including covariates can improve both MSE and interval length, sometimes dramatically, and moreover, the gains are in line with our theory: the gains are largest in Model 4 with the amplified residual correlation and least in Model 3 when that channel

is shut down. The results for Model 1 show that including an irrelevant covariate hardly changes empirical results and conclusions. Finally, Table 3 shows that the data-driven bandwidth selectors also work reasonably well.

7 Conclusion

We provided a formal framework for identification, estimation, and inference in RD designs when covariates are included in the estimation. We augmented the standard local polynomial estimator with covariates entering in an additive-separable, linear-in-parameters way. We showed that under minimal additional smoothness assumptions, the resulting covariate-adjusted RD estimator remains consistent for the standard RD treatment effect if the covariate adjustment is restricted to be equivalent above and below the cutoff. Furthermore, this estimator can achieve substantial efficiency gains relative to the unadjusted RD estimator. We also showed that relaxing the latter restriction with the inclusion of treatment-covariate interactions leads to a point estimator that is not generally consistent for the standard RD parameter of interest.

We also provided new MSE expansions, several optimal bandwidth choices and optimal point estimators, robust nonparametric inference procedures based on bias-correction, and heteroskedasticity-consistent and cluster-robust standard errors. All these results were obtained for sharp, fuzzy, and kink RD designs. Finally, we illustrated the practical implications of our results using two empirical applications and simulated data, and showed that including pre-intervention covariates in RD designs can lead to useful improvements in precision. All the results presented in this paper are implemented in companion general purpose `R` and `Stata` software packages.

8 Appendix: Sharp RD Design Main Formulas

We give a very succinct account of the main expressions for sharp RD designs, which were omitted in the main paper to avoid overwhelming notation. A detailed treatment of this and all other RD designs cases is given in the lengthy supplemental appendix.

Let $\mathbf{R}_p(h) = [(\mathbf{r}_p((X_1 - \bar{x})/h), \cdots, \mathbf{r}_p((X_n - \bar{x})/h))']$ be the $n \times (1+p)$ design matrix, and $\mathbf{K}_-(h) = \text{diag}(\mathbb{1}(X_i < \bar{x})K_h(X_i - \bar{x}) : i = 1, 2, \cdots, n)$ and $\mathbf{K}_+(h) = \text{diag}(\mathbb{1}(X_i \geq \bar{x})K_h(X_i - \bar{x}) : i = 1, 2, \cdots, n)$ be the $n \times n$ weighting matrices for control and treatment units, respectively. We

also define $\boldsymbol{\mu}_{S-}^{(a)} := (\mu_{Y-}^{(a)}, \boldsymbol{\mu}_{Z-}^{(a)\prime})'$, $\boldsymbol{\mu}_{S+}^{(a)} := (\mu_{Y+}^{(a)}, \boldsymbol{\mu}_{Z+}^{(a)\prime})'$, $a \in \mathbb{Z}_+$, and $\boldsymbol{\sigma}_{S-}^2 := \mathbb{V}[\mathbf{S}_i(0)|X_i = \bar{x}]$ and $\boldsymbol{\sigma}_{S+}^2 := \mathbb{V}[\mathbf{S}_i(1)|X_i = \bar{x}]$, recall that $\mathbf{S}_i(t) = (Y_i(t), \mathbf{Z}_i(t))'$, $t \in \{0, 1\}$. Let $\mathbf{e}_\nu$ denote a conformable $(1+\nu)$-th unit vector. Finally, recall that $\mathbf{s}(h) = (1, -\tilde{\boldsymbol{\gamma}}_Y(h))'$ and $\mathbf{s} = (1, -\boldsymbol{\gamma}_Y)'$.

