# REGRESSION DISCONTINUITY DESIGN WITH POTENTIALLY MANY COVARIATES

YOICHI ARAI, TAISUKE OTSU, AND MYUNG HWAN SEO

Abstract. This paper studies the case of possibly high-dimensional covariates in the regression discontinuity design (RDD) analysis. In particular, we propose estimation and inference methods for the RDD models with covariate selection which perform stably regardless of the number of covariates. The proposed methods combine the local approach using kernel weights with $\ell_1$-penalization to handle high-dimensional covariates, and the combination is new in the literature. We provide theoretical and numerical results which illustrate the usefulness of the proposed methods. Theoretically, we present risk and coverage properties for our point estimation and inference methods, respectively. Numerically, our simulation experiments and empirical example show the robust behaviors of the proposed methods to the number of covariates in terms of bias and variance for point estimation and coverage probability and interval length for inference.

## 1. Introduction

In causal or treatment effect analysis, discontinuities in regression functions induced by an assignment variable can provide useful information to identify certain causal effects. The regression discontinuity design (RDD) has been widely applied in observational studies to identify the average treatment effect at the discontinuity point. For the RDD, the causal parameters of interest are identified by some contrasts of the left and right limits of the conditional mean functions. See e.g. Imbens and Lemieux (2008), Cattaneo, Titiunik and Vazquez-Bare (2020), an edited volume Cattaneo and Escanciano (2017), and references therein.

In the growing literature on the RDD analysis, this paper focuses on the RDDs where covariates are included in the estimation, which is extensively studied by Calonico, Cattaneo, Farrell and Titiunik (2019) (hereafter, CCFT). See also Frölich and Huber (2019) for an alternative estimation method based on kernel smoothing after localization around the cutoff. In practice, researchers often augment the regression models for the RDD analysis with various additional predetermined covariates such as demographic or socioeconomic characteristics for data units. For several RDD estimators using covariates based on local polynomial regression methods, CCFT investigated the MSE expansion, asymptotic efficiency, and data-driven bandwidth selection methods. Furthermore, CCFT developed asymptotic distributional approximations for those

*Key words and phrases.* Regression discontinuity design; Covariate selection; Lasso.
This paper is a developed version of the previous manuscript (https://sticerd.lse.ac.uk/dps/em/em601.pdf) inspired by the discussion with Matias Cattaneo. We are grateful to Matias Cattaneo and Sebastian Calonico for helpful comments and discussions. This research was supported by Grants-in-Aid for Scientific Research 20K01598 from the Japan Society for the Promotion of Science (Arai) and the ERC Consolidator Grant (SNP 615882) (Otsu). Financial support from the Center for Distributive Justice in the Institute of Economic Research of Seoul National University and the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A01033487) is gratefully acknowledged (Seo).

estimators and proposed valid inference procedures by constructing bias and variance estimators with covariate adjustment. These results may be considered as extensions of the analyses in Calonico, Cattaneo and Titiunik (2014) (hereafter, CCT) combined with robust bias correction methods in Calonico, Cattaneo and Farrell (2018, 2020) to incorporate covariates in the RDD analysis. See also Calonico, Cattaneo, Farrell and Titiunik (2017) for the statistical package on these methods.

In randomized controlled trials, regression adjustment using covariates is a common practice since it is always helpful to improve asymptotic efficiency of the causal effect estimator as far as a full set of treatment-covariate interactions is included (Lin, 2013). Also a recent paper by Lei and Ding (2021) proposed a bias correction method for the regression adjustment estimator with a diverging number of covariates. On the other hand, in the RDD analysis, which is a quasi-experiment, the efficiency gain by introducing covariates is not necessarily guaranteed, and CCFT provided a concrete guideline by clarifying the conditions to achieve consistency and efficiency gain for the covariates adjusted RDD estimator. Typically the efficiency improves when the projection coefficients of the covariates on the outcome is equal for both control and treatment groups. Since practitioners also commonly incorporate covariates for the RDD analysis, CCFT's guideline has a large impact in applied research. When we use covariates, it is common to employ their transformations and interactions, and the number of terms can be pretty large. This paper adds a further guideline for practitioners who also face a large number of covariates. In the above scenario for efficiency improvement, it is beneficial to augment CCFT's procedure with covariate selection by the high-dimensional methods particularly when the projection coefficients satisfy certain sparsity. Furthermore, even if the practitioner is not convinced a priori with efficiency gain, the regression adjustment in the RDD analysis can be a useful complement to the unadjusted RDD estimates particularly when their standard errors are large.

For point estimation on the causal effect parameter identified by the RDD, we consider the Lasso estimator and its post-selection estimator based on the local linear regression (i.e., eq. (2) of CCFT). To the best of our knowledge, such a combination of the localization using kernel weights and $\ell_1$-penalization to deal with high-dimensional covariates is novel in the literature. Indeed this combination is particularly relevant for the RDD analysis, where the effective sample size would be typically small due to the localization so that the effect of dimensionality of covariates becomes severer. Theoretically, we derive the $\ell_1$-risk properties of our "local Lasso" estimators and its post-selection version. Practically, based on our simulation study, we recommend the CCFT estimator after selecting covariates by the $\ell_1$-penalization even for a relatively small number of covariates, which exhibits desirable MSE properties and stability across different setups.

For inference, we propose to select covariates with the local Lasso. We show that the inference based on the selected covariates can be implemented in the same manner as in CCFT. We also show under which circumstances our approach can lead to improved efficiency. Our simulation results demonstrate that the post-selection confidence interval exhibits robust performances in terms of both coverages and lengths, even for a relatively small number of covariates.

This paper also contributes to the large literature on high-dimensional methods in econometrics and statistics (see, e.g., Bühlmann and van de Geer, 2011, and Belloni *et al.*, 2018, for an overview) by combining the kernel localization with $\ell_1$-penalization to handle high-dimensional covariates. Our inference problem can be formulated as the one for low-dimensional parameters in high-dimensional models. In statistics literature, many papers investigated this issue, such as Belloni, Chernozhukov and Hansen (2014), van de Geer, *et al.* (2014), and Zhang and Zhang (2014). However, these approaches are not directly applicable to the RDD context because the current problem concerns the inference on a jump in a nonparametric regression model.[1]

This paper is organized as follows. Section 2.1 introduces our basic setup and local Lasso estimator, and presents the $\ell_1$-risk properties. In Section 2.2, we discuss the validity of CCFT's inference after selecting covariates by our Lasso procedure. Section 3 provides discussions on some extensions. A step-by-step procedure for implementation of our method is described in Section 4. To illustrate the proposed method, Section 5 conducts a simulation study, and Section 6 presents an empirical example based on the Head Start data.

## 2. Main result

### 2.1. Setup and local Lasso estimator for covariate selection.

In this subsection, we present our basic setup and introduce the local Lasso estimator for the RDD with possibly high-dimensional covariates. For each unit $i = 1, \ldots, n$, we observe an indicator variable $T_i$ for a treatment ($T_i = 1$ if treated and $T_i = 0$ otherwise), and outcome $Y_i = Y_i(0) \cdot (1 - T_i) + Y_i(1) \cdot T_i$, where $Y_i(0)$ and $Y_i(1)$ are potential outcomes for $T_i = 0$ and $T_i = 1$, respectively. Note that we cannot observe $Y_i(0)$ and $Y_i(1)$ simultaneously. Our purpose is to make inference on the causal effect of the treatment, or more specifically, some distributional aspects of the difference of potential outcomes $Y_i(1) - Y_i(0)$. The RDD analysis focuses on the case where the treatment assignment $T_i$ is completely or partly determined by some observable covariate $X_i$, called the running variable. For example, to study the effect of class size on pupils' achievements, it is reasonable to consider the following setup: the unit $i$ is school, $Y_i$ is an average exam score, $T_i$ is an indicator variable for the class size ($T_i = 0$ for one class and $T_i = 1$ for two classes), and $X_i$ is the number of enrollments. For more examples, see e.g. Imbens and Lemieux (2008), Cattaneo, Titiunik and Vazquez-Bare (2020), Cattaneo and Escanciano (2017), and references therein.

Depending on the assignment rule for $T_i$ based on $X_i$, we have two cases, called the sharp and fuzzy RDDs. In this section, we focus on the sharp RDD and discuss the fuzzy RDD in Section 3.1. In the sharp RDD, the treatment is deterministically assigned as $T_i = \mathbb{I}\{X_i \geq \bar{x}\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function and $\bar{x}$ is a known discontinuity (cutoff) point. Throughout the paper, we normalize $\bar{x} = 0$ to simplify the presentation. A parameter of interest, in this case, is the average causal effect at the discontinuity point,

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

Since the difference of potential outcomes $Y_i(1) - Y_i(0)$ is unobservable, we need a tractable representation of $\tau$ in terms of quantities that can be estimated by data. If the conditional mean functions $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ are continuous at the cutoff point $x = 0$, then the

[1]After submission of this paper, a recent paper by KreiSS and Rothe (2022) investigates a similar estimator to ours, and discusses an inference method based on the approach by Armstrong and Kolesár (2018).

average causal effect $\tau$ can be identified as a contrast of the left and right limits of the conditional mean $\mathbb{E}[Y_i|X_i = x]$ at $x = 0$, that is

$$\tau = \lim_{x \downarrow 0} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow 0} \mathbb{E}[Y_i|X_i = x]. \tag{2}$$

As argued in CCFT, it is usually the case that practitioners have access to additional covariates (denoted by $Z_i \in \mathbb{R}^p$) and augment their empirical models with $Z_i$ to estimate the causal effect $\tau$ of interest. This practically relevant setup is extensively studied in CCFT for the case where $Z_i$ is low-dimensional. In this paper, we consider the case of possibly high-dimensional $Z_i$, and propose a new point estimation method for $\tau$ and an adjustment of CCFT's inference method.

We examine the case where the additional covariates $Z_i$ are predetermined in the sense that $Z_i = Z_i(0) \cdot (1 - T_i) + Z_i(1) \cdot T_i$ but $Z_i(0) =_d Z_i(1)$ for the potential covariates $Z_i(0)$ and $Z_i(1)$ for $T_i = 0$ and $T_i = 1$, respectively. Motivated by CCFT's recommended model (in their eq. (2)), we propose the local Lasso estimator $\hat{\theta} = (\hat{\alpha}, \hat{\tau}, \hat{\beta}_-, \hat{\beta}_+, \hat{\gamma}')'$ that solves

$$\min_{\theta} \frac{1}{nh} \sum_{i=1}^{n} K_i (Y_i - \alpha - T_i \tau - X_i \beta_- - T_i X_i \beta_+ - Z_i' \gamma)^2 + \lambda_n |\theta|_1, \tag{3}$$

where $K_i = K(X_i/h)$ is the kernel weight to localize around the cutoff point $x = 0$ with a bandwidth denoted by $h$, $\theta = (\alpha, \tau, \beta_-, \beta_+, \gamma')'$ is a $(p + 4)$-dimensional vector of parameters, $|\theta|_1 = \sum_{j=1}^{p+4} |\theta^{(j)}|$ is the $\ell_1$-norm of the parameter vector ($\theta^{(j)}$ means the $j$-th element of $\theta$), and $\lambda_n$ is a penalty level. Popular choices for $K(\cdot)$ are the uniform and triangular kernels supported on $[-h, h]$. Then our point estimator for $\tau$ is given by $\hat{\tau}$.

It is often the case that researchers do not want to penalize some subset of parameters (particularly $\tau$ and perhaps $(\beta_-, \beta_+)$), denoted by $\theta_1$. In this case, we consider the partially penalized estimator $\tilde{\theta} = (\tilde{\alpha}, \tilde{\tau}, \tilde{\beta}_-, \tilde{\beta}_+, \tilde{\gamma}')'$ that solves

$$\min_{\theta} \frac{1}{nh} \sum_{i=1}^{n} K_i (Y_i - \alpha - T_i \tau - X_i \beta_- - T_i X_i \beta_+ - Z_i' \gamma)^2 + \lambda_n |\theta_2|_1, \tag{4}$$

where $\theta = (\theta_1', \theta_2')'$.

Our preliminary simulation results suggest that the local Lasso estimators for $\tau$ are somewhat biased in finite samples. Therefore, our recommendation for point estimation is to employ a post-selection method. Let $\bar{S} = \{j : \hat{\theta}^{(j)} \neq 0\}$ or $\{j : \tilde{\theta}^{(j)} \neq 0\}$ be the indices for selected covariates based on the local Lasso estimation in (3) or (4), respectively, $Z_{\bar{S}}$ be the vector of selected covariates, and $G_{\bar{S},i} = (1, T_i, X_i, T_i X_i, Z_{\bar{S},i}')'$. Then the local post-Lasso estimator $\bar{\theta}_{\bar{S}}$ is obtained by the local least square:

$$\min_{\theta_{\bar{S}}} \frac{1}{nh} \sum_{i=1}^{n} K_i (Y_i - G_{\bar{S},i}' \theta_{\bar{S}})^2, \tag{5}$$

and the estimator $\bar{\tau}$ for $\tau$ is given by the estimated coefficient of $T$.

