Regression Discontinuity Design under Self-selection

Sida Peng
Office of the Chief Economist,
Microsoft
sidpeng@microsoft.com

Yang Ning
Department of Statistics and Data Science,
Cornell University
yn265@cornell.edu

Abstract

Regression Discontinuity (RD) design is commonly used to estimate the causal effect of a policy. Existing RD relies on the continuity assumption of potential outcomes. However, self selection leads to different distributions of covariates on two sides of the policy intervention, which violates this assumption. The standard RD estimators are no longer applicable in such setting. We show that the direct causal effect can still be recovered under a class of weighted average treatment effects. We propose a set of estimators through a weighted local linear regression framework and prove the consistency and asymptotic normality of the estimators. We apply our method to a novel data set from Microsoft Bing on Generalized Second Price (GSP) auction and show that by placing the advertisement on the second ranked position can increase the click-ability by 1.91%.

1 Introduction

Regression discontinuity (RD) design is an important policy evaluation tool that has been widely used in empirical studies. Under the continuity assumptions (Lee, 2008), the RD design gives rise to many testable restrictions similar to a randomized control trial, and it allows for the identification of causal effects (Hahn et al., 2001). Standard nonparametric tools like series expansion method or kernel regression method can be applied to estimate this quantity under this minimal assumption, see Imbens and Lemieux (2008) and Cattaneo and Escanciano (2017).

An implication of the continuity assumption is that the distribution of the covariates conditional on the assignment variable is continuous at the cutoff. To validate this assumption, a variety of statistical tests and nonparametric inference procedures have been proposed, including Cattaneo et al. (2015) and Canay and Kamat (2018). However, this assumption may not hold in many applications. In particular, we consider the following two motivating examples.

Example 1 (Generalized Second Price (GSP) auction)**.** GSP is widely used by internet search engines like Google or Bing to allocate sponsored search advertisements. In reality, it is implemented through a reservation score system. Each bidder is assigned a score as a function of their bids and characteristics. The bidder's advertisement is displayed if her score passes the reservation. However, due to design of the score system, bidders with low quality may be selected below the reservation score while bidders with high quality may be selected above the reservation score. Given that low quality bidders may have a high willingness to pay, we often observe the distribution of bids to be different on two sides of the reservation score. We will further elaborate this example in the real data analysis.

Due to different distributions of the covariates at the cutoff in these examples, the standard RD estimand is no longer valid. We show in those cases the standard RD estimand can be decomposed into a direct treatment effect and an indirect treatment effect. The indirect effect is due to the unbalanced covariates near the cut-off. For example, the policy intervention may result more boys than girls to receive scholarship. If in general boys perform different from girls in SAT test, the difference in SAT

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

score due to gender will also be accounted into the standard RD estimand as if the policy intervention is selecting on genders. However as no direct causal mechanism is assumed between the unbalanced covariates and running variable, the selection could be generated due to an unknown equilibrium, completely reversed or purely spurious. For example, in a GSP auction, the reservation score is designed to separate the bidders by quality and in the meanwhile bidders may self-select such that low quality bidders may have incentive to bid higher. As a result, policy changes to move reservation score can be risky if the equilibrium between reservation score and quality of bidders are not disentangled.

In this paper, we propose a new framework to address this problem by adjusting unbalanced covariates due to self-selection. Consider the following sharp RD setup: T is a binary treatment variable, $Y(1)$ and $Y(0)$ are the potential outcomes under $T = 1$ and $T = 0$ and X is the running variable so that the treatment is fully determined by $T = \mathbf{1}(X > c)$ for a known threshold c. In the classical RD framework, it is assumed that $\mathbb{E}(Y(t)|X = x)$ is continuous at $x = c$ for $t = 0, 1$; see Assumption 2. This essentially assumes away self-selection based on both observed and unobserved covariates. To account for the self-selection effect, we assume that further covariate information can be collected. In particular, we define $Z(1)$ and $Z(0)$ as the potential covariates with or without treatment. By using the potential "outcome" formulation, we allow the distribution of the covariates on two sides of the threshold to be different, i.e., discontinuity of the covariate distribution. By controlling all unbalanced covariates, we assume that $\mathbb{E}(Y(t)|X = x, Z(t) = z)$ is continuous in x at the threshold; see Assumption 4. This is the main assumption made in this paper.