The pre-asymptotic bias $\mathcal{B}_{\tilde{\tau}}(h) = \mathcal{B}_{\tilde{\tau}+}(h) - \mathcal{B}_{\tilde{\tau}-}(h)$ and its asymptotic counterpart $\mathcal{B}_{\tilde{\tau}} := \mathcal{B}_{\tilde{\tau}+} - \mathcal{B}_{\tilde{\tau}-}$ are characterized by

$$\begin{aligned}
\mathcal{B}_{\tilde{\tau}-}(h) &:= \mathbf{e}_0'\boldsymbol{\Gamma}_{-,p}^{-1}(h)\boldsymbol{\vartheta}_{-,p}(h)\frac{\mathbf{s}'\boldsymbol{\mu}_{S-}^{(p+1)}}{(p+1)!} \;\to_{\mathbb{P}}\; \mathcal{B}_{\tilde{\tau}-} := \mathbf{e}_0'\boldsymbol{\Delta}_{p,-}\frac{\mathbf{s}'\boldsymbol{\mu}_{S-}^{(p+1)}}{(p+1)!} \\
\mathcal{B}_{\tilde{\tau}+}(h) &:= \mathbf{e}_0'\boldsymbol{\Gamma}_{+,p}^{-1}(h)\boldsymbol{\vartheta}_{+,p}(h)\frac{\mathbf{s}'\boldsymbol{\mu}_{S+}^{(p+1)}}{(p+1)!} \;\to_{\mathbb{P}}\; \mathcal{B}_{\tilde{\tau}+} := \mathbf{e}_0'\boldsymbol{\Delta}_{p,+}\frac{\mathbf{s}'\boldsymbol{\mu}_{S+}^{(p+1)}}{(p+1)!}
\end{aligned}$$

where, with the (slightly abusive) notation $\mathbf{v}^k = (v_1^k, v_2^k, \cdots, v_n^k)'$, $\boldsymbol{\iota}_n = (1, \cdots, 1)' \in \mathbb{R}^n$, $\boldsymbol{\Gamma}_{-,p}(h) = \mathbf{R}_p(h)'\mathbf{K}_-(h)\mathbf{R}_p(h)/n$ and $\boldsymbol{\vartheta}_{-,p}(h) = \mathbf{R}_p(h)'\mathbf{K}_-(h)(\mathbf{X} - \bar{x}\boldsymbol{\iota}_n/h)^{p+1}/n$, $\boldsymbol{\Gamma}_{+,p}(h)$ and $\boldsymbol{\vartheta}_{+,p}(h)$ defined analogously after replacing $\mathbf{K}_-(h)$ with $\mathbf{K}_+(h)$, and

$$\begin{aligned}
\boldsymbol{\Delta}_{p,-} &:= \left(\int_{-\infty}^0 \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)du\right)^{-1}\left(\int_{-\infty}^0 \mathbf{r}_p(u)u^{1+p}K(u)du\right), \\
\boldsymbol{\Delta}_{p,+} &:= \left(\int_0^\infty \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)du\right)^{-1}\left(\int_0^\infty \mathbf{r}_p(u)u^{1+p}K(u)du\right).
\end{aligned}$$

The pre-asymptotic variance $\mathcal{V}_{\tilde{\tau}}(h) = \mathcal{V}_{\tilde{\tau}-}(h) + \mathcal{V}_{\tilde{\tau}+}(h)$ and its asymptotic counterpart $\mathcal{V}_{\tilde{\tau}} := \mathcal{V}_{\tilde{\tau}-} + \mathcal{V}_{\tilde{\tau}+}$ are characterized by

$$\begin{aligned}
\mathcal{V}_{\tilde{\tau}-}(h) &:= [\mathbf{s}' \otimes \mathbf{e}_0'\mathbf{P}_{-,p}(h)]\boldsymbol{\Sigma}_{S-}[\mathbf{s} \otimes \mathbf{P}_{-,p}(h)\mathbf{e}_0] \;\to_{\mathbb{P}}\; \mathcal{V}_{\tilde{\tau}-} := \frac{\mathbf{s}'\boldsymbol{\sigma}_{S-}^2\mathbf{s}}{f}\mathbf{e}_0'\boldsymbol{\Lambda}_{p,-}\mathbf{e}_0 \\
\mathcal{V}_{\tilde{\tau}+}(h) &:= [\mathbf{s}' \otimes \mathbf{e}_0'\mathbf{P}_{+,p}(h)]\boldsymbol{\Sigma}_{S+}[\mathbf{s} \otimes \mathbf{P}_{+,p}(h)\mathbf{e}_0] \;\to_{\mathbb{P}}\; \mathcal{V}_{\tilde{\tau}+} := \frac{\mathbf{s}'\boldsymbol{\sigma}_{S+}^2\mathbf{s}}{f}\mathbf{e}_0'\boldsymbol{\Lambda}_{p,+}\mathbf{e}_0
\end{aligned}$$

