

Flexible Covariate Adjustments in Regression Discontinuity Designs

Claudia Noack

Tomasz Olma

Christoph Rothe

May 4, 2023

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: May 4, 2023. We thank Sebastian Calonico, Michal Kolesar, Thomas Lemieux, Jonathan Roth, Vira Semenova, Stefan Wager, Daniel Wilhelm, Andrei Zeleneev, and numerous conference and seminar participants for helpful comments and suggestions. 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 and Nuffield College, University of Oxford. E-Mail: claudia.noack@economics.ox.ac.uk. Website: <https://claudianoack.github.io>. Tomasz Olma, Department of Statistics, 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. These designs exploit 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 in such designs 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 in practice, 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).

An RD analysis generally does not require data beyond the outcome and the running variable, but in practice researchers often have access to additional covariate information, such as socio-demographic characteristics, that 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 linear adjustment estimator is consistent without functional form assumptions on the underlying conditional expectations if the covariates are predetermined, but generally does not exploit the available covariate information efficiently. We also argue that inference based on this estimator can be distorted if the number of covariates is too large relative to the sample size.

To address these issues, we propose a novel class of covariate-adjusted RD estimators that combine local linear regression techniques with flexible covariate adjustments, which can make use of modern machine learning techniques. To motivate the approach, let Y_i and Z_i denote the outcome and covariates, respectively, of observational unit i . Calonico et al. (2019) show that linear adjustment estimators are asymptotically equivalent to standard local linear RD regressions 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 such estimators with a modified 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 compatible with the notion of covariates being “predetermined”. We also show that their asymptotic variance 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 not known in practice,

but can be estimated from the data.

Our proposed estimators hence take the form of a local linear RD regression with the generated outcome $Y_i - \hat{\eta}(Z_i)$, where $\hat{\eta}$ 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.

Importantly, valid inference on the RD parameter in our setup does not require that η_0 is consistently estimated. Our theory only requires 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 analogues 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

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

¹A 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 if one of the two nuisance functions in a doubly robust moment (cf. Robins and Rotnitzky, 2001) is known.

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 are also rather straightforward to address in our setting. In particular, our results justify applying existing methods to a data set in which the outcome Y_i is replaced with the generated outcome $Y_i - \hat{\eta}(Z_i)$, ignoring that $\hat{\eta}$ has been estimated. Our approach can therefore easily be integrated into existing software packages.

Our 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 analogue 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.

Through simulations, we also show that our theoretical findings provide very good approximations to our estimators’ finite sample behavior. We also show that our approach can yield meaningful efficiency gains in empirical practice: we revisit the analysis of the effect of the antipoverty program Progresa/Oportunidades in Mexico on consumption, and find that a machine learning version of our flexible covariate adjustments reduce the standard error by up to 15.7% relative to an estimator that does not use covariate information.

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), who extend the approach in Calonico et al. (2019) to settings with high-dimensional covariates under sparsity conditions, and Frölich and Huber (2019), who propose to incorporate covariates into an RD analysis in a fully nonparametric fashion. The latter method 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. In Section 4, we present our main theoretical results. Further extensions are discussed in Section 5. Section 6 contains a simulation study and Section 7 an empirical application. Section 8 concludes. Proofs of our main results are given in Appendix A, and Appendices B–E give further theoretical, simulation and empirical results.

2. SETUP AND PRELIMINARIES

2.1. Model and Parameter of Interest. We begin by considering sharp RD designs. The data $\{W_i\}_{i \in [n]} = \{(Y_i, X_i, Z_i)\}_{i \in [n]}$, with $[n] = \{1, \dots, n\}$, 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:

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

²Throughout 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.

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 of interest is typically estimated by running a local linear regression (Fan and Gijbels, 1996) on each side of the cutoff. That is, the baseline “no covariates” RD estimator takes the form

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

where $S_i = (T_i, X_i, T_i X_i, 1)^\top$ collects the necessary covariates, $K(\cdot)$ is a kernel function with support $[-1, 1]$, $h > 0$ is a bandwidth, $K_h(v) = K(v/h)/h$, and $e_1 = (1, 0, 0, 0)^\top$ is the first unit vector. We will use repeatedly in this paper that the “no covariates” RD estimator can also be written as a weighted sum of the realizations of the outcome variable,

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

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 $\hat{\tau}_{base}(h)$ is approximately normally distributed in large samples, with bias of order h^2 and variance of order $(nh)^{-1}$:

$$\hat{\tau}_{base}(h) \stackrel{a}{\sim} N(\tau + h^2 B_{base}, (nh)^{-1} V_{base}) \quad (2.3)$$

where “ $\stackrel{a}{\sim}$ ” indicates a finite-sample distributional approximation justified by an asymptotic normality result, and the bias and variance terms are given, respectively, by

$$B_{base} = \frac{\bar{\nu}}{2} (\partial_x^2 \mathbb{E}[Y_i | X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i | X_i = x]|_{x=0^-}) \quad \text{and}$$

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

Here $\bar{\nu}$ 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 bandwidth choice, variance estimation, and the construction of confidence intervals based on approximations like (2.3) are discussed in Calonico et al. (2014) and Armstrong and Kolesár (2020), for example.

2.3. Linear Adjustment Estimator. To improve the accuracy of RD inference, empirical researchers often use a “linear adjustment” estimator that adds available covariates linearly and without kernel localization to the regression (2.2):

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

This estimator can equivalently be written as “no covariates” RD estimator with covariate-adjusted outcome $Y_i - Z_i^\top \hat{\gamma}_h$, where $\hat{\gamma}_h$ is the minimizer with respect to γ in (2.4):

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

Calonico et al. (2019) show that $\hat{\tau}_{lin}(h)$ 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 conditional distribution given the running variable varies smoothly around the cutoff. Specifically, if $\mathbb{E}[Z_i | X_i = x]$ is twice continuously differentiable around the cutoff, then

$$\hat{\tau}_{lin}(h) \stackrel{a}{\sim} N(\tau + h^2 B_{base}, (nh)^{-1} V_{lin})$$

under regularity conditions similar to those for the “no covariates” estimator, where the bias term B_{base} is as above and the new variance term is

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

with γ_0 , a non-random vector of projection coefficients, the probability limit of $\hat{\gamma}_h$.

The linear adjustment estimator generally has smaller asymptotic variance than the “no covariates” estimator, in the sense that $V_{lin} \leq V_{base}$ (Kreiß and Rothe, 2023, Remark 3.5). It also has the same asymptotic distribution as its infeasible counterpart

$$\tilde{\tau}_{lin}(h) = \sum_{i=1}^n w_i(h) (Y_i - Z_i^\top \gamma_0)$$

that uses the population projection coefficients γ_0 instead of their estimates $\hat{\gamma}_h$ to adjust the

outcome variable. Estimation uncertainty about $\hat{\gamma}_h$ does therefore not affect the (first-order) asymptotic properties of $\hat{\tau}_{in}(h)$. This insight can be used, as in Calonico et al. (2019) or Armstrong and Kolesár (2018), to adapt methods for bandwidth choice, variance estimation and the construction of confidence intervals for “no covariates” estimators to the case of linear adjustments.

3. FLEXIBLE COVARIATE ADJUSTMENTS

While linear adjustment estimators are easy to implement, their focus on linearity means that they generally do not exploit the available covariate information efficiently. Linear adjustment estimators might also not work well outside of low dimensional settings: they are not well-defined if the number of covariates exceeds the (local) sample size, and, as we illustrate in Section 3.4 below, the corresponding standard errors can be severely downward biased even if only a moderate number of covariates is used. In this paper, we propose “flexible covariate adjustment” estimators with cross-fitting to address these issues.

3.1. Motivation. 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 . We consider a more general class of estimators with covariate-adjusted outcomes based on potentially nonlinear adjustment functions η :

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

If the covariates are predetermined, in the sense that their values are not causally affected by the treatment, one would expect conditional expectations of transformations of the covariates given the running variable to vary smoothly around the cutoff in some appropriate sense. For our formal analysis, we specifically assume that $\mathbb{E}[\eta(Z_i)|X_i = x]$ is twice continuously differentiable with respect to x around the cutoff for (essentially) every adjustment function η .³ The continuity of the conditional expectation implied by this assumption means that

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

³By “essentially” we mean, for example, that we consider only functions for which the respective conditional expectations exist in the first place. Our assumptions are slightly stronger than those in Calonico et al. (2019), for example, who only assume smoothness of the conditional expectation of the covariates themselves given the running variable around the cutoff, as we require such smoothness to also hold for conditional expectations of transformations of the covariates. This additional assumption, however, is in line with the covariates being predetermined.

The estimator $\widehat{\tau}(h; \eta)$ can be seen as a sample analogue estimator based on the moment condition (3.2) that identifies τ . An important feature of this moment condition is that it is globally invariant with respect to the functional parameter η . Because of this invariance, $\widehat{\tau}(h; \eta)$ is consistent for the RD parameter τ for every η and satisfies

$$\widehat{\tau}(h; \eta) \stackrel{a}{\sim} N\left(\tau + h^2 B_{base}, (nh)^{-1} V(\eta)\right). \quad (3.3)$$

Due to the assumed continuity of second derivatives of $\mathbb{E}[\eta(Z_i)|X_i = x]$, the bias term B_{base} is again that of the baseline “no covariates” estimator, but the variance term is now

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

As the leading bias in (3.3) does not depend on the adjustment function, we ideally want to choose η such that $V(\eta)$ is as small as possible. Our Theorem 3 below implies that the optimal adjustment function η_0 that minimizes this asymptotic variance term is 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. That is, $V(\eta) \geq V(\eta_0)$ for all η , where

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

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

3.2. Proposed Estimator and its Theoretical Properties. Implementing our proposed estimation strategy requires choosing a first-stage estimator $\widehat{\eta}$ of η_0 . Our theoretical analysis below does not require this estimator to be of a particular type. Applied researchers can choose methods according to their assumptions about the shape of the optimal adjustment function, or simply focus on procedures that are convenient to implement; see Section 3.3 for details. We also employ a version of cross-fitting, analogous to the “DML2” procedure in Chernozhukov et al. (2018), in the construction of our proposed estimator. Cross-fitting is an efficient type of sample splitting that both prevents overfitting and allows a unified theoretical analysis under general conditions for the first-stage estimate of η_0 .

Specifically, our proposed procedure entails the following 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 s th 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 establishes that the estimator $\widehat{\tau}(h; \widehat{\eta})$ is asymptotically equivalent to the infeasible estimator $\widehat{\tau}(h; \bar{\eta}) = \sum_{i=1}^n w_i(h) M_i(\bar{\eta})$ that uses the variable $M_i(\bar{\eta})$ as the outcome, where $\bar{\eta}$ is a deterministic approximation of $\widehat{\eta}$ whose error vanishes in large samples in some appropriate sense. In view of (3.3), it then holds that

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

The asymptotic variance in the above expression is minimized if $\widehat{\eta}$ is consistent for η_0 , in the sense that $\bar{\eta} = \eta_0$. However, the distributional approximation is valid even if $\bar{\eta} \neq \eta_0$ because the moment condition (3.2) holds for (essentially) *all* adjustment functions, and not just the optimal one. In that sense, our procedure allows for misspecification in the first stage. Moreover, we show that even under misspecification $V(\bar{\eta})$ is typically smaller than V_{base} . We also demonstrate that one can easily construct valid confidence intervals for τ by applying standard methods developed for settings without covariates to a data set with running variable X_i and outcome $M_i(\widehat{\eta}_{s(i)})$, ignoring sampling uncertainty about the estimated adjustment function.