To the best of our knowledge, such a combination of the localization using kernel weights $K_i$ and $\ell_1$-penalization to deal with high-dimensional covariates $Z_i$ is novel in the literature and also practically relevant in the RDD analysis. Several points are worthy of remark for this estimator. First, without the $\ell_1$-penalization, our estimator reduces to the local linear-type estimator recommended by CCFT's eq. (2). Therefore, the proposed estimator is a natural generalization of CCFT's when the dimension of $Z_i$ is high. Second, without the kernel weights $K_i$ for localization, our estimator reduces to the conventional Lasso estimator. However, since our parameter of interest $\tau$ is identified as a local object in (2), it is crucial to introduce such localization to avoid misspecification bias of the conditional mean functions. Third, it is often the case that the kernel function $K(\cdot)$ has bounded support. In this case, the effective sample size would be typically of order $nh$. Thus even if the dimension of $Z_i$ is relatively small compared to the original sample size $n$, the $\ell_1$-penalization would be useful especially for small values of $h$.

We now present a risk property of the local Lasso estimators $\hat{\theta}$, $\tilde{\theta}$, and $\bar{\theta}$. Let $G_i = (1, T_i, X_i, T_i X_i, Z_i')'$ be the vector of regressors in (3), $G_{ij}$ be the $j$-th element of $G_i$, $\Theta_n = \arg\min_\theta \mathbb{E}[K_i(Y_i - G_i'\theta)^2]$ be an argmin set, and $\hat{M} = \frac{1}{nh}\sum_{i=1}^n K_i G_i G_i'$. We impose the following assumptions.

**Assumption 1.** (1) *There exists $\theta^* \in \Theta_n$ such that $|\theta^*|_0 \leq s^*$ for some sequence $s^* = o(n)$.*

(2) *Let $e_i = K_i^{1/2} Y_i - K_i^{1/2} G_i' \theta^*$. There exists some $C \in (0, \infty)$ such that*

$$\mathbb{E}[|K_i G_{ij} e_i|^m] \leq h \frac{m! C^{m-2}}{2},$$

*for all $j = 1, \ldots, p$ and $m = 2, 3, \ldots$.*

(3) *Let $\delta_A$ be the subvector of $\delta$ for an index set $A$. There exists some $\phi_* \in (0, \infty)$ for $S = S^* = \{j : \theta^{*(j)} \neq 0\}$ such that*

$$\min_{\delta : |\delta_{S^c}|_1 \leq 3|\delta_S|_1} \frac{\delta' \hat{M} \delta}{|\delta_S|_1^2} |S| \geq \phi_*^2,$$

*with probability approaching one.*

(4) *$K : \mathbb{R} \to \mathbb{R}$ is a bounded and symmetric second-order kernel function which is continuous with compact support. The bandwidth $h$ is a positive sequence satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$.*

The first condition defines $\theta^*$ as the best linear predictor with some sparsity feature. This sparsity assumption is not so restrictive as it appears because (i) the permitted sparsity or the number of relevant covariates is comparable to the one in CCFT, and (ii) in RDD analyses, high-dimensional covariates are often introduced to accommodate many generated covariates, such as transformations of initial covariates like polynomials, interactions, and various basis functions, without knowing which of them are relevant a priori.[2] Obviously this condition covers the case of finite dimensional covariates (i.e., $|\theta^*|_0$ is finite) as in CCFT. The second is a set of moment conditions, which appear in the Bernstein inequality in the high-dimensional literature except for the multiplicative factor by $h$. This extra factor is due to the presence of the kernel weight $K_i$. The third condition is a localized version of the compatibility condition. A sufficient condition

[2]To motivate the use of generated covariates, it is insightful to note that the asymptotic variance of CCFT's RDD estimator is proportional to $\mathrm{Var}(\{(Y_i(1) - Y_i(0)) - (Z_i(1) - Z_i(0))'\gamma\}^2 | X_i = 0)$ for some $\gamma$, which is considered as the (conditional) variance of the linear projection error. Although CCFT considered linear projection due to the constraint on dimensionality, it is clear that the asymptotic variance is minimized by employing the conditional expectation $\mathbb{E}[Y_i(1) - Y_i(0) | Z_i(1) - Z_i(0), X_i = 0]$ instead of the linear projection $(Z_i(1) - Z_i(0))'\gamma$. Therefore, it is natural to extend CCFT's approach to high-dimensional setting by employing generated covariates or series approximation for the conditional mean.

for this is the so-called restricted eigenvalue condition. More specifically, $\min_{\beta:|\beta|_0 \le s} \frac{\beta' \hat{M} \beta}{\beta' \beta}$, where $|\beta|_0$ denotes the cardinality of $\beta$, provides a lower bound for the compatibility constant $\phi_*^2$ (see, e.g., Section 6.13 of Bühlmann and van der Geer, 2011). If CCFT's method is feasible with each subset of covariates of dimension $s$, then the restricted eigenvalue condition is indeed satisfied. The fourth condition contains assumptions on the kernel $K$ and bandwidth $h$, which are standard in the literature of nonparametric methods.

The $\ell_1$-risk properties of the local Lasso estimators are obtained as follows.

**Theorem 1.** *Suppose* $\sqrt{\log p/(nh)} = o(\lambda_n)$.

**(i):** *Under Assumption 1, it holds*

$$|\hat{\theta} - \theta^*|_1 \le 4\lambda_n \frac{s^*}{\phi_*^2}, \tag{6}$$

*with probability approaching one.*

**(ii):** *Under Assumption 1 with* $S^*$ *containing all the indexes for* $\theta_1$*, it holds*

$$|\tilde{\theta} - \theta^*|_1 \le C\lambda_n \frac{s^*}{\phi_*^2}, \tag{7}$$

*for some* $C \in (0, \infty)$ *with probability approaching one.*

**(iii):** *Under Assumption 1 with* $S^*$ *containing all the indexes for* $\theta_1$*, it holds*

$$|\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}}|_1 \le \lambda_{\min} \left( \frac{1}{nh} \sum_{i=1}^{n} G_{\bar{S},i} G'_{\bar{S},i} \right)^{-1} \lambda_n |\bar{S}|, \tag{8}$$

*where* $\lambda_{\min}(A)$ *denotes the minimum eigenvalue of a matrix* $A$.

The proof of this theorem is presented in Appendix A.1. This theorem characterizes the risk properties of the estimators $\hat{\theta}$, $\tilde{\theta}$, and $\bar{\theta}$ around $\theta^*$, and the bounds depend on the tuning parameter $\lambda_n$, the number of non-zero coefficients $s^*$, and the compatibility constant $\phi_*$. Note that the decay rate of $\lambda_n$ is bounded from below by $\sqrt{\log p/(nh)}$. Thus, the risk bounds of $\hat{\theta}$ and $\tilde{\theta}$ get worse as the number of covariates $p$ increases or the effective sample size $nh$ for the kernel localization decreases. The result for the post-selection estimator $\bar{\theta}$ shows that the deviation from the original Lasso estimator $\tilde{\theta}$ is small when tuning parameter $\lambda_n$ or the number of selected covariates $|\bar{S}|$ is small, or the minimum eigenvalue of $\frac{1}{nh} \sum_{i=1}^{n} G_{\bar{S},i} G'_{\bar{S},i}$ is large.

The above theorem is on estimation of the coefficients of the best linear predictor $\theta^* = (\alpha^*, \tau^*, \beta_-^*, \beta_+^*, \gamma^{*\prime})'$. Additionally suppose that the assumptions of Lemma 1 of CCFT hold true, and the covariates $Z_i$ are predetermined. Then we can guarantee that $\tau^*$ coincides with the average causal effect $\tau$ in (1) so that Theorem 1 provides the conditions for the consistency and convergence rate of $\hat{\tau}$ to $\tau$. If $Z_i$ are not predetermined (i.e., $Z_i(0) \neq_d Z_i(1)$), then $\hat{\tau}$ typically converges to $\tau$ minus some bias component, which is obtained as a limit of CCFT's bias term in their Lemma 1.

Our estimators and above theorem can be extended to other regression models that contain the covariates $\{T_i Z_i, (1-T_i)Z_i\}$, $(Z_i - \bar{Z})$, or $\{T_i(Z_i - \bar{Z}), (1-T_i)(Z_i - \bar{Z})\}$ as in CCFT. However, as shown in Lemma 1 of CCFT, such estimators require more stringent conditions to guarantee the consistency for $\tau$. Furthermore, the local lasso regression (3) can be extended to incorporate the polynomials of $X_i$ and $T_iX_i$ even though this paper focuses on the local linear model.

Finally, we discuss the choices of the localization bandwidth $h$ and regularization parameter $\lambda_n$. We can use the MSE-optimal bandwidth based on the suggestion by CCFT and the regularization parameter $\lambda_n$ using cross-validation by Friedman, Hastie and Tibshirani (2010) and the data-driven choice by Belloni, Chernozhukov and Hansen (2014) among others. See Theorem 2 in the next subsection for their justification, and Section 4 for a detail on our practical recommendation.

2.2. **Inference.** We next consider interval estimation and hypothesis testing on the average causal effect $\tau$. For finite or low-dimensional $Z_i$, we recommend to use CCFT's inference methods. This subsection argues that we can still apply CCFT's inference procedures for high-dimensional $Z_i$, provided that CCFT's conditions remain valid for $S^*$ and $\theta^*_{S^*}$.

More precisely, based on the point estimator $\hat{\theta}$ in (3) or $\tilde{\theta}$ in (4), define the subsets of $\{1,\dots,p\}$ as

$$\hat{S}=\left\{j:|\hat{\gamma}^{(j)}|>\lambda_n\varrho_n\sum_{j=1}^{p}\mathbb{I}\{\hat{\gamma}^{(j)}\neq 0\}\right\},\qquad \tilde{S}=\left\{j:|\tilde{\gamma}^{(j)}|>\lambda_n\varrho_n\sum_{j=1}^{p}\mathbb{I}\{\tilde{\gamma}^{(j)}\neq 0\}\right\},$$

where we set $\varrho_n=\log\log\log n$. This choice of $\varrho_n$ is based on the simulation experiments in Section 5, and it is not optimal in any sense but works reasonably well.

Based on the selected covariates by $\hat{S}$ or $\tilde{S}$, we can apply CCFT's bias corrected t-ratio to conduct statistical inference on the causal effect parameter $\tau$. Consider the local post-Lasso estimator $\bar{\tau}$, which is the second element of $\bar{\theta}_{\bar{S}}$ in (5). As shown in Appendix A.2, the dominant term of $\bar{\tau}$ can be characterized as

$$\bar{\tau}=e_2'\left(\frac{1}{nh}\sum_{i=1}^{n}K_iG_{1i}G_{1i}'\right)^{-1}\frac{1}{nh}\sum_{i=1}^{n}K_iG_{1i}\xi_i+o_p((nh)^{-1/2}),\tag{9}$$

where $e_2=(0,1,0,0)'$, $G_{1i}=(1,T_i,X_i,T_iX_i)'$, and $\xi_i=T_i\xi_i(1)+(1-T_i)\xi_i(0)$ with $\xi_i(t)=Y_i(t)-Z_i(t)'\gamma_Y$ for some $\gamma_Y$ such that $\sum_{t=0}^{1}\mathrm{Cov}(\xi_i(t),Z_i(t)|X_i=0)=0$. Indeed, this term is analogous to the one derived for the case of fixed dimensional $Z_i$ in CCFT under which $\gamma_Y=\left[\sum_{t=0}^{1}\mathrm{Var}(Z_i(t)|X_i=0)\right]^{-1}\left(\sum_{t=0}^{1}\mathrm{Cov}(Z_i(t),Y_i(t)|X_i=0)\right)$ and $\gamma^*$ defined as a part of $\theta^*$ in Assumption 1 converges to $\gamma_Y$. Therefore, the pre-asymptotic bias and variance of $\bar{\tau}$ can be analogously written as

$$\begin{aligned}\mathcal{B}&=\frac{1}{2}e_1'(R'K_-R)^{-1}(R'K_-Q_-)(q'\mu_-^{(2)})+\frac{1}{2}e_1'(R'K_+R)^{-1}(R'K_+Q_+)(q'\mu_+^{(2)}),\\ \mathcal{V}&=(q'\otimes e_1'P_-)\Sigma_-(q\otimes P_-e_1)+(q'\otimes e_1'P_+)\Sigma_+(q\otimes P_+e_1),\end{aligned}\tag{10}$$

respectively, where $e_1 = (1, 0)$, $R = \begin{bmatrix} 1 & \dots & 1 \\ X_1/h & \dots & X_n/h \end{bmatrix}'$, $q = (1, -\gamma^{*\prime})'$, and

$$
\begin{aligned}
K_- &= \operatorname{diag}(\mathbb{I}\{X_1 < 0\}K_h(X_1), \dots, \mathbb{I}\{X_n < 0\}K_h(X_n)), \\
K_+ &= \operatorname{diag}(\mathbb{I}\{X_1 \geq 0\}K_h(X_1), \dots, \mathbb{I}\{X_n \geq 0\}K_h(X_n)), \\
Q_- &= R'K_-[(X_1 - \bar{X})^2/h^2, \dots, (X_n - \bar{X})^2/h^2], \\
Q_+ &= R'K_+[(X_1 - \bar{X})^2/h^2, \dots, (X_n - \bar{X})^2/h^2], \\
\mu_-^{(2)} &= \left[ \left.\frac{\partial^2 \mathbb{E}[Y_i(0)|X_i = x]}{\partial x^2}\right|_{x=0}, \left.\frac{\partial^2 \mathbb{E}[Z_i^*(0)'|X_i = x]}{\partial x^2}\right|_{x=0} \right]', \\
\mu_+^{(2)} &= \left[ \left.\frac{\partial^2 \mathbb{E}[Y_i(1)|X_i = x]}{\partial x^2}\right|_{x=0}, \left.\frac{\partial^2 \mathbb{E}[Z_i^*(1)'|X_i = x]}{\partial x^2}\right|_{x=0} \right]', \\
P_- &= \sqrt{nh}(R'K_-R)^{-1}R'K_-, \qquad P_+ = \sqrt{nh}(R'K_+R)^{-1}R'K_+, \\
\Sigma_- &= \operatorname{Var}((Y_i(0), Z_i^*(0)')|X_i = 0), \qquad \Sigma_+ = \operatorname{Var}((Y_i(1), Z_i^*(1)')|X_i = 0).
\end{aligned}
$$

