REGRESSION DISCONTINUITY DESIGNS USING COVARIATES

Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik*

Abstract—We study regression discontinuity designs when covariates are included in the estimation. We examine local polynomial estimators that include discrete or continuous covariates in an additive separable way, but without imposing any parametric restrictions on the underlying population regression functions. We recommend a covariate-adjustment approach that retains consistency under intuitive conditions and characterize the potential for estimation and inference improvements. We also present new covariate-adjusted mean-squared error expansions and robust bias-corrected inference procedures, with heteroskedasticity-consistent and cluster-robust standard errors. We provide an empirical illustration and an extensive simulation study. All methods are implemented in R and Stata software packages.

I.   Introduction

THE regression discontinuity (RD) design is widely used in economics, political science, and many other social, behavioral, biomedical, and statistical sciences. Within the causal inference framework, the RD design is considered to be one of the most credible nonexperimental strategies because it relies on weak and easy-to-interpret nonparametric identifying assumptions, which permit flexible and robust estimation and inference for local 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 group). For literature reviews and practical introductions, see Imbens & Lemieux, 2008; Lee & Lemieux, 2010; Cattaneo & Escanciano, 2017; Cattaneo, Titiunik, & Vazquez-Bare, 2017; and Cattaneo, Idrobo, & Titiunik, forthcoming.

Nonparametric identification of the RD treatment effect typically relies on continuity assumptions, which motivate nonparametric local polynomial methods tailored to flexibly approximate, above and below the cutoff, the unknown conditional mean function of the outcome variable given the score. In practice, researchers often choose a local linear polynomial and perform estimation using weighted linear least squares, giving higher weights to observations close to the cutoff. These estimates are then used to assess whether there is a discontinuity in levels, derivatives, or ratios thereof, at the cutoff. If present, this discontinuity is interpreted as some average response to the treatment (assignment) at the cutoff, depending on the setting and assumptions under consideration.

When practitioners employ weighted least squares to estimate RD effects, they often augment their estimation models with additional predetermined covariates such as demographic or socioeconomic characteristics of the units. Despite the pervasiveness of this practice, there has been little formal study of the consequences of covariate adjustment for identification, estimation, and inference of RD effects under standard smoothness conditions and when employing standard nonparametric local polynomial estimators (e.g., local linear regression). This has led to the proliferation of ad hoc covariate-adjustment practices that at best reduce the transparency of the estimation and at worst result in generally incomparable, or even inconsistent, estimators. We provide results that formalize and justify covariate adjustment in local polynomial analysis of RD designs, offering valid and practical estimation and inference procedures.

We augment the standard local polynomial regression to allow covariate adjustment in an additive separable, linear-in-parameters way, following very closely the estimation models typically employed by applied researchers. Our approach allows for the inclusion of continuous, discrete, and mixed regressors and does not require additional smoothing methods such as choosing other bandwidths or kernels, or parametric functional form assumptions. Importantly, although we consider estimation models where the covariates enter linearly (in parameters), this is purely a local linear projection and is in no way an assumption about the functional form of the underlying regression functions. In other words, we study the limiting behavior of local linear least squares estimators of regression functions in a fully nonparametric sense. Within this framework, there does not appear to be broad consensus in applied work as to how exactly the covariates should enter the model. We 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 or poorly behaved RD estimators. Thus, our formal results offer concrete guidelines for applied researchers who want to employ covariate-adjusted local regression methods that were previously unavailable.

We advocate the use of a simple covariate-adjusted RD estimator that imposes the same adjustment above and below the cutoff. This estimator is easy to implement and is consistent for the standard RD treatment effect under the mild,

intuitive condition that the treatment has no mean effect on the covariates at the cutoff. For example, in the sharp RD design, the only requirement is that the covariates have equal conditional expectation limits from above and below the cutoff, which is often conceived and presented as a falsification or "placebo" test in RD empirical studies (Lee, 2008). For kink RD designs, we show that additional conditions on the covariates are needed, suggesting new specification tests for empirical work (all details are provided in the supplemental appendix). The requirement of "balanced" covariates at the cutoff is the most natural and practically relevant sufficient condition. We also show that employing other covariate-adjusted estimators may lead to inconsistency or invalid inference, or both.

We offer a complete large sample analysis of our recommended covariate-adjusted RD estimator, including novel mean squared error (MSE) expansions, MSE-optimal bandwidths (with consistent data-driven implementations), 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. We characterize precisely the potential for efficiency gains, which are guaranteed when the best linear effect of the additional covariates on the outcome, at the cutoff, is equal for both control and treatment groups. These results have immediate practical use in any RD analysis and aid in interpreting prior results. We illustrate our methods numerically by revisiting the data of Ludwig and Miller (2007) to reanalyze the effect of Head Start on child mortality and with an extensive simulation study. We also provide `Stata` and `R` packages that implement our methods (Calonico et al., 2017).

Our framework for covariate adjustment is best suited for settings where the inclusion of covariates is intended to increase the precision of the RD treatment effect estimator, in the same spirit as (preintervention) covariates are often included in the analysis of randomized experiments. We require that the treatment have no effect on the covariates at the cutoff, and that the conditional expectations of potential outcomes and covariates given the score be continuous at the cutoff. We impose no restrictions on the "long" conditional expectation of potential outcomes given the score and the covariates. These conditions are in perfect agreement with nonparametric identification results in the RD literature and with standard regression adjustment arguments in the context of randomized experiments.

An alternative motivation for the inclusion of covariates in RD analysis is to increase the plausibility of the design in cases where researchers fear that the potential outcomes are discontinuous at the cutoff; in these cases, the inclusion of covariates is intended to restore identification of some RD parameter. This second motivation generally requires either parametric assumptions on the regression functions to enable extrapolation or other design assumptions that redefine the parameter of interest in a fixed local neighborhood rather than at the cutoff point. In this context, our estimation and inference results can continue to be used under stronger assumptions. For example, if the covariate specification in the "long" regression functions is assumed correct within a neighborhood around the cutoff, covariate adjustment leads to identification, estimation, and inference of the corresponding RD treatment effect, as it occurs in standard parametric linear regression settings. However, our main goal here is on nonparametric identification, estimation, and inference of the canonical RD parameter at the cutoff, which is assumed to be well defined from the start. We plan to investigate covariate adjustment for identification of other RD treatment effects in future work.

