Flexible Covariate Adjustments in Regression Discontinuity Designs

Claudia Noack Tomasz Olma Christoph Rothe

June 28, 2024

Abstract

Empirical regression discontinuity (RD) studies often use covariates to increase the precision of their estimates. In this paper, we propose a novel class of estimators that use such covariate information more efficiently than existing methods and can accommodate many covariates. It involves running a standard RD analysis in which a function of the covariates has been subtracted from the original outcome variable. We characterize the function that leads to the estimator with the smallest asymptotic variance, and consider feasible versions of such estimators in which this function is estimated, for example, through modern machine learning techniques.

First version: July 16, 2021. This version: June 28, 2024. We thank Sebastian Calonico, Michal Kolesár, Thomas Lemieux, Jonathan Roth, Vira Semenova, Stefan Wager, Daniel Wilhelm, Andrei Zeleneev, and numerous conference and seminar participants for helpful comments and suggestions. We thank Tobias Großbölting and Merve Öğretmek for excellent research assistance. The authors gratefully acknowledge financial support by the European Research Council (ERC) through grant SH1-77202. The second author also acknowledges support from the European Research Council through Starting Grant No. 852332. Author contact information: Claudia Noack, Department of Economics, University of Bonn. E-Mail: claudia.noack@uni-bonn.de. Website: https://claudianoack.github.io. Tomasz Olma, Department of Statistics, Ludwig Maximilian University of Munich. E-Mail: t.olma@lmu.de. Website: https://tomaszolma.github.io. Christoph Rothe, Department of Economics, University of Mannheim. E-Mail: rothe@vwl.uni-mannheim.de. Website: http://www.christophrothe.net.

1. INTRODUCTION

Regression discontinuity (RD) designs are widely used for estimating causal effects from observational data in economics and other social sciences. The design exploits that in many contexts a unit's treatment status is determined by whether its realization of a running variable exceeds some known cutoff value. For example, students might qualify for a scholarship if their GPA is above some threshold. Under continuity conditions on the distribution of potential outcomes the average treatment effect at the cutoff is identified by the jump in the conditional expectation of the outcome given the running variable at the cutoff. Methods for estimation and inference based on local linear regression are widely used, and their properties are by now well understood (e.g., Hahn et al., 2001; Imbens and Kalyanaraman, 2012; Calonico et al., 2014; Armstrong and Kolesár, 2020).

While an RD analysis generally does not require data beyond the outcome and the running variable, additional covariate information can be used to reduce the variance of empirical estimates. A common strategy is to include the covariates linearly and without separate localization in a local linear RD regression (Calonico et al., 2019). This conventional linear adjustment estimator is consistent without functional form assumptions on the underlying conditional expectations if the covariates are unaffected by the treatment in some appropriate sense. However, it generally does not exploit the available covariate information efficiently and it is also not well-suited for settings with many covariates.

To address these issues, we propose a novel class of flexible covariate-adjusted RD estimators. Our approach involves running a standard local linear RD regression after subtracting an (estimated) function of the covariates from the original outcome variable. We characterize the function that leads to the RD estimator with the smallest asymptotic variance, and show how this function can be estimated with modern machine learning techniques. We also show that existing methods for bandwidth choice and inference can directly be used with our adjusted outcomes, which means that our approach is easily implemented with existing software packages.

Our proposed flexible covariate adjustments can lead to substantial efficiency gains in practice. To illustrate this, we conducted a comprehensive literature survey and reanalyzed the available data from the papers that use RD estimation with covariates and appeared between 2018 and 2023 in the AEA journals that publish applied microeconomic research (AER, AER Insights, AEJ: Applied Economics, AEJ: Economic Policy, and AEA Papers and Proceedings). In total, we reanalyzed 56 specifications from 16 papers, and studied how different types of covariate adjustments affect the length of the confidence interval for the main RD parameter. While in about half of the specifications including covariates into the RD regression does not reduce the length of the confidence intervals, the reduction due to our proposed flexible adjustments reaches more than 35% in one setting. To put this into perspective, obtaining this reduction would require to triple the sample size if the covariates were not used. We also observe that the linear adjustments alone are unable to exhaust all the available covariate information as the largest reduction in the confidence interval

length from using our flexible adjustment relative to linear adjustments exceeds 21%.

To motivate our proposed procedure, let Y_i and Z_i denote the outcome and covariates, respectively, of observational unit i, and note that the conventional linear adjustment RD estimator is asymptotically equivalent to a local linear RD estimator with the modified outcome variable $Y_i - Z_i^\top \gamma_0$, where γ_0 is a vector of projection coefficients. We consider generalizations of sch estimators which replace the linearly adjusted outcome with a flexibly adjusted outcome of the form $Y_i - \eta(Z_i)$, for some generic function η. Such estimators are easily seen to be consistent for *any* fixed η if the distribution of the covariates varies smoothly around the cutoff in some appropriate sense (which is in line with the notion of covariates being unaffected by the treatment). We show that the asymptotic variance in this class of estimators is minimized if $\eta = \eta_0$ is the average of the two conditional expectations of the outcome variable given the running variable and the covariates just above and below the cutoff. This optimal adjustment function is generally nonlinear and unknown in practice but can be estimated from the data.

Our proposed estimators hence take the form of a local linear RD regression with generated outcome $Y_i - \widehat{\eta}(Z_i)$, where $\widehat{\eta}$ is some estimate of η_0 obtained in a preliminary stage. We implement such estimators with cross-fitting (e.g., Chernozhukov et al., 2018), which is an efficient form of sample splitting that removes some bias and allows us to accommodate a wide range of estimators of the optimal adjustment function. In particular, one can use modern machine learning methods like lasso regression, random forests, deep neural networks, or ensemble combinations thereof, to estimate the optimal adjustment function. However, in low-dimensional settings, researchers can also use classical nonparametric approaches like local polynomials or series regression, or estimators based on parametric specifications.

Our theory does not require that η_0 is consistently estimated for valid inference on the RD parameter in our setup. We only require that in large samples the first-stage estimates concentrate in a mean-square sense around some deterministic function $\bar{\eta}$, which could in principle be different from η_0. The rate of this convergence can be arbitrarily slow. Our setup allows for this kind of potential misspecification because our proposed RD estimators are "very insensitive" to estimation errors in the preliminary stage. This is because they are constructed as sample analogs of a moment function that contains η_0 as a nuisance function, but does not vary with it: as discussed above, our parameter of interest is equal to the jump in the conditional expectation of $Y_i - \eta(Z_i)$ given the running variable at the cutoff for *any* fixed function η. This insensitivity property is related to Neyman orthogonality, which features prominently in many modern two-stage estimation methods (e.g., Chernozhukov et al., 2018), but it is a global rather than a local property and is thus in effect substantially stronger.[1]

[1]A moment function is Neyman orthogonal if its first functional derivative with respect to the nuisance function is zero, but the (conditional) moment function on which our estimates are based is fully invariant with respect to the nuisance function. Chernozhukov et al. (2018) give several examples of setups in which such a property occurs, which include optimal instrument problems, certain partial linear models, and treatment effect estimation under unconfoundedness with known propensity score. Such global insensitivity is also easily seen to occur more generally

Our theoretical analysis shows that, under the conditions outlined above, our proposed RD estimator is first-order asymptotically equivalent to a local linear "no covariates" RD estimator with $Y_i - \bar{\eta}(Z_i)$ as the dependent variable. This result is then used to study its asymptotic bias and variance, and to derive an asymptotic normality result. The asymptotic variance of our estimator depends on the function $\bar{\eta}$ and achieves its minimum value if $\bar{\eta} = \eta_0$ (that is if η_0 is consistently estimated in the first stage), but the variance can be estimated consistently irrespective of whether or not that is the case. As our result does not require a particular rate of convergence for the first step estimate of η_0, our RD estimator can be seen as shielded from the "curse of dimensionality" to some degree, and can hence be expected to perform well in settings with many covariates.

Practical issues like bandwidth choice and construction of confidence intervals with good coverage properties can be addressed in a straightforward manner. Specifically, we prove that one can apply standard methods to a data set in which the outcome Y_i is replaced with the generated outcome $Y_i - \widehat{\eta}(Z_i)$, ignoring that $\widehat{\eta}$ has been estimated. Our approach can therefore easily be integrated into existing software packages.

In many empirical applications, the RD designs are fuzzy, meaning that the probability of treatment jumps at the cutoff but not necessarily from zero to one. In these settings, the parameter of interest is the ratio of two sharp RD parameters. Our methodology developed for sharp RD designs, can be easily extended to these settings. Specifically, the key insight is that it is optimal to use our proposed sharp RD estimator to estimate the numerator and denominator separately.

Our theoretical results are qualitatively similar to those that have been obtained for efficient influence function (EIF) estimators of the population average treatment effect in simple randomized experiments with known and constant propensity scores (e.g., Wager et al., 2016). Such parallels arise because EIF estimators are also based on a moment function that is globally invariant with respect to a nuisance function. In fact, we argue that our RD estimator is in many ways a direct analog of the EIF estimator, and that the variance it achieves under the optimal adjustment function is similar in structure to the semiparametric efficiency bound in simple randomized experiments.

We conduct simulations based on the data set from one of the papers from our empirical literature survey. In order to cover all types of settings from our empirical literature survey, we consider simulation setups of large sample sizes and a moderate number of covariates as well as small sample sizes and a varying number of covariates. Our proposed RD estimators perform very well in all these settings, in the sense that their standard errors are close to their standard deviations and the associated confidence intervals have simulated coverage rate close to the nominal one. RD estimators based on conventional linear adjustments also perform well if the number of covariates relative to the sample size is small. However, in settings of moderate and large numbers of covariates, their standard errors can be substantially downwards biased, so that the associated confidence intervals have simulated coverage substantially below their nominal one.

if one of the two nuisance functions in a doubly robust moment (cf. Robins and Rotnitzky, 2001) is known.

Related Literature. Our paper contributes to an extensive literature on estimation and inference in RD designs; see, e.g., Imbens and Lemieux (2008) and Lee and Lemieux (2010) for a surveys, and Cattaneo et al. (2019) for a textbook treatment. Different ad-hoc methods for incorporating covariates into an RD analysis have long been used in empirical economics (see, e.g., Lee and Lemieux, 2010, Section 3.2.3). Following Calonico et al. (2019), it has become common practice to include covariates without localization into the usual local linear regression estimator. We show that our approach nests this estimator as a special case, but is generally more efficient. Other closely related papers are Kreiß and Rothe (2023) and Arai et al. (2024), who extend the approach in Calonico et al. (2019) to settings with high-dimensional covariates under sparsity conditions using the lasso. In contrast, our approach allows for flexible use of other machine learning methods in the spirit of double-debiased machine learning of Chernozhukov et al. (2018). Moreover, even if one commits to lasso-based adjustments, there are two ways in which our approach can improve upon the methods of Kreiß and Rothe (2023) and Arai et al. (2024). First, we propose a different variant of (post-) lasso adjustments that can be more stable in finite samples (see "global adjustments" in Section 3.3). Second, cross-fitting yields a more precise standard error even if the number of selected covariates is not small relative to the effective sample size. Frölich and Huber (2019) propose to incorporate covariates into an RD analysis in a fully nonparametric fashion, but their approach is generally affected by the curse of dimensionality, and is thus unlikely to perform well in practice.

Our paper is also related in a more general way to the vast literature on two-step estimation problems with infinite-dimensional nuisance parameters (e.g., Andrews, 1994; Newey, 1994), especially the recent strand that exploits Neyman orthogonal (or debiased) moment functions and cross-fitting (e.g., Belloni et al., 2017; Chernozhukov et al., 2018). The latter literature focuses mostly on regular (root-n estimable) parameters, while our RD treatment effect is a non-regular (nonparametric) quantity. Some general results on non-regular estimation based on orthogonal moments are derived in Chernozhukov et al. (2019), and specific results for estimating conditional average treatment effects in models with unconfoundedness are given, for example, in Kennedy et al. (2017), Kennedy (2020) and Fan et al. (2020). Our results are qualitatively different because, as explained above, our estimator is based on a moment function that satisfies a property that is stronger than Neyman orthogonality.

Plan of the Paper. The remainder of this paper is organized as follows. In Section 2, we introduce the setup and review existing procedures. In Section 3, we describe our proposed covariate-adjusted RD estimator, and we present the results of our empirical literature survey and reanalysis in Section 4. In Section 5, we present our main theoretical results. Further extensions are discussed in Section 6. Section 7 contains a simulation study. Section 8 concludes. The proofs of our main results are given in Appendix A. Appendix B formally studies the proposed inference procedures and Appendix C gives details on our literature survey. The Online Supplement contains additional empirical and simulation results.

2. SETUP AND PRELIMINARIES

2.1. **Model and Parameter of Interest.** We begin by considering sharp RD designs. The data $\{W_i\}_{i\in[n]} = \{(Y_i, X_i, Z_i)\}_{i\in[n]}$, where $[n] = \{1, \ldots, n\}$, are an i.i.d. sample of size n from the distribution of $W = (Y, X, Z)$. Here, $Y_i \in \mathbb{R}$ is the outcome variable, $X_i \in \mathbb{R}$ is the running variable, and $Z_i \in \mathbb{R}^d$ is a (possibly high-dimensional) vector of covariates.[2] Units receive the treatment if and only if the running variable exceeds a known threshold, which we normalize to zero without loss of generality. We denote the treatment indicator by T_i, so that $T_i = \mathbf{1}\{X_i \geq 0\}$. The parameter of interest is the height of the jump in the conditional expectation of the observed outcome variable given the running variable at zero:

$$\tau = \mathbb{E}[Y_i|X_i = 0^+] - \mathbb{E}[Y_i|X_i = 0^-], \tag{2.1}$$

where we use the notation that $f(0^+) = \lim_{x\downarrow 0} f(x)$ and $f(0^-) = \lim_{x\uparrow 0} f(x)$ are the right and left limit, respectively, of a generic function $f(x)$ at zero. In a potential outcomes framework, the parameter τ coincides with the average treatment effect of units at the cutoff under certain continuity conditions (Hahn et al., 2001).

2.2. **Standard RD Estimator.** Without the use of covariates, the parameter τ is typically estimated by running separate local linear regressions (Fan and Gijbels, 1996) on each side of the cutoff. That is, the baseline no covariates RD estimator takes the form

$$\widehat{\tau}_{base}(h) = e_1^\top \operatorname*{argmin}_{\beta\in\mathbb{R}^4} \sum_{i=1}^{n} K_h(X_i)(Y_i - S_i^\top \beta)^2, \tag{2.2}$$

where $S_i = (T_i, X_i, T_i X_i, 1)^\top$, $K_h(v) = K(v/h)/h$ with $K(\cdot)$ a kernel function and $h > 0$ a bandwidth, and $e_1 = (1, 0, 0, 0)^\top$ is the first unit vector. This estimator is a linear smoother that can also be written as a weighted sum of the realizations of the outcome variable,

$$\widehat{\tau}_{base}(h) = \sum_{i=1}^{n} w_i(h) Y_i,$$

where the $w_i(h)$ are local linear regression weights that depend on the data through the realizations of the running variable only; see Appendix A.1 for an explicit expression.

Under standard conditions (e.g. Hahn et al., 2001), which include that the running variable is continuously distributed, and that the bandwidth h tends to zero at an appropriate rate, the estimator $\widehat{\tau}_{base}(h)$ is approximately normally distributed in large samples, with bias of order h^2 and

[2]Throughout the paper, we assume that the distribution of the running variable X_i is fixed, but we allow the conditional distribution of (Y_i, Z_i) given X_i to change with the sample size in our asymptotic analysis. In particular, we allow the dimension of Z_i to grow with n in order to accommodate high-dimensional settings, but we generally leave such dependence on n implicit in our notation.

variance of order $(nh)^{-1}$:

$$\widehat{\tau}_{base}(h) \overset{a}{\sim} N\left(\tau + h^2 B_{base}, (nh)^{-1} V_{base}\right). \tag{2.3}$$

Here "$\overset{a}{\sim}$" indicates a finite-sample distributional approximation justified by an asymptotic normality result, and the bias and variance terms are given, respectively, by

$$\begin{aligned} B_{base} &= \frac{\bar{\nu}}{2}\left(\partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^-}\right) \text{ and} \\ V_{base} &= \frac{\bar{\kappa}}{f_X(0)}(\mathbb{V}[Y_i|X_i = 0^+] + \mathbb{V}[Y_i|X_i = 0^-]). \end{aligned}$$

The terms $\bar{\nu}$ and $\bar{\kappa}$ are kernel constants, defined as $\bar{\nu} = (\bar{\nu}_2^2 - \bar{\nu}_1\bar{\nu}_3)/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)$ for $\bar{\nu}_j = \int_0^\infty v^j K(v)dv$ and $\bar{\kappa} = \int_0^\infty (K(v)(\bar{\nu}_1 v - \bar{\nu}_2))^2 dv/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)^2$, and f_X denotes the density of X_i. Practical methods for inference based on approximations like (2.3) are discussed, for instance, by Calonico et al. (2014) and Armstrong and Kolesár (2020).