By estimating the unknown components, the pre-asymptotic bias and variance can be estimated as

$$
\begin{aligned}
\bar{\mathcal{B}} &= \frac{1}{2}e_1'(R'K_-R)^{-1}(R'K_-Q_-)(\bar{q}'\bar{\mu}_-^{(2)}) + \frac{1}{2}e_1'(R'K_+R)^{-1}(R'K_+Q_+)(\bar{q}'\bar{\mu}_+^{(2)}), \\
\bar{\mathcal{V}} &= (\bar{q}' \otimes e_1'P_-)\bar{\Sigma}_-(\bar{q} \otimes P_-e_1) + (\bar{q}' \otimes e_1'P_+)\bar{\Sigma}_+(\bar{q} \otimes P_+e_1),
\end{aligned}
$$

respectively, where $\bar{q}' = (1, -\bar{\gamma}')'$, $\bar{\gamma}$ is the estimated coefficients of $Z_{\bar{S},i}$ in (5), $\bar{\mu}_-^{(2)}$ and $\bar{\mu}_+^{(2)}$ are local polynomial estimators of $\mu_-^{(2)}$ and $\mu_+^{(2)}$ for the elements corresponding to $Z_{\bar{S},i}$, respectively, and $\bar{\Sigma}_+$ and $\bar{\Sigma}_-$ are conditional variance estimators of $\Sigma_-$ and $\Sigma_+$, respectively, such as the nearest neighborhood or plug-in estimator in Section 7.9 of CCFT's supplement. Based on these estimators, the t-ratio for $\tau$ is obtained as

$$
T_\tau = \sqrt{\frac{nh}{\bar{\mathcal{V}}}}(\bar{\tau} - h^2\bar{\mathcal{B}} - \tau), \tag{11}
$$

which is exactly the same as the t-ratio in Theorem 2 of CCFT but using the selected covariates $Z_{\bar{S},i}$. By extending the theoretical developments in CCFT, we obtain the following result.

**Theorem 2.** *Suppose for all $x$ in a neighborhood of 0 and $t = 0, 1$, the density of $X_i$ is continuous and bounded away from zero, $\mathbb{E}[(Y_i(t), Z_i(t)')|X_i = x]$ is three times continuously differentiable, $\partial^2\mathbb{E}[Z_i(0)|X_i = x]/\partial x^2 = \partial^2\mathbb{E}[Z_i(1)|X_i = x]/\partial x^2$, $\mathbb{E}[Z_i(t)Y_i(t)|X_i = x]$ is continuously differentiable, $\operatorname{Var}((Y_i(t), Z_i(t)')|X_i = x)$ is continuously differentiable and invertible, and $\mathbb{E}[|(Y_i(t), Z_i(t)')|^4|X_i = x]$ is continuous. Finally, assume $(\sqrt{\log p} + \sqrt{nh}h^2)(h^2 + \lambda_n)s^* \to 0$.*

*Then the MSE expansion of the first term in (9) (denoted by $\bar{\tau}_1$) is obtained as*

$$
\operatorname{MSE}(\bar{\tau}_1) = h^4\mathcal{B}^2\{1 + o_p(1)\} + \frac{1}{nh}\mathcal{V}. \tag{12}
$$

*Furthermore, if we additionally assume $\sqrt{\frac{nh^5}{\mathcal{V}}}(\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$ and $\frac{\bar{\mathcal{V}}}{\mathcal{V}} \xrightarrow{p} 1$, then*

$$
T_\tau \xrightarrow{d} N(0, 1). \tag{13}
$$

The results in (12) and (13) are analogous to CCFT's Theorems 1 and 2, respectively. This theorem theoretically supports to employ the bias correction and bandwidth selection methods by CCFT based on the selected covariates $Z_{\bar{S},i}$. See Section 4 below for our practical recommendation. The assumption $\partial^2 \mathbb{E}[Z_i(0)|X_i = x]/\partial x^2 = \partial^2 \mathbb{E}[Z_i(1)|X_i = x]/\partial x^2$ is natural for predetermined covariates but may be relaxed by introducing additional regularity conditions (see, Kreiss and Rothe, 2021). Other assumptions except for the last one are also imposed in CCFT. The assumption $(\sqrt{\log p} + \sqrt{nh}h^2)\lambda_n s^* \to 0$ is used to control the remainder term in (9). Although the conditions $\sqrt{\frac{nh^5}{\mathcal{V}}}(\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$ and $\frac{\bar{\mathcal{V}}}{\mathcal{V}} \xrightarrow{p} 1$ are high level, these are typically satisfied for the bias and variance estimators discussed in CCFT.[3] [4]

It should be noted that the asymptotic variance $\mathcal{V}$ in (10) of the estimator $\bar{\tau}$ (or bias corrected version $\bar{\tau} - h^2\bar{\mathcal{B}}$) takes the same form as CCFT even when the dimension of $Z_{S^*}$ grows as $n$ increases. As investigated in CCFT, we can see that the relative efficiency $\bar{\tau}$ compared to the conventional estimator (say, $\hat{\tau}_{\text{unadjusted}}$) without covariates is

$$\frac{\mathcal{V}}{\mathcal{V}_{\text{unadjusted}}} = \frac{\sum_{t=0}^{1} \text{Var}(Y_i(t) - Z_{S^*,i}(t)'\gamma_Y|X_i = 0)}{\sum_{t=0}^{1} \text{Var}(Y_i(t)|X_i = 0)}, \tag{14}$$

where $\mathcal{V}_{\text{unadjusted}}$ is the asymptotic variance of $\hat{\tau}_{\text{unadjusted}}$. Generally there is no clear ranking for these asymptotic variances. However, in an important special case where

$$\gamma_Y = \text{Var}(Z_i(t)|X_i = 0)^{-1}\mathbb{E}[(Z_i(t) - \mathbb{E}[Z_i(t)|X_i])Y_i(t)|X_i = 0] \quad \text{for } t = 0 \text{ and } 1, \tag{15}$$

the coefficient vector $\gamma_Y$ becomes the best linear approximation for each group and thus $\bar{\tau}$ is asymptotically more efficient than $\hat{\tau}_{\text{unadjusted}}$.

The relative efficiency in (14) is also insightful to illustrate a merit of our Lasso approach compared to CCFT. Let $Z = [Z_{S^*} : Z_{S^{*c}}]$ and $Z_C$ be covariates employed to apply the CCFT estimator $\hat{\tau}_{\text{CCFT}}$ (without Lasso covariates selection). Consider the special case in (15). As far as $Z_C$ contains $Z_{S^*}$, $\bar{\tau}$ and $\hat{\tau}_{\text{CCFT}}$ achieve the same asymptotic efficiency $\mathcal{V}$. On the other hand, if $Z_C$ does not contain some elements of $Z_{S^*}$, then under (15), the CCFT estimator $\hat{\tau}_{\text{CCFT}}$ is asymptotically less efficient than the post-Lasso estimator $\bar{\tau}$. Therefore, when researchers are less certain whether all elements of $Z_{S^*}$ are included in $Z_C$ typically due to too many candidates

---

[3] Under the assumption $\partial^2 \mathbb{E}[Z_i(0)|X_i = x]/\partial x^2 = \partial^2 \mathbb{E}[Z_i(1)|X_i = x]/\partial x^2$, the components $q'\mu_-^{(2)}$ and $q'\mu_+^{(2)}$ in $\mathcal{B}$ become $\mu_{Y-}^{(2)} = \partial^2 \mathbb{E}[Y_i(0)|X_i = x]/\partial x^2\big|_{x=0}$ and $\mu_{Y+}^{(2)} = \partial^2 \mathbb{E}[Y_i(1)|X_i = x]/\partial x^2\big|_{x=0}$, respectively. Thus, in this case, the convergence rates of the conventional local polynomial estimators (see, e.g., Fan and Gijbels, 1992, and Ruppert and Wand, 1994) for $\mu_{Y-}^{(2)}$ and $\mu_{Y+}^{(2)}$ guarantee $\sqrt{\frac{nh^5}{\mathcal{V}}}(\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$.

[4] Since $\left|\frac{\bar{\mathcal{V}}}{\mathcal{V}} - 1\right| \leq \frac{||\bar{\Sigma}_- - \Sigma_-||_F}{\lambda_{\min}(\Sigma_-)} + \frac{||\bar{\Sigma}_+ - \Sigma_+||_F}{\lambda_{\min}(\Sigma_+)}$ for the Frobenius norm $||\cdot||_F$, the convergence rates of elements of $\bar{\Sigma}_- - \Sigma_-$ and $\bar{\Sigma}_+ - \Sigma_+$ (see, e.g., Section S.2.4 in the supplement of Calonico, Cattaneo and Titiunik, 2014) combined with assumptions for $\lambda_{\min}(\Sigma_-)$ and $\lambda_{\min}(\Sigma_+)$ can guarantee $\left|\frac{\bar{\mathcal{V}}}{\mathcal{V}} - 1\right| \xrightarrow{p} 0$.

in $Z$ or too small effective sample sizes used for estimation, our Lasso-based approach may be more attractive to achieve asymptotic efficiency $\mathcal{V}$ in more broader situations.[5] [6]

Although Theorem 2 is our main result on inference of the causal effect $\tau$, it is also possible to derive the consistency of the selection procedures $\hat{S}$ and $\tilde{S}$ under the additional $\beta$-min type condition, which is presented in the following theorem.

**Theorem 3.** *Suppose* $\sqrt{\log p/(nh)} = o(\lambda_n)$. *Under Assumption 1 and additionally* $|\theta^{*(j)}| > \lambda_n \varrho_n s^*(1+\varepsilon)$ *for each* $j \in S^*$ *for some* $\varepsilon > 0$*, it holds*

$$\mathbb{P}\{\hat{S} = S^*\} \to 1, \qquad \mathbb{P}\{\tilde{S} = S^*\} \to 1. \tag{16}$$

*(For the second statement,* $S^*$ *should contain* $\theta_1$*.)*

## 3. Discussion

3.1. **Fuzzy RDD.** Although the discussion so far focuses on the sharp RDD analysis, it is possible to extend our approach to the fuzzy RDD analysis, where the forcing variable $X_i$ is not informative enough to determine the treatment $W_i$ but still affects the treatment probability. In particular, the fuzzy RDD assumes that the conditional treatment probability $\mathbb{P}\{W_i = 1|X_i = x\}$ jumps at the cutoff point $\bar{x}$. As in the last section, we normalize $\bar{x} = 0$. To define a reasonable parameter of interest for the fuzzy case, let $W_i(x)$ be a potential treatment for unit $i$ when the cutoff level for the treatment was set at $x$, and assume that $W_i(x)$ is non-increasing in $x$ at $x = 0$. Using the terminology of Angrist, Imbens and Rubin (1996), unit $i$ is called a complier if her cutoff level is $X_i$ (i.e., $\lim_{x\downarrow X_i} W_i(x) = 0$ and $\lim_{x\uparrow X_i} W_i(x) = 1$). A parameter of interest in the fuzzy RDD, suggested by Hahn, Todd and van der Klaauw (2001), is the average causal effect for compliers at $x = 0$,

$$\tau_f = \mathbb{E}[Y_i(1) - Y_i(0)|i \text{ is complier}, X_i = 0].$$

Hahn, Todd and van der Klaauw (2001) showed that under mild conditions the parameter $\tau_f$ can be identified by the ratio of the jump in the conditional mean of $Y_i$ at $x = 0$ to the jump in the conditional treatment probability at $X_i = 0$, i.e.,

$$\tau_f = \frac{\lim_{x\downarrow 0} \mathbb{E}[Y_i|X_i = x] - \lim_{x\uparrow 0} \mathbb{E}[Y_i|X_i = x]}{\lim_{x\downarrow 0} \mathbb{P}\{W_i = 1|X_i = x\} - \lim_{x\uparrow 0} \mathbb{P}\{W_i = 1|X_i = x\}}. \tag{17}$$