Our paper contributes to the large and still rapidly expanding methodological literature on RD designs; we stay away from summarizing this literature due to space constraints and refer readers to the references cited in our opening paragraph. From the specific perspective of covariate adjustment, our paper is related to two strands of the literature. First, a portion of the RD literature reinterprets the data as being "as good as randomized" within a small window around the cutoff, though this requires stronger conditions than just continuity or smoothness of conditional expectations as we maintain here (Lee, 2008; Cattaneo, Frandsen, & Titiunik, 2015; Cattaneo, Titiunik, & Vazquez-Bare, 2017; Sekhon & Titiunik, 2017). Adopting a local randomization perspective, our work is related in obvious ways to the large literature on covariate adjustment in randomized experiments (e.g., Imbens & Rubin, 2015) and shows two interesting connections. On the one hand, we find that the conditions under which adjusting for pretreatment covariates can lead to efficiency gains parallel to those required in classical experiments (see section IVB), although we also find that some approaches have inferior estimation and inference properties because of the intrinsic smoothing in RD methods (lemma 1), which is absent in experimental methods. On the other hand, it follows from our results that adjusting for imbalanced covariates in order to restore identification is not possible without functional form assumptions, just as in randomized experiments it is not generally possible to remove imbalances by simply adjusting for the covariates that are correlated with treatment.

Second, Frölich and Huber (2018) develop a nonparametric kernel method to account for observed covariates for RD treatment effect estimation. This localize-then-smooth method of adding covariates to RD designs involves nonparametric estimation over both the running variable and the covariates, and thus requires $d + 2$ bandwidth choices (where d denotes the dimension of the additional covariates included). Their estimator recovers a weighted average of local treatment effects, which coincides with the standard RD treatment effect of interest under further balance-type conditions on the covariates. In contrast, our methods do not require any smoothness assumptions on, or nonparametric smoothing over, the "long" conditional expectation depending on the d additional covariates, and are in perfect

agreement with empirical practice (i.e., local regression methods adding covariates linearly and employing only linear projection arguments). Furthermore, we provide optimal bandwidth selectors and robust bias-corrected inference with valid variance estimators under both heteroskedasticity and clustering and also cover kink RD designs in addition to sharp and fuzzy designs.

The rest of the paper is organized as follows. Section II introduces the framework, sections III and IV summarize our main methodological results, section V briefly discusses numerical results, and section VI concludes. Due to space limitations, we discuss only local linear estimation in sharp RD designs, but in the supplemental appendix we extend our results to fuzzy and kink RD designs for any local polynomial order, in addition to reporting several other results of interest. Companion software, replication files, and other materials can be found at https://sites.google.com/site/rdpackages/.

II.   Sharp RD Designs Using Covariates

The observed data are assumed to be a random sample $(Y_i, T_i, X_i, \mathbf{Z}_i')'$, $i = 1, 2, \ldots, n$. The key feature of any RD design is the presence of a continuous score or running variable X_i, with known threshold $\bar{x}$, which determines treatment assignment for each unit in the sample. For simplicity, here we discuss only sharp RD designs, where all units with $X_i \geq \bar{x}$ are assigned to the treatment group and receive the treatment, and all units with $X_i < \bar{x}$ are assigned to the control group and do not receive the treatment. We also normalize $\bar{x} = 0$ to save notation. All other cases are discussed in the supplemental appendix.

We use the potential outcomes framework. The observed outcome is $Y_i = Y_i(0) \cdot (1 - T_i) + Y_i(1) \cdot T_i$, where $T_i = \mathbb{1}(X_i \geq \bar{x})$ denotes treatment status and $Y_i(1)$ and $Y_i(0)$ are 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:

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

Evaluation points of functions are dropped whenever possible throughout the paper. Hahn, Todd, and van der Klaauw (2001) give 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 $\mathbb{E}[Y_i(t)|X_i = x]$, $t \in \{0, 1\}$, are continuous at the cutoff $x = 0$.

The new feature studied in this paper is the presence of additional covariates, collected in the random vector $\mathbf{Z}_i \in \mathbb{R}^d$, which could be continuous, discrete, or mixed. We set $\mathbf{Z}_i = \mathbf{Z}_i(0) \cdot (1 - T_i) + \mathbf{Z}_i(1) \cdot T_i$, 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 near the cutoff or, more generally, that $\mathbf{Z}_i(1) =_d \mathbf{Z}_i(0)$, which is implied by the definition of a "pretreatment" or "predetermined" covariate in the causal inference literature.

A large portion of the literature on estimation and inference in RD designs focuses on local polynomial estimators. In practice, researchers first choose a neighborhood around the cutoff, usually via a bandwidth choice h, and then conduct local linear polynomial inference; that is, they rely on linear regression fits using only units whose scores X_i lie within that preselected neighborhood, with the weighting scheme determined by the choice of kernel function $K(\cdot)$. The role of the kernel and bandwidth is to localize the regression fit near the cutoff. The most popular choices are the uniform kernel (equally weighting observations $X_i \in [-h, h]$) and the triangular kernel (linearly downweighting observations $X_i \in [-h, h]$). To be specific, the standard local linear RD treatment effect estimator $\hat{\tau}$ is obtained by running the weighted least squares regression of Y_i on a constant, T_i, X_i, and $X_i T_i$ using only units with $X_i \in [-h, h]$ and applying weights $K(X_i/h)$, which in this paper we denote by

$$\hat{\tau} \quad : \quad \hat{Y}_i = \hat{\alpha} + T_i \hat{\tau} + X_i \hat{\beta}_- + T_i X_i \hat{\beta}_+. \tag{1}$$

We give a precise definition of $\hat{\tau}$ and all other estimators in the supplemental appendix. The estimator $\hat{\tau}$ is, of course, numerically equivalent to the difference in intercepts that would be obtained from two separate weighted least squares regressions using observations on each 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 covariate-adjusted RD estimators. We maintain the local linear specification for simplicity, but as we show in the supplemental appendix, all our results apply to any pth-order local polynomial fit.