2.3. **Conventional Linear Adjustment Estimator.** If covariates are available, they can be used to improve the accuracy of empirical RD estimates. The arguably most popular strategy (Calonico et al., 2019) is to include them linearly and without kernel localization in the local linear regression (2.2):

$$\widehat{\tau}_{lin}(h) = e_1^\top \underset{\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^n K_h(X_i)(Y_i - S_i^\top \beta - Z_i^\top \gamma)^2. \tag{2.4}$$

By simple least squares algebra, this "linear adjustment" estimator can be written as a no covariates estimator with covariate-adjusted outcome $Y_i - Z_i^\top \widehat{\gamma}_h$, where $\widehat{\gamma}_h$ is the minimizer with respect to γ in (2.4):

$$\widehat{\tau}_{lin}(h) = \sum_{i=1}^n w_i(h)(Y_i - Z_i^\top \widehat{\gamma}_h).$$

The linear adjustment estimator is consistent for the RD parameter without functional form assumptions on the underlying conditional expectations if the covariates are predetermined, in the sense that their values are not causally affected by the treatment, and thus their conditional expectation given the running variable varies smoothly around the cutoff. Moreover, if $\mathbb{E}[Z_i|X_i = x]$ is twice continuously differentiable around the cutoff, then

$$\widehat{\tau}_{lin}(h) \overset{a}{\sim} N\left(\tau + h^2 B_{base}, (nh)^{-1} V_{lin}\right)$$

under regularity conditions analogous to those for the no covariates estimator. Here the bias term B_{base} is the same as that of the no covariates estimator and the new variance term is

$$V_{lin} = \frac{\bar{\kappa}}{f_X(0)}(\mathbb{V}[Y_i - Z_i^\top \gamma_0|X_i = 0^+] + \mathbb{V}[Y_i - Z_i^\top \gamma_0|X_i = 0^-]),$$

where γ_0, a non-random vector of projection coefficients, is the probability limit of $\widehat{\gamma}_h$. The first-order asymptotic properties of $\widehat{\tau}_{lin}(h)$ are thus the same as that of its infeasible counterpart $\widetilde{\tau}_{lin}(h) = \sum_{i=1}^{n} w_i(h)(Y_i - Z_i^\top \gamma_0)$ that uses the population projection coefficients γ_0 instead of their estimates $\widehat{\gamma}_h$ to create the adjusted outcome variable. As $V_{lin} \leq V_{base}$ under standard conditions (Kreiß and Rothe, 2023, Remark 3.5), including a fixed number of covariates generally increases the precision of the estimator in large samples. To construct standard errors and confidence intervals, one can then use methods developed for the no covariates case, replacing the original outcome Y_i with the adjusted outcome $Y_i - Z_i^\top \widehat{\gamma}_h$ in the respective formulas (Calonico et al., 2019; Armstrong and Kolesár, 2018). For instance, one can construct a nearest-neighbor standard error $\widehat{\text{se}}_{lin}(h)$ of $\widehat{\tau}_{lin}(h)$ as

$$\widehat{\text{se}}^2_{lin}(h) = \sum_{i=1}^{n} w_i(h)^2 \widehat{\sigma}^2_{i,lin}, \qquad \widehat{\sigma}^2_{i,lin} = \frac{R}{R+1}\left((Y_i - Z_i^\top \widehat{\gamma}_h) - \frac{1}{R}\sum_{j \in \mathcal{R}_i}(Y_j - Z_j^\top \widehat{\gamma}_h)\right)^2. \tag{2.5}$$

Here $\widehat{\sigma}^2_{i,lin}$ is an estimate of $\sigma^2_{i,lin} = \mathbb{V}(Y_i - Z_i^\top \gamma_0 | X_i)$, $R \geq 1$ is a (small) fixed integer, and $\mathcal{R}_i$ is the set that contains the indices of the R nearest neighbors of unit i in terms of their realization of the running variable among units on the same side of the cutoff.

3. FLEXIBLE COVARIATE ADJUSTMENTS

3.1. **Motivation.** While linear adjustments are easy to implement, they might not exploit the available covariate information efficiently. Inference might also not be reliable with linear adjustments if the number of covariates is large relative to the effective sample size.[3] In this paper, we propose a new method to address this issue. It allows for general nonlinear covariate adjustments and can accommodate regularization methods in the estimation of the adjustment terms.

To motivate our flexible covariate adjustments, recall that the linear adjustment estimator is asymptotically equivalent to a no covariates RD estimator of the form in (2.2) that uses the covariate-adjusted outcome $Y_i - Z_i^\top \gamma_0$ instead of the original outcome Y_i. This can be generalized by considering a class of estimators with covariate-adjusted outcomes based on potentially nonlinear adjustment functions η:

$$\widehat{\tau}(h; \eta) = \sum_{i=1}^{n} w_i(h) M_i(\eta), \quad M_i(\eta) = Y_i - \eta(Z_i). \tag{3.1}$$

If the covariates are predetermined, one would arguably expect their entire conditional distribution given the running variable to vary smoothly around the cutoff. We formalize this notion by

[3] If there are many covariates relative to the number of observations that receive positive kernel weights in (2.4), the standard error in (2.5) is generally downward biased. This bias occurs because due to overfitting the local empirical variances $\widehat{\sigma}^2_{i,lin}$ are typically smaller than their population counterparts $\sigma^2_{i,lin}$ in such cases. If the number of covariates exceeds the number of observations with positive kernel weights, the estimator in equation (2.4) is of course not even well-defined in the first place.

assuming that for every adjustment function η the function $\mathbb{E}[\eta(Z_i)|X_i = x]$ is twice continuously differentiable around the cutoff.[4] This assumption implies that

$$\tau = \mathbb{E}[M_i(\eta)|X_i = 0^+] - \mathbb{E}[M_i(\eta)|X_i = 0^-] \text{ for all } \eta. \tag{3.2}$$

The estimator $\widehat{\tau}(h;\eta)$ can thus be seen as a sample analog estimator based on the moment condition (3.2), which identifies τ and is globally invariant with respect to the adjustment function η. Because of this global invariance, we expect that

$$\widehat{\tau}(h;\eta) \overset{a}{\sim} N\Big(\tau + h^2 B_{base}, (nh)^{-1} V(\eta)\Big). \tag{3.3}$$

for every η under standard regularity conditions. The bias term in (3.3) is again that of the baseline no covariates estimator. Because of the assumed smoothness of $\mathbb{E}[\eta(Z_i)|X_i = x]$, it does not depend on the adjustment function. On the other hand, the variance term in (3.3) does depend on η, and is given by

$$V(\eta) = \frac{\bar{\kappa}}{f_X(0)}(\mathbb{V}[M_i(\eta)|X_i = 0^+] + \mathbb{V}[M_i(\eta)|X_i = 0^-]).$$

To maximize the precision of the estimator $\widehat{\tau}(h;\eta)$ for any particular bandwidth h, we want to choose η such that $V(\eta)$ is as small as possible. Our analysis below shows that using the equally-weighted average of the left and right limits of the "long" conditional expectation function $\mathbb{E}[Y_i|X_i = x, Z_i = z]$ at the cutoff achieves this goal. That is, we show that

$$V(\eta) \geq V(\eta_0) \text{ for all } \eta,$$

where

$$\eta_0(z) = \frac{1}{2}\left(\mu_0^+(z) + \mu_0^-(z)\right), \quad \mu_0^\star(z) = \mathbb{E}[Y_i|X_i = 0^\star, Z_i = z] \text{ for } \star \in \{+,-\}. \tag{3.4}$$

As the optimal adjustment function η_0 is generally unknown in practice, we propose to estimate the RD parameter τ by a feasible version of $\widehat{\tau}(h;\eta_0)$.

3.2. **Proposed Estimator and its General Properties.** Our proposed estimator requires a first-stage estimate of the optimal adjustment function. This does not have to be of a particular type: practitioners can use classical nonparametric or modern machine learning methods to reduce the risk of model misspecification, or choose suitable parametric methods based on their domain knowledge (conventional linear adjustments can be seen as a special case of the latter type). Our

[4]Our analysis only rules out adjustment functions that do not satisfy certain technical regularity conditions, such as functions for which the respective conditional expectation does not exist in the first place. Assuming smoothness of $\mathbb{E}[\eta(Z_i)|X_i = x]$ for (essentially) all η is of course stronger than only assuming smoothness of $\mathbb{E}[Z_i|X_i = x]$, as in Calonico et al. (2019). Our stronger assumption, however, is still very much in line with the notion of covariates being predetermined.

proposed estimator also uses cross-fitting, which is an efficient form of sample splitting that prevents overfitting of the estimated adjustment function and avoids unrealistic empirical process conditions in the theoretical analysis (Chernozhukov et al., 2018). Specifically, our estimator is computed in two steps:

1. Randomly split the data $\{W_i\}_{i\in[n]}$ into S folds of equal size, collecting the corresponding indices in the sets I_s, for $s \in [S]$. In practice, $S = 5$ or $S = 10$ are common choices for the number of cross-fitting folds. Let $\widehat{\eta}(z) = \widehat{\eta}(z; \{W_i\}_{i\in[n]})$ be the researcher's preferred estimator of η_0, calculated on the full sample; and let $\widehat{\eta}_s(z) = \widehat{\eta}(z; \{W_i\}_{i\in I_s^c})$, for $s \in [S]$, be a version of this estimator that only uses data outside the sth fold.

2. Estimate τ by computing a local linear no covariates RD estimator that uses the adjusted outcome $M_i(\widehat{\eta}_{s(i)}) = Y_i - \widehat{\eta}_{s(i)}(Z_i)$ as the dependent variable, where $s(i)$ denotes the fold that contains observation i:
$$\widehat{\tau}(h; \widehat{\eta}) = \sum_{i=1}^{n} w_i(h) M_i(\widehat{\eta}_{s(i)}).$$

Our theoretical analysis below shows that under weak conditions the estimator $\widehat{\tau}(h; \widehat{\eta})$ is asymptotically equivalent to the infeasible estimator $\widehat{\tau}(h; \bar{\eta}) = \sum_{i=1}^{n} w_i(h) M_i(\bar{\eta})$, where $\bar{\eta}$ is a deterministic approximation of $\widehat{\eta}$ whose error vanishes in large samples in some appropriate sense. Importantly, our approach does not require the first-stage estimator of η_0 to be consistent, in the sense that we allow for the possibility that $\bar{\eta} \neq \eta_0$. The first-stage estimator also does not have to converge with a particularly fast rate. In view of (3.3), it then holds that

$$\widehat{\tau}(h; \widehat{\eta}) \overset{a}{\sim} N\left(\tau + h^2 B_{base}, (nh)^{-1} V(\bar{\eta})\right).$$

As mentioned above, the variance term $V(\bar{\eta})$ is minimized if $\bar{\eta} = \eta_0$. However, the distributional approximation is also valid if $\bar{\eta} \neq \eta_0$ because the moment condition (3.2) holds for *all* adjustment functions, and not just the optimal one. In that sense, our procedure is robust to misspecification or over-regularized estimation of the optimal adjustment function. Moreover, we argue that $V(\bar{\eta})$ is typically smaller than V_{base} or V_{lin} even if $\bar{\eta} \neq \eta_0$.

We also show that other common steps in an empirical RD analysis can easily be implemented by applying existing methods that are devised for settings without covariates to the generated data set $\{(X_i, M_i(\widehat{\eta}_{s(i)}))\}_{i\in[n]}$. For example, we can construct an estimator of the bandwidth that minimizes the asymptotic mean squared error of $\widehat{\tau}(h; \widehat{\eta})$ by using the procedures proposed by Calonico et al. (2014) or Imbens and Kalyanaraman (2012). Similarly, we can generalize the standard error (2.5) and construct a valid nearest-neighbor standard error $\widehat{\text{se}}(h; \widehat{\eta})$ as

$$\widehat{\text{se}}^2(h; \widehat{\eta}) = \sum_{i=1}^{n} w_i(h)^2 \widehat{\sigma}_i^2(\widehat{\eta}), \quad \widehat{\sigma}_i^2(\widehat{\eta}) = \frac{R}{R+1}\Big(M_i(\widehat{\eta}_{s(i)}) - \frac{1}{R}\sum_{j\in\mathcal{R}_i} M_j(\widehat{\eta}_{s(j)})\Big)^2, \tag{3.5}$$

and construct "bias-aware" and "robust bias correction" confidence intervals as in Calonico et al. (2014) and Armstrong and Kolesár (2020), respectively. To reduce the sensitivity of empirical findings to the particular data split in the cross-fitting step, we can proceed as in Chernozhukov et al. (2018, Section 3.4) by repeating the respective procedure several times and reporting a summary measure of the results, such as the median. We recommend proceeding like this especially when working with smaller sample sizes.

3.3. **Estimating the Adjustment Function.** We now discuss some implementation details for the first-stage estimator of the optimal adjustment function η_0. Our asymptotic analysis allows for a variety of different methods to be used in this context.

If one wishes to maintain the simplicity of the conventional linear adjustment, one can obtain a "cross-fitted" version of $\widehat{\tau}_{lin}(h)$ by setting $\widehat{\eta}_s(z) = z^\top \widehat{\gamma}_{s,h}$, for $s \in \{1, \dots, S\}$, where $\widehat{\gamma}_{s,h}$ is the minimizer w.r.t. γ in the minimization problem

$$\min_{\beta,\gamma} \sum_{i \in I_s^c} K_h(X_i)(Y_i - S_i^\top \beta - Z_i^\top \gamma)^2. \tag{3.6}$$

We refer to this procedure as the cross-fitted "localized" linear adjustment. The adjustment function, however, does not need to be estimated using the kernel weights from the second-stage regression. As an important variant of cross-fitted linear adjustments, we consider $\widehat{\eta}_s(z) = z^\top \widehat{\gamma}_{s,\infty}$, where $\widehat{\gamma}_{s,\infty}$ is obtained via (3.6) with "$h = \infty$", i.e., with equal kernel weights for all observations. Since all the observations outside of fold s are used to obtain $\widehat{\gamma}_{s,\infty}$, we refer to this procedure as the cross-fitted "global" linear adjustment. In finite samples, the global version can outperform the localized one in terms of the resulting standard deviation of the RD estimator due to the increased stability of the first-stage estimates. This approach can be naturally extended to other parametric models where the components involving S_i and Z_i are additively separable, and it can be combined with lasso regularization. The cross-fitted post-lasso adjustments are then obtained via (3.6) with the set of covariates restricted to those "selected" by the lasso.

More generally, our approach allows for any parametric, classical nonparametric as well as generic modern machine learning methods. To allow for such generality, we consider adjustment functions of the form

$$\widehat{\eta}_s(z) = \frac{1}{2}(\widehat{\mu}_s^+(z) + \widehat{\mu}_s^-(z)), \quad s \in \{1, \dots, S\},$$

where $\widehat{\mu}_s^+(z)$ and $\widehat{\mu}_s^-(z)$ are separate estimators of $\mu_0^+(z) = \mathbb{E}[Y_i | X_i = 0^+, Z_i = z]$ and $\mu_0^-(z) = \mathbb{E}[Y_i | X_i = 0^-, Z_i = z]$, respectively, using the data outside of fold s. With appropriate subject knowledge, one can then, for example, specify parametric models for $\mathbb{E}[Y_i | X_i = x, Z_i = z]$. If the number of covariates is small, the functions μ_0^+ and μ_0^- can be also estimated using classical nonparametric methods under smoothness conditions, with local polynomial regression being particularly suitable due to its good boundary properties. If the number of covariates is large, however, we recommend using modern machine learning methods, such as lasso or post-lasso regression,

random forests, deep neural networks, boosting, or ensemble combinations thereof.

One issue to consider is that the default implementations of generic machine learning estimators of $\mathbb{E}[Y_i|X_i = x, Z_i = z]$ will not automatically produce an estimate with a jump at the cutoff. As having this feature is potentially important in our context, we consider two simple variations of generic machine learning estimators. To define the first type, let

$$\widehat{\mathbb{E}}_s[Y_i|T_i = t, X_i = x, Z_i = z] = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i\in I_s^c} l(Y_i, f(T_i, X_i, Z_i)) \tag{3.7}$$

be a generic machine learning estimator of $\mathbb{E}[Y_i|T_i = t, X_i = x, Z_i = z]$, computed by minimizing some empirical loss function $L(f) = \sum_{i\in I_s^c} l(Y_i, f(T_i, X_i, Z_i))$ over a set of candidate functions $\mathcal{F}$. We can then estimate $\mu^+(z)$ by $\widehat{\mathbb{E}}_s[Y_i|T_i = 1, X_i = 0, Z_i = z]$ and $\mu^-(z)$ by $\widehat{\mathbb{E}}_s[Y_i|T_i = 0, X_i = 0, Z_i = z]$. Here including the seemingly superfluous treatment indicator $T_i = \mathbf{1}\{X_i \geq 0\}$ as a predictor allows the machine learner to create the "jump" in the estimated function at the cutoff value. We refer to this type of implementation as "global", as it uses all available observations.