where $\mathbf{P}_{-,p}(h) = \sqrt{h}\boldsymbol{\Gamma}_{-,p}^{-1}(h)\mathbf{R}_p(h)'\mathbf{K}_-(h)/\sqrt{n}$ and $\mathbf{P}_{+,p}(h) = \sqrt{h}\boldsymbol{\Gamma}_{+,p}^{-1}(h)\mathbf{R}_p(h)'\mathbf{K}_+(h)/\sqrt{n}$, and

$$\begin{aligned}
\boldsymbol{\Lambda}_{p,-} &:= \left(\int_{-\infty}^0 \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)du\right)^{-1}\left(\int_{-\infty}^0 \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)^2du\right)\left(\int_{-\infty}^0 \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)du\right)^{-1}, \\
\boldsymbol{\Lambda}_{p,+} &:= \left(\int_0^\infty \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)du\right)^{-1}\left(\int_0^\infty \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)^2du\right)\left(\int_0^\infty \mathbf{r}_p(u)\mathbf{r}_p(u)'K(u)du\right)^{-1}.
\end{aligned}$$

To construct pre-asymptotic estimates of the bias terms, we replace the only unknowns, $\boldsymbol{\mu}_{S-}^{(p+1)}$ and $\boldsymbol{\mu}_{S+}^{(p+1)}$, by q-th order $(p < q)$ local polynomial estimates thereof, using the preliminary band-

width b. This leads to the pre-asymptotic feasible bias estimate $\tilde{\mathcal{B}}_{\tilde{\tau}}(b) := \tilde{\mathcal{B}}_{\tilde{\tau}+}(b) - \tilde{\mathcal{B}}_{\tilde{\tau}-}(b)$ with

$$\tilde{\mathcal{B}}_{\tilde{\tau}-}(b) := \mathbf{e}_0'\boldsymbol{\Gamma}_{-,p}^{-1}(h)\boldsymbol{\vartheta}_{-,p}(h)\frac{\mathbf{s}(h)'\tilde{\boldsymbol{\mu}}_{S-,q}^{(p+1)}(b)}{(p+1)!} \quad \text{and} \quad \tilde{\mathcal{B}}_{\tilde{\tau}+}(b) := \mathbf{e}_0'\boldsymbol{\Gamma}_{+,p}^{-1}(h)\boldsymbol{\vartheta}_{+,p}(h)\frac{\mathbf{s}(h)'\tilde{\boldsymbol{\mu}}_{S+,q}^{(p+1)}(b)}{(p+1)!}$$

where $\tilde{\boldsymbol{\mu}}_{S-,q}^{(p+1)}(b)$ and $\tilde{\boldsymbol{\mu}}_{S+,q}^{(p+1)}(b)$ collect the q-th order local polynomial estimates of the $(p+1)$-th derivatives using as outcomes each of the variables in $\mathbf{S}_i = (Y_i, \mathbf{Z}_i')'$ for control and treatment units, that is, as in (1). Therefore, the bias-corrected covariate-adjusted sharp RD estimator is

$$\tilde{\tau}^{\mathsf{bc}}(h) = \frac{1}{\sqrt{nh}}[\mathbf{s}(h)' \otimes \mathbf{e}_0'(\mathbf{P}_{+,p}^{\mathsf{bc}}(h,b) - \mathbf{P}_{-,p}^{\mathsf{bc}}(h,b))]\mathbf{S},$$

with $\mathbf{S} = (\mathbf{Y}, \mathrm{vec}(\mathbf{Z})')'$, $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)'$, and

$$\mathbf{P}_{-,p}^{\mathsf{bc}}(h,b) = \sqrt{h}\boldsymbol{\Gamma}_{-,p}^{-1}(h)\left[\mathbf{R}_p(h)'\mathbf{K}_-(h) - \rho^{1+p}\boldsymbol{\vartheta}_{-,p}(h)\mathbf{e}_{p+1}'\boldsymbol{\Gamma}_{-,q}^{-1}(b)\mathbf{R}_q(b)'\mathbf{K}_-(b)\right]/\sqrt{n},$$