We propose a class of estimands for RD design in the framework of weighted average treatment effect (WATE). We show that our estimands can tease out the effect of discontinuity of the conditional distribution of covariates through re-weighting the marginal treatment effects. For instance, under the constant treatment effect model with unbalanced covariates, our estimands reduce to the direct treatment effect. One special case of our estimands is Frölich and Huber (2018). Another special case can be interpreted as the "global" average treatment effect (ATE) $\mathbb{E}(Y(1) - Y(0))$ under the conditional independence assumption (CIA) as in Angrist and Rokkanen (2015). The CIA assumes that the treatment is mean independent of the running variable near the cutoff conditional on the covariates. Intuitively, after projecting the outcome variable onto a rich set of covariates (excluding running variable), the residual should not depend on the running variable and can be viewed as an experiment with randomly assigned treatment. We show that our method identifies ATE under CIA assumption, whereas the standard RD estimand remains a "local" treatment effect at the cutoff (Lee, 2008).

We further provide the nonparametric identification for our estimands and propose nonparametric estimators based on the inverse propensity score weighted (IPW) approach, see Horvitz and Thompson (1952) and Abadie and Imbens (2016). However, notice that the treatment assignment is degenerate with respect to the running variable. We get around this problem by considering the marginal effect of the treatment and re-weighting on the other covariates first. The kernel method is applied to estimate the conditional mean function. The consistency and asymptotic normality of the proposed estimator are established. We further extend our method to the fuzzy RD design where the treatment compliance is imperfect. Similarly, we provide the nonparametric identification of the causal effect under the fuzzy RD design, and propose a nonparametric estimator. The proposed estimator is similar to that of the local average treatment effect (LATE) in a fractional format but with numerator being adjusted to incorporate additional selections.

This work is connected to the growing literature on the RD design with covariates. In particular, two recent papers provide insightful guidance on the subject. Calonico et al. (2019) estimated the marginal treatment effect by a local linear regression with the linear-in-parameters specification for the covariates. The main advantage of their method is that the nonparametric estimation of $\mathbb{E}(Y(t)|X = x, Z(t) = z)$ is avoided. In another paper, Frölich and Huber (2018) proposed a fully nonparametric estimator of the marginal treatment effect by estimating $\mathbb{E}(Y(t)|X = x, Z(t) = z)$ nonparametrically. They allowed the conditional density of $Z(t)$ given X to be discontinuous. Our work differs from the above papers by considering a different estimand that is less local under CIA and identifies the direct treatment effect under the constant treatment effect model. Unlike Calonico et al. (2019), we do not require the continuity of the conditional mean of $Z(t)$ given X. Instead, our Assumption 4 is similar to assumption 1 (iv) in Frölich and Huber (2018) under the sharp RD design. The proposed IPW estimator is also different from the above regression based estimators.

2 Sharp RD Design

2.1 Problem Setup and Continuity Assumption

In the standard RD design setting, we observe n i.i.d. random samples $\{Y_i, X_i, Z_i, T_i\}_{i=1}^n$, where Y_i is the outcome variable of interest for the ith sample, $T_i \in \{0,1\}$ is the binary treatment variable, $X_i \in \mathbb{R}$ is the running variable and $Z_i \in \mathbb{R}^p$ is the covariate. In the sharp RD design, the treatment T_i is perfectly assigned through the running variable X_i relative to a known cutoff c. For example, we have

$$T_i = \mathbf{1}(X_i > c).$$

Adopting a potential outcome framework, we can write the observed outcome variable Y_i as

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

where $Y_i(0)$ and $Y_i(1)$ represent the potential outcomes without or with treatment Rubin (2006). The average treatment effect is defined as $\mathbb{E}(Y_i(1) - Y_i(0))$. However, this estimand is not identifiable under the RD design as the treatment assignment T_i is a deterministic function of X_i. In this framework, Hahn et al. (2001), Lee (2008) and Cattaneo et al. (2015) showed that one can still identify the treatment effect at the cutoff

$$\tau_{SRD} = \mathbb{E}(Y_i(1) - Y_i(0)|X_i = c),$$

under the following continuity assumption on the potential outcomes.

Assumption 2.

$$\mathbb{E}(Y_i(1)|X_i = x) \quad \text{and} \quad \mathbb{E}(Y_i(0)|X_i = x)$$

are continuous in x at $x = c$.