[5] Although there is no gain of using $\bar{\tau}$ instead of $\hat{\tau}_{\mathrm{CCFT}}$ in terms of asymptotic efficiency as far as $Z_C$ contains $Z_{S^*}$, the estimator $\bar{\tau}$ may exhibit better finite sample performance even in such a scenario. To see this point, let $(\hat{\beta}_C', \hat{\gamma}_C')$ be the OLS estimator for the regression of $K^{1/2}Y$ on $K^{1/2}(1, T, X, TX)$ and $K^{1/2}Z_C$ so that $\hat{\tau}_{\mathrm{CCFT}}$ is the second element of $\hat{\beta}_C$. As can be seen from CCFT's supplement, one of the remainder terms of $\sqrt{\frac{nh}{\mathcal{V}}}(\hat{\tau}_{\mathrm{CCFT}} - \tau)$ involves a linear combination of the estimation error $\hat{\gamma}_C - (\gamma^{*\prime}, 0')'$. Since the $\ell_2$-convergence rate of $\hat{\gamma}_C - (\gamma^{*\prime}, 0')'$ is typically of order $\sqrt{\dim Z_C/n}$, this remainder term is of larger order than $\hat{\gamma}^* - \gamma_Y$, where $(\hat{\beta}^{*\prime}, \hat{\gamma}^{*\prime})$ is the OLS estimator for the regression of $K^{1/2}Y$ on $K^{1/2}(1, T, X, TX)$ and $K^{1/2}Z_{S^*}$. Even though these remainder terms do not appear in the first order asymptotic distribution, they contribute to the finite sample behaviors of $\bar{\tau}$ and $\hat{\tau}_{\mathrm{CCFT}}$.

[6] Another finite sample issue we encounter in our simulation study below is that as the dimension of $Z_C$ increases, the values of the MSE-optimal bandwidth tend to be smaller (due to larger bias estimates but relatively stable variance estimates). Thus the effective sample size used for the RD estimation tends to be smaller so that we observe larger variations in the resulting RD estimates across simulation draws.

In this case, letting $T_i = \mathbb{I}\{X_i \geq 0\}$, the numerator and denominator of (17) can be estimated by the local Lasso estimators $\hat{\theta}_Y$ and $\hat{\theta}_W$, which solve

$$\min_{\theta_Y} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{Y_i - \alpha_Y - T_i \tau_Y - X_i \beta_{Y-} - T_i X_i \beta_{Y+} - Z_i' \gamma_Y\right\}^2 + \lambda_n |\theta_Y|_1,$$

$$\min_{\theta_W} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{W_i - \alpha_W - T_i \tau_W - X_i \beta_{W-} - T_i X_i \beta_{W+} - Z_i' \gamma_W\right\}^2 + \lambda_n |\theta_W|_1,$$

respectively. Then, based on the union of the selected covariates from the above local Lasso regressions (i.e., $\bar{S} = \{j : |\hat{\theta}_Y^{(j)}| > 0, \text{ or } |\hat{\theta}_W^{(j)}| > 0\}$), we implement the local least squares as in CCFT:

$$\min_{\theta_{Y,\bar{S}}} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{Y_i - G'_{\bar{S},i} \theta_{Y,\bar{S}}\right\}^2, \qquad \min_{\theta_{W,\bar{S}}} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{W_i - G'_{\bar{S},i} \theta_{W,\bar{S}}\right\}^2,$$

where $G_{\bar{S}} = (1, T, X, TX, Z'_{\bar{S}})'$. The numerator and denominator of (17) are given by the estimated coefficients of $\tau_Y$ and $\tau_W$ for the above minimizations, respectively. If the treatment variable $W$ satisfies analogous conditions for Theorems 1 and 3 (by replacing $W$ with $Y$), we expect that analogous results to the sharp RDD case can be established.

3.2. **Regression kink design.** Our high-dimensional method can be extended for the RKDs. For each unit $i = 1, \ldots, n$, we observe continuous outcome and explanatory variables denoted by $Y_i$ and $X_i$, respectively. The RKD analysis is concerned with the following nonseparable model

$$Y = f(B, X, U),$$

where $U$ is an error term (possibly multivariate) and $B = b(X)$ is a continuous policy variable of interest with known $b(\cdot)$. In general, even though we know the function $b(\cdot)$, we are not able to identify the treatment effect by the policy variable $B$. However, it is often the case that the policy function $b(\cdot)$ has some kinks (but is continuous). For instance, suppose $Y$ is duration of unemployment and $X$ is earnings before losing the job. We are interested in the effect of unemployment benefits $B = b(X)$. In many unemployment insurance systems (e.g., the one in Austria), $b(\cdot)$ is specified by a piecewise linear function. In such a scenario, one may exploit changes of slopes in the conditional mean $\mathbb{E}[Y|X = x]$ to identify a treatment effect of $B$. Suppose $b(\cdot)$ is kinked at 0. Otherwise, we redefine $X$ by subtracting the kink point $c$ from $X$. In particular, Card, *et al.* (2015) have shown that a treatment on treated parameter $\tau_k = \int \frac{\partial f(b,x,u)}{\partial b} dF_{U|B=b,X=x}(u)$ is identified as

$$\tau_k = \frac{\lim_{x \downarrow 0} \frac{d}{dx} \mathbb{E}[Y|X = x] - \lim_{x \uparrow 0} \frac{d}{dx} \mathbb{E}[Y|X = x]}{\lim_{x \downarrow 0} \frac{d}{dx} b(x) - \lim_{x \uparrow 0} \frac{d}{dx} b(x)}. \tag{18}$$

To estimate $\tau_k$, we propose the following local lasso regression

$$\min_{\theta} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{Y_i - \alpha - T_i X_i \delta - X_i \beta - T_i X_i^2 \zeta - X_i^2 \eta - Z_i' \gamma\right\}^2 + \lambda_n |\theta|_1, \tag{19}$$

where $\theta = (\alpha, \delta, \beta, \zeta, \eta, \gamma')'$ is a vector of parameters. Let $\hat{\delta}$ be the lasso estimator of $\delta$ by (19). Since the denominator $b_0 = \lim_{x \downarrow c} \frac{d}{dx} b(x) - \lim_{x \uparrow c} \frac{d}{dx} b(x)$ in (18) is assumed to be known, the estimator of $\tau_k$ is given by $\hat{\tau}_k = \hat{\delta}/b_0$. Under analogous conditions to Theorems 1 and 3 (by setting $\beta_+ = 0$ in Assumption 1), we expect that analogous results to the sharp RDD case can be established.

## 4. Recommendation for Implementation

We recommend the following steps to implement our method in R. The most important difference from CCFT is a covariate selection step (Step 2). This procedure is employed in our numerical studies in the following sections.

(1) Without using data on covariates, obtain the MSE-optimal bandwidth $\tilde{h}$ developed by CCT. This can be implemented by an R package, `Rdrobust` (Calonico, Cattaneo and Titiunik, 2015b).
(2) By using the data on covariates, implement the Lasso estimation in (4) by setting $h = \tilde{h}$ and choosing $\lambda_n$ by the data-driven choice. This can be implemented by an R package, `hdm` (https://cran.r-project.org/web/packages/hdm/index.html) for example.
(3) Based on the selected covariates obtained in Step 2, implement CCFT's RD estimation and inference procedure including the bias correction and bandwidth selection. For this step, we recommend to follow the procedures detailed in CCFT's supplement by using `Rdrobust`.

The MSE-optimal bandwidth with the full set of covariates can be too narrow when the number of covariates is large relative to the number of observations and many of covariates are not relevant. Given that we do not know the exact identities of $S^*$ a priori, we instead employ the MSE-optimal bandwidth $\tilde{h}$ by CCT in Step 1, which does not include any covariate. This yields more robust results as shown in the simulation in the next section.

## 5. Simulation

In this section, we conduct simulation experiments to investigate finite sample properties of our covariate selection approach for estimation and inference on the sharp RDD analysis. We consider three simulation designs based on CCFT with introducing additional covariates. The additional covariates are generated based on the simulation designs in Belloni, Chernozhukov and Hansen (2014). Let $\mathcal{B}(a, b)$ be a beta distribution with parameters $a$ and $b$. The data generating process (DGP) is specified as follows

$$Y = \mu_1(X) + \mu_2(Z) + \mu_3(W) + \varepsilon_y, \quad X \sim 2\mathcal{B}(2,4) - 1, \quad Z = \mu_z(X) + \varepsilon_z,$$

$$\mu_z(x) = \begin{cases} 0.49 + 1.06x + 5.74x^2 + 17.14x^3 + 19.75x^5 + 7.47x^5 & \text{for } x < 0, \\ 0.49 + 0.61x + 0.23x^2 - 3.46x^3 + 6.43x^4 - 3.48x^5 & \text{for } x \geq 0, \end{cases}$$

$W = (W_1, W_2, \ldots, W_p)' \sim N(0, \Sigma_W)$ with $\mathbb{E}(W_h^2) = 1$ and $Cov(W_h, W_l) = 0.5^{|h-l|}$, and

$$\begin{pmatrix} \varepsilon_y \\ \varepsilon_z \end{pmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_z \\ \rho\sigma_y\sigma_z & \sigma_z^2 \end{pmatrix},$$

with $\sigma_y = 0.1295$ and $\sigma_z = 0.1353$.

For the functions $\mu_1$, $\mu_2$, and $\mu_3$, we consider three cases. For DGP1, we set $\rho = 0.2692$, all the coefficients of $\mu_2(z)$ and $\mu_3(w)$ to be zero, and

$$\mu_1(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{for } x < 0, \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{for } x \geq 0, \end{cases}$$

For DGP2, we set $\rho = 0.2692$,

$$\mu_1(x) = \begin{cases} 0.36 + 0.96x + 5.47x^2 + 15.28x^3 + 15.87x^4 + 5.14x^5 & \text{for } x < 0, \\ 0.38 + 0.62x - 2.84x^2 + 8.42x^3 - 10.24x^4 + 4.31x^5 & \text{for } x \geq 0, \end{cases}$$

$$\mu_2(z) = \begin{cases} 0.22z & \text{for } x < 0, \\ 0.28z & \text{for } x \geq 0, \end{cases}$$

and $\mu_3(w) = \sum_{h=1}^{p} \pi_h w_h$ with $\pi_h = 0.2^h$. For DGP3, we set $\mu_1(x)$ and $\mu_2(z)$ as in DGP2, and $\mu_3(w) = \sum_{h=1}^{p} \pi_h w_h$ with $\pi_h = 0.5^h$. The sample size is set as $n = 500$ for all cases. The number of the covariates $p$ varies from 5 to 500. The results are based on 1,000 Monte Carlo replications.

Table 1 shows the biases and RMSEs of the four point estimation methods. For the bandwidth $h$, the first two methods use the MSE-optimal bandwidth without covariates proposed by CCT. The third method uses the MSE-optimal bandwidth with covariates proposed by CCFT. The fourth method, called "Adaptive", is the bandwidth for our covariate selection approach which uses the MSE optimal bandwidth without covariates for the covariate selection stage and uses that with covariates in the estimation stage. For estimation methods, the first method uses the standard RD estimation method without covariates by CCT. The second and the third methods use the RD estimation with covariates by CCFT. The fourth method uses the RD estimation with the selected covariates. See Section 4 for more details on the fourth method.

Our findings are summarized as follows. First, the RMSEs of the covariate-adjusted estimation get larger irrespective of the bandwidths as the number of covariates increases across all DPGs. These increases in the RMSEs are due to inflated standard errors caused by a large number of covariates. This result clearly indicates the need for covariate selection. Second, the covariate selection approach shows excellent performances for all cases. Both the biases and RMSEs are stable for different values of $p$ for all designs. Finally, all methods work equally well for DGP1, where all the additional covariates are irrelevant. However, for DGP2 and DGP3, we find substantial efficiency loss of the standard method. Overall, we recommend the covariate selection even for relatively small $p$.

Table 2 reports the number of selected covariates for our covariate selection approach. It is quite natural that the number of selected covariates increases when the number of non-zero coefficients of covariates increases. It is interesting to note that the average number of selected covariates decreases as the number of covariates increase.

Table 3 shows the coverage probabilities and interval lengths of the robust confidence intervals for the causal effect. The nominal coverage level is 0.95. The following points are notable. First, the performances of our covariate selection approach are stable for all DGPs, although the coverage probabilities tend to be a little bit smaller than the nominal level. Second, for the covariate-adjusted approaches, the coverage probabilities decrease and the interval length gets shorter as $p$ increases. Third, the coverage of CCT is more stable and better than other methods, especially for DGP2 and DGP3. However, the average lengths of CCT are substantially longer than the other methods. Overall, the covariate selection approach is promising for inference as well since it exhibits robust performances in both coverages and lengths for different number of covariates and DGPs.