3.3. Estimating the Adjustment Function. A wide range of methods can be used to obtain a first-stage estimate of the adjustment function in our framework. We mostly focus on estimates of the optimal adjustment function η_0 that take the form

$$\widehat{\eta}(z) = \frac{1}{2}(\widehat{\mu}^+(z) + \widehat{\mu}^-(z)),$$

where $\widehat{\mu}^+(z)$ and $\widehat{\mu}^-(z)$ are separate estimates 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. As mentioned above, our theoretical analysis shows that consistent estimation of η_0 is not necessary in order for inference based $\widehat{\tau}(h; \widehat{\eta})$ to be valid, as we only need $\widehat{\eta}$ to be consistent for some deterministic function $\bar{\eta}$ in a particular

sense. Specifying a correct model for η_0 , or for the two conditional expectations μ_0^+ and μ_0^- , is therefore not a highly critical concern in our setup. For efficiency, however, it is of course desirable that $\bar{\eta}$ is as close to η_0 as possible.

If one wishes to maintain the simplicity of linear adjustments, one can for example set $\hat{\eta}(z) = z^\top \hat{\gamma}_h$, where $\hat{\gamma}_h$ is the minimizer with respect to γ in (2.4), and thus obtain a cross-fitting version of $\hat{\tau}_{lin}(h)$. That is, one could interpret the linear adjustments as an estimate of the optimal adjustment function under the implicit (and generally incorrect) specification that $\mu_0^+(z) + \mu_0^-(z) = c + z^\top \gamma_0$ for some vector of coefficients γ_0 and some constant c .⁴ We discuss the advantages of such an estimator in Section 3.4. One can in principle also obtain estimates of η_0 by specifying other simple parametric models for $\mathbb{E}[Y_i|X_i = x, Z_i = z]$, such as a global linear regression models. Under appropriate smoothness conditions, one can also use classical nonparametric methods to estimate μ_0^+ and μ_0^- , with local polynomial regression being particularly suitable due to their good boundary properties.

If the number of covariates is large, however, we recommend the use of modern machine learning methods, such as lasso or post-lasso regression, random forests, deep neural networks, boosting, or ensemble combinations thereof. Such methods might need a particular tuning though, as the underlying algorithms typically aim for a good overall estimate by optimizing an “integrated mean squared error” type criterion, and are not guaranteed to produce estimates that very accurate at any particular point. That is, a machine learning method that is given the task of learning the conditional expectation $\mathbb{E}[Y_i|X_i = x, Z_i = z]$ might not produce an estimator with good properties at $x = 0^+$ or $x = 0^-$. One can adapt many machine learning methods to our setup, however, by focusing the respective algorithm to units whose realization of the running variable is close to the cutoff. More formally, let

$$\hat{\mathbb{E}}[Y_i|Z_i = z] = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n l(Y_i, f(Z_i))$$

be a generic estimator of $\mathbb{E}[Y_i|Z_i = z]$ computed by minimizing some empirical loss function $L(f) = \sum_{i=1}^n l(Y_i, f(Z_i))$ over a set of candidate functions \mathcal{F} , and let $b > 0$ be some positive bandwidth. We can then define a version of this estimator that “targets” the area just to the right of the cutoff as

$$\hat{\mu}^+(z) = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n l(Y_i, f(Z_i)) K(X_i/b) \mathbf{1}\{X_i \geq 0\},$$

⁴We stress again that correct specification of μ_0^+ and μ_0^- , or the optimal adjustment function, is not required for our corresponding estimator of τ to be consistent.

and similarly for $\hat{\mu}^-(z)$. 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 “localized” machine learning estimators in settings in which $b \rightarrow 0$ as $n \rightarrow \infty$. However, 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.

3.4. Linear Adjustments and the Role of Cross-Fitting. The use of cross-fitting simplifies many arguments in our theoretical analysis,⁵ but is also key for good practical performance of covariate-adjusted RD estimators. To see this, it is instructive to compare the linear adjustment estimator $\hat{\tau}_{lin}(h)$, described in (2.4), with a version of our estimator $\hat{\tau}(h; \hat{\eta})$, described in Section 3.2, that uses the same linear adjustment function $\hat{\eta}(z) = \hat{\gamma}_h^\top z$, so that our use of cross-fitting becomes the only difference between the two procedures. We now illustrate through a small simulation experiment that conventional inference based on linear adjustment estimators can be meaningfully distorted even with a moderate number of covariates, and that cross-fitting by itself alleviates much of the issue.

We consider simple data generating processes (DGPs) in which we observe mutually independent standard normal covariates $Z_i = (Z_{1i}, \dots, Z_{di}) \sim \mathcal{N}(0, \mathbf{I}_d)$, for $d \in \{0, 10, 20, \dots, 50\}$, that are all irrelevant, in the sense that they are fully independent of the outcome and the running variable, which are in turn generated as

$$Y_i = \sin(X_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1), \quad X_i \sim \text{Uniform}(-\pi, \pi), \quad X_i \perp \varepsilon_i. \quad (3.5)$$

For each of 50,000 replications with sample size $n = 1,000$, we then compute the linear adjustment estimator together with its standard error and a conventional robust bias corrected (RBC) confidence intervals for the RD parameter $\tau = 0$ as in Calonico et al. (2019), using their R package `rdrobust`. We also compute our “cross-fitted” version of this estimator, and the analogue standard error and RBC confidence interval, as described in Section 5.1.

As inspection of the simulation results suggests that both procedures yield approximately unbiased estimates of τ for all values of d under consideration, we focus on differences in standard errors and confidence interval coverage. The left panel of Figure 1 shows that the standard error of the conventional linear adjustment estimator exhibits a downward bias that increases substantially with the number of covariates, from about 7% for $d = 0$ to almost 40% for $d = 50$. This effect is due to overfitting: with many covariates, the regression residuals

⁵Without cross-fitting, one would generally have to impose Donsker conditions on the space in which the first-stage estimator takes values, which imply severe limits on the complexity of the estimated functions that might not be palatable for many machine learning methods (Chernozhukov et al., 2018).

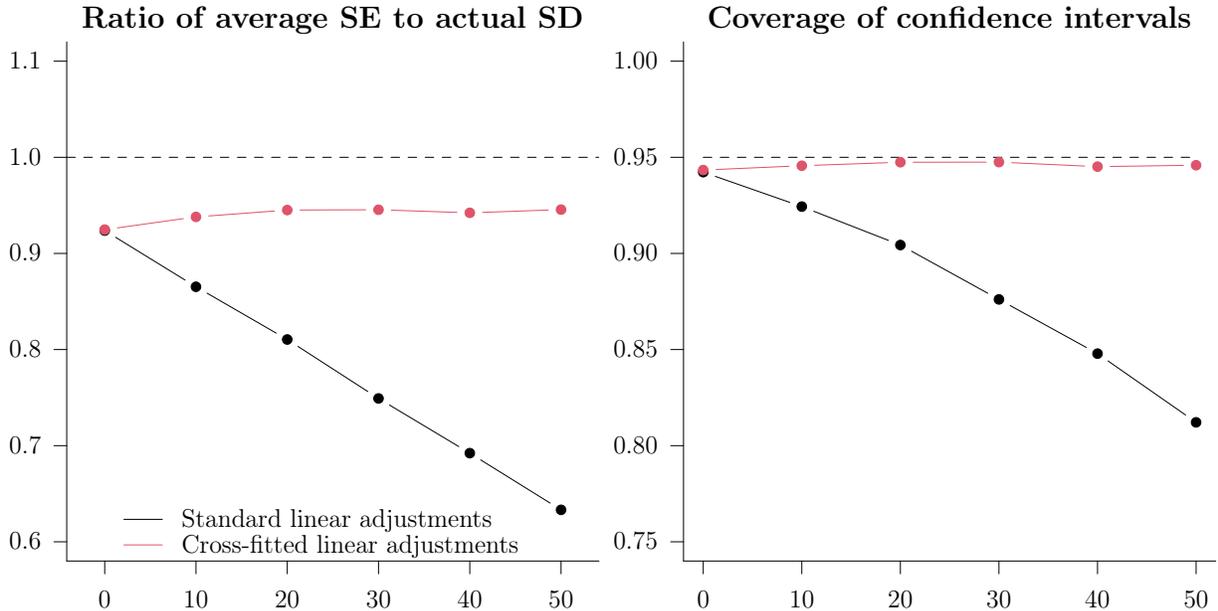


Figure 1: Ratio of simulated average standard error to actual standard deviation (left panel) and simulated coverage rates of RBC confidence intervals with nominal level 95% (right panel) under data generating process (3.5) for conventional linear adjustments (black line) and linear adjustments with cross-fitting (red line) for 0, 10, \dots , 50 (all irrelevant) covariates.

that enter the standard error formula become “too close to zero”, and standard errors therefore become “too small”. With cross-fitting the finite-sample bias of the standard error remains at a moderate 5 to 7% for all values of d under consideration. The right panel of Figure 1 shows that, due to increasingly biased standard errors, the coverage of linear adjustment RBC confidence intervals with nominal level 95% deteriorates from slightly below the nominal level for $d = 0$ to slightly above 80% for $d = 50$. With cross-fitting, RBC confidence intervals have close to nominal coverage for all numbers of covariates under consideration.

Our simulation results not only demonstrate the benefits of cross-fitting, but also suggests that practitioners should thus use caution when inference based on conventional linear adjustment estimators even if the number of covariates used in the analysis is only moderate to low (relative to the effective sample size), as standard errors can be severely downward biased. We revisit this issue in the context of our empirical application below.

4. THEORETICAL PROPERTIES

4.1. 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 $\hat{\eta}$. 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 the conditional distribution of the covariates given the running variable changes smoothly around the cutoff, the class \mathcal{E} contains essentially all functions of the covariates, subject only to technical integrability conditions.⁶

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) $\hat{\eta}_s$ belongs to \mathcal{T}_n with probability approaching 1 for all $s \in [S]$; (ii) it holds that:*

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

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 \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 $\hat{\eta}$ concentrates around $\bar{\eta}$. It also allows the function $\bar{\eta}$ to be different from the target function η_0 , which means that $\hat{\eta}$ 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 type of “localized” machine learning estimators of η_0 described in Section 3.3, existing results imply that for fixed $b > 0$ and K the uniform kernel

$$\sup_{\eta \in \mathcal{T}_n} \mathbb{E} [(\eta(Z_i) - \bar{\eta}(Z_i))^2 | X_i \in (-b, b)] = O(r_n^2), \quad (4.1)$$

⁶For 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 x in a neighborhood of the cutoff, some integrable functions g_j , and $j \in \{0, 1, 2\}$, then \mathcal{E} contains at least 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.