This assumption implies that the conditional mean of the potential outcomes near the cutoff $x = c$ are similar. There is no discontinuity of the conditional mean functions at the cutoff. This assumption enables us to identify τ_{SRD} in the RD design. We refer to Hahn et al. (2001), Lee (2008) and Cattaneo et al. (2015) for further discussion on this assumption.

Now let us consider the case that additional covariates Z_i are observed. Denote

$$Z_i = Z_i(0) \cdot (1 - T_i) + Z_i(1) \cdot T_i,$$

where $Z_i(0)$ and $Z_i(1)$ represent the potential covariates without or with treatment. In the presence of covariates Z_i, the causal parameter τ_{SRD} can be rewritten as

$$\tau_{SRD} = \mathbb{E}\Big(\mathbb{E}[Y_i(1)|X_i = c, Z_i(1)] - \mathbb{E}[Y_i(0)|X_i = c, Z_i(0)]\Big|X_i = c\Big). \tag{2.1}$$

In a recent work, Calonico et al. (2019) proposed a kernel based estimator of τ_{SRD} by accounting for the additional covariates Z_i. In addition to the continuity Assumption 2, it is also assumed that $\mathbb{E}(Z_i(1)|X_i = c) = \mathbb{E}(Z_i(0)|X_i = c)$ for the consistency of the resulting kernel estimator, that is the potential covariates $Z_i(1)$ and $Z_i(0)$ have the same conditional mean at the cutoff $X_i = c$.

However, in some applications we may observe $\mathbb{E}(Z_i(1)|X_i = c) \neq \mathbb{E}(Z_i(0)|X_i = c)$, when self-selection based on the covariates exists. For example, consider the classical scholarship example. Students with SAT score higher than a threshold will receive scholarship. The treatment effect of interest is the effect of scholarship on the students' first semester GPAs. If the cutoff is pre-released, it might be possible that students with some common characteristics (i.e. gender) may study harder to pass the bar. This leads to an ex-ante selection based on the covariates. Thus, one may observe that the conditional mean functions of Z_i given X_i right below or above the threshold are different, i.e.,

$$\lim_{x \to c^+} \mathbb{E}(Z_i|X_i = x) \neq \lim_{x \to c^-} \mathbb{E}(Z_i|X_i = x). \tag{2.2}$$

The following simple lemma essentially says that the self-selection based on the covariates (i.e., eq 2.2) implies $\mathbb{E}(Z_i(1)|X_i = c) \neq \mathbb{E}(Z_i(0)|X_i = c)$.

Lemma 3. If $\mathbb{E}(Z_i(t)|X_i = x)$ is continuous at $x = c$ for $t \in \{0,1\}$, then

$$\mathbb{E}(Z_i(1)|X_i = c) = \mathbb{E}(Z_i(0)|X_i = c) \iff \lim_{x \to c^+} \mathbb{E}(Z_i|X_i = x) = \lim_{x \to c^-} \mathbb{E}(Z_i|X_i = x).$$

The above lemma provides a convenient way to check whether $\mathbb{E}(Z_i(1)|X_i = c) = \mathbb{E}(Z_i(0)|X_i = c)$ holds in empirical studies. One may simply plot the observed covariates Z_i against X_i and examine whether there is a discontinuity of the trend around $x = c$. The method is applied in the real data analysis.

In the following, we investigate the consequence of $\mathbb{E}(Z_i(1)|X_i = c) \neq \mathbb{E}(Z_i(0)|X_i = c)$. To be specific, we consider the following constant treatment effect model

$$\mathbb{E}(Y_i(t)|X_i, Z_i(t)) = \alpha + \tau\mathbf{1}(t = 1) + g(X_i) + Z_i(t)\gamma, \tag{2.3}$$

for $t \in \{0,1\}$, where $g(\cdot)$ is an arbitrary continuous function. By (2.1), one can show that

$$\tau_{SRD} = \tau + \Big(\mathbb{E}(Z_i(1)|X_i = c) - \mathbb{E}(Z_i(0)|X_i = c)\Big)\gamma. \tag{2.4}$$

The estimand τ_{SRD} can be decomposed into two terms. The first term τ represents the direct treatment effect after controlling the running variable X_i and the covariates $Z_i(t)$. The second term in the right hand side of (2.4) can be interpreted as the indirect effect of the policy due to the unbalanced covariates near the cutoff or self-selection, which is nonzero if $\gamma \neq 0$ and $\mathbb{E}(Z_i(1)|X_i = c) \neq \mathbb{E}(Z_i(0)|X_i = c)$. In many applications, the direct treatment effect τ is usually more meaningful and interpretable than τ_{SRD}, as τ_{SRD} is confounded by the self-selection effect.