$$\mathbf{P}_{+,p}^{\mathsf{bc}}(h,b) = \sqrt{h}\boldsymbol{\Gamma}_{+,p}^{-1}(h)\left[\mathbf{R}_p(h)'\mathbf{K}_+(h) - \rho^{1+p}\boldsymbol{\vartheta}_{+,p}(h)\mathbf{e}_{p+1}'\boldsymbol{\Gamma}_{+,q}^{-1}(b)\mathbf{R}_q(b)'\mathbf{K}_+(b)\right]/\sqrt{n},$$

where $\tilde{\mathbf{P}}_{-,p}^{\mathsf{bc}}(h,b)$ and $\tilde{\mathbf{P}}_{-,p}^{\mathsf{bc}}(h,b)$ are directly computable from observed data, given the choices of bandwidth h and b, with $\rho = h/b$, and the choices of polynomial order p and q, with $p < q$.

The exact form of the (pre-asymptotic) heteroskedasticity-robust or cluster-robust variance estimator follows directly from the formulas above. All other details such preliminary bandwidth selection, plug-in data-driven MSE-optimal bandwidth estimation, and other extensions and results, are given in the supplemental appendix.

References

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

Abadie, A., and Imbens, G. W. (2008), "Estimation of the Conditional Variance in Paired Experiments," *Annales d'Economie et de Statistique*, 175–187.

Angrist, J., and Rokkanen, M. (2015), "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association*, 110, 1331–1344.

Armstrong, T. B., and Kolesar, M. (2015), "A Simple Adjustment for Bandwidth Snooping," arXiv:1412.0267.

Bartalotti, O., and Brummet, Q. (2016), "Regression Discontinuity Designs with Clustered Data: Mean Square Error and Bandwidth Choice," in *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, eds. M. D. Cattaneo and J. C. Escanciano, Emerald Group Publishing*, to appear*.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016a), "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," arXiv:1508.02973.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2016b), "`rdrobust`: Software for Regression Discontinuity Designs," working paper, University of Michigan.

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

Cameron, A. C., and Miller, D. L. (2015), "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50, 317–372.

Canay, I. A., and Kamat, V. (2015), "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," CeMMAP working paper CWP27/15.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015), "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica*, 83, 2453–2483.

Cattaneo, M. D., Frandsen, B., and Titiunik, R. (2015), "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," *Journal of Causal Inference*, 3, 1–24.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2016), "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," working paper, University of Michigan.

Chay, K. Y., McEwan, P. J., and Urquiola, M. (2005), "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools," *American Economic Review*, 95, 1237–1258.

de la Cuesta, B., and Imai, K. (2016), "Misunderstandings about the Regression Discontinuity Design in the Study of Close Elections," *Annual Review of Political Science*, forthcoming, 19.

Dong, Y., and Lewbel, A. (2015), "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Review of Economics and Statistics*, 97, 1081–1092.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, New York: Chapman & Hall/CRC.

Gelman, A., and Imbens, G. W. (2014), "Why High-Order Polynomials Should Not be Used in Regression Discontinuity Designs," NBER working paper 20405.

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

Imbens, G., and Lemieux, T. (2008), "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635.

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

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

Kamat, V. (2015), "On Nonparametric Inference in the Regression Discontinuity Design," arXiv:1505.06483.

Keele, L. J., Titiunik, R., and Zubizarreta, J. (2015), "Enhancing a Geographic Regression Discontinuity Design Through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout," *Journal of the Royal Statistical Society: Series A*, 178, 223–239.

Lee, D. S. (2008), "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, 142, 675–697.

Lee, D. S., and Lemieux, T. (2010), "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

Long, J. S., and Ervin, L. H. (2000), "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54, 217–224.

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

MacKinnon, J. G. (2012), "Thirty years of heteroskedasticity-robust inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, eds. X. Chen and N. R. Swanson, Springer.