While the standard estimator $\hat{\tau}$ is popular in empirical work, it is common to augment the specification with the additional covariates $\mathbf{Z}_i$. In practice, these covariates can be included in many different ways. We consider five linear covariate-adjustment specifications, some with centered covariates, to mimic the variety of current empirical practice:

$$\tilde{\tau} \quad : \quad \tilde{Y}_i = \tilde{\alpha} + T_i \tilde{\tau} + X_i \tilde{\beta}_- + T_i X_i \tilde{\beta}_+ + \mathbf{Z}_i' \tilde{\boldsymbol{\gamma}}, \tag{2}$$

$$\check{\tau} \quad : \quad \check{Y}_i = \check{\alpha} + T_i \check{\tau} + X_i \check{\beta}_- + T_i X_i \check{\beta}_+ + (1 - T_i)\mathbf{Z}_i' \check{\boldsymbol{\gamma}}_- \\ + T_i \mathbf{Z}_i' \check{\boldsymbol{\gamma}}_+, \tag{3}$$

$$\dot{\tau} \quad : \quad \dot{Y}_i = \dot{\alpha} + T_i \dot{\tau} + X_i \dot{\beta}_- + T_i X_i \dot{\beta}_+ + (\mathbf{Z}_i - \bar{\mathbf{Z}})' \dot{\boldsymbol{\gamma}}, \tag{4}$$

$$\ddot{\tau} \quad : \quad \ddot{Y}_i = \ddot{\alpha} + T_i \ddot{\tau} + X_i \ddot{\beta}_- + T_i X_i \ddot{\beta}_+ \\ + (1 - T_i)(\mathbf{Z}_i - \bar{\mathbf{Z}})' \ddot{\boldsymbol{\gamma}}_- + T_i (\mathbf{Z}_i - \bar{\mathbf{Z}})' \ddot{\boldsymbol{\gamma}}_+, \tag{5}$$

$$\dddot{\tau} \quad : \quad \dddot{Y}_i = \dddot{\alpha} + T_i \dddot{\tau} + X_i \dddot{\beta}_- + T_i X_i \dddot{\beta}_+ \\ + (1 - T_i)(\mathbf{Z}_i - \bar{\mathbf{Z}}_-)' \, \dddot{\boldsymbol{\gamma}}_- + T_i (\mathbf{Z}_i - \bar{\mathbf{Z}}_+)' \, \dddot{\boldsymbol{\gamma}}_+, \tag{6}$$

where all these weighted least squares regressions are computed only for observations with $X_i \in [-h, h]$ and weights $K(X_i/h)$, and $\bar{\mathbf{Z}}$, $\bar{\mathbf{Z}}_-$, and $\bar{\mathbf{Z}}_+$ correspond, respectively, to the sample averages of Z_i for $X_i \in [-h, h]$, $X_i \in [-h, 0)$, and $X_i \in [0, h]$.

Our recommended approach is $\tilde{\tau}$ in equation (2), which we call the *covariate-adjusted RD estimator*. This estimator broadly captures the common empirical practice of first choosing a neighborhood around the cutoff and then conducting local linear least squares estimation and inference with covariates. We formalize two important points regarding the way that the additional covariates $\mathbf{Z}_i$ are used in the local least squares fit: (a) additive separability between the running variable and the covariates and (b) linear-in-parameters specification for the 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. Furthermore, in practice, $\mathbf{Z}_i$ could include power expansions, interactions, and other "flexible" transformations of the original covariates. This approach to RD covariate adjustment allows for any type of additional regressors, including fixed effects or other discrete variables.

As discussed above, we focus on the common motivation for covariate adjustment based on improving the precision of the estimator of the RD treatment effect, τ, analogous to the standard justification for covariate adjustment in randomized experiments. We build on this intuition and make precise the conditions required for consistency of the covariate-adjusted RD estimator $\tilde{\tau}$ for τ. We also show that much more stringent conditions are required if the estimation model includes treatment-covariate interactions or centering, or both.

A. Notation and Regularity Conditions

The following assumption defines notation and collects the conditions required for sharp RD designs.

Assumption SRD. For $t \in \{0, 1\}$ 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. $\mu_{Y-}(x) = \mathbb{E}[Y_i(0)|X_i = x]$ and $\mu_{Y+}(x) = \mathbb{E}[Y_i(1)|X_i = x]$ are thrice continuously differentiable.
c. $\mu_{Z-}(x) = \mathbb{E}[\mathbf{Z}_i(0)|X_i = x]$ and $\mu_{Z+}(x) = \mathbb{E}[\mathbf{Z}_i(1)|X_i = x]$ are thrice continuously differentiable, and $\mathbb{E}[\mathbf{Z}_i(t)Y_i(t)|X_i = x]$ is continuously differentiable.
d. $\mathbb{V}[(Y_i(t), \mathbf{Z}_i(t)')|X_i = x]$ is continuously differentiable and invertible.
e. $\mathbb{E}[|(Y_i(t), \mathbf{Z}_i(t)')|^4|X_i = x]$, is continuous, where $|\cdot|$ denotes the Euclidean norm.

Assumption SRD imposes standard continuity or smoothness assumptions common to all nonparametric analyses of RD designs (parts a and b), plus a mild assumption to allow for the inclusion of additional covariates (part c). Parts d and e are standard restrictions on the conditional variances and higher moments. Indeed, if one simply ignores all statements involving the additional covariates, the conditions are exactly those found in the RD literature for local linear regression (the supplemental appendix shows the version for polynomial fits of degree $p \geq 1$).

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 SRD does not restrict in any way the long conditional expectations $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]$, $t \in \{0, 1\}$, 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, we allow for complete misspecification of $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]$ in any finite sample, and hence give a best linear approximation or local linear projection interpretation to the RD estimators in equations (2) to (6). In other words, the linearity in these specifications represents the empirical use of the covariates and does not impose parametric assumptions on the underlying regression functions.

III. Estimation in Sharp RD Designs Using Covariates

We now present the first main result of the paper, which constructs and gives an interpretation to the implicit estimands associated with the covariate-adjusted RD estimators. All limits are taken as $n \to \infty$, unless otherwise noted.