In this example, when the indirect effects in τ_{SRD} is assumed away by requiring the continuity of the conditional density of the covariates Z_i at the cutoff $X_i = c$, the data around the cutoff can be viewed as a natural experiment and continuity on the covariates implies a balanced design for this experiment so that we can estimate a local average treatment effect. However when the self-selection exits, the experiment is no longer balanced and the average treatment effect τ_{SRD} will typically differ from the direct effect τ.

2.2 Weighted Average Treatment Effect

As seen in (2.4), if $\mathbb{E}(Z_i(1)|X_i = c) \neq \mathbb{E}(Z_i(0)|X_i = c)$ (i.e., the covariates are unbalanced at the cutoff) and $\gamma \neq 0$, τ_{SRD} can be different from the causal parameter of interest. To overcome this difficulty, we propose a new class of causal parameters, called the weighted average treatment effect (WATE). Denote

$$\begin{aligned}\Delta(c,z) = &\mathbb{E}(Y(1)|X = c, Z(1) = z)\\ &- \mathbb{E}(Y(0)|X = c, Z(0) = z).\end{aligned}$$

Define the average treatment effect over entire population as

$$\tau^{w1}_{SRD} = \int \Delta(c,z) f_Z(z) dz,$$

where $f_Z(\cdot)$ is the p.d.f of the covariates Z. In τ^{w1}_{SRD}, we average the conditional mean difference over the entire population whose covariates follow from the *marginal distribution* $f_Z(z)$.

We can also define average treatment effect over locally randomized population τ^{w2}_{SRD} and average treatment effect via classical RD estimand τ^{w3}_{SRD} by choosing different weights. Those results are giving in the Appendix. In practice, which causal estimand in above examples to use should depend on the target population of interest and is often determined on a case-by-case basis. Indeed, our framework opens a door towards designing new causal parameters tailored to specific applications. Since the goal of the paper is to deal with unbalanced covariates, to fix the idea we will mainly focus on τ^{w1}_{SRD}.

In the following, we comment on two properties of our estimands τ^{w1}_{SRD}. First, under the constant treatment effect model (2.3), direct calculation shows that $\Delta(c,z) = \tau$ and thus τ^{w1}_{SRD} equals to the direct treatment effect τ without any further assumption. In contrast, the classical RD estimand τ_{SRD} reduces to τ under the extra assumption that $\gamma = 0$ or $\mathbb{E}(Z_i(1)|X_i = c) = \mathbb{E}(Z_i(0)|X_i = c)$.

Second, τ^{w1}_{SRD} generalizes to the overall average treatment effect (ATE) under the conditional independence assumption (CIA) proposed by Angrist and Rokkanen (2015). Assume that Z_i are the pretreatment covariates, i.e, $Z_i(1) = Z_i(0) = Z_i$. The CIA is defined as

$$\begin{aligned}\mathbb{E}(Y_i(1)|X_i, Z_i) &= \mathbb{E}(Y_i(1)|Z_i),\\ \mathbb{E}(Y_i(0)|X_i, Z_i) &= \mathbb{E}(Y_i(0)|Z_i),\end{aligned} \tag{2.5}$$

which implies that the potential outcomes are mean independent of the running variable conditional on the covariates. By controlling a rich set of covariates, CIA seems to be a reasonable assumption as the link between the running variable and outcomes can be blocked (Angrist and Rokkanen, 2015). Since the CIA (2.5) implies $\Delta(c,z) = \Delta(z)$, our estimand τ^{w1}_{SRD} reduces to

$$\tau^{w1}_{SRD} = \int \Delta(z) f_Z(z) dz = \mathbb{E}(Y_i(1) - Y_i(0)),$$

which is the overall ATE. Thus, the new estimand τ^{w1}_{SRD} can represent a causal effect that is less local than the standard RD estimand τ_{SRD}.