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 (4.1) can hold with $r_n^2 = n^{-2s/(2s+d)}$ for estimators that exploit smoothness. If $\bar{\eta}(z)$ is s -sparse, then (4.1) can hold with $r_n^2 = s \log(d)/n$ for estimators that exploit sparsity. Assumption 1 then follows from (4.1) 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 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 $b \rightarrow 0$ 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:

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

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 $\mathbb{E}[\eta(Z_i) - \bar{\eta}(Z_i)|X_i = x]$ 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 1.⁷

Assumption 3. (i) X_i is continuously distributed with density f_X , which is continuous and bounded away from zero over an open neighborhood of the cutoff; (ii) 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 $[-1, 1]$; (iii) The bandwidth satisfies $h \rightarrow 0$ and $nh \rightarrow \infty$ as $n \rightarrow \infty$.

⁷For example, this can easily be seen to be the case if $\hat{\eta}$ converges to $\bar{\eta}$ uniformly on \mathcal{Z} with rate r_n and the smoothness conditions for $f_{Z|X}(z|x)$ given in Footnote 6 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 $\mathbb{E}[(\partial_x^j f_{Z|X}(Z_i|x)/f_{Z|X}(Z_i|x))^2|X_i = x]$ is bounded for $j \in \{1, 2\}$ uniformly in x and the conditions from Footnote 6 hold.

Assumption 3 collects some standard conditions from the RD literature. Note that 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. *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 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 4 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 $\hat{\tau}(h; \bar{\eta})$ even in settings where the distribution of the data, and thus $\bar{\eta}$, might change with n .

4.2. Main Results. We give three main results in this subsection. The first shows that our proposed estimator $\hat{\tau}(h; \hat{\eta})$ is asymptotically equivalent to an infeasible analogue $\hat{\tau}(h; \bar{\eta})$ that replaces the estimator $\hat{\eta}$ with the deterministic sequence $\bar{\eta}$; the second shows the asymptotic normality of the estimator; and the third characterizes how the asymptotic variance changes with the adjustment function and shows that η_0 is indeed the optimal adjustment.

Theorem 1. *Suppose that Assumptions 1–3 hold. Then*

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

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

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

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

occurs because $\widehat{\tau}(h; \widehat{\eta})$ is a sample analogue of the following moment function

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

which is insensitive to variation in η over the set \mathcal{E} . Moment functions with a local form of insensitivity with respect to a nuisance function, 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 5.2 for further discussion.

Theorem 2. *Suppose that Assumptions 1–4 holds. Then*

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

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

Theorem 2 follows from Theorem 1 under the additional regularity conditions of Assumption 4. 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*

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

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

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

Theorem 3 introduces a function class \mathcal{V} that, similarly to the discussion of the class \mathcal{E} above, we expect to contain essentially all functions, subject to some integrability conditions. The theorem implies that a generic adjustment function $\eta^{(a)}$ leads to a lower asymptotic variance than another function $\eta^{(b)}$ if $\eta^{(a)}$ is closer to the optimal adjustment function

than $\eta^{(b)}$ in a particular sense: $V(\eta^{(a)}) < V(\eta^{(b)})$ if and only if $\mathbb{V}[\eta_0(Z_i) - \eta^{(a)}(Z_i)|X_i = 0] < \mathbb{V}[\eta_0(Z_i) - \eta^{(b)}(Z_i)|X_i = 0]$. We therefore obtain the lowest possible value of $V(\bar{\eta})$ for $\bar{\eta} = \eta_0$. Moreover, even if $\bar{\eta} \neq \eta_0$, our flexible covariate adjustments typically still yield efficiency gains relative to existing RD estimators. For example, $V(\bar{\eta}) < V_{base}$ 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. Similarly, $V(\bar{\eta}) < V_{lin}$ 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 -type sense than the population linear adjustment $z^\top \gamma_0$ is.

We remark that in practice a smaller asymptotic variance generally leads to a bias reduction through channel of bandwidth choice. If $V(\bar{\eta}) < V_{base}$, then for any given bandwidth sequence h that satisfies our assumptions the estimator $\hat{\tau}(h; \hat{\eta})$ has the same asymptotic bias and smaller variance than the “no covariates” estimator $\hat{\tau}_{base}(h)$. However, if we consider the bandwidths $h^*(\bar{\eta}) = n^{-1/5} (V(\bar{\eta})/4B_{base}^2)^{1/5}$ and $h_{base}^* = n^{-1/5} (V_{base}/4B_{base}^2)^{1/5}$ that minimize the respective first-order mean squared errors, the estimator $\hat{\tau}(h^*(\bar{\eta}); \hat{\eta})$ has both smaller asymptotic bias and smaller asymptotic variance than $\hat{\tau}_{base}(h_{base}^*)$.

5. FURTHER RESULTS AND DISCUSSIONS

5.1. Inference. Our result in Theorem 2 suggests that one should be able to construct valid confidence intervals for τ by applying standard methods for inference based on the “no covariates” RD estimator to the generated data set $\{(X_i, M_i(\hat{\eta}_{s(i)}))\}_{i \in [n]}$, ignoring the sampling uncertainty about the estimated adjustment function. As we show in more detail in Appendix B, this turns out to be correct.

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, we can construct a “bias-aware” confidence interval as in Armstrong and Kolesár (2020) as

$$CI_{1-\alpha}^{ba} = [\hat{\tau}(h; \hat{\eta}) \pm z_\alpha(\bar{b}(h)/\hat{se}(h; \hat{\eta})) \hat{se}(h; \hat{\eta})].$$

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, $\bar{b}(h)$ is an explicit bound on the finite sample bias of “no covariates” RD estimator, and $\hat{se}(h; \hat{\eta})$ is a nearest-neighbor standard error. Alternatively, we can construct a “robust bias correction” confidence interval as in Calonico et al. (2014) by subtracting an estimate of the first-order bias of $\hat{\tau}(h; \hat{\eta})$, based on local quadratic regression, from the estimator, and adjusting the standard error appropriately. This yields the confidence

interval

$$CI_{1-\alpha}^{rbc} = [\hat{\tau}^{rbc}(h; \hat{\eta}) \pm z_\alpha \widehat{\text{se}}^{rbc}(h; \hat{\eta})],$$

where z_α is the $1 - \alpha$ quantile of the standard normal distribution. We show in Appendix B that both $CI_{1-\alpha}^{ba}$ and $CI_{1-\alpha}^{rbc}$ have correct asymptotic coverage under standard conditions even if the bandwidth sequence $h \sim n^{-1/5}$ is such that the asymptotic bias of $\hat{\tau}(h; \hat{\eta})$ is of the same order as its standard deviation. We also derive results for confidence intervals based on undersmoothing, and methods for bandwidth selection.

5.2. Analogies with Randomized Experiments. The results in Section 4 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 analogue estimator of τ_{PATE} based on this moment function reaches has asymptotic variance V_{PATE} if \hat{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 \hat{m}^t is consistent for some other function \bar{m}^t , $t \in \{0, 1\}$. The convergence of \hat{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 4 arise because our covariate-adjusted RD estimator is in many ways a direct analogue of such EIF estimators. To show this, write $m(z) = (1 - p)m^1(z) + pm^0(z)$ for any two functions m^0 and m^1 , so that $m_0(z) = (1 - p)m_0^1(z) + pm_0^0(z)$. 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 $\hat{\tau}(h; \hat{\eta})$, as they are both sample analogues a moment function with the same basic properties.

5.3. 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.⁸ Under standard conditions (Hahn et al., 2001; Dong, 2017), 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:

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

In Appendix C, we show that this 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.⁹ We further show that this asymptotic variance is minimized if the

⁸Throughout this subsection, the notation is analogous to that used before, with the subscripts Y and T referencing the respective outcome variable.

⁹This result can be used to construct a confidence interval for θ based on the t-statistic. Alternatively, confidence sets for θ can be constructed via the Anderson-Rubin-type approach, which circumvents some

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 $\hat{\theta}(h; \eta_Y, \eta_T) - \theta$ is first-order asymptotically equivalent to a sharp RD estimator with the infeasible outcome $U_i(\eta_Y, \eta_T) = (Y_i - \theta T_i - (\eta_Y(Z_i) - \theta \eta_T(Z_i))) / \tau_T$. Applying the result of Theorem 3, it follows that the asymptotic variance of $\hat{\theta}(h; \eta_Y, \eta_T)$ 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 $\eta_T = \eta_{T,0}$.

5.4. Cross-Fitting. Two remarks on cross-fitting are in order. First, we note that the realization of our estimator $\hat{\tau}(h; \hat{\eta})$ can in principle depend on the particular random splits of the data into S folds in finite samples. To make the results more robust with respect to sample splitting we can proceed as suggested in Chernozhukov et al. (2018, Section 3.4) by repeating the estimation procedure a number of times, and reporting a summary measure of the estimates obtained in this fashion, such as the median. We proceed in this fashion in our empirical application below, for example.

As a second remark, 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 analogue 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

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

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

$$\hat{\tau}_{alt}(h; \hat{\eta}) - \hat{\tau}(h; \bar{\eta}) = O_P(r_n(nh)^{-1/2} + v_{2,n}h^2). \quad (5.1)$$

The estimators $\hat{\tau}(h; \hat{\eta})$ and $\hat{\tau}_{alt}(h; \hat{\eta})$ thus have the same first-order asymptotic distribution.¹⁰ Comparing the rate in (5.1) to that obtained in Theorem 1, we can see that the alternative implementation removes the term of order $O_P(v_{1,n}h(nh)^{-1/2})$. We still prefer our proposed implementation of cross-fitting because it allows existing routines for bandwidth

problems of the delta-method-based inference (Noack and Rothe, 2021).

¹⁰An analogous point is made by Chernozhukov et al. (2018) in their specific context for their methods DML1 and DML2.

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 5.1.

6. SIMULATIONS

6.1. Estimators. We consider a number of variations of our covariate-adjusted RD estimator that use different methods for estimating the conditional expectations μ_0^+ and μ_0^- in each fold of the data. Specifically, we compute “localized” estimators as described in Section 3.3 with a uniform kernel and $\widehat{\mathbb{E}}[Y_i|Z_i = z]$ obtained using one of the following methods: (i) global linear regression; (ii) shallow neural net; (iii) random forest; (iv) boosted tree; (v) post-lasso estimator of Belloni et al. (2012); (vi) an ensemble combination (SuperLearner) of the just-mentioned methods. We use the statistical software R. The choice of machine learning methods follows that of Chernozhukov et al. (2018). We use the default values of the tuning parameters in the respective packages and use cross-fitting with five folds in the simulations and ten folds in the empirical application.¹¹ In the second-stage RD regression, we select the bandwidth and conduct inference based on the covariate-adjusted outcomes $M(\widehat{\eta})$ using the bias-aware approach `RDHonest`.¹² In the second stage, a triangular kernel is used and nearest-neighbor standard errors are computed. For reference, we calculate an oracle version of our flexible covariate adjustments that uses the true optimal adjustment function instead of an estimate.