Lemma 1 (*Sharp RD with covariates*). *Let assumption SRD hold, and assume the weights obey $K(u) = \mathbb{1}(u < 0)k(-u) + \mathbb{1}(u \geq 0)k(u)$, with $k(\cdot) : [0, 1] \mapsto \mathbb{R}_+$ bounded, zero outside its support, and positive and continuous on (0,1). If $nh \to \infty$ and $h \to 0$, then*

$$\tilde{\tau} \to_\mathbb{P} \tau - [\mu_{Z+} - \mu_{Z-}]' \gamma_Y,$$

$$\check{\tau} \to_\mathbb{P} \tau - [\mu_{Z+}' \gamma_{Y+} - \mu_{Z-}' \gamma_{Y-}],$$

$$\dot{\tau} \to_\mathbb{P} \tau - [\mu_{Z+} - \mu_{Z-}]' \gamma_Y,$$

$$\ddot{\tau} \to_\mathbb{P} \tau - [(\mu_{Z+} - \bar{\mu}_Z)' \gamma_{Y+} - (\mu_{Z-} - \bar{\mu}_Z)' \gamma_{Y-}],$$

$$\dddot{\tau} \to_\mathbb{P} \tau,$$

where $\gamma_Y = (\sigma_{Z-}^2 + \sigma_{Z+}^2)^{-1} \mathbb{E}[(\mathbf{Z}_i(0) - \mu_{Z-}(X_i)) Y_i(0) + (\mathbf{Z}_i(1) - \mu_{Z+}(X_i)) Y_i(1)|X_i = \bar{x}]$, $\mu_{Z-} = \mu_{Z-}(\bar{x})$, $\gamma_{Y-} = (\sigma_{Z-}^2)^{-1} \mathbb{E}[(\mathbf{Z}_i(0) - \mu_{Z-}(X_i)) Y_i(0)|X_i = \bar{x}]$, $\sigma_{Z-}^2 = \mathbb{V}[\mathbf{Z}_i(0)|X_i = \bar{x}]$, and similarly for μ_{Z+}, γ_{Y+} and σ_{Z+}^2, and $\bar{\mu}_Z = \mu_{Z+}/2 + \mu_{Z-}/2$.

It is well known that under the same conditions imposed here, $\hat{\tau} \to_\mathbb{P} \tau$. The conclusion of this lemma gives a precise description of the probability limit of each covariate-adjusted sharp RD estimator in equations (2) to (6). Similar results for other RD designs are shown in the supplemental appendix.

Lemma 1 shows that our recommended covariate-adjusted RD estimator, $\tilde{\tau}$, 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 on the covariates, $\tau_Z := \mu_{Z+} - \mu_{Z-}$. It follows that a sufficient condition for $\tilde{\tau} \to_{\mathbb{P}} \tau$ is that $\mu_{Z+} = \mu_{Z-}$, that is, that there is no RD treatment effect on the covariates. This is weaker than assuming that the marginal distributions of $\mathbf{Z}_i(0)$ and $\mathbf{Z}_i(1)$ are equal near the cutoff, which is the usual definition of predetermined covariates in randomized experiments.

Further, lemma 1 shows that the treatment-interacted, covariate-adjusted RD estimator $\check{\tau}$, which corresponds to fitting two separate weighted linear regressions on each side of the cutoff, is consistent for τ under stronger conditions than $\tilde{\tau}$. The difference arises because including this interaction allows $\boldsymbol{\gamma}_- \neq \boldsymbol{\gamma}_+$ in the estimation, whereas our recommended covariate-adjusted RD estimator, $\tilde{\tau}$, forces equality. The additional term, $[\boldsymbol{\mu}'_{Z+}\boldsymbol{\gamma}_{Y+} - \boldsymbol{\mu}'_{Z-}\boldsymbol{\gamma}_{Y-}]$, can be interpreted as the difference of the best linear approximations at the cutoff of the unknown conditional expectations $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]$, $t \in \{0, 1\}$, based on the additional covariates included in the RD estimation. It follows that a necessary and sufficient condition for $\check{\tau} \to_{\mathbb{P}} \tau$ is that $\boldsymbol{\mu}'_{Z+}\boldsymbol{\gamma}_{Y+} = \boldsymbol{\mu}'_{Z-}\boldsymbol{\gamma}_{Y-}$. However, this is harder to justify in practice than the condition required for the model without the interaction, since we must assume $\boldsymbol{\gamma}_{Y+} = \boldsymbol{\gamma}_{Y-}$ in addition to $\mu_{Z+} = \mu_{Z-}$ ("covariate balance") above. From a linear least squares regression perspective, the discrepancy between the probability limits of $\tilde{\tau}$ and $\check{\tau}$ can be explained as a misinterpretation of the interaction term in the estimation model.

Finally, the last three results in lemma 1 study different demeaning approaches motivated by the literature on covariate adjustment in randomized experiments and by classical least squares regression with interactions. As a result of the demeaning of covariates, the probability limit of the estimators changes: $\dot{\tau}$ (common demeaning) behaves like $\tilde{\tau}$, $\ddot{\tau}$ (common demeaning, treatment interaction) has a new probability limit involving a recentering, and $\ddot{\tau}$ (group demeaning, treatment interaction) behaves like $\hat{\tau}$. For the former two, the same condition, $\mu_{Z+} = \mu_{Z-}$, is sufficient for consistency; $\ddot{\tau}$ does not require this condition because the demeaned covariates are essentially orthogonal to treatment status. These requirements and conclusions are conceptually similar to zero-correlation assumptions in standard least squares algebra. Despite these similarities, the nonparametric nature of the problem introduces important differences with standard least squares: formal derivations require nonparametric large-sample approximations (otherwise, misspecification biases appear for fixed n), and inference results do not follow from standard parametric least squares arguments.