2.3 Nonparametric Identification

In this subsection, we study the nonparametric identification of τ^{w1}_{SRD}. The identification results of τ^{w2}_{SRD} and τ^{w3}_{SRD} are given in the Appendix. Instead of Assumption 2, we impose the following continuity assumption.

Assumption 4. $\mathbb{E}(Y_i(1)|X_i = x, Z_i(1) = z)$ and $\mathbb{E}(Y_i(0)|X_i = x, Z_i(0) = z)$ are right and left continuous in x at $x = c$ for any $z \in \mathcal{Z}$ respectively.

Intuitively, this assumption says, once all unbalanced covariates are controlled, there is no further discontinuity between the running variable and outcomes at the threshold. This assumption is similar to assumption 1 (iv) in Frölich and Huber (2018) under the sharp RD design. In addition, this assumption is weaker than CIA, as (2.5) implies our Assumption 4 when the covariates are pre-determined.

The following theorem shows that τ_{SRD}^{w1} is identifiable based on the distribution of the observed data under Assumption 4.

Theorem 5 (Nonparametric Identification)**.** Under Assumption 4, τ_{SRD}^{w1} is identifiable:

$$\tau_{SRD}^{w1} = \int \big[\mathbb{E}(Y|X=c^+, Z=z) - \mathbb{E}(Y|X=c^-, Z=z)\big] f_Z(z)dz,$$

where $\mathbb{E}(Y|X=c^+, Z) = \lim_{x\to c^+}\mathbb{E}(Y|X=x, Z)$ and $\mathbb{E}(Y|X=c^-, Z) = \lim_{x\to c^-}\mathbb{E}(Y|X=x, Z)$.

An alternative proof of identification showing

$$\tau_{SRD}^{w1} = \lim_{\delta\to 0^+} E\Big(\frac{YT}{\pi_1(Z)}|X=c+\delta\Big) - \lim_{\delta\to 0^+} E\Big(\frac{Y(1-T)}{\pi_0(Z)}|X=c-\delta\Big),$$

is provided in the appendix.

3 Nonparametric Estimation

In the causal inference literature, inverse propensity score weighting (IPW) is one of the most widely used tools to handle the unbalanced covariates in the treatment and control groups (Horvitz and Thompson, 1952). However, the standard IPW method is not directly applicable because in the sharp RD design the treatment assignment is a deterministic function of the running variable and thus the propensity score is degenerate. In this section, we propose a class of nonparametric estimators of τ_{SRD}^{w1}, by modifying the inverse propensity score weighting approach.

To motivate our nonparametric estimator, we consider the following notation. Denote by $\pi_1(Z_i)$ a function of the covariate Z_i to be chosen later, $K(\cdot)$ a symmetric kernel function and h a bandwidth that shrinks to 0. The detailed conditions on the kernel function and bandwidth are deferred to the next section. Consider the following inverse weighted kernel estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_iT_i}{\pi_1(Z_i)}\cdot h^{-1}K\Big(\frac{X_i-c}{h}\Big), \tag{3.1}$$

for the estimand $\mathbb{E}\{Y(1)w_1(Z(1))|X=c\}$, where $\pi_1(Z)$ plays the same role as the propensity score in the IPW method. Since in the RD design the propensity score function is degenerate, in the following we will show that the choice of $\pi_1(Z_i)$ differs from the standard propensity score model. The rationale is to choose $\pi_1(Z_i)$ so that the estimator (3.1) is asymptotically unbiased,

$$\mathbb{E}\left(\frac{Y_iT_i}{\pi_1(Z_i)}\cdot h^{-1}K\Big(\frac{X_i-c}{h}\Big)\right) \approx \int \mathbb{E}(Y|X=c^+, Z=z)f_Z(z)dz. \tag{3.2}$$