To define the second type of implementation of machine learning we consider in this paper, let

$$\widehat{\mathbb{E}}_s[Y_i|T_i = t, Z_i = z] = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i\in I_s^c} K(X_i/b) l(Y_i, f(T_i, Z_i)) \tag{3.8}$$

be a generic machine learning estimator of $\mathbb{E}[Y_i|T_i = t, Z_i = z]$, where $b > 0$ is some positive bandwidth and K is again a kernel function. We can then estimate $\mu^+(z)$ as $\widehat{\mathbb{E}}_s[Y_i|T_i = 1, Z_i = z]$ and $\mu^-(z)$ as $\widehat{\mathbb{E}}_s[Y_i|T_i = 0, Z_i = z]$. We refer to this type of implementation as "localized", as it effectively only uses data points whose realization of the running variable is close to the cutoff. The idea is to produce an estimate with small empirical loss in the relevant area around the cutoff rather than one with small "overall" loss. The downside of this approach is the reduced effective sample size and the need to choose the tuning parameter b.[5]

3.4. **Our Proposed Flexible Adjustment.** The specific flexible covariate adjustment that we propose and implement in our empirical analysis and simulations is an ensemble of the following methods: (i) linear regression; (ii) post-lasso; (iii) boosted trees; and (iv) random forest. All four methods are implemented in localized and global versions discussed above. We use the cross-fitted linear and post-lasso adjustments specified in the discussion following (3.6), and the boosted trees and random forest adjustments are based on the formulations in (3.7) and (3.8). Our proposed

[5]The choice of b involves a bias-variance trade-off similar to the one encountered in classical nonparametric kernel regression problems. We are not aware of generic theoretical results for such estimators in settings with $b \to 0$ as $n \to \infty$. Specific results are given by Su et al. (2019) for the lasso, and by Colangelo and Lee (2022) for series estimators and deep neural networks. In our applications below, he simply use $b = h$. To make this simultaneous choice feasible, we use an iterative procedure. We first choose a reasonable preliminary first-stage bandwidth, like the one that would be optimal for RD estimation without covariates, and generate preliminary versions of the adjustment terms as described above. Next, we use the preliminary covariate-adjusted outcomes to pick an optimized second-stage bandwidth. Finally, we rerun both stages with this last bandwidth to obtain our empirical results.

flexible covariate adjustment is a convex combination of these eight adjustment functions and the trivial no-adjustment function that minimizes the mean squared error for predicting the outcome close to the cutoff. Specifically, we employ the super learning approach of Van der Laan et al. (2007), where the optimal weights are chosen via cross-validation.[6]

4. PRACTICAL PERFORMANCE

Between 2018 and 2023, the main AEA journals for applied microeconomic research published 16 empirical RD studies that use covariates, fit into our general framework, and have directly available public replication data. To illustrate the scope for efficiency improvements that flexible covariate adjustments can achieve in practically relevant settings, we compare the performance of our proposed method to that of existing ones on the 56 main empirical specifications considered in these papers. Specifically, we compute the length of bias-aware 95% confidence intervals (as described in Section 6.1) for the respective RD parameter in each specification based on local linear estimators that use our flexible covariate adjustments, linear adjustments, and no covariate adjustments, respectively. See Appendix C for further details on the implementation and the data collection process, and Table S2 in the Online Supplement for the complete list of papers and specifications used.

Before turning to the results of this exercise, we want to comment briefly on the empirical relevance of concerns about the bias of the usual standard error of the conventional linear adjustment estimator in settings with rather many covariates relative to the effective sample size mentioned in Section 3. From Table S2, we can see that the ratio of the effective sample size to the number of covariates ranges from 2.7 to 23,254, with a median value of 112, across the 56 specifications under consideration in this section. In about one-fifth of the specifications, the ratio falls below 20.[7] Empirical RD specifications with rather high-dimensional covariate adjustments are thus not uncommon in the recent literature. To take this into account, we use cross-fitting with both flexible and linear adjustments in this section.

Turning to the results of our empirical exercise, Figure 1 shows the distribution of the ratio of confidence interval lengths for flexible adjustments relative to no adjustments in its left panel, and for flexible adjustments relative to linear adjustments in its right panel. From the left panel, we

[6]We implemented our procedure in the `R` programming language. The boosted trees adjustments are obtained using the package `xgboost` with trees of depth 2 and shrinkage rate 0.1. The number of boosting iterations is chosen via cross-validation with a maximum of 1,000 iterations, separately for the localized and global versions. The random forest with 1,000 trees is implemented using the package `ranger` with the minimal node size set to the maximum of 10 and 0.1% of the sample size. All other parameters are set to the default values in the respective packages. For post-lasso estimation, we use the function `rlasso` from the package `hdm` with a data-driven penalty parameter. For choosing the cross-validated optimal weights, we use the package `SuperLearner`.

[7]In separate simulations in Section 7, we show that the conventional standard error of the linear adjustment estimator exhibits a moderate 10% downward bias if the ratio of effective sample size and the number of covariates is around 20, and a substantial bias of more than 20% if that ratio falls below a value around 8, and that cross-fitting can remove this bias almost completely.

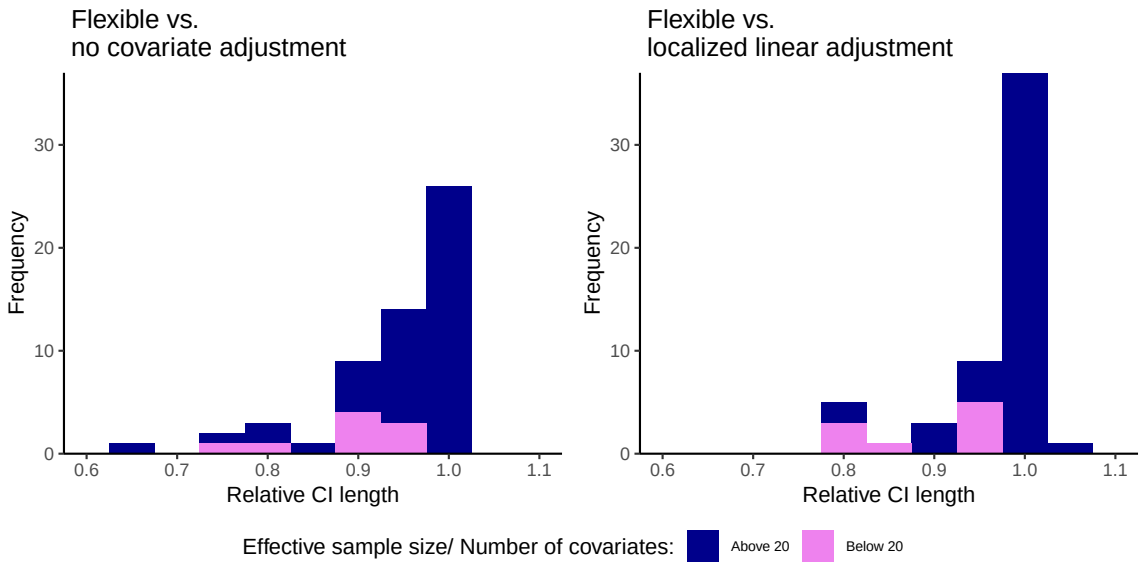

Figure 1: CI lengths with flexible covariate adjustment and with cross-fitted localized linear adjustment relative to CI length with no covariates for all specifications of the empirical literature survey. We choose the bandwidth and conduct inference using the bias-aware approach with smoothness constants calibrated via the rule of thumb of Armstrong and Kolesár (2020) for each of the specification separately. The effective sample size refers to the no covariates estimator of the respective specification.

can see that in about half of our specifications the flexible covariate adjustments yield confidence intervals that are not noticeably shorter than the ones obtained without covariate adjustments. Given the flexibility of our methods, this suggests that the covariates are not very informative about the outcome in these specifications, and that there is hence no scope for efficiency gains. In many specifications, however, flexible covariate adjustments lead to substantially shorter confidence intervals, with the biggest reduction being greater than 35%. To put this into perspective, note that one would have to increase the sample size used by an unadjusted "baseline" RD estimator by a factor of about 3 to receive a similar reduction in the length of the confidence interval. From the right panel of Figure 1, we can see that linear adjustments are typically unable to exhaust the available covariate information. Indeed, the confidence intervals based on flexible adjustments can be substantially shorter, with the biggest reduction amounting to 21% in our empirical exercise.

5. MAIN THEORETICAL RESULTS

5.1. **Assumptions.** We study the theoretical properties of our proposed estimator under a number of conditions that are either standard in the RD literature, or concern the general properties of the first-stage estimator $\widehat{\eta}$. To describe them, we denote the support of Z_i by $\mathcal{Z}$, and the support of X_i by $\mathcal{X}$. We write $\mathcal{X}_h = \mathcal{X} \cap [-h, h]$, and $\mathcal{Z}_h$ denotes the support of Z_i given $X_i \in \mathcal{X}_h$. We also

define the following class of admissible adjustment functions:

$$\mathcal{E} = \{\eta : \mathbb{E}[\eta(Z_i)|X_i = x] \text{ exists and is twice continuously differentiable around the cutoff}\}.$$

The class $\mathcal{E}$ implicitly depends on the underlying conditional distribution of the covariates given the running variable. If this conditional distribution changes smoothly around the cutoff, the class $\mathcal{E}$ contains essentially all functions of the covariates, subject only to technical integrability conditions.[8]

Assumption 1. *For all $n \in \mathbb{N}$, there exist a set $\mathcal{T}_n \subset \mathcal{E}$ and a function $\bar{\eta} \in \mathcal{T}_n$ such that: (i) $\widehat{\eta}_s$ belongs to $\mathcal{T}_n$ with probability approaching 1 for all $s \in [S]$; (ii) it holds that:*

$$\sup_{\eta \in \mathcal{T}_n} \sup_{x \in \mathcal{X}_h} \mathbb{E}\left[(\eta(Z_i) - \bar{\eta}(Z_i))^2 | X_i = x\right] = O(r_n^2)$$

for some deterministic sequence $r_n = o(1)$.

Assumption 1 states that with high probability the first-stage estimator belongs to some realization set $\mathcal{T}_n \subset \mathcal{E}$. As discussed above, this requirement seems weak as we generally expect the class $\mathcal{E}$ to be very large. The assumption also states that the sets $\mathcal{T}_n$ contract around a deterministic sequence of functions in a particular L_2-type sense. Note that taking the supremum in Assumption 1 over $\mathcal{X}_h$ instead of $\mathcal{X}$ suffices as the properties of the first stage estimator are only relevant for observations with non-zero kernel weights in the second-stage local linear regression. The assumption does not impose any restrictions on the speed at which $\widehat{\eta}$ concentrates around $\bar{\eta}$. It also allows the function $\bar{\eta}$ to be different from the target function η_0, which means that $\widehat{\eta}$ can be inconsistent for η_0.

Mean-square error consistency as prescribed in Assumption 1 follows under classical conditions for the parametric and nonparametric procedures for settings in which the number of covariates is fixed. For the "localized" versions of the machine learning methods described in Section 3.3, existing results imply that for fixed $b > 0$ and K the uniform kernel,

$$\sup_{\eta \in \mathcal{T}_n} \mathbb{E}\left[(\eta(Z_i) - \bar{\eta}(Z_i))^2 | X_i \in (-b, b)\right] = O(r_n^2), \tag{5.1}$$

with $\bar{\eta}(z) = (\mathbb{E}[Y_i|X_i \in (-b, 0), Z_i = z] + \mathbb{E}[Y_i|X_i \in (0, b), Z_i = z])/2$ and some $r_n = o(1)$, under general conditions. For example, if $\bar{\eta}(z)$ is contained in a Hölder class of order s, then (5.1) can hold with $r_n^2 = n^{-2s/(2s+d)}$ for estimators that exploit smoothness. If $\bar{\eta}(z)$ is s-sparse, then (5.1) can hold with $r_n^2 = s \log(d)/n$ for estimators that exploit sparsity. Assumption 1 then follows from (5.1) if the conditional distribution of the covariates does not change "too quickly" when moving away from the cutoff. For example, if the covariates are continuously distributed conditional on the running

[8] For example, if the conditional distribution of Z_i given X_i admits a density $f_{Z|X}(z|x)$ that is twice continuously differentiable in x and $|\partial_x^j f_{Z|X}(z|x)| \leq g_j(z)$ for all x in a neighborhood of the cutoff, some integrable functions g_j, and $j \in \{0, 1, 2\}$, then $\mathcal{E}$ contains all bounded Borel functions. The class $\mathcal{E}$ also contains all polynomials if the corresponding conditional moments of Z_i exist and are twice continuously differentiable.

variable, having that

$$\sup_{x\in\mathcal{X}_h}\sup_{z\in\mathcal{Z}_h}\frac{f_{Z|X}(z|x)}{f_{Z|X\in(-b,b)}(z)}<C,$$

for some constant C and all n sufficiently large, suffices. Similar conditions can be given for discrete conditional covariate distributions, or intermediate cases. If $\mathbb{E}[Y_i|X_i=x,Z_i=z]$ is sufficiently smooth in x on both sides of the cutoff, we can also expect that $\bar{\eta}$ is "close" to η_0 for "small" values of b. Formal rate results with $b\to 0$ are given by Su et al. (2019) for the Lasso, and by Colangelo and Lee (2022) for series estimators and deep neural networks.

Assumption 2. *For $j\in\{1,2\}$, it holds that:*

$$\sup_{\eta\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h\setminus\{0\}}\left|\partial_x^j\mathbb{E}\left[\eta(Z_i)-\bar{\eta}(Z_i)|X_i=x\right]\right|=O(v_{j,n}).$$

for some deterministic sequences $v_{j,n}=o(1)$.

Assumption 2 also concerns the first-stage estimator, and requires the first and second derivatives of $\mathbb{E}\left[\eta(Z_i)-\bar{\eta}(Z_i)|X_i=x\right]$ to be close to zero in large samples for all $\eta\in\mathcal{T}_n$. We generally expect this condition to hold with $v_{1,n}=v_{2,n}=r_n$, where r_n is as in Assumption 1.[9]

Assumption 3. *X_i is continuously distributed with density f_X, which is continuous and bounded away from zero over an open neighborhood of the cutoff.*

Assumption 3 is a standard condition from the RD literature. Continuity of the running variable's density f_X around the cutoff is strictly speaking not required for an RD analysis. However, a discontinuity in f_X is typically considered to be an indication of a design failure that prevents τ from being interpreted as a causal parameter (McCrary, 2008; Gerard et al., 2020). For this reason, we focus on the case of a continuous running variable density in this paper.

Assumption 4. *(i) The kernel function K is a bounded and symmetric density function that is continuous on its support, and equal to zero outside some compact set, say $[-1,1]$; (ii) The bandwidth satisfies $h\to 0$ and $nh\to\infty$ as $n\to\infty$.*

The conditions on the kernel and the bandwidth that are imposed in Assumption 4 are standard in the RD literature.

Assumption 5. *There exist constants C and L such that the following conditions hold for all $n\in\mathbb{N}$. (i) $\mathbb{E}[M_i(\bar{\eta})|X_i=x]$ is twice continuously differentiable on $\mathcal{X}\setminus\{0\}$ with L-Lipschitz continuous*

[9] For example, this can easily be seen to be the case if $\widehat{\eta}$ converges to $\bar{\eta}$ uniformly over $\mathcal{Z}$ with rate r_n and the smoothness conditions for $f_{Z|X}(z|x)$ given in footnote 8 hold. Similarly, under regularity conditions on $\mathbb{E}[Z_i|X_i=x]$, these three rates coincide if $\mathcal{T}_n$ contains only linear functions. Without any additional restrictions on first stage estimators or $\mathcal{T}_n$, except that it contains only bounded functions, Assumption 2 also follows from Assumption 1, again with $v_{1,n}=v_{2,n}=r_n$, under restrictions concerning solely the conditional density $f_{Z|X}(z|x)$. Specifically, it suffices that $\mathbb{E}\left[\left(\partial_x^j f_{Z|X}(Z_i|x)/f_{Z|X}(Z_i|x)\right)^2|X_i=x\right]$ is bounded for $j\in\{1,2\}$ uniformly in x and the conditions from footnote 8 hold.

second derivative bounded by C; (ii) For all $x \in \mathcal{X}$ and some $q > 2$ $\mathbb{E}[(M_i(\bar{\eta}) - \mathbb{E}[M_i(\bar{\eta})|X_i])^q|X_i = x]$ exists and is bounded by C; (iii) $\mathbb{V}[M_i(\bar{\eta})|X_i = x]$ is L-Lipschitz continuous and bounded from below by $1/C$ for all $x \in \mathcal{X} \setminus \{0\}$.