An important drawback of all three demeaning-based estimators ($\dot{\tau}, \ddot{\tau}, \ddot{\tau}$) is that they employ some form of local average of the covariates ($\bar{\mathbf{Z}}, \bar{\mathbf{Z}}_-, \bar{\mathbf{Z}}_+$), which is equivalent to a kernel regression estimator with a uniform kernel. Therefore, inference using the resulting demeaned estimators is severely affected: these estimators exhibit slower convergence rates, new misspecification biases, and additional asymptotic variability when compared to $\tilde{\tau}$ (or even $\check{\tau}$). Thus, in these cases,

the analogy with classical least squares regression breaks down. Furthermore, as shown in the supplemental appendix, simple demeaning does not work for other RD designs.

Putting the above together and from a practical perspective, not only does lemma 1 give general, precise, and intuitive characterizations of the probability limits of the various covariate-adjusted RD estimators; it also has interesting implications for the analysis and interpretation of RD designs using covariates. Most notable, the lemma shows the conditions under which a covariate-adjusted RD estimator is consistent for the standard (causal) RD treatment effect of interest, τ, and, by implication, establishes when estimators with and without covariate adjustment are valid (for estimating τ) and can be compared. We conclude that for estimating τ, $\tilde{\tau}$ requires the weakest assumption, $\mu_{Z+} = \mu_{Z-}$, while at the same time it is not affected by the poor behavior of demeaning. We therefore advocate this estimator for use in applications, and the rest of the paper gives a thorough analysis of its large sample properties and empirical performance.

In the supplemental appendix, we extend the results to other RD designs, where we show that a new condition for covariate balance emerges for kink RD designs; the additional requirement is that $\boldsymbol{\mu}_{Z+}^{(1)} = \boldsymbol{\mu}_{Z-}^{(1)}$, where $\boldsymbol{\mu}_{Z+}^{(1)}$ and $\boldsymbol{\mu}_{Z-}^{(1)}$ denote the first derivative of the conditional expectations of the covariates at the cutoff for treatment and control, respectively.

IV. 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 (references are given in section I). In this section, we study the large sample properties of the covariate-adjusted RD estimator $\tilde{\tau}(h) = \tilde{\tau}$, now making the dependence on h explicit, and assuming throughout that $\mu_{Z+} = \mu_{Z-}$ 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 and clustered data are given in the supplemental appendix.

We rely on the following representation (valid for each n): $\tilde{\tau}(h) = \hat{\tau}(h) - \hat{\boldsymbol{\tau}}_Z(h)' \tilde{\boldsymbol{\gamma}}_Y(h)$, where $\hat{\tau}(h) = \hat{\tau}$ and $\tilde{\boldsymbol{\gamma}}_Y(h) = \tilde{\boldsymbol{\gamma}}_Y$ are given in equations (1) and (2), respectively, and $\hat{\boldsymbol{\tau}}_Z(h)$ is a d-dimensional vector containing the standard RD treatment effect estimators for each covariate. In other words, each element of $\hat{\boldsymbol{\tau}}_Z(h)$ is constructed using the corresponding covariate as outcome variable in equation (1). 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 $\mathbf{s}(h) = (1, \tilde{\boldsymbol{\gamma}}_Y(h)')'$ and $\mathbf{s} = (1, \boldsymbol{\gamma}'_Y)'$, and because $\mathbf{s}(h) \to_{\mathbb{P}} \mathbf{s}$ using the results underlying lemma 1 (and using

$\mu_{Z+} = \mu_{Z-}$, so that $\tau_Z = \mathbf{0}$). The asymptotic analysis proceeds by studying the (joint) large-sample properties of the vector $\hat{\tau}_S(h) := (\hat{\tau}(h), \hat{\tau}_Z(h)')'$ and then taking the linear combination $\mathbf{s}(h)$ or $\mathbf{s}$, as appropriate. Note that $\hat{\tau}_S(h) \to_{\mathbb{P}} \tau_S := (\tau, \tau_Z')'$ under the conditions in lemma 1. We give exact details in the supplemental appendix.

A. MSE Expansion and Data-Driven Bandwidth Selection

We first establish a valid asymptotic MSE-type expansion for the covariate-adjusted RD estimator, based on the representation already given, which is useful to develop optimal bandwidth choices and optimal point estimators. Further, the bias expressions will be used to develop inference procedures based on robust bias correction. The object we study is defined as $\mathsf{MSE}[\tilde{\tau}(h)] = \mathbb{E}[(\mathbf{s}'\hat{\tau}_S(h) - \mathbf{s}'\tau_S)^2 | \mathbf{X}] = (\mathsf{Bias}[\tilde{\tau}(h)])^2 + \mathsf{Var}[\tilde{\tau}(h)]$, where $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$, $\mathsf{Bias}[\tilde{\tau}(h)] := \mathbb{E}[\mathbf{s}'\hat{\tau}_S(h) - \mathbf{s}'\tau_S | \mathbf{X}]$, and $\mathsf{Var}[\tilde{\tau}(h)] := \mathbb{V}[\mathbf{s}'\hat{\tau}_S(h) | \mathbf{X}]$.

Theorem 1 *(MSE expansion). Let the conditions of lemma 1 hold. Then*

$$\mathsf{MSE}[\tilde{\tau}(h)] = h^4 \mathcal{B}_{\tilde{\tau}}(h)^2 \{1 + o_{\mathbb{P}}(1)\} + \frac{1}{nh} \mathcal{V}_{\tilde{\tau}}(h),$$

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

The bias and variance expressions in theorem 1 are different from those available in the literature (Imbens & Kalyanaraman, 2012; Calonico, Cattaneo, & Titiunik, 2014; Arai & Ichimura, 2018) due to the presence of the covariates $\mathbf{Z}_i$. As a consequence, MSE-optimal bandwidth selection and point estimators are different when covariate adjustment is employed. Bias-correction techniques and standard error constructions are also different, as discussed below.