With some algebra we show that (3.2) holds provided

$$\pi_1(z) = \frac{f_{X,Z(1)}(c,z)}{2f_Z(z)}, \tag{3.3}$$

where $f_X(c)$ is the p.d.f of X at $x=c$. Following from the same argument, one can show that $\pi_0(z) = \frac{f_{X,Z(0)}(c,z)}{2f_Z(z)}$. Since the weight $\pi_1(z)$ and $\pi_0(z)$ depends on the unknown density functions, we propose to estimate those densities by the following kernel estimators

$$\widehat{f}_{X,Z(1)}(c,z) = 2\cdot(nh_1^2)^{-1}\sum_{x_i>c}K_1(\frac{c-x_i}{h_1}, \frac{z-z_i}{h_1}), \tag{3.4}$$

$$\widehat{f}_{X,Z(0)}(c,z) = 2\cdot(nh_1^2)^{-1}\sum_{x_i<c}K_1(\frac{c-x_i}{h_1}, \frac{z-z_i}{h_1}), \tag{3.5}$$

$$\widehat{f}_Z(z) = (nh_2)^{-1}\sum_{i=1}^{n}K(\frac{z-z_i}{h_2}), \quad \widehat{f}_X(c) = (nh_2)^{-1}\sum_{i=1}^{n}K(\frac{c-x_i}{h_2}), \tag{3.6}$$

where h_1 and h_2 are bandwidth parameters. Note that the kernel estimator is known to suffer from the curse of dimensionality and is only applicable when the dimension of Z_i is small. For the applications in which a large number of covariates can be collected, one may consider alternative parametric or semiparametric approaches for density estimation. In this work, we only focus on the above kernel estimators and leave the alternatives for future investigation.

Replacing the unknown density functions in $\pi_1(z)$ and $\pi_0(z)$ with the corresponding kernel estimators, we can obtain $\widehat{\pi}_1(z)$ and $\widehat{\pi}_0(z)$. While we can construct the final estimator by plugging $\widehat{\pi}_1(z)$ into (3.1), for practical use and theoretical analysis we

recommend the local linear estimator, since it has smaller asymptotic bias and better finite sample behavior near the boundary (Fan and Gijbels, 1996). Motivated by the formulation of the kernel estimator (3.1), we propose the following weighted local linear (WLL) estimator

$$\begin{aligned}\Big(\widehat{\alpha}_0, \widehat{\beta}_0\Big) &= \arg\min_{\alpha,\beta} \sum_{\{i:X_i<c\}} \frac{1-T_i}{\widehat{\pi}_0(Z_i)} \\ &\quad \cdot \Big(Y_i - \alpha - (X_i - c)\beta\Big)^2 K\Big(\frac{X_i - c}{h}\Big), \\ \Big(\widehat{\alpha}_1, \widehat{\beta}_1\Big) &= \arg\min_{\alpha,\beta} \sum_{\{i:X_i>c\}} \frac{T_i}{\widehat{\pi}_1(Z_i)} \\ &\quad \cdot \Big(Y_i - \alpha - (X_i - c)\beta\Big)^2 K\Big(\frac{X_i - c}{h}\Big),\end{aligned} \tag{3.7}$$

Thus, we can estimate the WATE τ^{w1}_{SDR} by

$$\widehat{\tau}^{w1}_{SDR} = \widehat{\alpha}_1 - \widehat{\alpha}_0. \tag{3.8}$$

4 Theoretical Results

In this section, we study the asymptotic properties of the proposed estimator. We focus on the local linear estimator (3.7) for the estimand τ^{w1}_{SRD}, due to the nice properties of τ^{w1}_{SRD} as explained in Section 2.2.

Let δ denote a small positive constant. For notational simplicity, define $\mathcal{F}^-$ as the class of functions of $x \in (c-\delta, c]$ and $z \in \mathcal{Z}$ such that for any $f \in \mathcal{F}^-$, $\frac{\partial^3}{\partial a \partial b \partial c} f(x,z)$ is continuous, where $a, b, c = \{x, z\}$. Here, the derivatives of $f(x,z)$ with respect to x at $x = c$ (say $\frac{\partial^3}{\partial x^3} f(c,z)$) are interpreted as left derivatives. Similarly, we define $\mathcal{F}^+$ as the class of functions of $x \in [c, c+\delta)$ and $z \in \mathcal{Z}$ such that for any $f \in \mathcal{F}^+$, $\frac{\partial^3}{\partial a \partial b \partial c} f(x,z)$ is continuous, where $a, b, c = \{x, z\}$. The derivatives at $x = c$ correspond to the right derivatives. For simplicity, we only consider the case that $\dim(Z) = 1$. The generalization to multivariate covariates follows from the similar argument.