Assumption 5 collects standard conditions for an RD analysis with $M_i(\bar{\eta})$ as the outcome variable. Part (i) imposes smoothness conditions on $\mathbb{E}[M_i(\bar{\eta})|X_i = x]$, and parts (ii) and (iii) impose restrictions on conditional moments of the outcome variable. Throughout, we use constants C and L independent of the sample size to ensure asymptotic normality of the infeasible estimator $\widehat{\tau}(h; \bar{\eta})$ even in settings where the distribution of the data, and thus $\bar{\eta}$, might change with n.

5.2. **Asymptotic Properties.** We give four main results in this subsection. The first shows that our proposed estimator $\widehat{\tau}(h; \widehat{\eta})$ is asymptotically equivalent to an infeasible analog $\widehat{\tau}(h; \bar{\eta})$ that replaces the estimator $\widehat{\eta}$ with the deterministic sequence $\bar{\eta}$; the second shows the asymptotic normality of the estimator; the third characterizes how the asymptotic variance changes with the adjustment function and shows that η_0 is indeed the optimal adjustment; and the fourth shows the impact of flexible covariate adjustments on the optimal bandwidth and the corresponding mean squared error.

Theorem 1. *Suppose that Assumptions 1–4 hold. Then*

$$\widehat{\tau}(h; \widehat{\eta}) = \widehat{\tau}(h; \bar{\eta}) + O_P(r_n(nh)^{-1/2} + v_{1,n}h(nh)^{-1/2} + v_{2,n}h^2).$$

Theorem 1 is easiest to interpret in what is arguably the standard case that $v_{1,n} = v_{2,n} = r_n$, in which it holds that

$$\widehat{\tau}(h; \widehat{\eta}) = \widehat{\tau}(h; \bar{\eta}) + O_P(r_n(h^2 + (nh)^{-1/2})) = \widehat{\tau}(h; \bar{\eta}) + O_P(r_n|\widehat{\tau}(h; \bar{\eta}) - \tau|).$$

The accuracy of the approximation that $\widehat{\tau}(h; \widehat{\eta}) \approx \widehat{\tau}(h; \bar{\eta})$ thus increases with the rate at which $\widehat{\eta}$ concentrates around $\bar{\eta}$, but first-order asymptotic equivalence holds even if the first-stage estimator converges arbitrarily slowly. This insensitivity of $\widehat{\tau}(h; \widehat{\eta})$ to sampling variation in $\widehat{\eta}$ occurs because $\widehat{\tau}(h; \widehat{\eta})$ is based on the moment condition

$$\tau = \mathbb{E}[M_i(\eta)|X_i = 0^+] - \mathbb{E}[M_i(\eta)|X_i = 0^-],$$

which is insensitive to variation in η. Moment conditions with a local form of insensitivity with respect to a nuisance function, often called Neyman orthogonality, are used extensively in the recent literature on two-stage estimators that use machine learning in the first stage (e.g. Belloni et al., 2017; Chernozhukov et al., 2018). The global insensitivity that arises in our RD setup is stronger, and allows us to work with weaker conditions on the first-stage estimates than those used in papers that work with Neyman orthogonality. Similarly, globally insensitive moment function exists, for example, in certain types of randomized experiments, and our proposed estimator is in many ways

analogous to efficient estimators in such setups; see Section 6.2 for further discussion.

Theorem 2. *Suppose that Assumptions 1–5 hold. Then*

$$\sqrt{nh}\,V(\bar{\eta})^{-1/2}\left(\widehat{\tau}(h;\widehat{\eta})-\tau-h^2B_n\right)\xrightarrow{d}\mathcal{N}(0,1),$$

for some $B_n = B_{base} + o_P(1)$, where B_{base} and $V(\cdot)$ are as defined in Sections 2.2 and 3.1, respectively.

Theorem 2 follows from Theorem 1 under the additional regularity conditions of Assumption 5. It shows that our estimator is asymptotically normal, gives explicit expressions for its asymptotic bias and variance, and justifies the distributional approximation given in Section 3.2.

Theorem 3. *Suppose $\mathbb{E}[Y_i^2|X_i=x]$ is uniformly bounded in x, the limit $\mathbb{V}[Y_i-\mu_0^\star(Z_i)|X_i=0^\star]$ exists for $\star\in\{+,-\}$, and $\eta_0\in\mathcal{V}$, where the function class $\mathcal{V}$ is defined as*

$$\mathcal{V}\equiv\left\{\eta:\mathbb{V}[\eta(Z_i)|X=x]\text{ and }\mathrm{Cov}[\eta(Z_i),\mu_0^\star(Z_i)|X_i=x]\text{ are continuous for }\star\in\{+,-\}\right\}.$$

Then, for any $\eta^{(a)},\eta^{(b)}\in\mathcal{V}$,

$$V(\eta^{(a)})-V(\eta^{(b)})=2\frac{\bar{\kappa}}{f_X(0)}\left(\mathbb{V}[\eta_0(Z_i)-\eta^{(a)}(Z_i)|X_i=0]-\mathbb{V}[\eta_0(Z_i)-\eta^{(b)}(Z_i)|X_i=0]\right).$$

Theorem 3 introduces a function class $\mathcal{V}$ that, similarly to the class $\mathcal{E}$ above, enforces some technical integrability conditions. The theorem shows that $V(\eta^{(a)})<V(\eta^{(b)})$ for generic adjustment functions $\eta^{(a)}$ and $\eta^{(b)}$ if and only if $\mathbb{V}[\eta_0(Z_i)-\eta^{(a)}(Z_i)|X_i=0]<\mathbb{V}[\eta_0(Z_i)-\eta^{(b)}(Z_i)|X_i=0]$. That is, the "closer" (in a particular L_2-sense) the adjustment function is to the optimal one, the smaller the asymptotic variance. In consequence, the lowest possible value of $V(\bar{\eta})$ is achieved for $\bar{\eta}=\eta_0$. Even if $\bar{\eta}\neq\eta_0$, our flexible covariate adjustments typically still have smaller asymptotic variances than the "no covariates" and "linear adjustments" RD estimators. Specifically, $V(\bar{\eta})<V_{base}$ if and only if $\mathbb{V}[\eta_0(Z_i)-\bar{\eta}(Z_i)|X_i=0]<\mathbb{V}[\eta_0(Z_i)|X_i=0]$, i.e. whenever $\bar{\eta}(Z_i)$ captures *some* of the variance of $\eta_0(Z_i)$ among units near the cutoff; and $V(\bar{\eta})<V_{lin}$ if and only if $\mathbb{V}[\eta_0(Z_i)-\bar{\eta}(Z_i)|X_i=0]<\mathbb{V}[\eta_0(Z_i)-Z_i^\top\gamma_0|X_i=0]$, i.e. whenever $\bar{\eta}$ is "closer" to η_0 in our particular L_2-sense than the population linear adjustment function.

Theorem 4. *Let $\mathrm{AMSE}(h,\eta)=h^4B_{base}^2+(nh)^{-1}V(\eta)$ be the approximate (first-order) mean squared error of $\widehat{\tau}(h;\eta)$, and $h_{AMSE}(\eta)=\mathrm{argmin}_h\,\mathrm{AMSE}(h,\eta)=n^{-1/5}\left(V(\eta)/4B_{base}^2\right)^{1/5}$ the corresponding optimal bandwidth. Then for any pair of adjustment functions $\eta^{(a)},\eta^{(b)}\in\mathcal{V}$ we have*

$$\frac{h_{AMSE}(\eta^{(a)})}{h_{AMSE}(\eta^{(b)})}=\left(\frac{v(\eta^{(a)})}{v(\eta^{(b)})}\right)^{1/5}\quad\text{and}\quad\frac{\mathrm{AMSE}(h_{AMSE}(\eta^{(a)}),\eta^{(a)})}{\mathrm{AMSE}(h_{AMSE}(\eta^{(b)}),\eta^{(b)})}=\left(\frac{v(\eta^{(a)})}{v(\eta^{(b)})}\right)^{4/5},$$

where $v(\eta)=\mathbb{V}[M_i(\eta)|X_i=0^+]+\mathbb{V}[M_i(\eta)|X_i=0^-]$.

Theorem 4 implies that flexible covariate adjustments can reduce the (approximate) mean squared error of our estimator not only directly through a smaller asymptotic variance term but also indirectly through a change in the optimal bandwidth and a corresponding reduction in bias. That is, if $V(\eta^{(a)}) < V(\eta^{(b)})$ for generic adjustment functions $\eta^{(a)}$ and $\eta^{(b)}$, then the optimal bandwidth $h_{AMSE}(\eta^{(a)})$ is smaller than $h_{AMSE}(\eta^{(b)})$, and the corresponding estimator $\widehat{\tau}(h_{AMSE}(\eta^{(a)}); \eta^{(a)})$ has both smaller asymptotic bias *and* smaller asymptotic variance than $\widehat{\tau}(h_{AMSE}(\eta^{(b)}); \eta^{(b)})$.

6. ADDITIONAL THEORETICAL RESULTS AND DISCUSSIONS

6.1. **Bandwidth Choice and Inference.** We formally show in Appendix B that standard methods for bandwidth choice and confidence interval construction based on the no covariates RD estimator maintain their general asymptotic properties when they are applied to the generated data set $\{(X_i, M_i(\widehat{\eta}_{s(i)}))\}_{i\in[n]}$ without any adjustment for the sampling uncertainty about the estimated adjustment function. Specifically, we derive three groups of results, all under conditions that are rather weak and analogous to those commonly imposed in setups without covariates.

First, we show that the nearest neighbor standard error (3.5) is consistent, in the sense that

$$nh\,\widehat{\mathrm{se}}^2(h;\widehat{\eta})/V(\bar{\eta}) \xrightarrow{p} 1.$$

Second, we show that commonly used methods for confidence interval construction achieve correct asymptotic coverage. For example, assuming a bound on $|\partial_x^2 \mathbb{E}[Y_i|X_i = x]|$, the absolute value of the second derivative of the conditional expectation of the outcome given the running variable, one can construct a "bias-aware" confidence interval as in Armstrong and Kolesár (2020) as

$$CI^{ba}_{1-\alpha} = \left[\widehat{\tau}(h;\widehat{\eta}) \pm z_\alpha(\bar{b}(h)/\widehat{\mathrm{se}}(h;\widehat{\eta}))\,\widehat{\mathrm{se}}(h;\widehat{\eta})\right].$$

Here $z_\alpha(r)$ is the $1-\alpha/2$ quantile of $|N(r,1)|$, the absolute value of the normal distribution with mean r and variance one and $\bar{b}(h)$ is an explicit bound on the finite sample bias of the no covariates RD estimator given in the Appendix. Alternatively, one can also construct a "robust bias correction" confidence interval as in Calonico et al. (2014) by subtracting a local quadratic estimate of the first-order bias of $\widehat{\tau}(h;\widehat{\eta})$ from the estimator, and adjusting the standard error appropriately. This yields a confidence interval of the form

$$CI^{rbc}_{1-\alpha} = \left[\widehat{\tau}^{rbc}(h;\widehat{\eta}) \pm z_\alpha \widehat{\mathrm{se}}^{rbc}(h;\widehat{\eta})\right],$$

where $z_\alpha = z_\alpha(0)$ and the other terms are formally defined in the appendix. Third, we show that the "MSE-optimal" bandwidth selector $\widehat{h}_n$ of Calonico et al. (2014), which is similar to that of Imbens and Kalyanaraman (2012), consistently estimates the AMSE optimal bandwidth $h_{AMSE}(\bar{\eta})$ defined in Theorem 4, in the sense that

$$\widehat{h}_n / h_{AMSE}(\bar{\eta}) \xrightarrow{p} 1.$$

RD estimation and inference with flexible covariate adjustments are thus easy to implement with existing software packages.

6.2. **Analogies with Randomized Experiments.** The results in Section 5 are qualitatively similar to ones obtained for efficient influence function (EIF) estimators of the population average treatment effect (PATE) in randomized experiments with known and constant propensity scores (e.g., Wager et al., 2016; Chernozhukov et al., 2018). To see this, consider a randomized experiment with unconfounded treatment assignment and a known and constant propensity score p. Using our notation in an analogous fashion, the EIF of the PATE in such a setup is typically given in the literature (e.g., Hahn, 1998) in the form

$$\psi_i(m_0^0, m_0^1) = m_0^1(Z_i) - m_0^0(Z_i) + \frac{T_i(Y_i - m_0^1(Z_i))}{p} - \frac{(1-T_i)(Y_i - m_0^0(Z_i))}{1-p},$$

where $m_0^t(z) = \mathbb{E}[Y_i | Z_i = z, T_i = t]$ for $t \in \{0,1\}$. The minimum variance any regular estimator of the PATE can achieve is thus $V_{\text{PATE}} = \mathbb{V}(\psi_i(m_0^0, m_0^1))$. By randomization, it also holds that $\tau_{\text{PATE}} = \mathbb{E}[\psi_i(m^0, m^1)]$ for all (suitably integrable) functions m^0 and m^1, and thus the PATE is identified by a moment function that satisfies a global invariance property. A sample analog estimator of τ_{PATE} based on this moment function reaches has asymptotic variance V_{PATE} if $\widehat{m}^t$ is a consistent estimator of m_0^t for $t \in \{0,1\}$, but remains consistent and asymptotically normal with asymptotic variance $\mathbb{V}(\psi_i(\bar{m}^0, \bar{m}^1))$ if $\widehat{m}^t$ is consistent for some other function $\bar{m}^t$, $t \in \{0,1\}$. The convergence of $\widehat{m}^t$ to $\bar{m}^t$ can be arbitrarily slow for these results (e.g. Wager et al., 2016; Chernozhukov et al., 2018).

The qualitative parallels between these findings and ours in Section 5 arise because our covariate-adjusted RD estimator is in many ways a direct analog of such EIF estimators. To show this, write $m(z) = (1-p)m^1(z) + p\, m^0(z)$ for any two functions m^0 and m^1, so that $m_0(z) = (1-p)m_0^1(z) + p\, m_0^0(z)$. The PATE's influence function can then be expressed as

$$\psi_i(m_0^0, m_0^1) = \frac{T_i(Y_i - m_0(Z_i))}{p} - \frac{(1-T_i)(Y_i - m_0(Z_i))}{1-p},$$

and it holds that

$$\mathbb{E}[\psi_i(m^0, m^1)] = \mathbb{E}[Y_i - m(Z_i) | T_i = 1] - \mathbb{E}[Y_i - m(Z_i) | T_i = 0],$$

which is the difference in average covariate-adjusted outcomes between treated and untreated units. This last equation is fully analogous to our equation (3.2), with $p = 1/2$, and conditioning on $T_i = 1$ and $T_i = 0$ replaced by conditioning on X_i in infinitesimal right and left neighborhoods of the cutoff (the value $p = 1/2$ is appropriate here because continuity of the running variable's density implies that an equal share of units close to the cutoff can be found on either side). An EIF estimator of τ_{PATE} is thus analogous to our estimator $\widehat{\tau}(h; \widehat{\eta})$, as they are both sample analogs

a moment function with the same basic properties.

6.3. **Fuzzy RD Designs.** In fuzzy RD designs, units are assigned to treatment if their realization of the running variable falls above the threshold value, but might not comply with their assignment. The conditional treatment probability given the running variable hence changes discontinuously at the cutoff, but in contrast to sharp RD designs it does not jump from zero to one. The parameter of interest in fuzzy RD designs is

$$\theta = \frac{\tau_Y}{\tau_T} \equiv \frac{\mathbb{E}[Y_i|X_i = 0^+] - \mathbb{E}[Y_i|X_i = 0^-]}{\mathbb{E}[T_i|X_i = 0^+] - \mathbb{E}[T_i|X_i = 0^-]},$$

which is the ratio of two sharp RD estimands (throughout this subsection, the notation is analogous to that used before, with the subscripts Y and T referencing the respective outcome variable). Under standard conditions (Hahn et al., 2001; Dong, 2018), one can interpret θ as the average causal effect of the treatment among units at the cutoff whose treatment decision is affected by whether their value of the running variable is above or below the cutoff.