The leading bias and variance formulas in theorem 1 are derived in preasymptotic form. For the bias, the random term $\mathcal{B}_{\tilde{\tau}}(h)$ gives a preasymptotic 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, Cattaneo, and Farrell (2018) prove, using valid Edgeworth expansions, that employing preasymptotic approximations when conducting asymptotic inference in nonparametrics can lead to superior performance. Furthermore, fewer unknown features 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. The bias is entirely due to estimating the unknown functions $\mu_{Y-}(\cdot)$ and $\mu_{Z-}(\cdot)$ for the control group and $\mu_{Y+}(\cdot)$ and $\mu_{Z+}(\cdot)$ for the treatment group. When the covariates are not included, these constants reduce exactly

to those already available in the literature. In the supplemental 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 local linear covariate-adjusted RD estimator $\tilde{\tau}(h)$ is

$$\mathfrak{h}_{\tilde{\tau}} = \left[\frac{\mathcal{V}_{\tilde{\tau}}/n}{4\mathcal{B}_{\tilde{\tau}}^2}\right]^{1/5}.$$

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 $\tau_Z = \mathbf{0}$. The convergence rate is not affected by $d = \dim(\mathbf{Z}_i)$ because no nonparametric smoothing is done on the additional covariates.

To construct feasible MSE-optimal bandwidth choices, we proceed in the familiar way. For pilot bandwidths $b \to 0$ and $v \to 0$, we can implement $\tilde{\mathfrak{h}}_{\tilde{\tau}} = \left[\frac{\tilde{\mathcal{V}}_{\tilde{\tau}}(v)/n}{4\tilde{\mathcal{B}}_{\tilde{\tau}}(b)^2}\right]^{1/5}$, where the exact forms 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 (we also show results for a generic degree $p \geq 1$, where the optimal bandwidth decays as $n^{-1/(3+2p)}$). Heuristically, these estimators are formed as plug-in versions of the preasymptotic formulas obtained in theorem 1. In the supplemental appendix, we show that these feasible versions of the optimal bandwidths are consistent for their infeasible analogs, $\tilde{\mathfrak{h}}_{\tilde{\tau}}/\mathfrak{h}_{\tilde{\tau}} \to_{\mathbb{P}} 1$.

Finally, in the supplemental appendix, we also discuss other MSE-optimal bandwidth selectors, including (a) separate MSE optimizations on either side of the cutoff, (b) the MSE for the sum rather than the difference of the one-sided estimators, and (c) 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, and these choices may be more stable in finite samples or more robust to situations where the smoothing bias may be small.

B. 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 effect of introducing covariates in RD estimation on asymptotic efficiency. Using explicit results proven in the supplemental appendix, we can compare the asymptotic variance of the covariate-adjusted estimator $\tilde{\tau}$, denoted by $\mathcal{V}_{\tilde{\tau}}$, to that of the standard RD estimator $\hat{\tau}$, denoted by $\mathcal{V}_{\hat{\tau}}$. This comparison reduces to studying

$$\frac{\mathcal{V}_{\tilde{\tau}}}{\mathcal{V}_{\hat{\tau}}} = \frac{\begin{array}{c}\mathbb{V}[Y_i(0) - \mathbf{Z}_i(0)'\boldsymbol{\gamma}_Y | X_i = \bar{x}] \\ + \mathbb{V}[Y_i(1) - \mathbf{Z}_i(1)'\boldsymbol{\gamma}_Y | X_i = \bar{x}]\end{array}}{\mathbb{V}[Y_i(0) | X_i = \bar{x}] + \mathbb{V}[Y_i(1) | X_i = \bar{x}]},$$

where $\boldsymbol{\gamma}_Y$ is given in lemma 1. In general, a definitive ranking is not available because of the linear combination $\boldsymbol{\gamma}_Y$, which may not be equal to either $\boldsymbol{\gamma}_{Y_-}$ or $\boldsymbol{\gamma}_{Y_+}$ given in lemma 1. However, an interesting special case is $\boldsymbol{\gamma}_Y = \boldsymbol{\gamma}_{Y_-} = \boldsymbol{\gamma}_{Y_+}$, in which case $\boldsymbol{\gamma}_Y$ reduces to the best linear approximation for each group and therefore $\mathbb{V}[Y_i(t) - \mathbf{Z}_i(t)'\boldsymbol{\gamma}_Y | X_i = \bar{x}] \leq \mathbb{V}[Y_i(t)|X_i = \bar{x}]$. This result implies that whenever the effect of the additional covariates on the potential outcomes near the cutoff (via their local linear projections) is (roughly) the same for both control and treatment units, including the covariates may lead to efficiency gains.

The efficiency results above are based on large sample nonparametric approximations using only local linear projections on the additional covariates. Remarkably, however, these results are in perfect agreement with those in the literature on analysis of experiments obtained using Neyman's repeated sampling (Freedman, 2008; Lin, 2013), where it is also found that incorporating covariates in randomized controlled trials using linear regression leads to efficiency gains only under particular assumptions—for example, using our notation, when $\mathbb{E}[Y_i(t)|\mathbf{Z}_i(t), X_i = \bar{x}] = a_t + \mathbf{Z}_i(t)'\mathbf{b}_t$, $t = 0, 1$, and $\mathbf{b}_0 = \mathbf{b}_1$ (note that this parametric assumption implies $\boldsymbol{\gamma}_Y = \boldsymbol{\gamma}_{Y_-} = \boldsymbol{\gamma}_{Y_+}$).

These results also show that $\tilde{\tau}(\mathfrak{h}_{\tilde{\tau}})$ can be a better point estimator in an MSE sense than its 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 & Kalyanaraman, 2012; Calonico et al., 2014; Arai & Ichimura, 2018). Using the explicit formulas, 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 point estimator because it relies on minimal identifying assumptions, and thus researchers can incorporate covariates to increase precision relative to it.

C. *Asymptotic Distribution and Valid Inference*

To develop valid asymptotic distributional approximations and inference procedures, we employ nonparametric robust bias correction (see Calonico et al., 2014; Calonico, Cattaneo, & Farrell, 2018, 2019). Inference based on large-sample distribution theory using MSE-optimal bandwidths will suffer from a first-order bias, leading to invalid hypothesis tests 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.