Assumption 6 (Smoothness condition)**.** Assume $f_{X,Z(0)}(x,z) \in \mathcal{F}^-$ and $f_{X,Z(1)}(x,z) \in \mathcal{F}^+$. Denote $m_t(x,z) = \mathbb{E}(Y_i(t)|X_i = x, Z_i(t) = z)$ for $t = \{0,1\}$. Assume $m_0(x,z) \in \mathcal{F}^-$ and $m_1(x,z) \in \mathcal{F}^+$.

Since the proposed method requires to estimate the unknown density functions $f_{X,Z(0)}(x,z)$ and $f_{X,Z(1)}(x,z)$, we assume they are sufficiently smooth. In addition, we also need the smoothness of $m_t(x,z)$ in order to study the local linear estimator (3.7). This assumption implies Assumption 4, which is required for identification purpose.

Assumption 7. Assume the following conditions hold:

1. $X \in [x_l, x_u]$ such that $x_l < c < x_u$ and Z also has bounded support.
2. The kernel $K(u)$ is a non-negative, symmetric and bounded function with compact support which satisfies
$$\int K(z)\,dz = 1, \quad \text{and} \quad K(u) = K(-u).$$
The bivariate kernel $K_1(u,v)$ is a product kernel $K_1(u,v) = K(u)K(v)$, where $K(\cdot)$ satisfies the above conditions.
3. $f_{X,Z(1)}(x,z)$ is bounded way from 0 by a constant for $x \in [c, c+\delta)$ and $z \in \mathcal{Z}$ and $f_{X,Z(0)}(x,z)$ is bounded way from 0 by a constant for $x \in (c-\delta, c]$ and $z \in \mathcal{Z}$.
4. $\sigma^2_t = \mathbb{E}[(Y(t) - m_t(X, Z(t)))^2 | X = x, Z(t) = z] < \infty$ for $t \in \{0,1\}$.

Assumption 7 has four parts. The first and second parts are standard assumptions in kernel density estimation problems. Since our RD estimator applies local linear estimators at the boundary, the uniform kernel or triangular kernel has been shown to have good performance under such scenario (Calonico et al., 2019). The third part requires $f_{X,Z(1)}(x,z)$ and $f_{X,Z(0)}(x,z)$ to be bounded away from 0 so that the inverse weights in the local linear estimator can be well controlled. This is similar to IPW estimator which requires the propensity score to be bounded away from 0. However, it also implies the overlapping of the support for $(X, Z(1))$ and $(X, Z(0))$ near the cut-off, which can be a strong assumption if the covariates $Z(1)$ and $Z(0)$ are high-dimensional (see D'Amour et al. (2020)). The last part assumes that the (homoscedastic) noise has finite variance. Denote $\alpha_t = \int \mathbb{E}(Y(t)|X = c, Z(t) = z) f_Z(z) dz$, and recall that $\tau^{w1}_{SRD} = \alpha_1 - \alpha_0$. The main theorem in this section shows the rate of convergence of the local linear estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_0$ and their limiting distributions.

Theorem 1. Assume that assumptions 6 and 7 hold, and let h be the bandwidth in estimator (3.7), and h_1, h_2 be the bandwidth choice in estimator (3.4), (3.5) and (3.6). We choose $h \asymp \sqrt{h_1} \asymp h_2$. Then for $t = 0, 1$

$$|\widehat{\alpha}_t - \alpha_t| = O_p(h^2 + (nh^2)^{-1/2}). \tag{4.1}$$

Denote $d_t(x_i, z_i) = m_t(x_i, z_i) - \alpha_t$. Furthermore, if $h^3 n^{1/2} = o(1)$ holds, we have

$$\sqrt{nh^2}(\widehat{\alpha}_1 - \alpha_1) \sim N\left(0, C_v \cdot \mathbb{E}_Z\left(\frac{f_Z(z_i)}{f_{X,Z(1)}(c^+, z_i)} d_1(c^+, z_i)^2\right)\right),$$

and

$$\sqrt{nh^2}(\widehat{\alpha}_0 - \alpha_0) \sim N\left(0, C_v \cdot \mathbb{E}_Z\left(\frac{f_Z(z_i)}{f_{X,Z(0)}(c^-, z_i)} d_0(c^-, z_i)^2\right)\right),$$

where $\mathbb{E}_Z$ denotes the expectation under the marginal distribution of Z, and

$$C_v = \frac{\kappa_2^2 \kappa_{20} + \kappa_1^2 \kappa_{22} - 2\kappa_1 \kappa_2 \kappa_{21}}{\left(\frac{1}{2}\kappa_2 - \kappa_1^2\right)^2},$$

with

$$\kappa_q = \int_{u>0} K(u) u^q du \quad \text{and} \quad \kappa_{2q} = \int_{u>0} K(u)^2 u^q du,$$

for $q = 0, 1, 2$.