Similarly to sharp RD designs, predetermined covariates can be used in fuzzy RD designs to improve efficiency. Building on our proposed method, we consider estimating θ by the ratio of two generic flexible covariate-adjusted sharp RD estimators:

$$\widehat{\theta}(h; \widehat{\eta}_Y, \widehat{\eta}_T) = \frac{\widehat{\tau}_Y(h; \widehat{\eta}_Y)}{\widehat{\tau}_T(h; \widehat{\eta}_T)} = \frac{\sum_{i=1}^n w_i(h)(Y_i - \widehat{\eta}_{Y,s(i)}(Z_i))}{\sum_{i=1}^n w_i(h)(T_i - \widehat{\eta}_{T,s(i)}(Z_i))}.$$

Proposition 1. *Suppose that Assumptions 1–5 hold also with T_i replacing Y_i, mutatis mutandis.*

(i) It holds that

$$\sqrt{nh}\, V_\theta(\bar{\eta}_Y, \bar{\eta}_T)^{-1/2} \left(\widehat{\theta}(h; \widehat{\eta}_Y, \widehat{\eta}_T) - \theta - B_\theta(\bar{\eta}_Y, \bar{\eta}_T) h^2 \right) \xrightarrow{d} \mathcal{N}(0, 1),$$

where

$$B_\theta(\bar{\eta}_Y, \bar{\eta}_T) = \frac{\bar{\nu}}{2\tau_T} \left(\partial_x^2 \mathbb{E}[Y_i - \theta T_i | X_i = x]\big|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i - \theta T_i | X_i = x]\big|_{x=0^-} \right) + o_P(1),$$
$$V_\theta(\bar{\eta}_Y, \bar{\eta}_T) = \frac{\bar{\kappa}}{f_X(0)} \left(\mathbb{V}[U_i(\bar{\eta}_Y, \bar{\eta}_T) | X_i = 0^+] + \mathbb{V}[U_i(\bar{\eta}_Y, \bar{\eta}_T) | X_i = 0^-] \right),$$

and $U_i(\bar{\eta}_Y, \bar{\eta}_T) = (Y_i - \theta T_i - (\bar{\eta}_Y(Z_i) - \theta \bar{\eta}_T(Z_i)))/\tau_T$.

(ii) Suppose additionally that the assumptions of Theorem 3 hold, mutatis mutandis, also with T_i replacing Y_i and the definition of $\mathcal{V}$ adjusted accordingly. Then, for any $\eta_Y^{(a)}, \eta_Y^{(b)}, \eta_T^{(a)}, \eta_T^{(b)} \in \mathcal{V}$,

it holds that

$$
\begin{aligned}
&V_\theta(\eta_Y^{(a)}, \eta_T^{(a)}) - V_\theta(\eta_Y^{(b)}, \eta_T^{(b)}) \\
&\quad = \frac{2\bar{\kappa}}{\tau_T^2 f_X(0)} \Big(\mathbb{V}[\eta_{Y,0}(Z_i) - \theta\eta_{T,0}(Z_i) - (\eta_Y^{(a)}(Z_i) - \theta\eta_T^{(a)}(Z_i))|X_i = 0] \\
&\qquad - \mathbb{V}[\eta_{Y,0}(Z_i) - \theta\eta_{T,0}(Z_i) - (\eta_Y^{(b)}(Z_i) - \theta\eta_T^{(b)}(Z_i))|X_i = 0] \Big).
\end{aligned}
$$

The first part of the proposition shows that our flexible covariate-adjusted fuzzy RD estimator is asymptotically normal, with asymptotic variance that depends on the population counterparts $\bar{\eta}_Y$ and $\bar{\eta}_T$ of the two estimated adjustment functions. This result can then be used to construct a confidence interval for θ based on the t-statistic. Alternatively, confidence sets for θ can be constructed via an Anderson-Rubin-type approach, which circumvents certain problems of ratio estimators (Noack and Rothe, 2024).

The second part of the proposition shows that the asymptotic variance of our estimator is minimized if the estimated adjustment functions concentrate around $\bar{\eta}_Y = \eta_{Y,0}$ and $\bar{\eta}_T = \eta_{T,0}$, respectively. That is, the optimal adjustment functions for fuzzy RD designs can be obtained by separately considering two covariate-adjusted sharp RD problems with outcomes Y_i and T_i, respectively. This holds because for fixed adjustment functions η_Y and η_T we have that $\widehat{\theta}(h; \eta_Y, \eta_T) - \theta$ is first-order asymptotically equivalent to a sharp RD estimator with the infeasible outcome $U_i(\eta_Y, \eta_T) = (Y_i - \theta T_i - (\eta_Y(Z_i) - \theta\eta_T(Z_i)))/\tau_T$. By our Theorem 3, the asymptotic variance of $\widehat{\theta}(h; \eta_Y, \eta_T)$ is minimized if $(\eta_Y(Z_i) - \theta\eta_T(Z_i))/\tau_T$ equals the optimal adjustment function for the outcome $(Y_i - \theta T_i)/\tau_T$. By linearity of conditional expectations, this holds if $\eta_Y = \eta_{Y,0}$ and $\eta_T = \eta_{T,0}$.

6.4. **Variants of Cross-Fitting.** We note that instead of the type of cross-fitting described in Section 3.2, which is analogous to the "DML2" method in Chernozhukov et al. (2018), one could also consider an analog of their "DML1" method, which creates an overall estimate by averaging separate estimates from each data fold. In our context, this would yield an estimator of the form

$$
\widehat{\tau}_{alt}(h; \widehat{\eta}) = \frac{1}{S} \sum_{s \in [S]} \sum_{i \in I_s} w_{i,s}(h) M_i(\widehat{\eta}_s),
$$

where $w_{i,s}(h)$ is the local linear regression weight of unit i using only data from the s-th fold; see Appendix A.1. Under the conditions of theorem one, one can see from its proof that

$$
\widehat{\tau}_{alt}(h; \widehat{\eta}) - \widehat{\tau}(h; \bar{\eta}) = O_P(r_n(nh)^{-1/2} + v_{2,n}h^2). \tag{6.1}
$$

The estimators $\widehat{\tau}(h; \widehat{\eta})$ and $\widehat{\tau}_{alt}(h; \widehat{\eta})$ thus have the same first-order asymptotic distribution. However, comparing the rate in (6.1) to that in Theorem 1 shows that the alternative implementation removes a term of order $O_P(v_{1,n}h(nh)^{-1/2})$. We still prefer our proposed implementation of cross-fitting despite this improvement in second-order asymptotic properties because it allows existing

routines for bandwidth selection and confidence interval construction to be applied directly to the generated data set $\{(X_i, M_i(\widehat{\eta}_{s(i)}))\}_{i\in[n]}$, as discussed in Section 6.1.

7. SIMULATIONS

In this section, we investigate the finite sample properties of our proposed flexible covariate adjustedment RD estimators under realistic conditions in two simulation studies. The first study's purpose is to show that our theoretical results provide accurate approximations to our estimator's actual finite properties, whereas the second study's purpose is to document how the properties of our estimator and that of existing methods are affected if the number of covariates becomes large relative to the effective sample size.

7.1. **General Setup.** Our simulations are based on real data from Londoño-Vélez et al. (2020), who study the impact of merit-based college financial aid for low-income students in a sharp RD design. Their data contain the outcome variable, a dummy for immediate enrollment in any post secondary education, the running variable, a test score,[10] and 21 covariates, namely age, family size, indicators for gender, ethnicity, employment status, parent's education, household residential stratum, high school schedule, and high school type. Our simulations involve repeatedly drawing random samples from a version of the data that is restricted to the $n = 259,419$ observations with test scores below the original treatment threshold (so that none of the students remaining in the data set are actually assigned to treatment), and then estimating the effect of a placebo treatment "received" by students with test scores above the median test score value. We use either the original outcome (enrollment in any post secondary education) or age (one of the original covariates) as the dependent variable. These two outcomes correspond to settings in which covariate adjustments achieve almost no and quite substantial efficiency gains, respectively. See the RD estimates from the entire restricted data in Table S1 for details. In the main text, we conduct inference using the bias-aware approach with smoothness constants calibrated via the rule of thumb of Armstrong and Kolesár (2020) on the full restricted dataset.[11] All simulations are based on 10,000 Monte Carlo draws.

7.2. **Simulation I: Moderate number of covariates.** In this simulation study, we evaluate the finite-sample performance of our methods in a typical RD setting with a moderate number of covariates and a relatively large number of observations. Specifically, we consider estimation with the original baseline covariates and samples of size 5000, which is around the median of the sample sizes of the empirical applications of our literature survey. We apply our flexible covariate adjustment discussed in Section 3.4. Additionally, we consider the deterministic approximations of

[10] Londoño-Vélez et al. (2020) consider two different test scores as running variables. We focus on the *SABER 11* test score in this section as it is available for a larger number of data points.

[11] The smoothness constants selected via the rule of thumb are 0.00024 and 0.01034 for the original outcome and the age as the dependent variable, respectively.

Table 1: Main results for Simulation I with bias-aware inference.

Adjustment Method		SE x100	SD x100	Bias x100	RMSE x100	Band-width	CI Cov in %	CI Length x100	CI Length Reduction in %
Panel A - Original Outcome									
No Covariates		2.57	2.56	0.38	2.59	23.31	97.10	11.25	0.00
Conventional Linear		2.50	2.52	0.44	2.56	23.16	96.95	10.97	2.49
Localized Linear	Feasible	2.55	2.53	0.47	2.57	23.21	96.98	11.13	1.04
	Oracle	2.53	2.51	0.47	2.56	23.15	96.90	11.06	1.68
Flexible	Feasible	2.53	2.52	0.44	2.56	23.15	96.98	11.05	1.71
	Oracle	2.52	2.51	0.42	2.54	23.12	97.07	11.03	1.93
Panel B - Age									
No Covariates		38.39	38.79	-7.53	39.51	14.66	97.85	173.82	0.00
Conventional Linear		33.41	34.19	-6.66	34.83	13.92	97.54	152.55	12.24
Localized Linear	Feasible	34.47	34.52	-6.64	35.15	13.96	97.86	156.46	9.99
	Oracle	34.04	34.10	-6.64	34.74	13.91	97.85	154.66	11.02
Flexible	Feasible	33.52	33.69	-6.25	34.26	13.85	97.93	152.58	12.22
	Oracle	33.06	33.18	-5.66	33.66	13.71	98.00	150.30	13.53

Notes: Results are based on 10,000 Monte Carlo draws and a sample size of $n = 5000$ (see Section 7 for details). Data generating process is based on Londoño-Vélez et al. (2020). The bandwidth is chosen and the confidence sets are constructed based on bias-aware inference. The columns show results for simulated mean standard error (SE), standard deviation (SD); bias (Bias); root mean squared error (RMSE); the average bandwidth (Bandwidth), coverage of confidence intervals with 95% nominal level (CI Cov); the average confidence interval length (CI Length); and the mean CI length relative to the no covariates CI length (CI Length Reduction). Estimators are described in Section 3.

all the feasible adjustment methods, which were obtained by running the respective method on the restricted dataset. By comparing the feasible adjustments and their respective deterministic approximation, we can assess the quality of the approximation in our equivalence result of Theorem 1. For comparison, we also report the results without covariate adjustments and with conventional linear adjustments. For each adjustment method, we select the bandwidth and construct a confidence interval using the bias-aware approach.[12] The results are based on one random data split for each Monte Carlo draw.

Table 1 reports the main results of this simulation study. Our methods work very well for both dependent variables and all types of adjustments in that the mean simulated standard errors are close to the simulated standard deviations and the confidence intervals have simulated coverage rates close to the nominal one. The confidence intervals are slightly conservative, which is typical in bias-aware inference. The changes in the mean bias for different types of adjustments are negligible relative to the standard deviation, which is consistent with our key insight that covariate adjustments have no first-order effect on the leading bias constant.

[12]The number of effective observations used in the second stage is on average around 2500 for the original outcome and around 1600 for age as the dependent variable.

In Panel A, the covariates have essentially no impact on the dependent variable, and so none of the methods leads to noticeable reductions in the standard deviation. In Panel B, where the covariates have some explanatory power for the dependent variable, the cross-fitted RD estimator with localized linear adjustment yields a confidence interval that is on average 10% shorter than the no covariates confidence interval. The flexible adjustment improves this performance even further. As can be expected in a setting with a small number of covariates relative to the sample size, the conventional and cross-fitted localized linear covariate adjustments yield similar results.

In Appendix B of the Online Supplement, we present additional simulation results for all individual adjustment methods described in Section 3.4. We further investigate the asymptotic equivalence result presented in Theorem 1 within this simulation design and we illustrate that the estimation uncertainty of the adjustment functions is indeed almost negligible relative to the overall estimation uncertainty of the RD estimators. Additionally, we present simulation results for covariate-adjusted RD estimators that conduct estimation and inference based on robust bias corrections. The qualitative conclusions remain very similar to those presented above.

7.3. **Simulation II: Many covariates.** Our literature survey documents that the effective sample size relative to the number of covariates is small in many empirical applications; see Table S2 in the Online Supplement. We investigate the performance of our proposed approach in such settings within a simulation study with a fixed effective sample size and varying number of covariates, using the original outcome as the dependent variable. We create additional covariates by generating all second-order interaction terms of the original covariates, and we consider different settings by including the first 2, 10, 50, 100, and 150 of these covariates.[13] To mimic settings where there are many covariates relative to the effective sample size, we sample 500 observations without replacement within a distance of 25 from the placebo cutoff and use all of them in the RD regressions, i.e. we use a fixed bandwidth $h = 25$.[14] In this setting, the ratio of the effective sample size to the number of covariates lies in the range between 250 and 3.33, which corresponds to the settings in our literature analysis with small values of this ratio. For each subsample, we estimate the RD parameter using the no covariates, the conventional linear adjustment, and our cross-fitted RD estimator with localized linear adjustments and localized random forest adjustments.[15] The results are based on the bias-aware inference approach and $B = 11$ data splits for each Monte Carlo draw.

First, we illustrate that the conventional linear adjustment RD estimator can be less efficient than the no covariates RD estimator and the conventional standard errors can be severely distorted even in settings with a moderate number of covariates. Second, the cross-fitted linear adjustment

[13]The first 21 of the technical covariates correspond to the original covariates, followed by the interaction terms. Since the covariates have essentially no explanatory power for the original outcome, the exact order of inclusion does not affect the results in this section.

[14]This fixed bandwidth is close to the average optimal bandwidth selected in Simulation I in a data-driven way.

[15]We chose the random forest to represent the machine learning adjustments here, but the qualitative results are similar when employing other methods. In this section, we focus on the individual adjustment methods, rather than on the flexible ensemble, to offer more direct insights into the mechanics of the linear and regularized adjustments.

RD estimator can also be less efficient than the no covariates RD estimators, but its standard error is not downward biased in this simulation, which results in valid inference. Third, cross-fitted RD estimators employing machine learning methods and regularization mitigate both of these issues in our simulation study, i.e. they are at least as efficient as the no covariates RD estimator and the standard errors are unbiased.

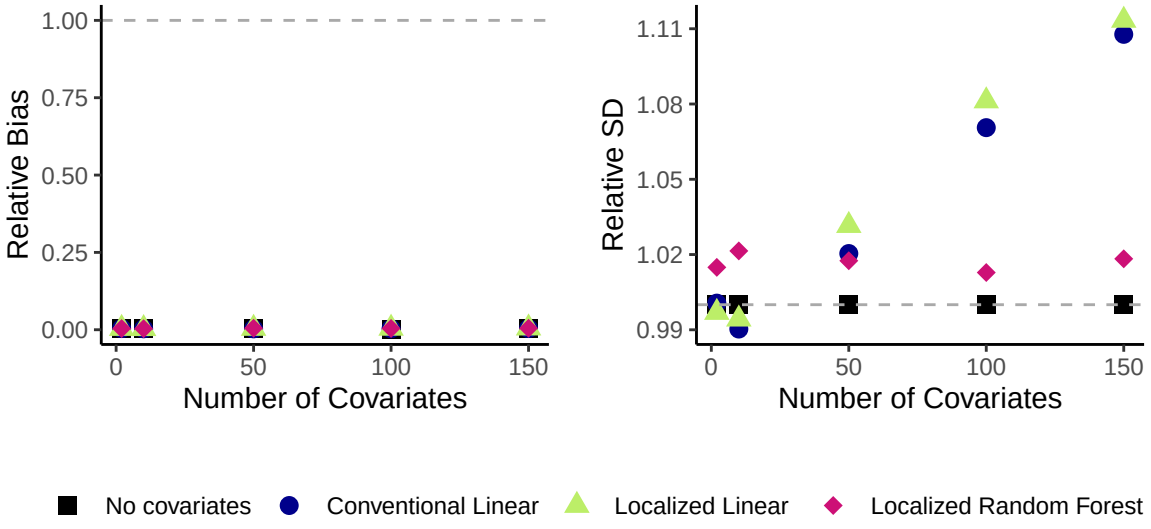

Figure 2: Results for Simulation II - Bias and standard deviation of the respective estimator relative to the standard deviation of the no covariates RD estimator in the left and right panel, respectively. The results are based on samples of size $n = 500$ within the estimation window with $h = 25$ and 10,000 Monte Carlo draws.

7.3.1. *Results on efficiency.* Figure 2 shows the bias and the standard deviation of the four estimation methods, normalized by the standard deviation of the no covariates RD estimator, for a varying number of covariates. We note that the simulated bias is insensitive to including many covariates, and we therefore focus on the standard deviation. In this setting, the covariates seem to have essentially no explanatory power for the dependent variable, and so adjustments based on them cannot lead to a reduction in the asymptotic variance of the RD estimator; see estimation results in Table S1. As predicted by our theory, when the number of covariates remains moderate, all estimators perform very similarly, meaning that all the adjustments concentrate around the optimal function of no adjustment. However, as the number of covariates increases, the standard deviations of both the conventional and cross-fitted localized linear adjustment estimators become substantially larger than that of the no covariates RD estimator. The reason for that is that the linear regression with a large number of covariates is very variable, such that the estimated adjustments are no longer close to a constant[16] and the high-dimensional linear adjustments effectively

[16]Such finite-sample behavior renders our asymptotic theory as well as the results of Calonico et al. (2019) inapplicable in this setting.

add non-negligible noise to the outcome variable in this setting.