Furthermore, we report the “no covariates” RD estimator, and we adapt the conventional linear adjustment estimator, with and without cross-fitting, to the bias-aware inference framework. The conventional linear adjustment estimator without cross-fitting here is obtained as follows. In the first step, we run a local linear RD regression with covariates included linearly (without localization) as in (2.4) and the bandwidth equal to the optimal “no covariates” bandwidth. Based on this regression, we generate the adjustment terms $Z_i^\top \widehat{\gamma}^{(1)}$, where $\widehat{\gamma}^{(1)}$ is the vector of estimated coefficients on Z_i . Next, we select the optimal bandwidth with $Y_i - Z_i^\top \widehat{\gamma}^{(1)}$ as the outcome variable, and rerun the first regression using this new bandwidth to obtain a new vector of estimated coefficients $\widehat{\gamma}^{(2)}$. In the second stage, we

¹¹The only exceptions are the neural nets, where the decay parameter has to be selected by the user. Specifically, we use shallow neural with one hidden layer, two nodes, and decay parameter set to 0.1 (package `mnet`); random forest with minimal leaf size set to five that averages over 500 trees (package `randomForest`); boosted trees implemented with 100 boosting rounds (package `gbm`); and post-lasso regression with data-driven penalty terms (package `hdm`). For estimation with neural nets we normalize all variables to lie between zero and one. The ensemble method is implemented using the package `SuperLearner` with ten-fold cross-validation.

¹²In Appendix D, we also present results based on robust bias correction and a version of undersmoothing in the second stage.

use the RDHonest command with the modified outcome $Y_i - Z_i^\top \hat{\gamma}^{(2)}$, i.e. we select a new bandwidth and obtain a point estimate and confidence interval based on it. In the cross-fitted version of this procedure, the coefficients in the i th adjustment term are obtained using data outside of the fold of observation i , similarly to the procedure described in Section 3.2.

6.2. Data-Generating Processes. We consider three different data generating processes (DGPs) indexed by $L \in \{0, 4, 16\}$ in our simulations. In each DGP, we have four independent baseline covariates, (Z_{1i}, \dots, Z_{4i}) , that are distributed uniformly over $[-1, 1]$, and we vary the degree of complexity of their association with the outcome. In DGP 1, the covariates do not enter the optimal adjustment function; in DGP 2, they enter linearly; and in DGP 3, they enter nonlinearly. Specifically, in each DGP, the running variable X_i follows the uniform distribution over $[-1, 1]$ and the outcome is generated as:

$$Y_i = \mathbf{1}\{X_i \geq 0\} + \phi_X(X_i) + \phi_L(X_i, Z_i) + \varepsilon_i,$$

$$\phi_L(X_i, Z_i) = \mathbf{1}\{L > 0\} (\phi_X(X_i) + \mathbf{1}\{X_i \geq 0\}) \cdot \left(\sum_{l=1}^L b_l(Z_i) + 0.25 \cdot \sum_{j=1}^4 Z_{ji} \right),$$

where $\phi_X(X_i) = \text{sign}(X_i)(X_i^2 + .5X)$, $\varepsilon_i \sim \mathcal{N}(0, 0.25)$, and the terms $b_l(Z_i)$, are Hermite polynomials of the covariates, where the first four Hermite polynomials are the baseline covariates. We supply the four baseline covariates to all estimation methods of the first stage, and for lasso estimation we additionally supply 100 Hermite polynomials. The second stage smoothness constant required by RDHonest is set to its population value in each DGP. We conduct 50,000 replications with samples of size $n = 2,000$.

6.3. Results. Table 1 reports the main results of our simulation study. We first note that here all CIs have simulated coverage rates close to the nominal one. We next compare the “no covariates” RD estimator and the oracle estimator. In DGP 1, these two estimators are numerically equal. In DGPs 2–3, the covariates have some explanatory power for the outcome, and the oracle estimator has a substantially lower standard deviation than the “no covariates” estimator.

We now turn to our RD estimators with flexible covariate adjustments. In DGP 1 and 2, the feasible estimators perform similarly to the oracle, with only a minor increase in the standard deviation. In DGP 3, where the linear adjustments are not optimal, all the adjustments based on machine learning methods perform better than the linear adjustments. The ensemble method is closest to the oracle in terms of mean squared error. We emphasize that we have chosen the tuning parameters of the flexible methods not necessarily optimally

and choosing them via cross-validation might improve the overall performance of the methods.

In Appendix D, we present additional simulation results for estimates based on different first-stage adjustment functions that all use the bandwidth that is optimal for the “no covariates” RD estimator in the second stage. We also consider covariate-adjusted RD estimators that use the bandwidth selected in the second stage, and conduct inference using the robust bias corrections and a version of undersmoothing. The qualitative conclusions remain very similar to those presented above.

7. EMPIRICAL APPLICATION

7.1. Setup. To illustrate the use of our proposed methods in empirical practice, we revisit the analysis of the effect on consumption of the antipoverty, conditional cash transfer program Progresa/Oportunidades in Mexico in the early 2000s. Eligibility for the program was determined based on a pre-intervention household poverty-index, which led to a regression discontinuity design. We use a dataset assembled by Calonico et al. (2014) and focus on urban localities. We consider four outcome variables, namely food and non-food consumption, one year and two years after the implementation of the program, and conduct an intention-to-treat analysis where eligibility for the cash transfer constitutes the treatment.

The data set contains 1,944 observations for which full covariate information is available. The 85 baseline covariates, recorded prior to program implementation, include: the households size, household head’s age, sex, years of education and employment status, spouse’s age and years of education, number of children not older than five years and their sex, house characteristics: whether the house has cement floors, water connection, water connection inside the house, a bathroom, electricity, number of rooms, pre-intervention consumption, and an identifier of the urban locality in which the house is located.¹³

7.2. Results. We compute estimates of the RD parameter for the methods considered in our simulations above (see Section 6.1), and also use the bias-aware approach for the second stage.¹⁴ For brevity, we focus on a single outcome, the effect of the cash transfer on food consumption one year after the program was introduced, and the bias-aware approach of e.g. Armstrong and Kolesár (2020) in this section. Results for the other outcome variables and different second-stage inference methods are reported in Appendix E.

¹³Calonico et al. (2014) use these covariates for their falsification tests (see Table S.A.XI in their supplementary materials) but do not to obtain their empirical estimates.

¹⁴We supply the 85 covariates to all methods, and for lasso estimation additionally include interaction terms between all baseline covariates other than the location dummies, which results in a total of 238 covariates.

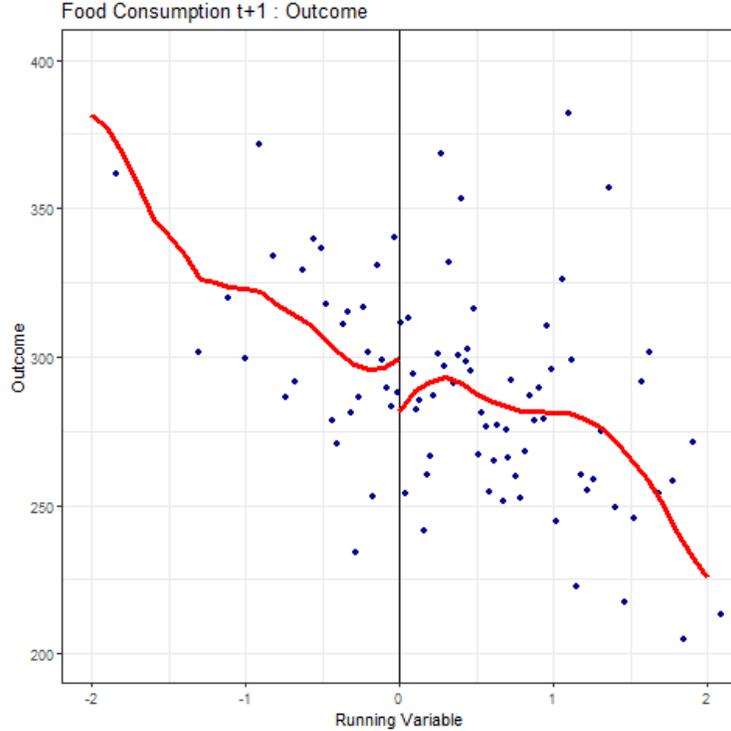


Figure 2: Outcome variable in the empirical application.

Notes: The outcome is the food consumption one year after the program was introduced. Each blue dot represents the mean of approximately 20 observations. The solid lines represents the local linear fit with triangular kernel and bandwidth $h = 0.6$.

Our results are reported in Table 2, where the upper panel presents the results using no covariates and conventional linear adjustments. In the lower panel, the rows correspond to different first-stage methods of constructing adjustment terms in the approach proposed in this paper. We select a conservative bound on the maximal second derivative of the conditional expectation of the outcome variable given the running variable equal to $M = 300$.¹⁵ The relative performance of different covariate adjustments is very similar over a wide range of choices for the smoothness constant, as predicted by the theory. We report estimation results with $M = 100$ and $M = 500$ in the appendix. The results are based on 100 random splits of the data and we report the median point estimates, standard errors accounting for the variation introduced by sample splitting (Chernozhukov et al., 2018, Section 3.4), the percentage reduction of the standard error relative to the “no covariates” RD estimator, and the median bandwidths.