Finally, Table 4 presents the properties of the MSE-optimal bandwidths. We observe that the MSE-optimal bandwidth without covariates and the adaptive one are very stable while the MSE optimal bandwidth with covariates shrinks as $p$ increases. This is possibly the main source of the increased RMSE and under-coverages.

Table 1. Simulation: Point estimation

| MSE-Optimal bandwidths: | | w/o Covariates | | | | w/ Covariates | | Adaptive | |
|---|---|---|---|---|---|---|---|---|---|
| Estimation methods: | | Standard | | Covariate adjusted | | Covariate adjusted | | Covariate selection | |
| | $p$ | Bias | RMSE | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| DGP1 | 5 | 0.019 | 0.063 | 0.023 | 0.064 | 0.019 | 0.065 | 0.019 | 0.064 |
| | 10 | 0.020 | 0.063 | 0.022 | 0.068 | 0.019 | 0.069 | 0.019 | 0.061 |
| | 20 | 0.020 | 0.061 | 0.019 | 0.066 | 0.014 | 0.073 | 0.016 | 0.064 |
| | 30 | 0.018 | 0.063 | 0.022 | 0.074 | 0.011 | 0.093 | 0.020 | 0.062 |
| | 40 | 0.015 | 0.064 | 0.022 | 0.080 | 0.004 | 0.141 | 0.019 | 0.064 |
| | 50 | 0.022 | 0.065 | 0.027 | 0.101 | 0.012 | 0.397 | 0.019 | 0.060 |
| | 100 | 0.019 | 0.062 | 0.038 | 0.387 | 0.000 | 0.130 | 0.023 | 0.065 |
| | 250 | 0.023 | 0.064 | 0.026 | 0.082 | 0.012 | 0.108 | 0.019 | 0.064 |
| | 500 | 0.022 | 0.065 | 0.028 | 0.072 | 0.012 | 0.096 | 0.022 | 0.059 |
| DGP2 | 5 | 0.004 | 0.526 | 0.000 | 0.058 | 0.030 | 0.082 | 0.029 | 0.095 |
| | 10 | -0.007 | 0.527 | -0.001 | 0.061 | 0.027 | 0.087 | 0.028 | 0.100 |
| | 20 | 0.032 | 0.507 | -0.007 | 0.064 | 0.031 | 0.098 | 0.025 | 0.097 |
| | 30 | 0.017 | 0.529 | -0.006 | 0.070 | 0.026 | 0.194 | 0.026 | 0.101 |
| | 40 | 0.008 | 0.532 | -0.012 | 0.080 | 0.032 | 0.223 | 0.027 | 0.100 |
| | 50 | 0.001 | 0.519 | -0.009 | 0.087 | 0.011 | 0.420 | 0.026 | 0.101 |
| | 100 | 0.021 | 0.516 | -0.032 | 0.462 | 0.047 | 0.509 | 0.025 | 0.107 |
| | 250 | 0.003 | 0.542 | -0.042 | 0.310 | -0.114 | 4.310 | 0.024 | 0.109 |
| | 500 | 0.021 | 0.543 | -0.044 | 0.402 | 0.021 | 0.806 | 0.020 | 0.114 |
| DGP3 | 5 | -0.005 | 0.673 | -0.004 | 0.058 | 0.030 | 0.082 | 0.027 | 0.109 |
| | 10 | -0.012 | 0.689 | -0.005 | 0.060 | 0.027 | 0.087 | 0.020 | 0.177 |
| | 20 | 0.033 | 0.672 | -0.011 | 0.064 | 0.031 | 0.098 | 0.024 | 0.178 |
| | 30 | 0.015 | 0.701 | -0.011 | 0.070 | 0.026 | 0.194 | 0.015 | 0.185 |
| | 40 | -0.005 | 0.674 | -0.017 | 0.080 | 0.032 | 0.222 | 0.015 | 0.185 |
| | 50 | -0.004 | 0.695 | -0.015 | 0.085 | -0.005 | 0.460 | 0.027 | 0.195 |
| | 100 | -0.007 | 0.689 | -0.036 | 0.301 | 0.009 | 0.621 | 0.019 | 0.190 |
| | 250 | -0.013 | 0.689 | -0.055 | 0.347 | 0.009 | 0.861 | 0.012 | 0.201 |
| | 500 | 0.031 | 0.663 | -0.050 | 0.485 | 0.095 | 1.030 | 0.014 | 0.216 |

Table 2. Simulation: Number of selected covariates

| | $p$ | Average | Min | Max |
|---|---|---|---|---|
| DGP1 | 5 | 0.409 | 0 | 1 |
| | 10 | 0.371 | 0 | 1 |
| | 20 | 0.379 | 0 | 1 |
| | 30 | 0.377 | 0 | 1 |
| | 40 | 0.347 | 0 | 2 |
| | 50 | 0.338 | 0 | 1 |
| | 100 | 0.342 | 0 | 2 |
| | 250 | 0.322 | 0 | 1 |
| | 500 | 0.310 | 0 | 1 |
| DGP2 | 5 | 2.869 | 2 | 3 |
| | 10 | 2.789 | 2 | 3 |
| | 20 | 2.717 | 1 | 3 |
| | 30 | 2.647 | 1 | 3 |
| | 40 | 2.573 | 1 | 3 |
| | 50 | 2.564 | 1 | 3 |
| | 100 | 2.461 | 1 | 3 |
| | 250 | 2.392 | 1 | 3 |
| | 500 | 2.325 | 1 | 3 |
| DGP3 | 5 | 3.915 | 2 | 5 |
| | 10 | 3.080 | 2 | 5 |
| | 20 | 3.013 | 2 | 4 |
| | 30 | 2.980 | 2 | 5 |
| | 40 | 2.944 | 1 | 4 |
| | 50 | 2.940 | 1 | 5 |
| | 100 | 2.866 | 1 | 5 |
| | 250 | 2.725 | 1 | 4 |
| | 500 | 2.548 | 0 | 4 |

Table 3. Simulation: Inference

| MSE-Optimal bandwidths: | | w/o Covariates | | | | w/ Covariates | | Adaptive | |
|---|---|---|---|---|---|---|---|---|---|
| Estimation methods: | | Standard | | Covariate adjusted | | Covariate adjusted | | Covariate selection | |
| | $p$ | CP | Length | CP | Length | CP | Length | CP | Length |
| DGP1 | 5 | 0.923 | 0.205 | 0.898 | 0.246 | 0.892 | 0.241 | 0.915 | 0.231 |
| | 10 | 0.910 | 0.226 | 0.855 | 0.211 | 0.851 | 0.219 | 0.930 | 0.188 |
| | 20 | 0.930 | 0.172 | 0.815 | 0.259 | 0.804 | 0.173 | 0.911 | 0.190 |
| | 30 | 0.915 | 0.215 | 0.734 | 0.148 | 0.644 | 0.165 | 0.911 | 0.226 |
| | 40 | 0.911 | 0.214 | 0.649 | 0.182 | 0.458 | 0.168 | 0.923 | 0.335 |
| | 50 | 0.898 | 0.309 | 0.540 | 0.152 | 0.258 | 0.074 | 0.919 | 0.223 |
| | 100 | 0.918 | 0.218 | 0.194 | 0.342 | 0.193 | 0.176 | 0.896 | 0.161 |
| | 250 | 0.908 | 0.344 | 0.108 | 0.040 | 0.172 | 0.068 | 0.906 | 0.225 |
| | 500 | 0.912 | 0.195 | 0.138 | 0.020 | 0.188 | 0.097 | 0.931 | 0.262 |
| DGP2 | 5 | 0.924 | 2.472 | 0.661 | 0.232 | 0.852 | 0.278 | 0.842 | 0.548 |
| | 10 | 0.925 | 2.449 | 0.627 | 0.209 | 0.786 | 0.234 | 0.845 | 0.466 |
| | 20 | 0.929 | 2.456 | 0.591 | 0.261 | 0.653 | 0.167 | 0.872 | 0.490 |
| | 30 | 0.928 | 2.305 | 0.517 | 0.136 | 0.462 | 0.178 | 0.880 | 0.465 |
| | 40 | 0.926 | 2.006 | 0.472 | 0.170 | 0.286 | 0.128 | 0.890 | 0.438 |
| | 50 | 0.927 | 2.276 | 0.407 | 0.127 | 0.191 | 0.762 | 0.876 | 0.548 |
| | 100 | 0.933 | 2.563 | 0.156 | 0.069 | 0.172 | 0.754 | 0.883 | 0.284 |
| | 250 | 0.933 | 1.647 | 0.135 | 0.103 | 0.177 | 0.195 | 0.890 | 0.640 |
| | 500 | 0.919 | 1.820 | 0.137 | 0.112 | 0.174 | 0.431 | 0.900 | 0.509 |
| DGP3 | 5 | 0.934 | 3.350 | 0.615 | 0.219 | 0.852 | 0.278 | 0.861 | 0.329 |
| | 10 | 0.911 | 2.965 | 0.601 | 0.202 | 0.786 | 0.234 | 0.891 | 0.704 |
| | 20 | 0.925 | 3.395 | 0.564 | 0.236 | 0.653 | 0.167 | 0.905 | 0.721 |
| | 30 | 0.920 | 2.951 | 0.495 | 0.151 | 0.462 | 0.178 | 0.899 | 0.713 |
| | 40 | 0.940 | 2.848 | 0.456 | 0.174 | 0.288 | 0.128 | 0.903 | 0.740 |
| | 50 | 0.928 | 2.654 | 0.397 | 0.132 | 0.188 | 0.071 | 0.885 | 0.697 |
| | 100 | 0.933 | 2.954 | 0.182 | 0.073 | 0.169 | 0.407 | 0.918 | 0.910 |
| | 250 | 0.927 | 3.446 | 0.119 | 0.139 | 0.176 | 0.529 | 0.895 | 0.765 |
| | 500 | 0.939 | 2.820 | 0.130 | 0.158 | 0.173 | 0.389 | 0.908 | 0.751 |

Table 4. Simulation: MSE-Optimal Bandwidths

| MSE-Optimal bandwidths: | | w/o Covariates | | w/ Covariates | | Adaptive | |
|---|---|---|---|---|---|---|---|
| | $p$ | Mean | SD | Mean | SD | Mean | SD |
| DGP1 | 5 | 0.194 | 0.044 | 0.187 | 0.042 | 0.194 | 0.046 |
| | 10 | 0.195 | 0.043 | 0.181 | 0.041 | 0.195 | 0.045 |
| | 20 | 0.195 | 0.045 | 0.162 | 0.037 | 0.196 | 0.045 |
| | 30 | 0.196 | 0.045 | 0.146 | 0.031 | 0.196 | 0.045 |
| | 40 | 0.195 | 0.045 | 0.127 | 0.026 | 0.196 | 0.046 |
| | 50 | 0.195 | 0.045 | 0.108 | 0.024 | 0.196 | 0.045 |
| | 100 | 0.198 | 0.044 | 0.073 | 0.023 | 0.197 | 0.047 |
| | 250 | 0.199 | 0.045 | 0.070 | 0.020 | 0.195 | 0.046 |
| | 500 | 0.197 | 0.045 | 0.070 | 0.020 | 0.198 | 0.044 |
| DGP2 | 5 | 0.176 | 0.025 | 0.108 | 0.011 | 0.111 | 0.014 |
| | 10 | 0.176 | 0.025 | 0.107 | 0.011 | 0.112 | 0.014 |
| | 20 | 0.176 | 0.025 | 0.104 | 0.011 | 0.114 | 0.015 |
| | 30 | 0.176 | 0.024 | 0.102 | 0.012 | 0.116 | 0.015 |
| | 40 | 0.176 | 0.026 | 0.098 | 0.012 | 0.117 | 0.015 |
| | 50 | 0.176 | 0.024 | 0.092 | 0.013 | 0.117 | 0.015 |
| | 100 | 0.177 | 0.025 | 0.079 | 0.020 | 0.119 | 0.015 |
| | 250 | 0.176 | 0.025 | 0.077 | 0.022 | 0.120 | 0.015 |
| | 500 | 0.175 | 0.024 | 0.076 | 0.021 | 0.121 | 0.015 |
| DGP3 | 5 | 0.182 | 0.028 | 0.108 | 0.011 | 0.117 | 0.014 |
| | 10 | 0.182 | 0.028 | 0.107 | 0.011 | 0.138 | 0.017 |
| | 20 | 0.182 | 0.028 | 0.104 | 0.011 | 0.139 | 0.016 |
| | 30 | 0.181 | 0.028 | 0.102 | 0.012 | 0.139 | 0.015 |
| | 40 | 0.182 | 0.029 | 0.098 | 0.012 | 0.140 | 0.016 |
| | 50 | 0.183 | 0.028 | 0.093 | 0.014 | 0.140 | 0.016 |
| | 100 | 0.183 | 0.029 | 0.078 | 0.020 | 0.141 | 0.016 |
| | 250 | 0.182 | 0.029 | 0.076 | 0.022 | 0.142 | 0.017 |
| | 500 | 0.182 | 0.028 | 0.074 | 0.021 | 0.144 | 0.019 |

## 6. Empirical illustration: Head Start data

To illustrate our variable selection approach, we revisit the problem of the Head Start program first studied by Ludwig and Miller (2007) where they investigate the effect of the Head Start program on various outcomes related to health and schooling. The federal government provided grant-writing assistance to the 300 poorest counties based on the poverty index to apply for the Head Start program. This leads to the RD design with the poverty index as a running variable where the cut-off value is set as $\bar{x} = 59.1984$. Ludwig and Miller (2007) conducted their RDD analysis using no covariate, and CCFT examined the impact of the covariance-adjustment. CCFT employed nine pre-intervention covariates from the U.S. Census, which include total population, percentage of population, percentages of black and urban population, and levels and percentages of population in three age groups (children aged 3 to 5, children aged 14 to 17, and adults older than 25). The main finding by CCFT is that the covariate-adjusted RD inference yields shorter confidence intervals while the RD point estimates remain stable.