The bias term $\mathcal{B}_{\tilde{\tau}}(h)$ is known up to the higher-order derivatives of the unknown regression functions, $\boldsymbol{\mu}_{Y-}(\cdot)$, $\boldsymbol{\mu}_{Z-}(\cdot)$, $\boldsymbol{\mu}_{Y+}(\cdot)$, 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 supplemental appendix; we replace

$\mathbf{s}$ by $\mathbf{s}(h)$ for implementation. For our presentation, we simply take as given the bias estimator $\tilde{\mathcal{B}}_{\tilde{\tau}}(b)$ based on a local quadratic regression (in general, a higher-order polynomial than used to form $\tilde{\tau}(h)$) and a preliminary bandwidth $b \to 0$, possibly different from h. Then the bias-corrected covariate-adjusted sharp RD estimator is

$$\tilde{\tau}^{\mathsf{bc}}(h, b) = \tilde{\tau}(h) - h^2 \tilde{\mathcal{B}}_{\tilde{\tau}}(b). \tag{7}$$

An empirically useful choice is $b = h$, which is both allowed by our asymptotic theory and has some optimality properties (see Calonico et al., 2018, 2019, for theoretical results on robust bias correction). This bias correction approach captures "flexible" regression adjustments to account for misspecification in finite samples (Calonico et al., 2014, Remark 7).

The 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 by

$$\begin{aligned}
\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h, b) = {} & \left[\mathbf{s}' \otimes \mathbf{P}_-^{\mathsf{bc}}(h, b) \right] \boldsymbol{\Sigma}_{S-} \left[\mathbf{s}' \otimes \mathbf{P}_-^{\mathsf{bc}}(h, b) \right]' \\
& + \left[\mathbf{s}' \otimes \mathbf{P}_+^{\mathsf{bc}}(h, b) \right] \boldsymbol{\Sigma}_{S+} \left[\mathbf{s}' \otimes \mathbf{P}_+^{\mathsf{bc}}(h, b) \right]',
\end{aligned}$$

where the n-vectors $\mathbf{P}_-^{\mathsf{bc}}(h, b)$ and $\mathbf{P}_+^{\mathsf{bc}}(h, b)$ 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 the only unknowns. The supplemental appendix collects details and specific formulas.

The (infeasible) variance formula $\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h, b)$ differs from that presented in theorem 1, $\mathcal{V}_{\tilde{\tau}}(h)$, because it also accounts for the additional variability introduced by the bias estimation, $h^2 \tilde{\mathcal{B}}_{\tilde{\tau}}(b)$. By virtue of the variance formula being computed both conditionally and preasymptotically, up to the linear combination term $\mathbf{s}$, it involves only one unknown feature, $\boldsymbol{\Sigma}_{S-}$ and $\boldsymbol{\Sigma}_{S+}$, which must be estimated, thereby considerably simplifying implementation.

To operationalize the variance formula, we replace unknown quantities by plug-in estimators thereof, which must account for the specific data structure at hand, such as heteroskedasticity or clustering. In the supplemental appendix, we discuss two different 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. Here, we let $\tilde{\mathcal{V}}_{\tilde{\tau}}^{\mathsf{bc}}(h, b)$ denote a generic, feasible estimator of $\mathcal{V}_{\tilde{\tau}}^{\mathsf{bc}}(h, b)$.

Putting together all the pieces, we obtain the following distributional approximation result, which provides valid local polynomial inference in sharp RD designs using covariates:

TABLE 1.—EMPIRICAL ILLUSTRATION: HEAD START DATA

MSE-Optimal bandwidths:	Not Using Covariates		Using Covariates
	Standard	Covariate Adjusted	Covariate Adjusted
RD treatment effect	−2.41	−2.51	−2.47
Inference with 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 $h/b = 1$			
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

All estimates are computed using a triangular kernel and nearest neighbor heteroskedasticity-robust variance estimators. Columns "Standard" and "Cov Adjusted" correspond to, respectively, standard and covariate-adjusted RD estimation and inference methods, given a choice of bandwidths. Bandwidths used (h and b) are data-driven MSE-optimal for either standard RD estimator or covariate-adjusted RD estimator (depending on the group of columns). Specifically, in the first two columns ("Not Using Covariates") the bandwidths are selected to be MSE optimal for $\tilde{\tau}$ (standard RD estimation), while in the third column ("Using Covariates") the bandwidths are selected to be MSE optimal for $\tilde{\tau}$ (covariate-adjusted RD estimation).

Theorem 2 (asymptotic normality). *Let the conditions of theorem 1 hold, and assume* $\tau_Z = 0$. *If* $nh^7 \to 0$ *and* $\overline{\lim}(h/b) < \infty$, *then*

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

$$\tilde{\mathcal{V}}_{\tilde{\tau}}^{\mathrm{bc}}(h, b)/\mathcal{V}_{\tilde{\tau}}^{\mathrm{bc}}(h, b) \to_{\mathbb{P}} 1.$$

Extensions of this result to all other popular RD designs are available in the supplemental appendix. One of the strengths of theorem 2 is that the distributional approximation is valid even when the MSE-optimal bandwidth choice is used, which is not true of standard inference procedures. 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 a common bandwidth $h = b$ is given by

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

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

A particularly attractive alternative to MSE-optimal bandwidth selection is to develop coverage error rate (CER) optimal bandwidth choices. Following the valid Edgeworth expansions by Calonico et al. (2018, 2019), we also propose the following plug-in bandwidth selector $\tilde{\mathfrak{h}}_{\mathrm{CER},\tilde{\tau}} = n^{-1/20} \times \tilde{\mathfrak{h}}_{\tilde{\tau}}$. This bandwidth choice minimizes the coverage error rate for confidence intervals based on theorem 2 and may be preferred for inference purposes (the supplemental appendix gives the rate scaling for generic degree $p \geq 1$.) (See Cattaneo & Vazquez-Bare, 2016, for an introductory discussion on bandwidth selection for RD designs.)

Theorems 1 and 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 clus-tering. In this case, asymptotics are conducted under the standard assumptions: (a) each unit i belongs to exactly one of G clusters, and (b) $G \to \infty$ and $Gh \to \infty$ (see Cameron & Miller, 2015, for a review of cluster-robust inference and Bartalotti & Brummet, 2017, 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 supplemental appendix. Our companion software in R and Stata also includes cluster-robust options for bandwidth selection, MSE-optimal point estimation, and robust bias-corrected inference.

V. Numerical Results

We briefly summarize the main numerical findings from an empirical illustration and a Monte Carlo study. Additional simulation results can be found in the supplemental appendix.