¹⁵The bound on the smoothness constant can be calibrated using the following method suggested by Kolesár and Rothe (2018): “if a researcher believes, for example, that the CEF differs by no more than S from a straight line between the CEF values at the endpoints of any interval of length one in the support of the

Table 1: Simulation Results

	Bias $\times 100$	Std Dev $\times 100$	RMSE $\times 100$	CI Coverage $\times 100$	Avg CI Length $\times 100$	Avg Bandwidth $\times 100$
DGP L=0						
No covariates	3.69	7.49	8.35	94.90	32.58	43.20
Linear Adjustments	3.69	7.51	8.37	94.68	32.44	43.17
Linear Adjustments with CF	3.71	7.52	8.38	94.95	32.84	43.31
<i>Flexible Adjustments</i>						
Oracle	3.69	7.49	8.35	94.90	32.58	43.20
Linear Regression	3.73	7.50	8.38	94.82	32.64	43.24
Neural Nets	3.73	7.50	8.38	94.84	32.63	43.23
Boosted Tree	3.78	7.61	8.50	94.77	33.10	43.54
Random Forest	3.75	7.54	8.42	94.83	32.81	43.36
rLasso	3.73	7.49	8.36	94.87	32.59	43.21
SuperLearner	3.73	7.50	8.38	94.86	32.64	43.24
DGP L=4						
No covariates	8.52	17.03	19.04	94.95	74.29	65.59
Linear Adjustments	3.84	7.75	8.65	94.66	33.56	44.03
Linear Adjustments with CF	3.85	7.76	8.67	94.92	33.96	44.17
<i>Flexible Adjustments</i>						
Oracle	3.94	7.95	8.88	94.84	34.68	44.62
Linear Regression	3.89	7.76	8.68	94.77	33.85	44.14
Neural Nets	3.89	7.76	8.68	94.75	33.84	44.14
Boosted Tree	4.30	8.51	9.54	94.86	37.13	46.33
Random Forest	4.14	8.22	9.20	94.87	35.85	45.51
rLasso	3.91	7.78	8.71	94.80	33.95	44.20
SuperLearner	3.89	7.75	8.67	94.78	33.82	44.13
DGP L=16						
No covariates	10.45	20.96	23.42	94.87	91.01	72.23
Linear Adjustments	7.35	14.67	16.40	94.70	63.43	60.42
Linear Adjustments with CF	7.38	14.70	16.45	94.85	64.07	60.59
<i>Flexible Adjustments</i>						
Oracle	3.99	7.93	8.88	94.95	34.68	44.62
Linear Regression	7.39	14.69	16.44	94.74	63.82	60.52
Neural Nets	4.95	9.73	10.92	95.00	42.44	49.37
Boosted Tree	5.04	9.97	11.17	94.79	43.33	49.90
Random Forest	6.58	13.12	14.67	94.73	56.91	57.17
rLasso	4.06	7.98	8.95	94.84	34.83	44.76
SuperLearner	4.06	7.98	8.95	94.89	34.83	44.76

Notes: Results are based on 50,000 Monte Carlo draws. Columns show results for the simulated mean bias (Bias); the simulated standard deviation of the estimator (Std. Dev.); the simulated root mean squared error (RMSE); simulated coverage of confidence intervals with 95% nominal level (CI Coverage); the average simulated confidence interval length (Avg CI Length); and the average simulated bandwidth (Avg Bandwidth). For flexible adjustments, in the first stage, the sample is restricted to the observations that lie in the window $(-b, b)$ with b being twice the “no covariates” bandwidth.

Table 2: Estimation results for the empirical application.

	Estimate	SE	Δ SE	Bandwidth
No covariates	-18.6	16.6	-	0.591
Linear Adjustments	-14.8	13.7	-21.0%	0.567
Linear Adjustments with CF	-13.3	17.1	2.8%	0.579
<i>Flexible Adjustments</i>				
Linear Regression	-16.0	15.7	-5.4%	0.578
Neural Nets	-16.2	15.5	-6.9%	0.576
Boosted Trees	-17.2	14.5	-14.7%	0.569
Random Forest	-19.1	14.4	-15.1%	0.570
rLasso	-21.5	14.4	-15.3%	0.575
SuperLearner	-18.3	14.3	-15.7%	0.569

Notes: Covariate adjustments are obtained using the methods listed. In the second stage, RDHonest with smoothness constant $M = 300$ is used. We report the median estimate across 100 splits of the data, the nearest-neighbor standard error taking into account the variation introduced by sample splitting (SE), the percentage change in the standard error relative to the “no covariates” standard error (Δ SE), and the median bandwidth. For flexible adjustments, in the first stage, the sample is restricted to the observations that lie in the window $(-b, b)$ with $b = 1.182$ being twice the “no covariates” bandwidth.

All point estimates are very similar considering the magnitude of the standard errors. All confidence intervals contain zero, but their length varies. Our estimator with linear adjustments reduces the standard error by 5.4% relative the standard RD estimator with no covariates. All our nonparametric methods improve upon linear regression adjustments. In particular, our preferred approach based on the ensemble of all ML methods achieves the reduction in the standard error of 15.7%. We note that the conventional linear adjustments appear to yield larger efficiency gains, but with 85 covariates the standard error is likely to be severely downward biased here; see Section 3.4. In fact, the cross-fitted version of this estimator exhibits a slight increase in the standard error relative to the “no covariates” standard error.

7.3. Adjustment Terms. As explained above, our analysis assumes that the conditional expectation functions of transformations of the covariates given the running variable vary smoothly around the cutoff. Most importantly, this must be satisfied for $\mathbb{E}(\bar{\eta}(Z_i)|X_i = x)$,

running variable, a reasonable choice for the bound on the second derivative is $M = 8S$.” In our application, the difference between the conditional expectation function and a straight line can be very conservatively bounded by $S = 25$. To ensure valid inference, we select a conservative bound on the smoothness constant equal to $M = 300$.

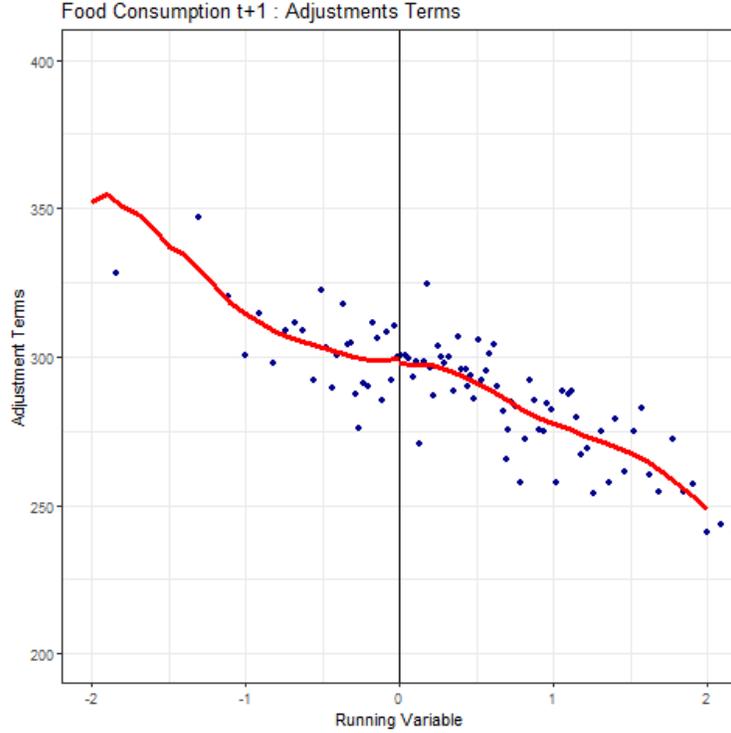


Figure 3: Adjustment terms in the empirical application.

Notes: The figure presents the adjustment terms (for one split of the data) obtained using SuperLearner for the food consumption one year after the program was introduced. Each blue dot represents the mean of approximately 20 observations. The solid lines represents the local linear fit with triangular kernel and bandwidth $h = 0.6$.

where $\bar{\eta}$ is the population function targeted by the estimated adjustment. To investigate the plausibility of this assumption, we can plot the estimated adjustment terms against the running variable and see whether there are any visually apparent discontinuities around the cutoff. Put differently, if we compute a “no covariates” RD estimator with $\hat{\eta}_{s(i)}(Z_i)$ as the outcome variable, we would expect to see an estimated jump with only small deviations from zero that are compatible with statistical noise if our assumptions hold.¹⁶ In Figure 3 shows such a plot for our empirical application. Each blue dot represents the mean of approximately 20 adjustment terms based on SuperLearner binned based on the running variable, and the red line shows the local linear regression fit with bandwidth $h = 0.6$. We can see that the fit is very smooth, with essentially no jump at the cutoff, lending credibility to our assumptions.

¹⁶If this jump would happens to be exactly zero in a specific data set, our flexible covariate adjusted RD estimator would be numerically equal to the “no covariates” RD estimate, but the standard errors of both estimators would still generally be different.

8. CONCLUSIONS

We have proposed a novel class of estimators that can make use of covariate information more efficiently than the 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. For this reason, we expect it to be attractive for a wide range of economic applications.

A. PROOFS OF THE MAIN RESULTS

In this section, we prove Theorems 1–3. To this end, we show a more general result that allows for a local polynomial regression of an arbitrary order p . We also use this result in Appendix B to establish the validity of the inference methods discussed in Section 5.1.

A.1. Additional Notation. Denote the realizations of the running variable by $\mathbb{X}_n = (X_i)_{i \in [n]}$. 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 at the cutoff using the p -th order local polynomial regression as:

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

with $\tilde{X}_{p,i} = (1, X_i, \dots, X_i^p)^\top$, $K_h(v) = K(v/h)/h$, $K_h^+(v) = K_h(v) \mathbf{1}\{v \geq 0\}$, $K_h^-(v) = K_h(v) \mathbf{1}\{v < 0\}$. Corresponding estimates of $\beta_v^*(\bar{\eta}) = \partial_x^v \mathbb{E}[M_i(\bar{\eta}) | X_i = x]_{x=0^*}$ are:

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

A.2. General Result. In this subsection, we state and prove two theorems that generalize Theorems 1 and 2.

Theorem A.1. *Suppose that Assumptions 1–3 hold with $j \in \{1, \dots, p+1\}$ 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 Assumption 3 holds. For $s \in [S]$ and $\star \in \{-, +\}$, it holds that:*

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

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

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

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

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

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 $s \in \{1, \dots, S\}$. 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

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

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

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

for some $\tilde{x}_{i,p}$ 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 $j \in \{1, \dots, p\}$, 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

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

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

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

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). \square

Theorem A.2. *Suppose that the assumptions of Theorem A.1 hold, Assumption 4 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*

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

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

$$B_{p,n} = \frac{\bar{\nu}_p}{2} (\partial_x^{p+1} \mathbb{E}[M_i(\bar{\eta})|X_i = x]|_{x=0^+} + (-1)^p \partial_x^{p+1} \mathbb{E}[M_i(\bar{\eta})|X_i = x]|_{x=0^-}) + o_P(1),$$

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

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

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

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

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

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

A.3. Proofs of Theorems 1–3. Theorems 1 and 2 follow directly from the general results in Theorems A.1 and A.2 with $p = 1$. 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} . □

B. DETAILS ON SECTION 5.1: INFERENCE

In this section, we formally show that, under suitable assumptions, existing procedures for bandwidth selection and construction of confidence intervals based on “no covariates” estimators can be directly applied to the modified data $\{(X_i, M_i(\hat{\eta}))\}_{i \in [n]}$.