An important aspect of the Head start example is that it is unclear which covariates become useful to improve efficiency mainly due to the lack of economic theories behind the problem. We conduct the empirical exercises of CCFT by applying our variable selection approach with two extensions. First, we introduce 36 interaction terms in addition to the nine original covariates. Second, we also implement those estimation and inference for subsamples to see the effect of changes in the ratio of the number of covariates ($p$) to that of observations ($n$). Hereafter, as in CCFT, we focus on child mortality among many outcome variables.

Table 6 shows the results of our empirical illustration. Four columns correspond to four estimation procedures which are the same as those used in the simulation experiments. The first panel shows the full sample results ($n = 2799$ and $p/n = 0.016$). The RD causal effect estimates are presented in the first row. The next three rows show 95% confidence intervals, their percentage length changes relative to the one in the first column, and their associated $p$-values where these are obtained without restriction on the MSE optimal bandwidth for the local linear regression ($h$) and the pilot bandwidth ($b$). See CCT and CCFT for more details on the robust inference methods. These results are also obtained under the restriction $h/b = 1$, which are reported in the following three rows. The next two rows in the same panel present the bandwidths $(h, b)$, and effective sample sizes $(n_-, n_+)$ used for the RD estimates. The effective sample sizes are the number of observations with the running variable in the intervals $[\bar{x} - h, \bar{x}]$ and $[\bar{x}, \bar{x} + h]$. We also report the selected covariates for our covariate selection approach. We use subsamples of the first 1000 and 500 observations for the second and third panels, leading to $p/n = 0.045$ and .0.090, respectively.

For the full sample case, the covariate-adjusted estimates mildly deviate from the standard one while our estimate based on the variable selection is identical to the standard one. Although the confidence intervals of the covariate-adjusted approaches are shorter than the standard one, this might induce under-coverages for the case of many covariates as illustrated in the simulation experiment. As the sample size gets smaller, the observations made here are amplified. In contrast, we can see the stable performance of the variable selection approach and its mild contribution to shorten the confidence intervals.

Table 5. Empirical illustration: Head Start data

| | MSE-Optimal bandwidths: | w/o Covariates | | w/ Covariates | Adaptive |
|---|---|---|---|---|---|
| | Estimation methods: | Standard | Cov-adjusted | Cov-adjusted | Variable selection |
| $n = 2779$ | Point estimate | $-2.41$ | $-2.19$ | $-3.14$ | $-2.41$ |
| $p/n = 0.016$ | $h/b$ unrestricted | | | | |
| | Robust 95% CI | $[-5.46, -0.1]$ | $[-4.7, -0.27]$ | $[-5.59, -0.56]$ | $[-5.46, -0.1]$ |
| | CI length change (%) | | $-17.42$ | $-6.18$ | 0 |
| | Robust p-value | 0.042 | 0.028 | 0.017 | 0.042 |
| | $h/b = 1$ | | | | |
| | Robust 95% CI | $[-6.41, -1.09]$ | $[-5.75, -1.07]$ | $[-6.37, -0.31]$ | $[-6.41, -1.09]$ |
| | CI length change (%) | | $-12.05$ | $-13.82$ | 0 |
| | Robust p-value | 0.006 | 0.004 | 0.031 | 0.006 |
| | $h$, $b$ | 6.81, 10.73 | 6.81, 10.73 | 3.26, 6.05 | 6.81, 10.73 |
| | $n_-$, $n_+$ | 234, 180 | 234, 180 | 99, 94 | 234, 180 |
| | Selected covariates | | | | None |
| $n = 1000$ | Point estimate | $-1.68$ | $-3.1$ | $-4.1$ | $-1.48$ |
| $p/n = 0.045$ | $h/b$ unrestricted | | | | |
| | Robust 95% CI | $[-5.45, 1.75]$ | $[-6.35, -1.1]$ | $[-7.7, -2.28]$ | $[-5.08, 1.79]$ |
| | CI length change (%) | | $-26.92$ | $-24.6$ | $-4.49$ |
| | Robust p-value | 0.314 | 0.005 | 0.000 | 0.347 |
| | $h/b = 1$ | | | | |
| | Robust 95% CI | $[-8.26, 0.22]$ | $[-8.27, -1.87]$ | $[-8.74, -1.85]$ | $[-7.94, 0.27]$ |
| | CI length change (%) | | $-24.52$ | $-18.78$ | $-3.28$ |
| | Robust p-value | 0.063 | 0.002 | 0.003 | 0.070 |
| | $h$, $b$ | 6.52, 10.23 | 6.52, 10.23 | 5.26, 8.07 | 5.26, 8.07 |
| | $n_-$, $n_+$ | 74, 77 | 79, 79 | 64, 69 | 79, 79 |
| | Selected covariates | | | | % of adult population |
| $n = 500$ | Point estimate | $-2.35$ | $-4.51$ | $-7.1$ | $-2.22$ |
| $p/n = 0.090$ | $h/b$ unrestricted | | | | |
| | Robust 95% CI | $[-7.25, 2.48]$ | $[-8.59, -2.38]$ | $[-11.57, -5.28]$ | $[-6.93, 2.25]$ |
| | CI length change (%) | | $-36.26$ | $-35.47$ | $-5.74$ |
| | Robust p-value | 0.337 | 0.001 | 0.000 | 0.317 |
| | $h/b = 1$ | | | | |
| | Robust 95% CI | $[-10.22, 1.42]$ | $[-10.67, -3.35]$ | $[-12.33, -4.45]$ | $[-10, 1.07]$ |
| | CI length change (%) | | $-37.1$ | $-32.34$ | $-4.94$ |
| | Robust p-value | 0.139 | 0.000 | 0.000 | 0.110 |
| | $h$, $b$ | 6.37, 9.16 | 6.37, 9.16 | 4.31, 6.82 | 4.31, 6.82 |
| | $n_-$, $n_+$ | 60, 56 | 61, 56 | 42, 42 | 61, 56 |
| | Selected covariates | | | | % of adult population |

Table 6. Head Start Example: Summary of Setups

| Setup | Number of Covariates | Sample Size |
|---|---|---|
| 1 | 9 | 1500 |
| 2 | 9 | 1000 |
| 3 | 9 | 500 |
| 4 | 45 | 1500 |
| 5 | 45 | 1000 |
| 6 | 45 | 500 |

Figure 1. Comparisons of covariate-adjusted approach ($h$: w/o covariates) with standard one for point estimates (left panel) and CI lengths (right panel)

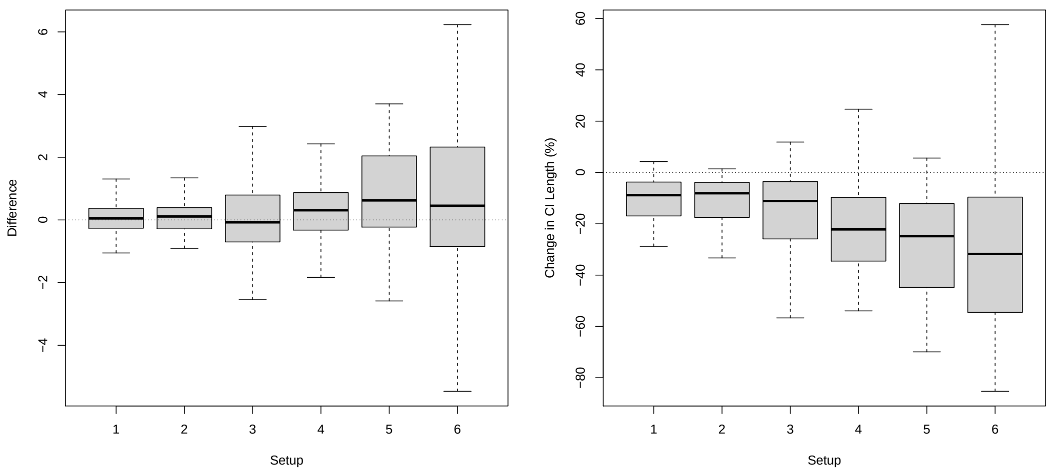



Figure 2. Comparisons of covariate-adjusted approach ($h$: w covariates) with standard one for point estimates (left panel) and CI lengths (right panel)

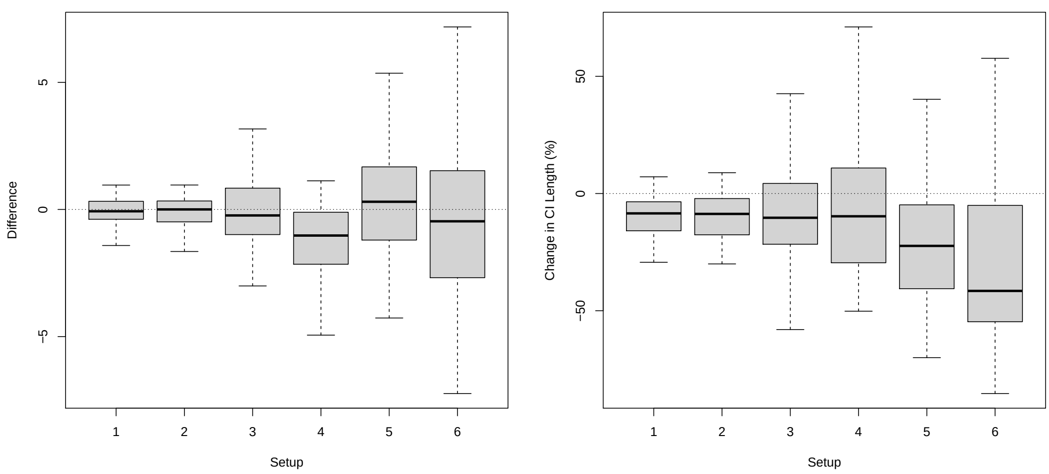
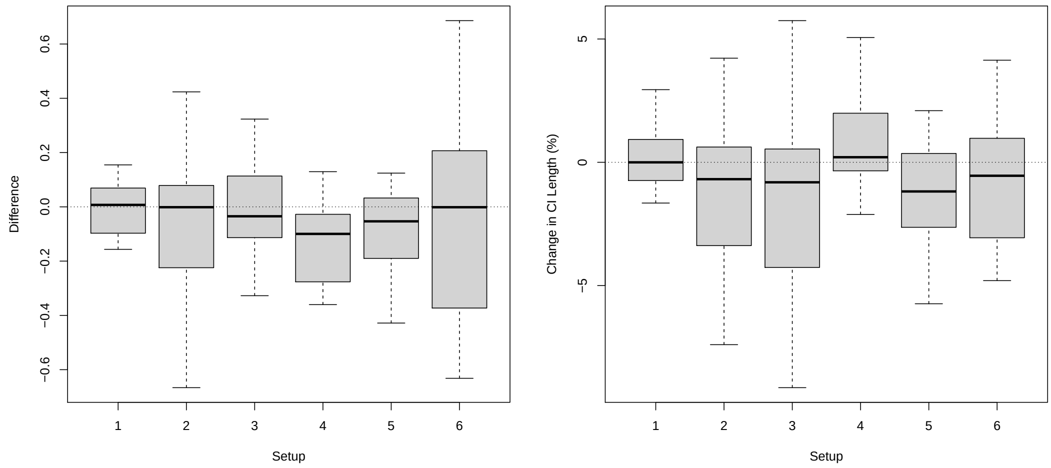


FIGURE 3. Comparisons of variable selection approach ($h$: adaptive) with standard one for point estimates (left panel) and CI lengths (right panel)




To better understand the proposed method in this paper, we look at behaviors of three approaches, two covariance-adjusted and variable selection approaches, through the Head start example where the approaches are the same as those considered in the simulation and Table 6. We construct 50 subsamples of $k$ observations out of the Head start data where $k$ is set to 1500, 1000, and 500. We estimate the RD treatment effects based on the subsamples using nine or forty-five covariates where nine covariates are those considered in CCFT and forty-five are those considered in Table 6. It amounts to consider six setups, and they are summarized in Table 6. Figures 1-3 show boxplots of the results on estimation and inference.[7] The top and bottom of the boxes are the third and first quartiles, and the top- and bottom-bars show the maximum and minimum values less than (the third quartiles + 1.5 times the interquartile range) and greater than (the first quartile - 1.5 times the interquartile range), respectively. The left panels in Figures 1-3 show the differences in the RD treatment effects between the three approaches and the standard one where the standard one is based on CCT. The right panel in each figure shows the corresponding differences in the confidence interval lengths.[8]

Before discussing results, we note that the scales of the vertical axes for the covariate-adjusted approaches in Figures 1 and 2 are about ten times as large as that of the variable selection approach in Figure 3. First, Figures 1 and 2 show the point estimates by the covariate-adjusted approaches are stable when the number of covariates is small, while these deviate from the standard approach by a large extent when the number of covariates is large, especially when the sample size is relatively small. Second, we note in Figure 3 that the point estimates based on the standard and variable selection approaches are very close and they are stable for all setups. Third, we observe that the covariate-adjusted approaches tend to produce large decreases in the

[7]The results shown in Figures 1-3 are based on the bandwidths with $h/b$ unrestricted, and those with the bandwidths with $h/b = 1$ are qualitatively similar.