Muller, H.-G., and Stadtmuller, U. (1987), "Estimation of Heteroscedasticity in Regression Analysis," *The Annals of Statistics*, 15, 610–625.

Porter, J. (2003), "Estimation in the Regression Discontinuity Model," working paper, University of Wisconsin.

Skovron, C., and Titiunik, R. (2016), "A Practical Guide to Regression Discontinuity Designs in Political Science," working paper, University of Michigan.

Table 1: Empirical Illustrations

MSE-optimal bandwidths:	**not using covariates**		**using covariates**
	Standard	Cov-adjusted	Cov-adjusted
Head Start Data			
RD treatment effect	−2.41	−2.51	−2.47
Inference with $\rho = h/b$ unrestricted			
Robust 95% CI	[−5.46 , −0.10]	[−5.37 , −0.45]	[−5.21 , −0.37]
CI length change (%)		−8.25	−9.76
Robust p-value	0.042	0.021	0.024
Inference with $\rho = h/b = 1$			
Robust 95% CI	[−6.41 , −1.09]	[−6.64 , −1.46]	[−6.54 , −1.39]
CI length change (%)		−2.86	−3.23
Robust p-value	0.006	0.002	0.003
$h \mid b$	6.81 | 10.72	6.81 | 10.72	6.98 | 11.64
$n_- \mid n_+$	234 | 180	234 | 180	240 | 184
Education Data			
RD treatment effect	3.45	3.42	3.49
Inference with $\rho = h/b$ unrestricted			
Robust 95% CI	[1.53 , 6.19]	[1.56 , 6.02]	[1.61 , 6.21]
CI length change (%)		−4.19	−1.31
Robust p-value	0.001	0.001	0.001
Inference with $\rho = h/b = 1$			
Robust 95% CI	[1.45 , 7.36]	[1.21 , 6.80]	[1.10 , 6.87]
CI length change (%)		−5.43	−2.47
Robust p-value	0.004	0.005	0.007
$h \mid b$	3.62 | 6.64	3.62 | 6.64	3.36 | 6.11
$n_- \mid n_+$	385 | 280	385 | 280	362 | 259

Notes:

(i) All estimates are computed using a triangular kernel and nearest neighbor heteroskedasticity-robust variance estimators.

(ii) Columns under "Standard" and "Cov-adjusted" correspond to, respectively, standard and covariate-adjusted RD estimation and inference methods, given a choice of bandwidths.

(iii) Bandwidths used (h and b) are data-driven MSE-optimal for either standard RD estimator or covariated-adjusted RD estimator (depending on the group of columns). Specifically, in the first two columns the bandwidths are selected to be MSE-optimal for $\hat{\tau}$ (standard RD estimation), while in the third column the bandwidths are selected to be MSE-optimal for $\tilde{\tau}$ (covariate-adjusted RD estimation).

Table 2: Simulation Results (MSE, Bias, Empirical Coverage and Interval Length)

	$\hat{\tau}$				$\tilde{\tau}$				**Change** (%)			
	$\sqrt{MSE}$	Bias	EC	IL	$\sqrt{MSE}$	Bias	EC	IL	$\sqrt{MSE}$	Bias	EC	IL
Model 1												
MSE-POP	0.045	0.012	0.938	0.199	0.045	0.012	0.934	0.198	0.2	0.1	−0.4	−0.6
MSE-EST	0.045	0.018	0.924	0.171	0.045	0.018	0.927	0.171	0.0	−1.0	0.3	−0.2
CER-POP	0.052	0.006	0.934	0.242	0.052	0.006	0.929	0.240	0.4	1.2	−0.5	−0.9
CER-EST	0.049	0.010	0.940	0.207	0.049	0.010	0.933	0.206	0.5	−1.5	−0.7	−0.5
Model 2												
MSE-POP	0.047	0.013	0.935	0.213	0.041	0.008	0.941	0.185	−13.5	−33.6	0.6	−13.4
MSE-EST	0.048	0.017	0.929	0.188	0.041	0.011	0.932	0.163	−15.1	−34.8	0.3	−13.4
CER-POP	0.054	0.006	0.933	0.258	0.048	0.004	0.931	0.223	−11.7	−34.1	−0.2	−13.6
CER-EST	0.053	0.009	0.941	0.227	0.046	0.006	0.940	0.196	−13.1	−34.1	−0.2	−13.5
Model 3												
MSE-POP	0.044	0.013	0.935	0.200	0.043	0.010	0.938	0.193	−3.3	−19.6	0.3	−3.5
MSE-EST	0.046	0.017	0.926	0.177	0.043	0.014	0.929	0.170	−5.5	−17.2	0.3	−4.0
CER-POP	0.051	0.006	0.933	0.243	0.050	0.005	0.930	0.234	−1.8	−20.9	−0.3	−3.8
CER-EST	0.050	0.009	0.939	0.213	0.048	0.008	0.939	0.205	−4.0	−16.8	0.0	−4.2
Model 4												
MSE-POP	0.050	0.013	0.938	0.225	0.035	0.007	0.938	0.160	−29.3	−46.6	0.1	−28.8
MSE-EST	0.051	0.017	0.931	0.199	0.035	0.008	0.938	0.142	−30.5	−52.1	0.8	−28.4
CER-POP	0.058	0.006	0.934	0.273	0.042	0.003	0.926	0.194	−27.6	−43.6	−0.9	−29.0
CER-EST	0.056	0.009	0.942	0.240	0.040	0.005	0.934	0.171	−27.9	−51.2	−0.8	−28.5