B.1. Standard Errors. We propose a consistent standard error for $\hat{\tau}_{v,p}(h; \hat{\eta})$ of the form

$$\widehat{se}_{v,p}^2(h; \hat{\eta}) = \sum_{i=1}^n w_{i,v,p}^2(h) \widehat{\sigma}_i^2(\hat{\eta}),$$

where $\widehat{\sigma}_i^2(\hat{\eta})$ is a nearest-neighbor estimator of the variance $\sigma_i^2(\bar{\eta}) = \mathbb{V}[M_i(\bar{\eta})|X_i]$:

$$\widehat{\sigma}_i^2(\hat{\eta}) = \left(M_i(\hat{\eta}_{s(i)}) - \frac{1}{R} \sum_{j \in \mathcal{R}_i} M_j(\hat{\eta}_{s(j)}) \right)^2,$$

with \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 $\sum_{i \in [n]} w_{i,v,p}^2(h) \iota_i(\hat{\eta}) = o_P((nh^{1+2v})^{-1})$ for $0 \leq v \leq p$, where*

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

Proposition B.1. *Suppose that Assumptions 1–4 and B.1 hold. Moreover, suppose that Assumption 1 also holds with \mathcal{X}_h replaced by $\tilde{\mathcal{X}}_h$ that is an open set s.t. $\mathcal{X}_h \subset \tilde{\mathcal{X}}_h$, and $\sup_{\eta \in \mathcal{T}_n} \sup_{x \in \tilde{\mathcal{X}}_h} \mathbb{E}[(M_i(\eta) - \mathbb{E}[M_i(\eta)|X_i])^4 | X_i = x]$ 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 $0 \leq v \leq p$, it holds that*

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

where $se_{v,p}^2(h; \bar{\eta}) = \sum_{i=1}^n w_{i,v,p}^2(h) \sigma_i^2(\bar{\eta})$.

We note that Assumption B.1 could be dropped if we were to study a slight variation of $\widehat{se}_{v,p}^2(h; \hat{\eta})$ 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. We discuss three of types of confidence intervals for the RD parameter τ .

B.2.1. Confidence intervals with 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

$$CI_{1-\alpha}^{us} = [\widehat{\tau}(h; \widehat{\eta}) \pm z_\alpha \widehat{\text{se}}(h; \widehat{\eta})],$$

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

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

B.2.2. Robust bias-corrected confidence intervals. 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

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

Proposition B.3 shows that $CI_{1-\alpha}^{rbc}$ 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 $\mathbb{P}(\tau \in CI_{1-\alpha}^{rbc}) \geq 1 - \alpha + o_p(1)$.*

B.2.3. Bias-aware confidence intervals. We consider a simplified version of the bias-aware approach of Armstrong and Kolesár (2018) that accounts for the asymptotic bias in our setting. Suppose that the researcher is willing to assume that the second derivative of the conditional expectation function of the outcome Y_i is bounded by B_Y . Then it follows from the results of Armstrong and Kolesár (2020) and our Theorem 2 that the smoothing bias of

our covariate-adjusted RD estimator is bounded in absolute value by $\bar{b}(h) + o_P(h^2)$, where

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

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

$$CI_{1-\alpha}^{ba} = [\hat{\tau}(h; \hat{\eta}) \pm z_\alpha(\bar{b}(h)/\hat{\text{se}}_{0,1}(h; \hat{\eta})) \hat{\text{se}}_{0,1}(h; \hat{\eta})].$$

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 $CI_{1-\alpha}^{ba}$ is asymptotically valid.

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

In contrast to the results of Armstrong and Kolesár (2018, 2020), we only show that this confidence interval is valid for a fixed sequence of DGPs, rather than uniformly over a larger set of DGPs. We leave providing inference that is uniformly valid over how the covariates affect the outcome variable for future research.

B.3. Consistent estimation of the MSE-optimal bandwidth. From Theorem 2, it follows that the bandwidth that minimizes the Asymptotic Mean Squared Error (AMSE) is given by

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

This optimal bandwidth can be consistently estimated by applying the procedure of Calonico et al. (2014, s.6) to the modified data $\{(X_i, M_i(\hat{\eta}_{s(i)}))\}_{i \in [n]}$ using the following three steps.¹⁷
Step 0. Initial bandwidths.

- (i) Take any sequence v_n such that $v_n \rightarrow 0$ and $nv_n \rightarrow \infty$. In practice, set $\hat{v}_n = 2.58 \min\{S_X, IQR/1.349\}n^{-1/5}$, where S_X^2 and IQR_X denote, respectively, the sample variance and interquantile range of $\{X_i : 1 \leq i \leq n\}$.

¹⁷We recognize that, similarly to the original proposal of Calonico et al. (2014), the proposed bandwidth selector is subject to the criticism of Armstrong and Kolesár (2020, Section 4.1).

(ii) Choose c_n s.t. $c_n \rightarrow 0$ and $nc_n^7 \rightarrow \infty$. Specifically, let

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

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

Step 1. Choose a pilot bandwidth b_n such that $b_n \rightarrow 0$ and $nb_n^5 \rightarrow \infty$. In practice, estimate the AMSE-optimal bandwidth for the second derivative in the local quadratic regression:

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

Step 2. The main bandwidth h_{AMSE} is estimated as

$$\begin{aligned}\widehat{h}_n &= \widehat{H}_n^{1/5} n^{-1/5}, \\ \widehat{H}_n &= \frac{nv_n \widehat{se}_{0,1}(v_n; \widehat{\eta})}{4\mathcal{B}_{0,1}^2 \left(\left(\widehat{\beta}_{2,2}^+(b_n; \widehat{\eta}) - \widehat{\beta}_{2,2}^-(b_n; \widehat{\eta}) \right)^2 + 3\widehat{se}_{2,2}(b_n; \widehat{\eta}) \right)}.\end{aligned}$$

Proposition B.5. *Suppose that the assumptions of Theorem A.2 and Proposition B.1 hold for $p = 3$, \mathcal{X} is bounded, $\mathbb{P}[1/C \leq |\widehat{\gamma}_{4,4}^+(\widehat{\eta}) - \widehat{\gamma}_{4,4}^-(\widehat{\eta})| \leq C] \rightarrow 1$ 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, we note that the weights $w_{i,v,p}(h)$ satisfy:

(i) $\sum_{i \in [n]} w_{i,v,p}(h)^2 = O_P((nh^{1+2v})^{-1})$ and (ii) $\max_{i \in [n]} w_{i,v,p}(h)^2 = o_P((nh^{1+2v})^{-1})$. This can be shown using standard kernel calculations.

The proof of Proposition B.1 consists in showing that $\widehat{se}_{v,p}^2(h; \widehat{\eta})$ is asymptotically equivalent

to its infeasible version that uses the deterministic function $\bar{\eta}$, given by

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

Using arguments as in the proof of Theorem 4 in Noack and Rothe (2021), one can show that $\widehat{\text{se}}_{v,p}^2(h; \bar{\eta}) - \text{se}_{v,p}^2(h; \bar{\eta}) = o_P((nh^{1+2v})^{-1})$. It therefore remains to show that $\widehat{\text{se}}_{v,p}^2(h; \widehat{\eta}) - \widehat{\text{se}}_{v,p}^2(h; \bar{\eta}) = o_P((nh^{1+2v})^{-1})$. 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:

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

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

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

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

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

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:

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

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

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

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. \square

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

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. \square

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 bias is bounded in absolute value by $\bar{b}(h) + o_P(h^2)$. \square

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) $\hat{\gamma}_{4,4}^*(\hat{\eta}) - \hat{\gamma}_{4,4}^*(\bar{\eta}) = o_P(1)$,
- (ii) $\hat{\beta}_{3,3}^+(c_n; \hat{\eta}) + \hat{\beta}_{3,3}^-(c_n; \hat{\eta}) = \beta_3^+(\bar{\eta}) + \beta_3^-(\bar{\eta}) + o_P(1)$,
- (iii) $\hat{\beta}_{2,2}^+(b_n; \hat{\eta}) - \hat{\beta}_{2,2}^-(b_n; \hat{\eta}) = \beta_2^+(\bar{\eta}) - \beta_2^-(\bar{\eta}) + o_P(1)$.

Part (i). First, note that

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

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

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

Note that, with probability approaching one,

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

It follows that $\left| \frac{S}{n} \sum_{i \in I_s} X_i^j (\bar{\eta}(Z_i) - \hat{\eta}_{s(i)}(Z_i)) \right| = o_p(1)$. 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 $\widehat{\beta}_{p,p}^*(h; \widehat{\eta}) - \widehat{\beta}_{p,p}^*(h, \bar{\eta}) = o_P(1)$. Moreover, under the assumptions made, $\widehat{\beta}_{p,p}^*(h, \bar{\eta}) - \beta_p^*(\bar{\eta}) = O_P(h + (nh^{1+2p})^{-1/2})$. The claims follow using the conditions on b_n and c_n . \square

C. DETAILS ON SECTION 5.3: FUZZY RD DESIGNS

We show asymptotic normality of the fuzzy covariate-adjusted RD estimator introduced in Section 5.3 and characterize the dependence of the asymptotic variance on the population analogues of the adjustment functions.

Proposition C.1. *Suppose that Assumptions 1–4 hold also with T_i replacing Y_i , mutatis mutandis.*

(i) *It holds that*

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

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*

$$V_\theta(\eta_Y^{(a)}, \eta_T^{(a)}) - V_\theta(\eta_Y^{(b)}, \eta_T^{(b)})$$

$$= \frac{2\bar{\kappa}}{\tau_T^2 f_X(0)} \left(\mathbb{V}[\eta_{Y,0}(Z_i) - \theta \eta_{T,0}(Z_i) - (\eta_Y^{(a)}(Z_i) - \theta \eta_T^{(a)}(Z_i)) | X_i = 0] \right.$$

$$\left. - \mathbb{V}[\eta_{Y,0}(Z_i) - \theta \eta_{T,0}(Z_i) - (\eta_Y^{(b)}(Z_i) - \theta \eta_T^{(b)}(Z_i)) | X_i = 0] \right).$$

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

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

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

D. ADDITIONAL SIMULATION RESULTS

In this section, we present further simulation results. Table E.1 extends the results in Table 1. We present results for the bias-aware approach discussed in the main text with a bandwidth that is chosen optimally for the standard RD estimator without covariates. By doing so, we focus on the effect that our covariate adjustments have on the constants in the bias and variance expressions of the asymptotic distribution while holding the bandwidth fixed. We can therefore see that the bias is the same across all methods, and we can still draw the same qualitative conclusions about the standard deviations of the covariate-adjusted RD estimators using different first-stage estimators as discussed in Section 6. We emphasize that the procedure of holding the bandwidth fixed is useful to understand the mechanics of our variance reduction procedure, but in practice we recommend choosing the bandwidth optimally for the given adjustment terms as it leads to narrower confidence intervals.

In Table E.1, we also consider bandwidth choice and confidence intervals constructions based on robust bias corrections and undersmoothing (the bandwidth for undersmoothing is chosen as $n^{-1/20}$ times the MSE-optimal bandwidth estimated using the `rdrobust` package). The linear adjustment estimators with cross-fitting are constructed analogously to the procedure described in Section 6.1. For flexible covariate adjustments, the bandwidth is chosen optimally for each adjustment function. The qualitative conclusions about the relative performance of different first-stage estimators in different models remain the same as discussed in the main text. The simulated average bandwidth of robust bias corrections is typically smaller than that of the bias-aware approach, and the confidence intervals are larger. This feature is known in the nonparametric literature.