A. Empirical Illustration: Head Start Data

To illustrate our methods, we first reanalyze the effect of Head Start assistance on child mortality in the United States, which was first studied by Ludwig and Miller (2007). 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 by the federal government based on 1960 census information, with cutoff $\bar{x} = 59.1984$. The outcome is the child mortality rate (for children ages 5 to 9) due to causes affected by Head Start's health services component. We compare the standard RD estimator to the covariate-adjusted RD estimator employing heteroskedasticity-robust nearest-neighbor variance estimation. There are nine preintervention covariates from the 1960 U.S. Census: total population, percentage of black and urban population, and levels and percentages of population in three age groups (children aged 3 to 5, children aged 14 to 17, and adults older than 25). (Full replication code in both R and Stata is available at https://sites.google.com/site/rdpackages/replication/.)

Table 1 presents the main results. The first row reports the local linear point estimate using the corresponding

MSE-optimal bandwidth h as described in each 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 the 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, and then when $b = h$. Finally, the last two rows report, respectively, the estimated 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 empirical findings are consistent with our theoretical results: employing covariate-adjusted RD inference leads to precision improvements while the point estimators remain stable. The point estimate ranges from -2.41 to -2.51, and it is statistically different from 0 at 5% level in all cases. As should be expected when the additional covariates are truly predetermined, 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 covariates leads to sizable efficiency gains: the rows labeled "CI length change (%)" show a nearly 10% efficiency gain when the bandwidths are unrestricted and optimally chosen using covariates.

B.   Simulation Evidence

We also investigate the finite sample performance of our methods using realistic simulated data. We consider four data-generating processes constructed using the data of Lee (2008), where all parameters were obtained from real data unless explicitly noted otherwise. This simulation model has been used extensively in the literature, which facilitates the comparison across studies. All the models include a predetermined covariate (previous democratic vote share), and they vary in the importance of this covariate: (a) in model 1, the covariate is irrelevant; (b) in model 2, it enters the conditional expectation of the potential outcomes $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]$, $t \in \{0, 1\}$ according to the real data; (c) model 3 takes model 2 but sets the residual correlation between the outcome and covariate to 0; and (d) model 4 takes model 2 but doubles the residual correlation between the outcome and covariate equations. Models 3 and 4 do not imply $\mathbb{C}ov[Y_i(t), \mathbf{Z}_i(t)|X_i] = 0$, $t \in \{0, 1\}$.

The constructions allow $\mathbb{E}[Y_i(t)|X_i, \mathbf{Z}_i(t)]$ to have different coefficients on each side of the cutoff, while the conditional expectation of the potential covariates $\mathbb{E}[\mathbf{Z}_i(t)|X_i]$, $t \in \{0, 1\}$, are constructed assuming they are continuous at the cutoff (but still with different coefficients on either side). Therefore, our covariate-adjusted RD estimator is "misspecified" when viewed as a local weighted least squares fit. To conserve space, all details and results of our Monte Carlo study are given in part V of the supplemental appendix. All findings are consistent with our large sample theory. We

find that covariate-adjusted local polynomial analysis can improve both MSE and interval length, sometimes dramatically. The gains are largest in model 4, with the amplified residual correlation, and least in model 3, when that channel is shut down, as the theory predicts. The results for model 1 show that including an irrelevant covariate hardly changes empirical results and conclusions. Finally, we find that our data-driven bandwidth selectors work reasonably well.

VI.   Conclusion

We provided a formal framework for identification, estimation, and inference in RD designs when covariates are included in local polynomial estimation. We augmented the standard local polynomial estimator with covariates entering in an additive-separable, linear-in-parameters way and showed that 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. Thus, we are able to characterize precisely the potential for point estimation and inference improvements and, in particular, efficiency gains. 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. Our results and practical methods cover sharp, fuzzy, and kink RD designs, and we also discuss extensions to clustered data. Finally, we illustrated the practical implications of our results using both an application and simulated data.

REFERENCES

Arai, Y., and H. Ichimura, "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator," *Quantitative Economics* 9 (2018), 441–482.

Bartalotti, O., and Q. Brummet, "Regression Discontinuity Designs with Clustered Data" (pp. 383–420), in M. D. Cattaneo and J. C. Escanciano, eds., *Regression Discontinuity Designs: Theory and Applications* (Bingley, U.K.: Emerald Group, 2017).

Calonico, S., M. D. Cattaneo, and M. H. Farrell, "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association* 113:523 (2018), 767–779.

———— "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression" (2019), arXiv:1808.01398.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik, "rdrobust: Software for Regression Discontinuity Designs," *Stata Journal* 17 (2017), 372–404.

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

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

Cattaneo, M. D., and J. C. Escanciano, *Regression Discontinuity Designs: Theory and Applications* (Berlin, U.K.: Emerald Group, 2017).

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

Cattaneo, M. D., N. Idrobo, and R. Titiunik, *A Practical Introduction to Regression Discontinuity Designs: Foundations* (Cambridge: Cambridge University Press, forthcoming).

——— *A Practical Introduction to Regression Discontinuity Designs: Extensions* (Cambridge: Cambridge University Press, forthcoming).

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare, "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management* 36 (2017), 643–681.

Cattaneo, M. D., and G. Vazquez-Bare, "The Choice of Neighborhood in Regression Discontinuity Designs," *Observational Studies* 2 (2016), 134–146.

Freedman, D. A., "On Regression Adjustments in Experiments with Several Treatments," *Annals of Applied Statistics* 2 (2008), 176–196.

Frölich, M., and M. Huber, "Including Covariates in the Regression Discontinuity Design," *Journal of Business and Economic Statistics* (2018), https://doi.org/10.1080/07350015.2017.1421544.

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

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

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

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

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

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

Lin, W., "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *Annals of Applied Statistics* 7 (2013), 295–318.

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

Sekhon, J., and R. Titiunik, "On Interpreting the Regression Discontinuity Design as a Local Experiment" (pp. 1–28), in M. D. Cattaneo and J. C. Escanciano, eds., *Regression Discontinuity Designs: Theory and Applications* (Bingley, U.K.: Emerald Group, 2017).