In contrast, the RD estimator with random forest adjustments does not become more variable as the number of covariates increases, meaning that the estimated adjustment function remains close to the optimal function of no adjustment. This simulation illustrates the general point that by means of regularization, the machine learning methods produce stable adjustments in a much wider range of settings than the linear regression does, and they typically do not perform worse than the no covariates RD estimator. It is therefore advisable to always rely on regularized adjustments in high-dimensional settings.

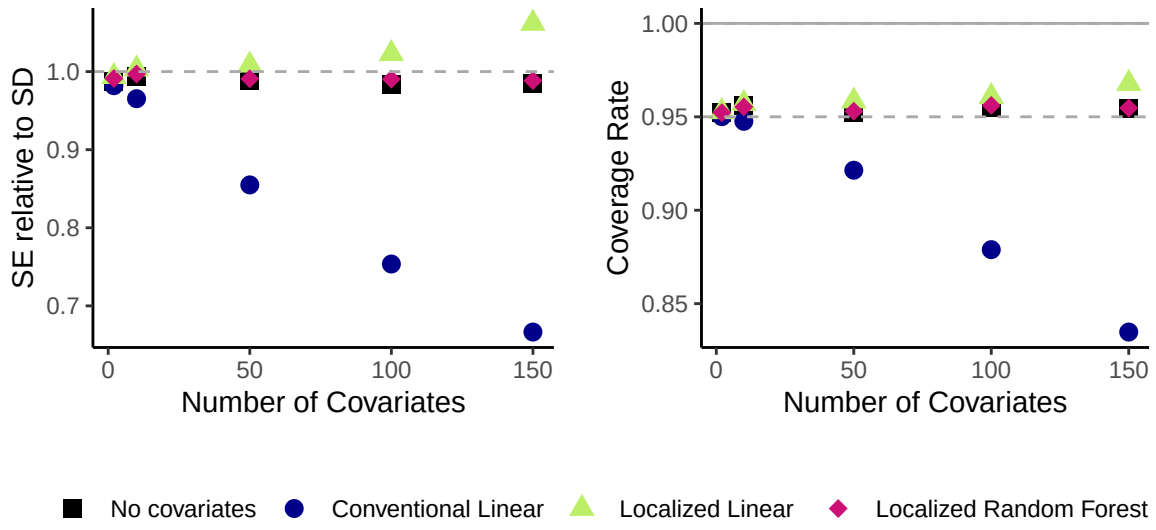

Figure 3: Results for Simulation II - Mean standard error relative to the standard deviation of the respective estimator and simulated confidence interval coverage for nominal confidence level 95%. We consider bias-aware confidence intervals and nearest neighbor standard errors. The results are based on samples of size $n = 500$ within the estimation window with $h = 25$ and 10,000 Monte Carlo draws.

7.3.2. *Results on inference.* We now turn to the standard error and coverage of the confidence intervals for the respective methods. The left panel of Figure 3 shows that the standard error of the conventional linear adjustment estimator exhibits a downward bias that increases substantially with the number of covariates, reaching 35% for 150 covariates. This effect is due to overfitting: with many covariates, the regression residuals that enter the standard error formula become "too close to zero", and standard errors therefore become "too small". With cross-fitting, this issue occurs neither for the linear adjustment nor for the random forest adjustment.

The right panel of Figure 3 shows that, due to increasingly biased standard errors, the coverage of linear adjustment bias-aware confidence intervals with nominal level 95% falls below 85% for 150 covariates. With cross-fitting, bias-aware confidence intervals have close to nominal coverage for both adjustment methods and all numbers of covariates under consideration. These simulation re-

sults demonstrate the benefits of cross-fitting and suggest that practitioners should exercise caution when doing inference based on the conventional linear adjustment estimators even if the number of covariates is only moderate to low (relative to the effective sample size), as the corresponding standard errors can be severely downward biased.

8. CONCLUSIONS

We have proposed a novel class of estimators that can make use of covariate information more efficiently than the conventional linear adjustment estimators that are currently used widely in practice. In particular, our approach allows the use of modern machine learning tools to adjust for covariates, and is at the same time largely unaffected by the "curse of dimensionality". Our estimator is also easy to implement in practice, and can be combined in a straightforward manner with existing methods for bandwidth choice and the construction of confidence intervals. In our reanalysis of the literature, we show that our proposed estimator yields shorter confidence intervals in almost all empirical applications, and in some cases, these reductions can be substantial. We therefore expect our proposed estimator to be very attractive for a wide range of future economic applications.

A. PROOFS OF THE MAIN RESULTS

In this section, we prove Theorems 1–4 and Proposition 1. To this end, we show a more general result that allows for a local polynomial regression of an arbitrary order p. We use this result also in Appendix B to establish the validity of the inference methods discussed in Section 6.1.

A.1. **Additional Notation.** Let $\mathbb{X}_n = (X_i)_{i\in[n]}$ denote the realizations of the running variable. For $0 \leq v \leq p$, we define feasible and infeasible estimators of the jump in the v-th derivative of the conditional expectation of the modified outcome $M(\bar{\eta})$ at the cutoff using the p-th order local polynomial regression as

$$\begin{aligned}
&\widehat{\tau}_{v,p}(h;\widehat{\eta}) = \sum_{i=1}^{n} w_{i,v,p}(h) M_i(\widehat{\eta}_{s(i)}) \quad \text{and} \quad \widehat{\tau}_{v,p}(h;\bar{\eta}) = \sum_{i=1}^{n} w_{i,v,p}(h) M_i(\bar{\eta}),\\
&w_{i,v,p}(h) = w^{+}_{i,v,p}(h) - w^{-}_{i,v,p}(h),\\
&w^{\star}_{i,v,p}(h) = e_v^{\top} \left(\sum_{i=1}^{n} K^{\star}_h(X_i) \widetilde{X}_{p,i} \widetilde{X}_{p,i}^{\top} \right)^{-1} K^{\star}_h(X_i) \widetilde{X}_{p,i} \quad \text{for } \star \in \{+,-\},
\end{aligned}$$

where $\widetilde{X}_{p,i} = (1, X_i, ..., X_i^p)^{\top}$, $K_h(v) = K(v/h)/h$, $K^{+}_h(v) = K_h(v)\mathbf{1}\{v \geq 0\}$, $K^{-}_h(v) = K_h(v)\mathbf{1}\{v < 0\}$. The corresponding estimates of $\beta^{\star}_v(\bar{\eta}) = \partial^v_x \mathbb{E}[M_i(\bar{\eta})|X_i = x]|_{x=0^{\star}}$ are given by

$$\widehat{\beta}^{\star}_{v,p}(h;\widehat{\eta}) = \sum_{i=1}^{n} w^{\star}_{i,v,p}(h) M_i(\widehat{\eta}_{s(i)}) \quad \text{and} \quad \widehat{\beta}^{\star}_{v,p}(h;\bar{\eta}) = \sum_{i=1}^{n} w^{\star}_{i,v,p}(h) M_i(\bar{\eta}) \quad \text{for } \star \in \{+,-\}.$$

A.2. **General Result.** We state and prove two theorems that generalize our Theorems 1 and 2.

Theorem A.1. *Suppose that Assumptions 1–4 hold with $j \in \{1, ..., p+1\}$ in Assumption 2. Then:*

$$\widehat{\tau}_{0,p}(h; \widehat{\eta}) = \widehat{\tau}_{0,p}(h; \bar{\eta}) + O_P(t_p),$$

where $t_p = r_n(nh)^{-1/2} + \sum_{j=1}^{p} v_{j,n} h^j (nh)^{-1/2} + v_{p+1,n} h^{p+1}$.

In the proof of Theorem A.1, we will use the following lemma that collects some standard intermediate steps in the analysis of local polynomial estimators, taking into account cross-fitting.

Lemma A.1. *Suppose that Assumptions 3 and 4 hold. For $s \in [S]$ and $\star \in \{-, +\}$, it holds that:*

(i) $\dfrac{S}{n} \displaystyle\sum_{i \in I_s} K_h^\star(X_i)(X_i/h)^j = \mathbb{E}[K_h^\star(X_i)(X_i/h)^j] + O_p((nh)^{-1/2})$ *for* $j \in \mathbb{N}$,

(ii) $\displaystyle\sum_{i \in I_s} w_{i,0,p}^\star(h) = 1/S + O_p((nh)^{-1/2})$,

(iii) $\displaystyle\sum_{i \in I_s} w_{i,0,p}^\star(h) X_i^j = O_p(h^j (nh)^{-1/2})$ *for* $1 \le j \le p$,

(iv) $\displaystyle\sum_{i \in I_s} |w_{i,0,p}^\star(h) X_i^j| = O_P(h^j)$ *for* $j \in \mathbb{N}$,

(v) $\displaystyle\sum_{i \in I_s} w_{i,0,p}^\star(h)^2 = O_P((nh)^{-1})$.

Proof. The results follow from standard kernel calculations. □

Proof of Theorem A.1. To begin with, note that

$$\widehat{\tau}_{0,p}(h; \bar{\eta}) - \widehat{\tau}_{0,p}(h; \widehat{\eta}) = \sum_{s=1}^{S} G_s(p), \quad G_s(p) \equiv \sum_{i \in I_s} w_{i,0,p}(h)(\widehat{\eta}_s(Z_i) - \bar{\eta}(Z_i)).$$

Since S is a fixed number, it suffices to show that $G_s(p) = O_p(t_p)$ for $s \in \{1, ..., S\}$. We analyze the expectation and variance of $G_s(p)$ conditional on $\mathbb{X}_n$ and $(W_j)_{j \in I_s^c}$. We begin with the expectation. It holds with probability approaching one that

$$\begin{aligned} \left|\mathbb{E}[G_s(p) | \mathbb{X}_n, (W_j)_{j \in I_s^c}]\right| &= \left| \sum_{i \in I_s} w_{i,0,p}(h) \mathbb{E}[\widehat{\eta}_s(Z_i) - \bar{\eta}(Z_i) | X_i, (W_j)_{j \in I_s^c}] \right| \\ &\le \sup_{\eta \in \mathcal{T}_n} \left| \sum_{i \in I_s} w_{i,0,p}(h) \mathbb{E}[\eta(Z_i) - \bar{\eta}(Z_i) | X_i] \right|. \end{aligned}$$

Let $m(x; \eta) = \mathbb{E}[\eta(Z_i) - \bar{\eta}(Z_i) | X_i = x]$. Taylor's theorem yields

$$m(X_i; \eta) = m(0; \eta) + \sum_{j=1}^{p} \frac{1}{j!} \partial_x^j m(0; \eta) X_i^j + \frac{1}{(p+1)!} \partial_x^{p+1} m(\widetilde{x}_{i,p}; \eta) X_i^{p+1}$$

for some $\widetilde{x}_{i,p}$ between 0 and X_i. We analyze the three terms associated with different terms of Taylor's expansion separately. We make use of Lemma A.1 in each step.

First, using the Cauchy-Schwarz inequality, we obtain that

$$\sup_{\eta\in\mathcal{T}_n}\left|m(0;\eta)\sum_{i\in I_s}w_{i,0,p}(h)\right| = \sup_{\eta\in\mathcal{T}_n}|m(0;\eta)|O_p((nh)^{-1/2}) = O_p(r_n(nh)^{-1/2}).$$

Second, for $j\in\{1,...,p\}$, we have that

$$\sup_{\eta\in\mathcal{T}_n}\left|\partial_x^j m(0;\eta)\sum_{i\in I_s}w_{i,0,p}(h)X_i^j\right| = \sup_{\eta\in\mathcal{T}_n}|\partial_x^j m(0;\eta)|h^j O_p((nh)^{-1/2}) = O_p(h^j(nh)^{-1/2}v_{j,n}).$$

Third, we note that

$$\sup_{\eta\in\mathcal{T}_n}\left|\sum_{i\in I_s}w_{i,0,p}(h)\partial_x^{p+1}m(\widetilde{x}_i;\eta)X_i^{p+1}\right| \le \sum_{i\in I_s}|w_{i,0,p}(h)X_i^{p+1}|\sup_{\eta\in\mathcal{T}_n}|\partial_x^{p+1}m(\widetilde{x}_i;\eta)| = O_p(h^{p+1}v_{p+1,n}).$$

Next, we consider the conditional variance. It holds with probability approaching one that

$$\begin{aligned}
\mathbb{V}\left[G_s(p)|\mathbb{X}_n,(W_j)_{j\in I_s^c}\right] &= \sum_{i\in I_s}w_{i,0,p}(h)^2\mathbb{V}\left[\bar{\eta}(Z_i)-\widehat{\eta}_s(Z_i)|\mathbb{X}_n,(W_j)_{j\in I_s^c}\right]\\
&\le \sup_{\eta\in\mathcal{T}_n}\sum_{i\in I_s}w_{i,0,p}(h)^2\mathbb{E}[(\bar{\eta}(Z_i)-\eta(Z_i))^2|X_i]\\
&\le \sup_{\eta\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}[(\bar{\eta}(Z_i)-\eta(Z_i))^2|X_i=x]\sum_{i\in I_s}w_{i,0,p}(h)^2\\
&= O_p(r_n^2(nh)^{-1}),
\end{aligned}$$

where we use Lemma A.1 and Assumption 1 in the last step. The conditional convergence then implies the unconditional one (see Chernozhukov et al., 2018, Lemma 6.1). □

Theorem A.2. *Suppose that the assumptions of Theorem A.1 hold, Assumption 5 holds, and* $\mathbb{E}[M_i(\bar{\eta})|X_i=x]$ *is* $p+1$ *times continuously differentiable with L-Lipschitz continuous* $p+1$ *derivative bounded by* C*. Then*

$$\sqrt{nh}\,V_p(\bar{\eta})^{-1/2}\left(\widehat{\tau}_{0,p}(h;\widehat{\eta})-\tau-h^{p+1}B_p\right)\xrightarrow{d}\mathcal{N}(0,1),$$

where, for some kernel constants $\bar{\nu}_p$ *and* $\bar{\kappa}_p$*,*

$$\begin{aligned}
B_{p,n} &= \frac{\bar{\nu}_p}{2}\left(\partial_x^{p+1}\mathbb{E}[M_i(\bar{\eta})|X_i=x]\big|_{x=0^+}+(-1)^p\partial_x^{p+1}\mathbb{E}[M_i(\bar{\eta})|X_i=x]\big|_{x=0^-}\right)+o_P(1),\\
V_p(\bar{\eta}) &= \frac{\bar{\kappa}_p}{f_X(0)}\left(\mathbb{V}[M_i(\bar{\eta})|X_i=0^+]+\mathbb{V}[M_i(\bar{\eta})|X_i=0^-]\right).
\end{aligned}$$

Proof of Theorem A.2. By the conditional version of Lyapunov's CLT, we obtain that

$$\mathrm{se}_{0,p}(h;\bar{\eta})^{-1}\left(\widehat{\tau}_{0,p}(h;\bar{\eta})-\mathbb{E}[\widehat{\tau}_{0,p}(h;\bar{\eta})|\mathbb{X}_n]\right)\to\mathcal{N}(0,1).$$

where $\mathrm{se}^2_{0,p}(h;\bar{\eta})=\sum_{i=1}^n w_{i,0,p}(h)^2\mathbb{V}[M_i(\bar{\eta})|X_i]$. By L-Lipschitz continuity of $\mathbb{V}[M_i(\bar{\eta})|X_i=x]$ in x, we obtain that

$$\mathrm{se}^2_{0,p}(h;\bar{\eta})=\sum_{i=1}^n w^-_{i,0,p}(h)^2\mathbb{V}[M_i(\bar{\eta})|X_i=0^-]+\sum_{i=1}^n w^+_{i,0,p}(h)^2\mathbb{V}[M_i(\bar{\eta})|X_i=0^+]+o_p((nh)^{-1}).$$

It then follows from standard kernel calculations that $nh\,\mathrm{se}^2_{0,p}(h;\bar{\eta})-V_p(\bar{\eta})=o_P(1)$ and $\mathbb{E}[\widehat{\tau}_{0,p}(h;\bar{\eta})|\mathbb{X}_n]-\tau=B_ph^{p+1}+o_p(h^{p+1})$ for some constant B_p. □