[8]The variable selection approach produced the identical result to the standard one 23, 23, 23, 18, 14, and 20 times for the setups 1-6, respectively. The boxplots are drawn based on non-identical results.

interval length particularly when the number of covariates is large. Fourth, the variable selection approach leads to reductions of around 5% which are nonnegligible, although the changes are mild compared to the covariate-adjusted approaches. The simulation in Table 3 shows that the interval lengths of the covariate-adjusted approaches tend to be too short and result in under-coverages when the number of covariates is large.

We often encounter situations where a number of potentially useful covariates are available. Furthermore, It is common to employ transformations of covariates such as interaction and quadratic terms. However, we are typically uncertain about which covariates contribute to improving efficiency possibly due to the lack of economic theories. The RD analysis is local in nature, and the number of covariates relative to the effective sample size can be pretty large. We have seen that covariate-adjusted approaches can become misleading in several examples. The steady performance of the variable selection approach under various circumstances is noteworthy, and it can provide an essential second opinion for the standard and covariate-adjusted approaches.

## Appendix A. Mathematical appendix

A.1. **Proof of Theorem 1 .** We use the following modification of Bernstein's inequality.

**Lemma A.1.** *Under Assumption 1, it holds*

$$\mathbb{E}\left[\max_{1\le j\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_iG_{ij}e_i\right|^m\right]\le 2h^{m/2}\log^{m/2}p,$$

*for* $m\le 1+\log p$.

**Proof of Lemma A.1.** Note that $\mathbb{E}[K_iG_ie_i]=0$ by construction. From Bernstein's inequality (e.g., Lemma 14.12 in BG)

$$\mathbb{E}\left[\max_{1\le j\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_iG_{ij}e_i\right|^m\right]\le 2h^{m/2}\log^{m/2}p,$$

for $m\le 1+\log p$. $\square$

A.1.1. *Proof of (i).* Under Assumption 1-2, Lemma A.1 implies

$$\mathbb{E}\left[\max_{1\le j\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_iG_{ij}e_i\right|\right]\le 2h^{1/2}\log^{1/2}p.$$

Thus, we have

$$\mathbb{P}\{\mathcal{A}_n\}:=\mathbb{P}\left\{\frac{4}{nh}\left|\sum_{i=1}^{n}K_iG_ie_i\right|_\infty\le\lambda_n\right\}\to 1, \tag{20}$$

as $n\to\infty$, provided that $\sqrt{\log p/(nh)}=o(\lambda_n)$.

Let $\mathbf{Y}=(Y_1K_1^{1/2},\dots,Y_nK_n^{1/2})'$, $\mathbf{e}=(e_1K_1^{1/2},\dots,e_nK_n^{1/2})'$, and $\mathbf{G}=(G_1K_1^{1/2},\dots,G_nK_n^{1/2})'$. Since $\hat\theta$ is a minimizer, we have

$$\frac{1}{nh}|\mathbf{Y}-\mathbf{G}\hat\theta|_2^2+\lambda_n|\hat\theta|_1\le\frac{1}{nh}|\mathbf{Y}-\mathbf{G}\theta^*|_2^2+\lambda_n|\theta^*|_1.$$

By plugging $\mathbf{Y} = \mathbf{G}\theta^* + \mathbf{e}$ into the above, we obtain

$$\begin{aligned}
\frac{2}{nh}|\mathbf{G}(\hat{\theta}-\theta^*)|_2^2 &\leq \frac{4}{nh}\mathbf{e}'\mathbf{G}(\hat{\theta}-\theta^*) + 2\lambda_n|\theta^*|_1 - 2\lambda_n|\hat{\theta}|_1 \\
&\leq \frac{4}{nh}|\mathbf{e}'\mathbf{G}|_\infty|\hat{\theta}-\theta^*|_1 + 2\lambda_n(|\theta^*|_1 - |\hat{\theta}|_1), \\
&\leq 3\lambda_n|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 - \lambda_n|\hat{\theta}_{S_c^*}|_1, \qquad (21)
\end{aligned}$$

conditionally on $\mathcal{A}_n$, where $S_c^*$ is the complement of $S^*$, the second inequality follows from the Hölder inequality, and the third inequality follows from the definition of $\mathcal{A}_n$ and the following facts

$$\begin{aligned}
|\hat{\theta}-\theta^*|_1 &= |\hat{\theta}_{S_*} - \theta^*_{S_*}|_1 + |\hat{\theta}_{S_c^*}|_1, \qquad (22) \\
|\theta^*|_1 - |\hat{\theta}|_1 &= |\theta^*_{S^*}|_1 - |\hat{\theta}_{S^*}|_1 - |\hat{\theta}_{S_c^*}|_1 \leq |\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 - |\hat{\theta}_{S_c^*}|_1,
\end{aligned}$$

due to the triangle inequality. Thus, (21) implies $3|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 \geq |\hat{\theta}_{S_c^*}|_1$ and

$$\frac{2}{nh}|\mathbf{G}(\hat{\theta}-\theta^*)|_2^2 + \lambda_n|\hat{\theta}-\theta^*|_1 \leq 4\lambda_n|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1, \qquad (23)$$

by using (22).

Now, Assumption 1-3 implies

$$4\lambda_n|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 \leq 4\lambda_n\sqrt{\frac{s^*}{nh\phi_*^2}}|\mathbf{G}(\hat{\theta}-\theta^*)|_2 \leq \frac{1}{nh}|\mathbf{G}(\hat{\theta}-\theta^*)|_2^2 + 4\lambda_n^2\frac{s^*}{\phi_*^2},$$

with probability approaching one, where we note that $2ab \leq a^2 + b^2$ for the second inequality. Combining this with (23), we have

$$\frac{1}{nh}|\mathbf{G}(\hat{\theta}-\theta^*)|_2^2 + \lambda_n|\hat{\theta}-\theta^*|_1 \leq 4\lambda_n^2\frac{s^*}{\phi_*^2},$$

and the conclusion in (6) follows.

A.1.2. *Proof of (ii).* Let $G_i = (G'_{1i}, G'_{2i})'$, where $G_{1i}$ and $G_{2i}$ correspond to covariates for $\theta_1$ and $\theta_2$ in (4), respectively. Define $\mathbf{G}_1 = (G_{11}, \ldots, G_{1n})'$ and $\mathbf{G}_2 = (G_{21}, \ldots, G_{2n})'$. The compatibility condition for $\mathbf{G}_1$ is implied by the usual full column rank condition in the classical linear regression since $|\theta_1|_1^2 \leq \dim(\theta_1)|\theta_1|_2^2$.

Note that the result in (20) still holds. Since $\tilde{\theta}$ is a minimizer, we have

$$\frac{1}{nh}|\mathbf{Y} - \mathbf{G}\tilde{\theta}|_2^2 + \lambda_n|\tilde{\theta}_2|_1 \leq \frac{1}{nh}|\mathbf{Y} - \mathbf{G}\theta^*|_2^2 + \lambda_n|\theta_2^*|_1.$$

By plugging $\mathbf{Y} = \mathbf{G}\theta^* + \mathbf{e}$ into the above,

$$\begin{aligned}
\frac{2}{nh}|\mathbf{G}(\tilde{\theta}-\theta^*)|_2^2 &\leq \frac{4}{nh}\mathbf{e}'\mathbf{G}(\tilde{\theta}-\theta^*) + 2\lambda_n|\theta_2^*|_1 - 2\lambda_n|\tilde{\theta}_2|_1 \\
&\leq \frac{4}{nh}|\mathbf{e}'\mathbf{G}|_\infty|\tilde{\theta}-\theta^*|_1 + 2\lambda_n(|\theta_2^*|_1 - |\tilde{\theta}_2|_1) \\
&\leq \lambda_n|\tilde{\theta}_1 - \theta_1^*|_1 + 3\lambda_n|\tilde{\theta}_{2,S^*} - \theta^*_{2,S^*}|_1 - \lambda_n|\tilde{\theta}_{2,S_c^*}|_1, \qquad (24)
\end{aligned}$$

conditionally on $\mathcal{A}_n$, where the second inequality follows from the Hölder inequality, and the third inequality follows from the definition of $\mathcal{A}_n$ and the following facts

$$\begin{aligned} |\tilde{\theta} - \theta^*|_1 &= |\tilde{\theta}_{S^*} - \theta^*_{S^*}|_1 + |\tilde{\theta}_{S^*_c}|_1, \\ |\theta^*|_1 - |\tilde{\theta}|_1 &= |\theta^*_{S^*}|_1 - |\tilde{\theta}_{S^*}|_1 - |\tilde{\theta}_{S^*_c}|_1 \le |\tilde{\theta}_{S^*} - \theta^*_{S^*}|_1 - |\tilde{\theta}_{S^*_c}|_1, \end{aligned} \tag{25}$$

due to the triangle inequality. Thus, (24) implies $3|\tilde{\theta}_{S_*} - \theta^*_{S_*}|_1 \ge |\tilde{\theta}_{S^c_*}|_1$ and

$$\frac{2}{nh}|\mathbf{G}(\tilde{\theta} - \theta^*)|_2^2 + \lambda_n|\tilde{\theta}_2 - \theta_2^*|_1 \le \lambda_n|\tilde{\theta}_1 - \theta_1^*|_1 + 4\lambda_n|\tilde{\theta}_{2,S^*} - \theta^*_{2,S^*}|_1, \tag{26}$$

by using (25).

Now Assumption 1-3 implies

$$\begin{aligned} 4\lambda_n|\tilde{\theta}_{2,S^*} - \theta^*_{2,S^*}|_1 &\le 4\lambda_n\sqrt{\frac{s^*}{nh\phi_*^2}}|\mathbf{G}_2(\tilde{\theta}_2 - \theta_2^*)|_2 \le \frac{1}{nh}|\mathbf{G}_2(\tilde{\theta}_2 - \theta_2^*)|_2^2 + 4\lambda_n^2\frac{s^*}{\phi_*^2}, \\ \lambda_n|\tilde{\theta}_1 - \theta_1^*|_1 &\le \lambda_n\sqrt{\frac{s^*}{nh\phi_*^2}}|\mathbf{G}_1(\tilde{\theta}_1 - \theta_1^*)|_2 \le \frac{1}{nh}|\mathbf{G}_1(\tilde{\theta}_1 - \theta_1^*)|_2^2 + \lambda_n^2\frac{s^*}{\phi_*^2}, \end{aligned}$$

with probability approaching one. Combining these inequalities with (26), we have

$$\frac{1}{nh}|\mathbf{G}(\tilde{\theta} - \theta^*)|_2^2 + \lambda_n|\tilde{\theta}_2 - \theta_2^*|_1 \le 5\lambda_n^2\frac{s^*}{\phi_*^2},$$

which implies

$$|\tilde{\theta}_2 - \theta_2^*|_1 \le 5\lambda_n\frac{s^*}{\phi_*^2}. \tag{27}$$

Turning to the finite dimensional component $\tilde{\theta}_1$, note that

$$\begin{aligned} \tilde{\theta}_1 - \theta_1^* &= \left(\frac{1}{nh}\mathbf{G}_1'\mathbf{G}_1\right)^{-1}\left(\frac{1}{nh}\mathbf{G}_1'e - \frac{1}{nh}\mathbf{G}_1'\mathbf{G}_2(\tilde{\theta}_2 - \theta_2^*)\right) \\ &= O_p((nh)^{-1/2}) + O_p(|\tilde{\theta}_2 - \theta_2^*|_1). \end{aligned}$$

Combining this with (27) yields the conclusion in (7).