Notes:

(i) All estimators are computed using the triangular kernel, NN variance estimation, and two bandwidths (h and b).
(ii) Columns $\hat{\tau}$ and $\tilde{\tau}$ correspond to, respectively, standard RD estimation and covariate-adjusted RD estimation; columns "$\sqrt{MSE}$" report the square root of the mean square error of point estimator; columns "Bias" report average bias relative to target population parameter; and columns "EC" and "IL" report, respectively, empirical coverage and interval length of robust bias-corrected 95% confidence intervals.
(iii) Rows correspond to bandwidth method used to construct the estimator and inference procedures. Rows "MSE-POP" and "MSE-EST" correspond to, respectively, procedures using infeasible population and feasible data-driven MSE-optimal bandwidths (without or with covariate adjustment depending on the column). Rows "CER-POP" and "CER-EST" correspond to, respectively, procedures using infeasible population and feasible data-driven CER-optimal bandwidths (without or with covariate adjustment depending on the column).

Table 3: Simulation Results (Data-Driven Bandwidth Selectors)

	Pop.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
Model 1								
$\hat{\mathfrak{h}}_{\hat{\tau}}$	0.144	0.097	0.168	0.191	0.195	0.217	0.338	0.041
$\tilde{\mathfrak{h}}_{\tilde{\tau}}$	0.144	0.094	0.166	0.189	0.194	0.217	0.338	0.040
Model 2								
$\hat{\mathfrak{h}}_{\hat{\tau}}$	0.156	0.092	0.170	0.193	0.198	0.222	0.335	0.041
$\tilde{\mathfrak{h}}_{\tilde{\tau}}$	0.158	0.095	0.171	0.197	0.201	0.227	0.336	0.042
Model 3								
$\hat{\mathfrak{h}}_{\hat{\tau}}$	0.156	0.091	0.169	0.193	0.197	0.221	0.335	0.040
$\tilde{\mathfrak{h}}_{\tilde{\tau}}$	0.154	0.095	0.170	0.194	0.198	0.223	0.334	0.041
Model 4								
$\hat{\mathfrak{h}}_{\hat{\tau}}$	0.156	0.093	0.170	0.194	0.198	0.223	0.334	0.041
$\tilde{\mathfrak{h}}_{\tilde{\tau}}$	0.161	0.088	0.172	0.199	0.203	0.231	0.336	0.043

Notes:

(i) All estimators are computed using the triangular kernel, and NN variance estimation.
(ii) Column "Pop." reports target population bandwidth, while the other columns report summary statistics of the distribution of feasible data-driven estimated bandwidths.
(iii) Rows $\hat{\mathfrak{h}}_{\hat{\tau}}$ and $\tilde{\mathfrak{h}}_{\tilde{\tau}}$ corresponds to feasible data-driven MSE-optimal bandwidth selectors without and with covariate adjustment, respectively.