A.3. **Proofs of Theorems 1–4.** Theorems 1 and 2 follow directly from the general results in Theorems A.1 and A.2 with $p=1$; and Theorem 4 follows from simple calculations. It remains to prove Theorem 3. For any $\eta\in\mathcal{V}$, it holds that

$$\frac{2f_X(0)}{\bar{\kappa}}V(\eta)=\mathbb{V}[Y_i-\mu_0^+(Z_i)|X_i=0^+]+\mathbb{V}[Y_i-\mu_0^-(Z_i)|X_i=0^-]+R(\eta),$$

where the first two terms on the right-hand side do not depend on η, and

$$R(\eta)=\mathbb{V}[\mu_0^+(Z_i)-\eta(Z_i)|X_i=0^+]+\mathbb{V}[\mu_0^-(Z_i)-\eta(Z_i)|X_i=0^-].$$

Further, it holds that

$$\begin{aligned}R(\eta)=R(\eta_0+\eta-\eta_0)&=\mathbb{V}\left[\frac{1}{2}(\mu_0^+(Z_i)-\mu_0^-(Z_i))-(\eta(Z_i)-\eta_0(Z_i))|X_i=0^+\right]\\&\quad+\mathbb{V}\left[-\frac{1}{2}(\mu_0^+(Z_i)-\mu_0^-(Z_i))-(\eta(Z_i)-\eta_0(Z_i))|X_i=0^-\right]\\&=R(\eta_0)+2\mathbb{V}[\eta(Z_i)-\eta_0(Z_i)|X_i=0],\end{aligned}$$

where in the last step we use the assumption on continuity of conditional covariances. The theorem follows from the above decomposition by taking the difference $V(\eta^{(a)})-V(\eta^{(b)})$ for arbitrary $\eta^{(a)}$ and $\eta^{(b)}$ in $\mathcal{V}$. □

A.4. **Proof of Proposition 1.** We first note that

$$\widehat{\theta}(h;\widehat{\eta}_Y,\widehat{\eta}_T)-\widehat{\theta}(h;\bar{\eta}_Y,\bar{\eta}_T)=O_P((r_n(nh)^{-1/2}+v_{1,n}h(nh)^{-1/2}+v_{2,n}h^2)^2).$$

This equality is an immediate consequence of Theorem 1 and an application of the continuous mapping theorem as $|\tau_T|>0$. Further, using a mean-value expansion, it follows that

$$\widehat{\theta}(h;\bar{\eta}_Y,\bar{\eta}_T)-\theta=\frac{1}{\tau_T}(\widehat{\tau}_Y(h;\bar{\eta}_Y)-\tau_Y)-\frac{\tau_Y}{\tau_T^2}(\widehat{\tau}_T(h;\bar{\eta}_T)-\tau_T)+\widehat{\rho}(\bar{\eta}_T,\bar{\eta}_Y)$$

with

$$\widehat{\rho}(\bar{\eta}_T, \bar{\eta}_Y) = \frac{\widehat{\tau}_Y(h;\bar{\eta}_Y)(\widehat{\tau}_T(h;\bar{\eta}_T) - \tau_T)^2}{2\widehat{\tau}_T^*(h;\bar{\eta}_T)^3} - \frac{(\widehat{\tau}_Y(h;\bar{\eta}_Y) - \tau_Y)(\widehat{\tau}_T(h;\bar{\eta}_T) - \tau_T)}{\tau_T^2},$$

where $\widehat{\tau}_T^*(h;\eta_T)$ is some intermediate value between τ_T and $\widehat{\tau}_T(h;\bar{\eta}_T)$. Given our assumptions, it follows that

$$\widehat{\rho}(\widehat{\eta}_T, \widehat{\eta}_Y) = O_P\big(((nh)^{-1/2} + h^2)^2\big).$$

Part (i) follows analogously to Theorems 1 and 2 and Part (ii) follows from Theorem 3. □

B. DETAILS ON SECTION 6.1

In this section, we formally show that, under suitable assumptions, existing procedures for bandwidth selection and construction of confidence intervals devised for settings without covariates can be directly applied to the modified data $\{(X_i, M_i(\widehat{\eta}))\}_{i\in[n]}$.

B.1. **Standard Errors.** We generalize the nearest-neighbors standard error from the main text to the local polynomial regression of an arbitrary order p. Let

$$\widehat{se}_{v,p}^2(h;\widehat{\eta}) = \sum_{i=1}^n w_{i,v,p}^2(h)\widehat{\sigma}_i^2(\widehat{\eta}), \quad \widehat{\sigma}_i^2(\widehat{\eta}) = \Big(M_i(\widehat{\eta}_{s(i)}) - \frac{R}{R+1}\sum_{j\in\mathcal{R}_i} M_j(\widehat{\eta}_{s(j)})\Big)^2,$$

where $\mathcal{R}_i$ is the set of R nearest neighbors of unit i in terms of their running variable realization on the respective side of the cutoff. Establishing consistency of this standard error requires the following technical assumption on the first stage estimator, which is implied by our main assumptions, for example, if $M_i(\bar{\eta})$ is bounded.

Assumption B.1. *For all $s \in [S]$, it holds that $\sum_{i\in[n]} w_{i,v,p}^2(h)\iota_i(\widehat{\eta}) = o_P((nh^{1+2v})^{-1})$ for $0 \le v \le p$, where*

$$\iota_i(\widehat{\eta}) = \sum_{\substack{(j,l)\in\mathcal{R}_i^2 \\ (j,l)\notin I_{s(i)}^2}} \big((\widehat{\eta}_{s(i)}(Z_i) - \bar{\eta}(Z_i)) - (\widehat{\eta}_{s(j)}(Z_j) - \bar{\eta}(Z_j))\big)\,(M_i(\bar{\eta}) - M_l(\bar{\eta})).$$

Proposition B.1. *Suppose that Assumptions 1–5 and B.1 hold. Moreover, suppose that Assumption 1 also holds with $\mathcal{X}_h$ replaced by $\widetilde{\mathcal{X}}_h$ that is an open set s.t. $\mathcal{X}_h \subset \widetilde{\mathcal{X}}_h$, and $\sup_{\eta\in\mathcal{T}_n}\sup_{x\in\widetilde{\mathcal{X}}_h}\mathbb{E}[(M_i(\eta) - \mathbb{E}[M_i(\eta)|X_i])^4|X_i = x]$ is bounded by B for all $n \in \mathbb{N}$. Let further $\mathbb{E}[(M_i(\bar{\eta})|X_i = x]$ and $\mathbb{V}[(M_i(\bar{\eta})|X_i = x]$ be L-Lipschitz continuous. Then for all $0 \le v \le p$, it holds that*

$$nh^{1+2v}\left(\widehat{se}_{v,p}^2(h;\widehat{\eta}) - se_{v,p}^2(h;\bar{\eta})\right) = o_P(1),$$

where $se_{v,p}^2(h;\bar{\eta}) = \sum_{i=1}^n w_{i,v,p}^2(h)\sigma_i^2(\bar{\eta})$ and $\sigma_i^2(\bar{\eta}) = \mathbb{V}[M_i(\bar{\eta})|X_i]$.

We note that Assumption B.1 could be dropped if we were to study a slight variation of $\widehat{se}_{v,p}^2(h;\widehat{\eta})$ in which we take the R nearest neighbors of unit i in terms of running variable values *among units*

in the same fold to compute $\widehat{\sigma}_i^2(\widehat{\eta})$. However, proceeding like this would mean that existing software packages that compute nearest neighbor standard errors would have to be adapted, and could not be applied directly to the modified data $\{(X_i, M_i(\widehat{\eta}))\}_{i\in[n]}$.

B.2. **Confidence intervals.** In this subsection, we discuss three types of confidence intervals for the RD parameter, based on undersmoothing, robust bias correction, and bias-aware critical values, respectively.

B.2.1. *Undersmoothing.* We first consider confidence intervals that are based on an undersmoothing bandwidth of order $o(n^{-1/5})$. This choice of bandwidth implies that the smoothing bias shrinks to zero at a faster rate than the standard deviation and can hence be ignored when constructing confidence intervals. Let

$$CI^{us}_{1-\alpha} = \big[\widehat{\tau}(h;\widehat{\eta}) \pm z_\alpha \widehat{\mathrm{se}}(h;\widehat{\eta})\big],$$

where z_α is the $1-\alpha/2$ quantile of the standard normal distribution. Proposition B.2 shows that $CI^{us}_{1-\alpha}$ is asymptotically valid.

Proposition B.2. *Suppose that the assumptions of Proposition B.1 hold for $p=1$. If $nh^5 = o(1)$, then $\mathbb{P}\big(\tau \in CI^{us}_{1-\alpha}\big) \geq 1-\alpha+o_p(1)$.*

B.2.2. *Robust bias correction.* We now adapt the robust bias corrections of Calonico et al. (2014) to our setting. To keep the exposition transparent, we focus on the important special case where the bandwidth used to obtain the bias correction is the same as the main bandwidth. In this case, the local linear estimator with a bias correction is numerically equal to the local quadratic estimator (with the same bandwidth), i.e. $\widehat{\tau}_{0,2}(h;\widehat{\eta})$. Let

$$CI^{rbc}_{1-\alpha} = \big[\widehat{\tau}_{0,2}(h;\widehat{\eta}) \pm z_\alpha \widehat{\mathrm{se}}_{0,2}(h;\widehat{\eta})\big].$$

Proposition B.3 shows that $CI^{rbc}_{1-\alpha}$ is asymptotically valid.

Proposition B.3. *Suppose that the assumptions of Theorem A.2 and Proposition B.1 hold for $p=2$. If $nh^7 = o(1)$, then $\mathbb{P}\big(\tau \in CI^{rbc}_{1-\alpha}\big) \geq 1-\alpha+o_p(1)$.*

B.2.3. *Bias-awareness.* We consider a simplified version of the bias-aware approach of Armstrong and Kolesár (2018), which adjusts critical values to account for possible (asymptotic) bias. Suppose that the second derivative of the conditional expectation function of the outcome Y_i given the running variable X_i is bounded in absolute value by some constant B_Y on either side of the cutoff. Then it follows from the results of Armstrong and Kolesár (2020) and our Theorem 2 that the asymptotic bias of our covariate-adjusted RD estimator is bounded in absolute value by $\bar{b}(h) + o_P(h^2)$, where

$$\bar{b}(h) = -\frac{B_Y}{2}\sum_{i=1}^{n} w_i(h) X_i^2 \operatorname{sign}(X_i).$$

We note that this bound is independent of the chosen adjustment function. The proposed confidence interval is

$$CI^{ba}_{1-\alpha} = \left[\widehat{\tau}(h;\widehat{\eta}) \pm z_\alpha(\bar{b}(h)/\widehat{\mathrm{se}}_{0,1}(h;\widehat{\eta}))\,\widehat{\mathrm{se}}_{0,1}(h;\widehat{\eta})\right].$$

where $z_\alpha(r)$ is the $1-\alpha/2$ quantile of the absolute value of the normal distribution with mean r and variance one. Proposition B.4 shows that $CI^{ba}_{1-\alpha}$ is asymptotically valid.

Proposition B.4. *Suppose that the assumptions of Proposition B.1 hold for $p=1$. If $nh^5 = O(1)$, then $\mathbb{P}\big(\tau \in CI^{ba}_{1-\alpha}\big) \geq 1-\alpha+o_p(1)$.*

B.3. **Optimal bandwidth.** In our Theorem 4, we show that the bandwidth that minimizes the Asymptotic Mean Squared Error (AMSE) of our proposed estimator is given by

$$h_{AMSE} = \left(\frac{V(\bar{\eta})}{4B^2_{\text{base}}}\right)^{1/5} n^{-1/5}.$$

This optimal bandwidth can be consistently estimated by applying the procedure of Calonico et al. (2014, s.6) to the modified data $\{(X_i, M_i(\widehat{\eta}_{s(i)}))\}_{i\in[n]}$ using the following three steps.
Step 0. Initial bandwidths.

(i) Let v_n be such that $v_n \to 0$ and $nv_n \to \infty$. In practice, set $\widehat{v}_n = 2.58\min\{S_X, IQR_X/1.349\}n^{-1/5}$, where S^2_X and IQR_X denote, respectively, the sample variance and interquartile range of $\{X_i : 1 \leq i \leq n\}$.

(ii) Choose c_n such that $c_n \to 0$ and $nc^7_n \to \infty$. In practice, let

$$\widehat{c}_n = \widehat{C}^{1/9}_n n^{-1/9}, \qquad \widehat{C}_n = \frac{7nv^7_n \widehat{se}_{3,3}(v_n;\widehat{\eta})}{2\mathcal{B}^2_{3,3}\left(\widehat{\gamma}^+_{4,4}(\widehat{\eta}) - \widehat{\gamma}^-_{4,4}(\widehat{\eta})\right)^2},$$

where $\widehat{\gamma}^\star_{4,4}(\widehat{\eta})$ is the coefficient on $(1/4!)X^4_i$ in the fourth-order global polynomial regression of $M_i(\widehat{\eta}_{s(i)})$ on a constant, X_i, $(1/2!)X^2_i$, $(1/3!)X^3_i$, and $(1/4!)X^4_i$, using the data on the respective side of the cutoff, and $\mathcal{B}^\star_{v,p}$ for $\star \in \{+,-\}$ is the kernel constant in the leading bias term of $\widehat{\beta}^\star_{v,p}(h;\widehat{\eta})$.

Step 1. Choose a pilot bandwidth b_n such that $b_n \to 0$ and $nb^5_n \to \infty$. In practice, use the following estimate of the bandwidth that minimizes the AMSE of the estimates of the second derivative terms in a local quadratic regression:

$$\widehat{b}_n = \widehat{B}^{1/7}_n n^{-1/7}, \qquad \widehat{B}_n = \frac{5nv^5_n \widehat{se}_{2,2}(v_n;\widehat{\eta})}{2\mathcal{B}^2_{2,2}\Big(\left(\widehat{\beta}^+_{3,3}(c_n,\widehat{\eta}) + \widehat{\beta}^-_{3,3}(c_n;\widehat{\eta})\right)^2 + 3\widehat{se}_{3,3}(c_n;\widehat{\eta})\Big)}.$$

Step 2. Estimate h_{AMSE} by

$$\widehat{h}_n = \widehat{H}_n^{1/5} n^{-1/5}, \qquad \widehat{H}_n = \frac{n v_n \widehat{se}_{0,1}(v_n;\widehat{\eta})}{4\mathcal{B}_{0,1}^2\Big(\Big(\widehat{\beta}_{2,2}^+(b_n;\widehat{\eta}) - \widehat{\beta}_{2,2}^-(b_n;\widehat{\eta})\Big)^2 + 3\widehat{se}_{2,2}(b_n;\widehat{\eta})\Big)}.$$

Proposition B.5. *Suppose that the assumptions of Theorem A.2 and Proposition B.1 hold for $p = 3$, $\mathcal{X}$ is bounded, $\mathbb{P}\big[1/C \leq |\widehat{\gamma}_{4,4}^+(\bar{\eta}) - \widehat{\gamma}_{4,4}^-(\bar{\eta})| \leq C\big] \to 1$ for some $C > 0$, and Assumption 1 holds with $\mathcal{X}_h$ replaced by $\mathcal{X}$. Suppose that $\beta_v^+(\bar{\eta}) - (-1)^{v+1}\beta_v^-(\bar{\eta})$ is bounded and bounded away from zero for $v \in \{2,3\}$. Then $\widehat{c}_n \xrightarrow{p} 0$, $n\widehat{c}_n^7 \xrightarrow{p} \infty$, $\widehat{b}_n \xrightarrow{p} 0$, $n\widehat{b}_n^5 \xrightarrow{p} \infty$, and $\widehat{h}_n/h_{AMSE} \xrightarrow{p} 1$.*

B.4. **Proofs of Propositions B.1–B.5.**

B.4.1. *Proof of Proposition B.1.* To begin with, note that standard kernel calculations show that: (i) $\sum_{i\in[n]} w_{i,v,p}(h)^2 = O_P((nh^{1+2v})^{-1})$ and (ii) $\max_{i\in[n]} w_{i,v,p}(h)^2 = o_P((nh^{1+2v})^{-1})$. The proof of Proposition B.1 then requires showing that $\widehat{\mathrm{se}}_{v,p}^2(h;\widehat{\eta})$ is asymptotically equivalent to the following infeasible version of itself, which uses the deterministic function $\bar{\eta}$:

$$\widehat{\mathrm{se}}_{v,p}^2(h;\bar{\eta}) = \sum_{i\in[n]} w_{i,v,p}^2(h)\bigg(M_i(\bar{\eta}) - \frac{1}{R}\sum_{j\in\mathcal{R}_i} M_j(\bar{\eta})\bigg)^2.$$