A.1.3. *Proof of (iii).* Let $\mathbf{G}_{\bar{S}} = (G_{\bar{S},1}, \ldots, G_{\bar{S},n})'$. Since $\bar{\theta}_{\bar{S}}$ is the weighted OLS estimate that minimizes the sum of the squared residuals in the regression of $\mathbf{Y}$ on $\mathbf{G}_{\bar{S}}$ and $\bar{\theta}_{\bar{S}^c} = \tilde{\theta}_{\bar{S}^c} = 0$, it holds

$$|\mathbf{Y} - \mathbf{G}_{\bar{S}}\bar{\theta}_{\bar{S}}|_2^2 \le |\mathbf{Y} - \mathbf{G}_{\bar{S}}\tilde{\theta}_{\bar{S}}|_2^2,$$

and thus,

$$\frac{1}{nh}|\mathbf{G}_{\bar{S}}(\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}})|_2^2 \le \frac{2}{nh}|\hat{\mathbf{e}}'\mathbf{G}_{\bar{S}}(\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}})| \le \lambda_n|\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}}|_1, \tag{28}$$

due to the Hölder inequality and KKT condition. On the other hand,

$$\lambda_{\min}\left(\frac{1}{nh}\mathbf{G}_{\bar{S}}'\mathbf{G}_{\bar{S}}\right)|\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}}|_2^2 \le \frac{1}{nh}|\mathbf{G}_{\bar{S}}(\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}})|_2^2.$$

Since $|a|_1 \le \sqrt{s}|a|_2$ for an $s$-dimensional vector $a$, we conclude that

$$|\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}}|_1 \le \lambda_{\min}\left(\frac{1}{nh}\mathbf{G}_{\bar{S}}'\mathbf{G}_{\bar{S}}\right)^{-1}\lambda_n|\bar{S}|.$$

A.2. **Proof of Theorem 2 .** First, we show the asymptotic expansion in (9). Let $\bar{\gamma}$ be a coefficient vector of $Z_i$, where the elements correspond to $Z_{\bar{S},i}$ is $\bar{\gamma}_{\bar{S}}$ in (5), and other elements are zero. Observe that

$$\bar{\tau} = e_2' \left( \sum_{i=1}^n K_i G_{1i} G_{1i}' \right)^{-1} \sum_{i=1}^n K_i G_{1i} (Y_i - Z_i' \bar{\gamma}) = \hat{\tau}_\xi - e_2' \left( \sum_{i=1}^n K_i G_{1i} G_{1i}' \right)^{-1} \sum_{i=1}^n K_i G_{1i} Z_i' (\bar{\gamma} - \gamma_Y),$$

where $\hat{\tau}_\xi = e_2' \left( \sum_{i=1}^n K_i G_{1i} G_{1i}' \right)^{-1} \sum_{i=1}^n K_i G_{1i} \xi_i$. Thus it is sufficient for (9) to show that the second term is of order $o_p((nh)^{-1/2})$. Note that

$$\begin{aligned} & \left\| e_2' \left( \frac{1}{nh} \sum_{i=1}^n K_i G_{1i} G_{1i}' \right)^{-1} \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_i G_{1i} Z_i' \right\|_\infty \\ \leq \;& \lambda_{\min} \left( \frac{1}{nh} \sum_{i=1}^n K_i G_{1i} G_{1i}' \right)^{-1} \left\| \frac{1}{\sqrt{nh}} \sum_{i=1}^n \{ K_i G_{1i} Z_i' - \mathbb{E}[K_i G_{1i} Z_i'] \} \right\|_\infty + O(\sqrt{nh} h^2) \\ = \;& O_p(\sqrt{\log p}) + O(\sqrt{nh} h^2), \end{aligned}$$

where the inequality follows from the fact that $( \sum_{i=1}^n K_i G_{1i} G_{1i}' )^{-1} \sum_{i=1}^n K_i G_{1i} Z_i'$ is a vector of the local linear RDD estimator for the outcome $Z_i$ and its bias is of order $O(h^2)$ from Lemma SA-2 of CCFT, and the equality follows from the local Bernstein's inequality in Lemma A.1. Now, recall that $\theta^*_{S^*} = \arg\min E[K_i (Y_i - G'_{S^* i} \theta_{S^*})^2]$ while $\gamma_Y = \arg\min_\gamma E[(\tilde{Y} - Z'_{S^*} \gamma)^2 | X = 0]$, where $\tilde{Y} = Y(1) - Y(0) - E[Y(1) - Y(0) | X = 0]$ and $Z_{S^*} = Z_{S^*}(1) - Z_{S^*}(0)$. Due to van der Vaart and Wellner (1996, Theorem 3.4.1), we can bound the contrast $|\gamma_Y - \gamma^*|$ by the difference in the criterions, which is of order $O(s^* h^2)$ by the standard bias calculation. Since $|\bar{\gamma} - \gamma^*|_1 = O_p(\lambda_n s^*)$ by Theorem 1 and $|\gamma_Y - \bar{\gamma}|_1 = O_p(s^* h^2 + \lambda_n s^*)$. Hölder's inequality and the assumption $(\sqrt{\log p} + \sqrt{nh} h^2)(s^* h^2 + \lambda_n s^*) \to 0$ guarantee (9).

Second, we note that $\hat{\tau}_\xi$ is the conventional local linear RDD estimator without covariates for the outcome variable $\xi_i$. Thus, the proof of CCFT's Theorem 1 is directly applicable and the MSE expansion in (12) follows.

Finally, we show (13). By (9), we have

$$T_\tau = \sqrt{\frac{\mathcal{V}}{\bar{\mathcal{V}}}} \sqrt{\frac{nh}{\mathcal{V}}} (\hat{\tau}_\xi - h^2 \mathcal{B} - \tau) - \sqrt{\frac{\mathcal{V}}{\bar{\mathcal{V}}}} \sqrt{\frac{nh}{\mathcal{V}}} h^2 (\bar{\mathcal{B}} - \mathcal{B}).$$

Since $\hat{\tau}_\xi$ is the conventional local linear RDD estimator without covariates for the outcome variable $\xi_i$, Lemma SA-10 of CCFT yields $\sqrt{\frac{nh}{\mathcal{V}}} (\hat{\tau}_\xi - h^2 \mathcal{B} - \tau) \xrightarrow{d} N(0,1)$. Therefore, the assumptions $\sqrt{\frac{nh^5}{\mathcal{V}}} (\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$ and $\frac{\bar{\mathcal{V}}}{\mathcal{V}} \xrightarrow{p} 1$ imply the conclusion in (13).

A.3. **Proof of Theorem 3 .** Let $a_n = \lambda_n \varrho_n \sum_{j=1}^p \mathbb{I}\{|\hat{\gamma}^{(j)}| > 0\}$. To prove that $\mathbb{P}\{\hat{S} = S^*\} \to 1$, we note that the deviation bound in Theorem 1 is asymptotically negligible to the threshold $a_n$. This implies that if $|\gamma^{*(j)}| = 0$, then $|\hat{\gamma}^{(j)}| \leq 4 s^* \lambda_n / \phi_*^2 = o(a_n)$ with probability approaching one, and otherwise $|\hat{\gamma}^{(j)}|$ must exceed the threshold $a_n$. The proof of $\mathbb{P}\{\tilde{S} = S^*\} \to 1$ is similar.

## References


[1] Angrist, J. D., Imbens, G. W. and D. B. Rubin (1996) Identification of causal effects using instrumental variables, *Journal of the American Statistical Association*, 91, 444-455.

[2] Armstrong, T. B. and M. Kolesár (2018) Optimal inference in a class of regression models, *Econometrica*, 86, 655-683.

[3] Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C. and K. Kato (2018) High-dimensional econometrics and regularized GMM, Working paper.

[4] Belloni, A., Chernozhukov, V. and C. Hansen (2014) Inference on treatment effects after selection amongst high-dimensional controls, *Review of Economic Studies*, 81, 608-650.

[5] Bickel, P. J., Ritov, Y. A. and A. B. Tsybakov (2009) Simultaneous analysis of Lasso and Dantzig selector, *Annals of Statistics*, 37, 1705-1732.

[6] Bühlmann, P. and S. van de Geer (2011) *Statistics for High-Dimensional Data*, Springer.

[7] Calonico, S., Cattaneo, M. D. and M. H. Farrell (2018) On the effect of bias estimation on coverage accuracy in nonparametric inference, *Journal of the American Statistical Association*, 113, 767-779.

[8] Calonico, S., Cattaneo, M. D. and M. H. Farrell (2020) Coverage error optimal confidence intervals for local polynomial regression, arXiv:1808.01398.

[9] Calonico, S., Cattaneo, M. D., Farrell, M. H., and R. Titiunik (2017) rdrobust: Software for regression discontinuity designs, *Stata Journal*, 17, 372-404.

[10] Calonico, S., Cattaneo, M. D., Farrell, M. H., and R. Titiunik (2019) Regression discontinuity designs using covariates, *Review of Economics and Statistics,* 101, 442-451.

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

[12] Calonico, S., Cattaneo, M. D. and R. Titiunik (2015a) Optimal data-driven regression discontinuity plots, *Journal of the American Statistical Association*, 110, 1753-1769.

[13] Calonico, S., Cattaneo, M. D. and R. Titiunik (2015b) Rdrobust: An R package for robust inference in regression discontinuity design, R Journal, 7, 38-51.

[14] Card, D., Lee, D. S., Pei, Z. and A. Weber (2015) Inference on causal effects in a generalized regression kink design, *Econometrica*, 83, 2453-2483.

[15] Card, D., Mas, A. and J. Rothstein (2008) Tipping and the dynamics of segregation, *Quarterly Journal Economics*, 123, 177–218.

[16] Cattaneo, M. D. and J. C. Escanciano (2017) *Regression Discontinuity Designs: Theory and Applications*, in *Advances in Econometrics*, vol. 38, Emerald Group Publishing.

[17] Cattaneo, M. D., Titiunik, R. and G. Vazquez-Bare (2020) The regression discontinuity design, in Curini, L. and R. J. Franzese (eds.), *Handbook of Research Methods in Political Science and International Relations*, ch. 44, pp. 835-857, Sage Publications.

[18] Chatterjee, A. and S. N. Lahiri (2011) Bootstrapping lasso estimators, *Journal of the American Statistical Association*, 106-608-625.

[19] Fan, J., and I. Gijbels. (1992): "Variable bandwidth and local linear regression smoothers," Annals of Statistics, 20, 2008–36.

[20] Fan, J. and I. Gijbels (1996) *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

[21] Friedman, J. H., Hastie, T. and R. Tibshirani (2010) Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, 33, 1-22.

[22] Frölich, M. and M. Huber (2019) Including covariates in the regression discontinuity design, *Journal of Business & Economic Statistics*, 37, 736-748.

[23] Ganong, P. and S. Jäger (2018) A permutation test for the regression kink design, *Journal of American Statistical Association*, 113, 494-504.

[24] Gelman, A. and G. W. Imbens (2018) Why high-order polynomials should not be used in regression discontinuity designs, forthcoming in *Journal of Business & Economic Statistics*.

[25] Hahn, J., Todd, P. and W. van der Klaauw (2001) Identification and estimation of treatment effects with a regression-discontinuity design, *Econometrica*, 69, 201-209.

[26] Imbens, G. W. and K. Kalyanaraman (2012) Optimal bandwidth choice for the regression discontinuity estimator, *Review of Economic Studies*, 79, 933-959.

[27] Imbens, G. W. and T. Lemieux (2008) Regression discontinuity designs: a guide to practice, *Journal of Econometrics*, 142, 615-635.

[28] KreiSS, A. and C. Rothe (2022) Inference in regression discontinuity designs with high-dimensional covariates, arXiv:2110.13725.

[29] Lee, S., Seo, M. H. and Y. Shin (2016) The lasso for high dimensional regression with a possible change point, *Journal of the Royal Statistical Society*, B, 78, 193-210.

[30] Lei, L. and P. Ding (2021) Regression adjustment in completely randomized experiments with a diverging number of covariates, *Biometrika*, 108, 815-828.

[31] Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique, *Annals of Applied Statistics*, 7, 295-318.

[32] Ludwig, J. and D. L. Miller (2007) Does Head Start improve children's life chances? Evidence from a regression discontinuity design, *Quarterly Journal of Economics*, 122, 159-208.

[33] Porter, J. and P. Yu (2015) Regression discontinuity designs with unknown discontinuity points: Testing and estimation, *Journal of Econometrics*, 189, 132-147.

[34] Ruppert, D. and M. P. Wand (1994) Multivariate locally weighted least squares regression, *Annals of Statistics*, 22, 1346-1370.

[35] van de Geer, S., Bühlmann, P., Ritov, Y. and R. Dezeure (2014) On asymptotically optimal confidence regions and tests for high-dimensional models, *Annals of Statistics*, 42, 1166-1202.

[36] Zhang, C.-H. and S. S. Zhang (2014) Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society*, B, 76, 217-242.

School of Social Sciences, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan.

*Email address*: yarai@waseda.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

*Email address*: t.otsu@lse.ac.uk

Department of Economics, Seoul National University, 1 Gwankro Gwanakgu, Seoul, 08826, Korea.

*Email address*: myunghseo@snu.ac.kr