In Figure E.1, we plot the empirical CDFs of the estimators considered in Section 6. These graphs illustrate the quality of the approximation obtained in Theorem 1. As predicted by our theory, for all DGPs, the entire distribution of the covariate-adjusted RD estimator using the ensemble combination of all feasible first-stage estimation methods is very similar to the one of the oracle estimator.

E. ADDITIONAL EMPIRICAL ESTIMATION RESULTS

In this section, we extend the empirical results from Section 7 by considering additional outcome variables and second-stage inference methods. Specifically, we consider four different outcome variables: food and non-food consumption, one year and two years after the program implementation. For the second stage, we employ bias-aware inference with different smoothness constants, robust bias corrections and a version of undersmoothing. As in Section 7, the results are based on 100 different data splits.

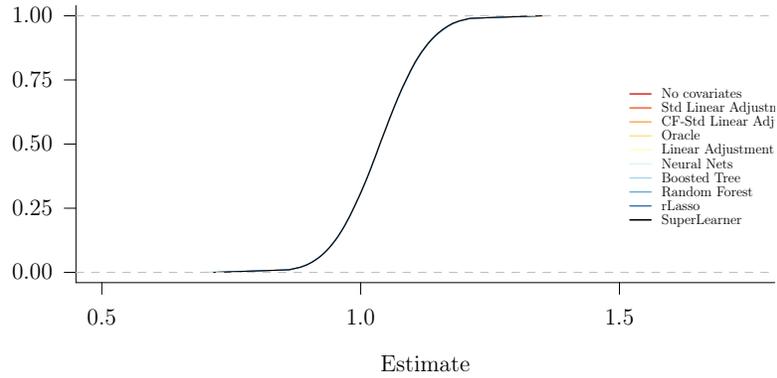
The magnitude of gains from using covariate adjustments varies across outcome variables, but the relative patterns remain the same in that our covariate adjustments based on the linear regression provide improvements upon the standard RD estimator, but SuperLearner adjustments generally perform better.

Table E.1: Simulation results for different second-stage specifications.

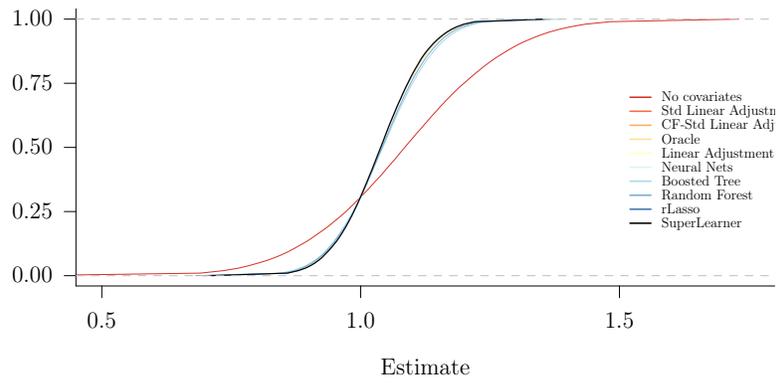
	Optimal “no covariates” bandwidth						Optimal covariate-adjusted bandwidth											
	RDHonest M=2						rdrobust						Undersmoothing					
	Bias	SD	RMSE	CI Cov	CI Length	h	Bias	SD	RMSE	CI Cov	CI Length	h	Bias	SD	RMSE	CI Cov	CI Length	h
DGP L=0																		
No covariates	3.69	7.49	8.35	94.90	32.58	43.20	1.40	9.78	9.88	94.84	52.18	29.78	0.57	11.49	11.50	94.47	43.03	20.36
Linear Adjustments	3.70	7.51	8.37	94.68	32.44	43.20	1.39	9.85	9.94	94.73	51.95	29.56	0.57	11.60	11.61	93.98	42.71	20.21
Linear Adjustments with CF	3.69	7.53	8.38	94.96	32.85	43.20	1.41	9.85	9.95	94.99	52.96	29.87	0.58	11.61	11.63	94.75	43.91	20.46
<i>Flexible Adjustments</i>																		
Oracle	3.69	7.49	8.35	94.90	32.58	43.20	1.40	9.78	9.88	94.84	52.18	29.78	0.57	11.49	11.50	94.47	43.03	20.36
Linear Regression	3.73	7.51	8.38	94.83	32.64	43.20	1.44	9.80	9.90	94.82	52.31	29.79	0.60	11.51	11.53	94.39	43.14	20.37
Neural Nets	3.72	7.50	8.38	94.84	32.63	43.20	1.44	9.80	9.90	94.82	52.28	29.79	0.60	11.51	11.52	94.42	43.12	20.37
Boosted Tree	3.72	7.64	8.50	94.77	33.11	43.20	1.43	9.96	10.07	94.80	53.20	29.81	0.60	11.72	11.73	94.50	43.88	20.39
Random Forest	3.73	7.56	8.42	94.83	32.82	43.20	1.43	9.86	9.97	94.87	52.64	29.79	0.61	11.60	11.61	94.44	43.42	20.37
rLasso	3.72	7.49	8.36	94.87	32.59	43.20	1.43	9.78	9.89	94.84	52.22	29.78	0.60	11.50	11.51	94.49	43.06	20.37
SuperLearner	3.73	7.51	8.38	94.86	32.64	43.20	1.44	9.80	9.91	94.88	52.31	29.78	0.61	11.52	11.53	94.51	43.14	20.37
DGP L=4																		
No covariates	8.52	17.03	19.04	94.95	74.29	65.59	1.54	26.96	27.01	94.90	144.63	30.55	0.78	31.80	31.81	94.57	119.28	20.89
Linear Adjustments	8.58	6.37	10.69	94.89	37.97	65.59	1.40	10.24	10.33	94.68	54.17	29.53	0.61	12.11	12.13	94.00	44.55	20.20
Linear Adjustments with CF	8.56	6.38	10.68	95.03	38.18	65.59	1.42	10.29	10.39	94.96	55.25	29.83	0.62	12.17	12.19	94.58	45.82	20.41
<i>Flexible Adjustments</i>																		
Oracle	8.58	6.58	10.81	94.98	38.79	65.59	1.41	10.59	10.68	94.84	56.28	29.91	0.59	12.46	12.47	94.48	46.42	20.45
Linear Regression	8.61	6.39	10.72	94.95	38.16	65.59	1.45	10.27	10.37	94.83	54.71	29.78	0.65	12.10	12.11	94.39	45.12	20.37
Neural Nets	8.61	6.39	10.72	94.94	38.15	65.59	1.45	10.26	10.36	94.86	54.70	29.77	0.65	12.09	12.11	94.41	45.12	20.36
Boosted Tree	8.61	7.19	11.22	94.82	40.74	65.59	1.50	11.49	11.59	94.89	61.37	29.81	0.72	13.53	13.55	94.36	50.62	20.39
Random Forest	8.61	6.88	11.02	94.91	39.73	65.59	1.48	11.03	11.13	94.88	58.77	29.78	0.70	12.99	13.01	94.36	48.47	20.36
rLasso	8.61	6.41	10.74	94.94	38.23	65.59	1.45	10.31	10.42	94.83	54.89	29.79	0.65	12.14	12.16	94.47	45.27	20.37
SuperLearner	8.61	6.38	10.72	94.98	38.14	65.59	1.45	10.26	10.36	94.84	54.66	29.77	0.65	12.08	12.10	94.44	45.09	20.36
DGP L=16																		
No covariates	10.45	20.96	23.42	94.87	91.01	72.23	1.59	34.61	34.65	94.98	185.82	30.78	0.74	40.94	40.95	94.62	153.28	21.05
Linear Adjustments	10.48	13.42	17.02	94.70	64.80	72.23	1.54	22.39	22.45	94.77	117.99	30.27	0.69	26.48	26.49	93.92	97.07	20.70
Linear Adjustments with CF	10.46	13.46	17.05	94.86	65.26	72.23	1.55	22.49	22.54	95.10	120.56	30.60	0.69	26.66	26.67	94.66	100.16	20.95
<i>Flexible Adjustments</i>																		
Oracle	10.42	6.24	12.15	95.02	41.44	72.23	1.50	10.51	10.61	94.86	56.31	29.88	0.69	12.35	12.37	94.39	46.45	20.43
Linear Regression	10.50	13.45	17.06	94.78	65.10	72.23	1.59	22.34	22.39	95.00	118.98	30.55	0.71	26.35	26.36	94.51	98.25	20.89
Neural Nets	10.51	8.06	13.24	94.92	47.35	72.23	1.57	13.49	13.58	94.96	72.09	30.16	0.75	15.90	15.92	94.56	59.48	20.62
Boosted Tree	10.48	8.30	13.37	94.76	48.05	72.23	1.58	13.82	13.91	94.89	73.90	30.21	0.76	16.29	16.31	94.57	60.98	20.66
Random Forest	10.47	11.68	15.68	94.82	59.12	72.23	1.58	19.42	19.48	94.90	103.36	30.47	0.72	22.88	22.89	94.37	85.33	20.84
rLasso	10.47	6.29	12.21	94.82	41.56	72.23	1.55	10.56	10.68	94.90	56.65	29.82	0.72	12.42	12.44	94.44	46.73	20.39
SuperLearner	10.47	6.29	12.22	94.77	41.56	72.23	1.55	10.57	10.68	94.88	56.65	29.82	0.73	12.43	12.45	94.44	46.73	20.39

Notes: Results are based on 50,000 Monte Carlo draws. The first panel uses the bandwidth that is optimal for the “no covariates” RD estimator and chooses the bandwidth and constructs confidence sets based on bias-aware inference. The second panel uses for each specification the optimal bandwidth and confidence sets based on robust bias correct and undersmoothing. The columns show results for simulated mean bias (Bias); the simulated standard deviation of the estimator (SD); the simulated root mean squared error (RMSE); simulated coverage of confidence intervals with 95% nominal level (CI Cov); the average simulated confidence interval length (CI Length); and the average simulated bandwidth (h). For flexible adjustments, in the first stage, the sample is restricted to the observations that lie in the window $(-b, b)$ with b being twice the “no covariates” bandwidth.

ECDFs for DGP1



ECDFs for DGP2



ECDFs for DGP3

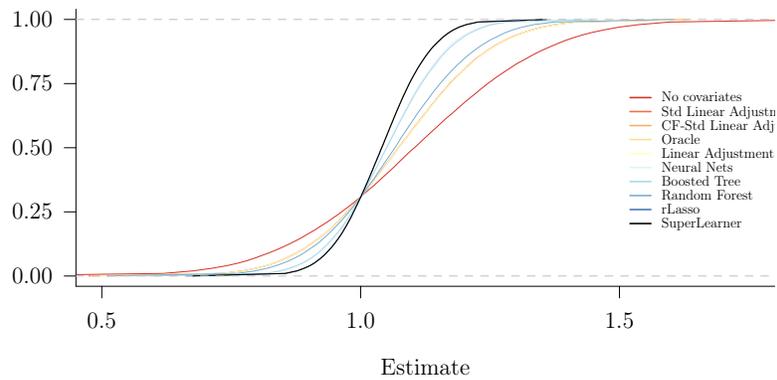


Figure E.1: Empirical cumulative distribution functions of covariate-adjusted RD estimates based on different adjustment methods and the different DGPs. The estimates are based on the data generating process explained in Section 6.2 and the estimators are implemented as explained in Section 6.1.