Using arguments as in the proof of Theorem 4 in Noack and Rothe (2024), one can show that $\widehat{\mathrm{se}}_{v,p}^2(h;\bar{\eta}) - \mathrm{se}_{v,p}^2(h;\bar{\eta}) = o_P((nh^{1+2v})^{-1})$. It therefore remains to show that $\widehat{\mathrm{se}}_{v,p}^2(h;\widehat{\eta}) - \widehat{\mathrm{se}}_{v,p}^2(h;\bar{\eta}) = o_P((nh^{1+2v})^{-1})$. We express this difference as the sum of terms that are linear in $M_i(\widehat{\eta}_{s(i)}) - M_i(\bar{\eta}) = \bar{\eta}(Z_i) - \widehat{\eta}_{s(i)}(Z_i)$ and a quadratic remainder:

$$\begin{aligned}
&\widehat{\mathrm{se}}_{v,p}^2(h;\widehat{\eta}) - \widehat{\mathrm{se}}_{v,p}^2(h;\bar{\eta}) \\
&\quad= 2\sum_{i\in[n]} w_{i,v,p}^2(h)\bigg(M_i(\bar{\eta}) - \frac{1}{R}\sum_{j\in\mathcal{R}_i} M_j(\bar{\eta})\bigg)\bigg(M_i(\widehat{\eta}_{s(i)}) - M_i(\bar{\eta}) - \frac{1}{R}\sum_{j\in\mathcal{R}_i}\big(M_j(\widehat{\eta}_{s(j)}) - M_j(\bar{\eta})\big)\bigg) \\
&\qquad+ \sum_{i\in[n]} w_{i,v,p}^2(h)\bigg(M_i(\widehat{\eta}_{s(i)}) - M_i(\bar{\eta}) - \frac{1}{R}\sum_{j\in\mathcal{R}_i}\big(M_j(\widehat{\eta}_{s(j)}) - M_j(\bar{\eta})\big)\bigg)^2 \\
&\quad\equiv A_1 + 2A_2.
\end{aligned}$$

We first consider A_2. Let C denote a generic constant that might change from line to line. It

holds that

$$\begin{aligned}\frac{1}{C}A_2 &\leq \sum_{i=1}^{n} w_{i,v,p}^2(h)\bigg((\widehat{\eta}_{s(i)}(Z_i)-\bar{\eta}(Z_i))^2+\frac{1}{R}\sum_{j\in\mathcal{R}_i}(\widehat{\eta}_{s(j)}(Z_j)-\bar{\eta}(Z_j))^2\bigg)\\ &\leq \sum_{i=1}^{n}\bigg(w_{i,v,p}^2(h)+\frac{C}{R}\sum_{j:\ i\in R_j} w_{j,v,p}^2(h)\bigg)(\widehat{\eta}_{s(i)}(Z_i)-\bar{\eta}(Z_i))^2\\ &= \sum_{s\in[S]}\sum_{i\in I_s}\bigg(w_{i,v,p}^2(h)+\frac{C}{R}\sum_{j:\ i\in R_j} w_{j,v,p}^2(h)\bigg)(\widehat{\eta}_{s(i)}(Z_i)-\bar{\eta}(Z_i))^2\\ &\equiv \sum_{s\in[S]} A_{2,s}.\end{aligned}$$

For all $s\in[S]$, it holds with probability approaching one that

$$\begin{aligned}&\mathbb{E}[A_{2,s}|\mathbb{X}_n,\{W_i\}_{i\in I_s^c}]\\ &\leq \sum_{i\in I_s}\left(w_{i,v,p}^2(h)+\frac{C}{R}\sum_{j:\ i\in R_j} w_{j,v,p}^2(h)\right)\sup_{\eta\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}\left[(\eta(Z_i)-\bar{\eta}(Z_i))^2|X_i=x\right]\\ &\leq C\sum_{i=1}^{n} w_{i,v,p}^2(h)\sup_{\eta\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}\left[(\eta(Z_i)-\bar{\eta}(Z_i))^2|X_i=x\right]=O_P((nh^{1+2v})^{-1}r_n^2).\end{aligned}$$

As S is finite and $A_{2,s}$ is a positive random variable, it follows that $A_2=o_P((nh^{1+2v})^{-1})$.

To show that A_1 is of order $o_P((nh^{1+2v})^{-1})$, we separate the terms involving the nearest neighbors in the fold of unit i and those that involve at least one neighbor from a different fold. Specifically, we have that:

$$\begin{aligned}A_1 &= \frac{1}{R^2}\sum_{i\in[n]} w_{i,v,p}^2(h)\bigg(\sum_{j,l\in\mathcal{R}_i}(M_i(\bar{\eta})-M_l(\bar{\eta}))\left((\widehat{\eta}_{s(i)}(Z_i)-\bar{\eta}(Z_i))-(\widehat{\eta}_{s(j)}(Z_j)-\bar{\eta}(Z_j))\right)\bigg)\\ &= \frac{1}{R^2}\sum_{i\in[n]} w_{i,v,p}^2(h)\left(\sum_{\substack{(j,l)\in\mathcal{R}_i^2\\(j,l)\notin I_{s(i)}^2}}(M_i(\bar{\eta})-M_l(\bar{\eta}))\left((\widehat{\eta}_{s(i)}(Z_i)-\bar{\eta}(Z_i))-(\widehat{\eta}_{s(j)}(Z_j)-\bar{\eta}(Z_j))\right)\right)\\ &\quad+\frac{1}{R^2}\sum_{s\in[S]}\sum_{i\in I_s} w_{i,v,p}^2(h)\bigg(\sum_{j,l\in\mathcal{R}_i\cap I_s}(M_i(\bar{\eta})-M_l(\bar{\eta}))\left((\widehat{\eta}_{s(i)}(Z_i)-\bar{\eta}(Z_i))-(\widehat{\eta}_{s(j)}(Z_j)-\bar{\eta}(Z_j))\right)\bigg)\\ &\equiv A_{1,1}+\frac{1}{R^2}\sum_{s\in[S]}A_{1,2,s}.\end{aligned}$$

By Assumption B.1, it holds that $A_{1,1}=o_P((nh^{1+2v})^{-1})$. For all $s\in[S]$, it holds with probability

approaching one that

$$\begin{aligned}
&\mathbb{E}[|A_{1,2,s}||\mathbb{X}_n, \{W_i\}_{i\in I_s^c}] \\
&\quad\leq \sum_{i\in I_s} w_{i,v,p}^2(h) \sum_{j,l\in(\mathcal{R}_i\cap I_s)\cup\{i\}} \mathbb{E}[|(M_i(\bar{\eta}) - M_l(\bar{\eta}))(\widehat{\eta}_{s(j)}(Z_j) - \bar{\eta}(Z_j))||\mathbb{X}_n, \{W_i\}_{i\in I_s^c}] \\
&\quad\leq \sum_{i\in I_s} w_{i,v,p}^2(h) \sum_{j,l\in(\mathcal{R}_i\cap I_s)\cup\{i\}} \sup_{\eta\in\mathcal{T}_n} \mathbb{E}[|(M_i(\bar{\eta}) - M_l(\bar{\eta}))(\eta(Z_j) - \bar{\eta}(Z_j))||\mathbb{X}_n] \\
&\quad\leq \sum_{i\in I_s} w_{i,v,p}^2(h) \sum_{j,l\in(\mathcal{R}_i\cap I_s)\cup\{i\}} \left(\mathbb{E}[(M_i(\bar{\eta}) - M_l(\bar{\eta}))^2|\mathbb{X}_n] \sup_{\eta\in\mathcal{T}_n} \mathbb{E}[(\eta(Z_j) - \bar{\eta}(Z_j))^2|\mathbb{X}_n]\right)^{1/2} \\
&\quad= O_P((nh^{1+2v})^{-1} r_n),
\end{aligned}$$

where the last equality follows from Assumption 1 and the assumption of bounded second moments. Hence, $A_{1,2,s} = o_p((nh^{1+2v})^{-1})$, which concludes this proof. □

B.4.2. *Proof of Proposition B.2.* The validity of the CI follows directly from the asymptotic normality of the local linear estimator established in Theorem A.2 and the fact that the standard error is consistent. □

B.4.3. *Proof of Proposition B.3.* Validity of the CI follows directly from asymptotic normality of the local quadratic estimator established in Theorem A.2 and the fact that the standard error is consistent. □

B.4.4. *Proof of Proposition B.4.* Validity of the CI follows directly from asymptotic normality of the local linear estimator established in Theorem A.2, the fact that the standard error is consistent, and that the asymptotic bias is bounded in absolute value by $\bar{b}(h) + o_P(h^2)$. □

B.4.5. *Proof of Proposition B.5.* The proposition follows, using the consistency of the standard error established in Proposition B.1, if the following claims hold:

(i) $\widehat{\gamma}_{4,4}^{\star}(\widehat{\eta}) - \widehat{\gamma}_{4,4}^{\star}(\bar{\eta}) = o_P(1)$,

(ii) $\widehat{\beta}_{3,3}^{+}(c_n;\widehat{\eta}) + \widehat{\beta}_{3,3}^{-}(c_n;\widehat{\eta}) = \beta_3^{+}(\bar{\eta}) + \beta_3^{-}(\bar{\eta}) + o_P(1)$,

(iii) $\widehat{\beta}_{2,2}^{+}(b_n;\widehat{\eta}) - \widehat{\beta}_{2,2}^{-}(b_n;\widehat{\eta}) = \beta_2^{+}(\bar{\eta}) - \beta_2^{-}(\bar{\eta}) + o_P(1)$.

Part (i). First, note that

$$\widehat{\gamma}_{4,4}^{\star}(\widehat{\eta}) - \widehat{\gamma}_{4,4}^{\star}(\bar{\eta}) = e_4' \left(\sum_{i=1}^{n} \widetilde{X}_{4,i}^{\star} \widetilde{X}_{4,i}^{\star\top}\right)^{-1} \sum_{i=1}^{n} \widetilde{X}_{4,i}^{\star}(\bar{\eta}(Z_i) - \widehat{\eta}_{s(i)}(Z_i)),$$

where $\widetilde{X}_{4,i}^{+} = \widetilde{X}_{4,i}\mathbf{1}\{X_i \geq 0\}$ and $\widetilde{X}_{4,i}^{-} = \widetilde{X}_{4,i}\mathbf{1}\{X_i < 0\}$. Further, for $s \in [S]$, we have that

$$\left|\frac{S}{n}\sum_{i\in I_s} X_i^j(\bar{\eta}(Z_i) - \widehat{\eta}_s(Z_i))\right| \leq \sqrt{\frac{S}{n}\sum_{i\in I_s} X_i^{2j}} \sqrt{\frac{S}{n}\sum_{i\in I_s}(\bar{\eta}(Z_i) - \widehat{\eta}_s(Z_i))^2}.$$

Note that, with probability approaching one,

$$\mathbb{E}\left[\frac{S}{n}\sum_{i\in I_s}(\bar{\eta}(Z_i)-\widehat{\eta}_s(Z_i))^2\Big|\mathbb{X}_n,(W_j)_{j\in I_s^c}\right] \leq \sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left[\frac{S}{n}\sum_{i\in I_s}(\bar{\eta}(Z_i)-\eta(Z_i))^2\Big|\mathbb{X}_n\right]$$
$$\leq \sup_{\eta\in\mathcal{T}_n}\sup_{x\in\mathcal{X}}\mathbb{E}[(\bar{\eta}(Z_i)-\eta(Z_i))^2|X_i=x]=o(1).$$

It follows that $\left|\frac{S}{n}\sum_{i\in I_s}X_i^j(\bar{\eta}(Z_i)-\widehat{\eta}_{s(i)}(Z_i))\right|=o_p(1)$. Since $\mathcal{X}$ is bounded, the claim follows.
Part (ii) and (iii). Using steps as in the proof of Theorem A.1, for $p\in\{2,3\}$, we obtain that $\widehat{\beta}^\star_{p,p}(h;\widehat{\eta})-\widehat{\beta}^\star_{p,p}(h,\bar{\eta})=o_P(1)$. Moreover, under the assumptions made, $\widehat{\beta}^\star_{p,p}(h,\bar{\eta})-\beta^\star_p(\bar{\eta})=O_P(h+(nh^{1+2p})^{-1/2})$. The claims follow using the conditions on b_n and c_n. □

C. DETAILS ON THE LITERATURE REANALYSIS

In this section, we provide additional details on the practical performance described in Section 4.

C.1. **Data Collection.** We conducted an extensive literature search in order to document how covariates are used in empirical RD designs and to collect data sets on which to compare our proposed method to the existing approaches. We focused on the publications in AER, AER Insights, AEJ: Applied Economics, AEJ: Economic Policy, and AEA Papers and Proceedings between 2018 and 2023. Starting from a Google Scholar search for the keywords "regression discontinuity",[17] we first identified 74 articles that appear to fit into our theoretical framework,[18] and then retained those 16 papers for which the journal's replication package contained all the data used in the empirical analysis. In 14 of these papers, covariates were used in at least one of the reported RD regressions, while in two papers the available covariates were used only for balance checks but could in principle have been used in the RD regressions, too. For each paper, we identified the main specification (or a version thereof) that includes covariates. These specifications often involve multiple outcomes or running variables, which yielded a total of 56 specifications. The details on all of them are given in Table S2 in the Online Supplement. In our reanalysis of these papers, we focus on these main specifications. In the two cases where only a no covariates RD analysis is reported, we included the covariates that were used for covariates balance checks.

C.2. **Implementation Details.** We apply our flexible adjustment RD estimator proposed in Section 3, and we contrast it with the no covariates, conventional linear, and cross-fitted localized linear adjustment RD estimators. In general, the flexible adjustment is implemented as an ensemble of eight learners listed in Section 3.4 and we use $B=25$ data splits. For three specifications with more

[17] A majority of papers found through the Google Scholar search did not conduct an original RD analysis, but only cited other RD papers, and were hence excluded.

[18] We excluded geographic RD designs where boundary fixed effects were included as part of the identification strategy, and a small number of other nonstandard RD analyses where the outcome variable is measured at a higher level of aggregation than the running variable and a donut design was used.

than 100,000 observations, we speed up the computations by considering only the local versions of machine learning methods and using $B = 5$ data splits. For one specification where the number of observations times covariates exceeds 500,000,000, we use only the local version of random forest as our flexible adjustment and consider $B = 1$ data split. For the bias-aware approach, we calibrated smoothness constants via the rule of thumb of Armstrong and Kolesár (2020). This choice was dictated by practical considerations, as it would not be possible to separately discuss the choice of smoothness bound for each of the 56 specifications. While one can argue whether the resulting smoothness bound is always appropriate, the qualitative conclusions about the relative reductions in the confidence interval length are not too sensitive to the choice of the smoothness bound.

REFERENCES

Andrews, D. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 62, 43–72.

Arai, Y., T. Otsu, and M. H. Seo (2024): "Regression Discontinuity Design with Potentially Many Covariates," *arXiv preprint arXiv:2109.08351*.

Armstrong, T. B. and M. Kolesár (2018): "Optimal inference in a class of regression models," *Econometrica*, 86, 655–683.

——— (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 11, 1–39.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression Discontinuity Designs Using Covariates," *Review of Economics and Statistics*, 101, 442–451.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019): *A practical introduction to regression discontinuity designs: Foundations*, Cambridge University Press.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., W. Newey, J. Robins, and R. Singh (2019): "Double/de-biased machine learning of global and local parameters using regularized Riesz representers," *Working Paper*.

Colangelo, K. and Y.-Y. Lee (2022): "Double debiased machine learning nonparametric inference with continuous treatments," *Working Paper*.

Dong, Y. (2018): "Alternative Assumptions to Identify LATE in Fuzzy Regression Discontinuity Designs," *Oxford Bulletin of Economics and Statistics*, 80, 1020–1027.

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2020): "Estimation of Conditional Average Treatment Effects With High-Dimensional Data," *Journal of Business & Economic Statistics*, 0, 1–15.

Frölich, M. and M. Huber (2019): "Including Covariates in the Regression Discontinuity Design," *Journal of Business & Economic Statistics*, 37, 736–748.

Gerard, F., M. Rokkanen, and C. Rothe (2020): "Bounds on treatment effects in regression discontinuity designs with a manipulated running variable," *Quantitative Economics*, 11, 839–870.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

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

Imbens, G. and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *Review of Economic Studies*, 79, 933–959.

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

Kennedy, E. H. (2020): "Optimal doubly robust estimation of heterogeneous causal effects," *arXiv preprint arXiv:2004.14497*.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017): "Nonparametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79, 1229.

Kreiss, A. and C. Rothe (2023): "Inference in regression discontinuity designs with high-dimensional covariates," *Econometrics Journal*.

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

Londoño-Vélez, J., C. Rodríguez, and F. Sánchez (2020): "Upstream and downstream impacts of college merit-based financial aid for low-income students: Ser Pilo Paga in Colombia," *American Economic Journal: Economic Policy*, 12, 193–227.

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

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Noack, C. and C. Rothe (2024): "Bias-aware inference in fuzzy regression discontinuity designs," *Econometrica*, 92, 687–711.

Robins, J. M. and A. Rotnitzky (2001): "Comment on "Inference for semiparametric models: some questions and an answer" by P. Bickel and J. Kwon," *Statistica Sinica*, 11, 920–936.

Su, L., T. Ura, and Y. Zhang (2019): "Non-separable models with high-dimensional data," *Journal of Econometrics*, 212, 646–677.

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007): "Super Learner," *Statistical applications in genetics and molecular biology*, 6.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016): "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 12673–12678.