Table E.2: Full results for the empirical application.

	RDHonest M=100				RDHonest M=300				RDHonest M=500				rdrobust				Undersmoothing			
	Est	SE	Δ SE(%)	h	Estimate	SE	Δ SE(%)	h	Est	SE	Δ SE(%)	h	Est	SE	Δ SE(%)	h	Est	SE	Δ SE(%)	h
Food Consumption in t+1																				
No covariates	0.1	14.1	-	0.961	-18.6	16.6	-	0.591	-24.7	18.1	-	0.478	-22.2	27.4	-	0.372	-4.1	23.3	-	0.254
Linear Adjustments	-0.2	11.9	-18.6	0.914	-14.8	13.7	-21.0	0.567	-21.1	14.6	-24.0	0.462	-23.1	20.5	-33.9	0.342	-7.9	16.4	-41.8	0.234
Linear Adjustments with CF	-0.6	13.4	-5.2	0.930	-13.3	17.1	2.8	0.579	-16.3	19.9	9.1	0.471	-16.7	30.2	9.3	0.415	-11.8	25.2	7.8	0.255
<i>Flexible Adjustments</i>																				
Linear Regression	-2.5	13.3	-6.5	0.933	-16.0	15.7	-5.4	0.578	-19.9	17.2	-5.0	0.469	-17.7	25.7	-6.7	0.378	-6.0	21.8	-6.9	0.259
Neural Nets	-2.1	13.1	-7.7	0.930	-16.2	15.5	-6.9	0.576	-19.8	17.0	-6.5	0.467	-17.3	25.1	-9.1	0.378	-4.5	21.4	-9.0	0.259
Boosted Trees	-3.1	12.4	-13.7	0.920	-17.2	14.5	-14.7	0.569	-18.9	15.7	-15.2	0.461	-15.0	21.7	-26.4	0.386	1.6	18.8	-23.9	0.264
Random Forest	-5.1	12.4	-14.0	0.921	-19.1	14.4	-15.1	0.570	-21.0	15.7	-15.6	0.462	-17.1	22.0	-24.3	0.384	-0.4	18.9	-22.8	0.263
rLasso	-5.4	12.5	-12.9	0.928	-21.5	14.4	-15.3	0.575	-24.9	15.5	-16.9	0.466	-21.4	21.9	-24.9	0.372	-7.3	18.7	-24.1	0.255
SuperLearner	-3.9	12.4	-14.4	0.921	-18.3	14.3	-15.7	0.569	-20.7	15.6	-16.3	0.462	-17.0	21.8	-25.5	0.379	-1.2	18.8	-23.8	0.260
Food Consumption in t+2																				
No covariates	56.7	23.5	-	1.096	52.9	30.2	-	0.678	52.1	34.2	-	0.539	54.9	48.1	-	0.444	55.0	44.1	-	0.304
Linear Adjustments	59.6	20.9	-12.0	1.078	62.9	25.8	-17.1	0.671	61.4	28.3	-20.8	0.538	62.0	38.6	-24.5	0.425	32.0	33.8	-30.5	0.291
Linear Adjustments with CF	58.5	22.7	-3.4	1.097	63.2	29.5	-2.2	0.685	65.3	34.1	-0.3	0.548	68.3	52.4	8.3	0.446	38.7	54.4	18.8	0.292
<i>Flexible Adjustments</i>																				
Linear Regression	51.9	22.9	-2.4	1.101	55.5	29.6	-2.1	0.680	57.4	33.5	-1.9	0.541	60.2	47.6	-0.9	0.450	61.8	43.6	-1.3	0.308
Neural Nets	54.2	22.6	-3.6	1.092	56.7	29.3	-3.0	0.673	58.0	33.2	-2.8	0.535	61.3	46.6	-3.1	0.450	62.4	42.7	-3.3	0.308
Boosted Trees	50.8	22.8	-3.1	1.093	52.3	29.4	-2.7	0.674	51.9	33.3	-2.7	0.536	54.4	46.1	-4.2	0.451	54.9	42.6	-3.6	0.309
Random Forest	50.5	22.8	-2.7	1.091	51.5	29.6	-2.0	0.673	52.4	33.6	-1.7	0.535	56.1	46.7	-3.0	0.452	59.0	43.0	-2.7	0.310
rLasso	52.5	22.7	-3.6	1.093	51.0	29.3	-3.1	0.675	50.4	33.2	-3.0	0.537	53.3	46.4	-3.5	0.443	54.1	42.9	-3.0	0.303
SuperLearner	54.0	22.6	-4.0	1.089	54.0	29.2	-3.5	0.671	54.0	33.1	-3.2	0.533	56.9	45.9	-4.7	0.448	57.3	42.4	-4.2	0.307
Non-Food Consumption in t+1																				
No covariates	5.0	12.0	-	0.881	-3.7	14.1	-	0.542	-9.7	15.5	-	0.440	-9.1	22.0	-	0.344	-7.4	19.2	-	0.236
Linear Adjustments	3.7	10.4	-15.1	0.861	0.3	11.7	-20.9	0.536	-6.3	12.7	-22.5	0.436	-5.1	17.0	-29.3	0.350	-6.0	13.8	-39.0	0.240
Linear Adjustments with CF	2.1	11.5	-4.2	0.878	-2.1	13.6	-3.8	0.543	-7.9	15.2	-1.8	0.443	-8.3	20.7	-5.8	0.421	-10.3	18.7	-2.5	0.300
<i>Flexible Adjustments</i>																				
Linear Regression	-0.8	12.0	0.2	0.880	-6.4	14.2	0.4	0.542	-10.7	15.6	0.5	0.440	-10.4	21.1	-4.2	0.391	-12.4	18.7	-2.7	0.268
Neural Nets	-0.5	11.9	-0.6	0.876	-6.9	14.0	-0.5	0.539	-11.5	15.4	-0.5	0.438	-11.4	20.9	-5.1	0.375	-15.2	18.6	-3.4	0.257
Boosted Trees	1.4	12.2	1.8	0.865	-3.8	14.6	3.2	0.535	-7.7	16.1	3.5	0.435	-7.3	21.7	-1.1	0.404	-6.5	19.1	-0.4	0.277
Random Forest	5.6	12.0	0.7	0.870	0.5	14.3	1.7	0.537	-2.9	15.8	1.8	0.436	-2.3	21.5	-2.1	0.403	1.8	18.9	-1.6	0.276
rLasso	6.9	11.8	-1.3	0.876	-0.9	13.8	-2.2	0.539	-5.5	15.2	-2.3	0.438	-4.3	21.2	-3.4	0.351	-4.5	18.6	-3.0	0.240
SuperLearner	3.9	11.8	-1.1	0.865	-2.5	14.0	-0.7	0.535	-6.8	15.4	-0.6	0.435	-6.1	21.1	-4.2	0.374	-6.3	18.6	-3.2	0.256
Non-Food Consumption in t+2																				
No covariates	37.9	17.3	-	0.986	41.7	21.7	-	0.605	42.6	24.4	-	0.487	43.8	32.3	-	0.440	37.5	29.7	-	0.301
Linear Adjustments	42.4	15.5	-11.6	0.975	46.4	18.3	-18.2	0.604	43.2	20.0	-22.3	0.490	40.3	26.9	-20.3	0.400	15.5	23.4	-27.0	0.274
Linear Adjustments with CF	43.4	17.1	-1.1	0.998	50.4	21.8	0.5	0.617	48.8	24.9	2.1	0.500	48.8	36.4	11.3	0.435	27.0	37.9	21.6	0.294
<i>Flexible Adjustments</i>																				
Linear Regression	41.3	17.2	-0.3	1.014	49.7	21.5	-0.6	0.617	51.1	24.2	-0.9	0.497	52.0	32.5	0.5	0.436	44.6	30.0	0.9	0.299
Neural Nets	39.5	17.0	-1.4	0.990	46.8	21.3	-1.6	0.608	47.6	24.0	-1.8	0.490	48.6	31.8	-1.6	0.439	41.9	29.4	-1.1	0.300
Boosted Trees	34.9	16.8	-2.4	0.988	41.0	21.1	-2.7	0.606	42.2	23.7	-3.0	0.487	43.3	31.2	-3.7	0.441	37.0	28.8	-3.2	0.302
Random Forest	34.7	17.2	-0.2	0.985	40.7	21.8	0.4	0.603	43.0	24.5	0.3	0.486	44.4	32.0	-0.9	0.448	40.9	29.5	-0.6	0.307
rLasso	38.3	17.1	-1.0	0.988	42.2	21.4	-1.2	0.605	42.9	24.1	-1.3	0.487	44.1	31.8	-1.8	0.436	37.9	29.4	-1.1	0.299
SuperLearner	38.3	16.8	-2.6	0.982	43.9	21.1	-2.5	0.602	45.5	23.8	-2.6	0.485	46.7	31.1	-3.8	0.444	41.0	28.7	-3.3	0.304

Notes: The table presents results for four outcome variables and five second-stage specifications. In each case, we consider various types of covariate adjustments. We report the median estimate across 100 splits of the data (Est), the nearest-neighbor standard error taking into account the variation introduced by sample splitting (SE), the percentage change in the standard error relative to the “no covariates” standard error (Δ SE(%)), and the median bandwidth (h). For flexible adjustments, in the first stage, the sample is restricted to the observations that lie in the window $(-b, b)$ where b is twice the “no covariates” bandwidth for RDHonest with smoothness constant $M = 300$ calculated for the respective outcome variable.

REFERENCES

- ANDREWS, D. (1994): “Asymptotics for semiparametric econometric models via stochastic equicontinuity,” *Econometrica*, 62, 43–72.
- 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., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): “Sparse models and methods for optimal instruments with an application to eminent domain,” *Econometrica*, 80, 2369–2429.
- 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,” *The 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,” *The 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. (2017): “Alternative Assumptions to Identify LATE in Fuzzy Regression Discontinuity Designs,” *Working Paper*.
- 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 KLAUW (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.
- KOLESÁR, M. AND C. ROTHE (2018): “Inference in Regression Discontinuity Designs with a Discrete Running Variable,” *American Economic Review*, 108, 2277–2304.
- 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.
- MCCRARY, J. (2008): “Manipulation of the running variable in the regression discontinuity design: A density test,” *Journal of econometrics*, 142, 698–714.
- NEWBY, W. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–1382.
- NOACK, C. AND C. ROTHE (2021): “Bias-aware inference in fuzzy regression discontinuity designs,” *arXiv preprint arXiv:1906.04631*.
- 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.

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.