We note that from (4.1) the optimal choice of h is of order $O(n^{-1/6})$ and the corresponding convergence rate is $|\widehat{\alpha}_t - \alpha_t| = O_p(n^{-1/3})$ due to the boundary effect. Specifically, the estimated weights $\widehat{\pi}_1(z)$ and $\widehat{\pi}_0(z)$ depend on the density estimators at the boundary. Since we only require $f_{X,Z(0)}(x, z) \in \mathcal{F}^-$ and $f_{X,Z(1)}(x, z) \in \mathcal{F}^+$ to be smooth from one side, the corresponding density estimators have a slower rate. So, the plug-in error becomes the dominant term when establishing the rate of $\widehat{\alpha}_t$. However, if Z_i are the pre-treatment covariates, i.e, $Z_i(1) = Z_i(0) = Z_i$, we can estimate the density $f_{X,Z}(c, z)$ by

$$\widehat{f}_{X,Z}(c, z) = (nh_1^2)^{-1} \sum_{i=1}^{n} K_1\left(\frac{c - x_i}{h_1}, \frac{z - z_i}{h_1}\right). \quad (4.2)$$

The following corollary shows that in this case $\widehat{\alpha}_t$ has an improved rate. With an optimal choice of the bandwidth parameters, we prove that $|\widehat{\alpha}_t - \alpha_t| = O_p(n^{-2/5})$, that is the boundary effect for estimating α_t is automatically removed without applying any additional bias correction procedures.

Corollary 1. Assume that assumptions 6 and 7 hold and Z_i are the pre-treatment covariates. Choosing $h \asymp h_1 \asymp h_2$, we have

$$|\widehat{\alpha}_t - \alpha_t| = O_p(h^2 + (nh)^{-1/2}).$$

If $h^5 n = o(1)$, we have

$$\sqrt{nh}(\widehat{\alpha}_t - \alpha_t) \sim N\left(0, C_v \cdot \sigma_t^2 \int \frac{f_Z(z)^2}{f_{X,Z}(c, z)} dz\right).$$

The asymptotic results proved in theorem 1 and corollary 1 require undersmoothing to justify the intervals constructed (see Armstrong and Kolesár, 2020; Calonico et al., 2014). In our simulation below, we did not employ undersmoothing but used standard 10-fold cross-validation to select bandwidth. We consider bias correction for our estimator as a future research topic.

5 Simulation and Empirical Examples

5.1 Simulation Study

We consider the following data generating process:

$$y_i(1) = 3 + x_i + z_i + \epsilon_{1i},$$

$$y_i(0) = 1 + x_i + z_i + \epsilon_{0i},$$

where x_i and ϵ_i are generated independently from $N(0, 1)$ distribution, however, z_i is generated from another independent $N(0, 1)$ process with a discontinuity at $X > 0$, i.e. $z_i = \gamma \cdot \mathbf{1}(x_i > 0) + z_i^*$, where $z_i^* \sim N(0, 1)$. The treatment T_i is assigned at the cutoff 0: $T_i = \mathbf{1}(x_i > 0)$. When $\gamma = 0$, again there is no discontinuity of the conditional distribution of z_i given $x_i = 0$. Both our estimaind τ_{SRD}^{w1} and the standard RD estimand τ_{SRD} are equal to 2. However, as γ differs from 0, the conditional distribution of z_i given x_i is discontinuous at $x_i = 0$. Our estimand τ_{SRD}^{w1} still equals to 2, which is the direct causal effect of interest. If we adopt the standard RD framework and ignore the discontinuity of the conditional distribution of z_i given x_i, we would expect that the standard RD estimator is biased for estimating the direct causal effect of interest (which is 2 in this example). In the data generating process, we vary γ from 0 to 1 and compare our estimator with the standard RD estimator Lee (2008). The results are shown in Table 1. The standard RD estimator has large bias and very poor coverage probability when γ is close to 1, which agrees with our expectation. In contrast, the proposed estimator has relatively small MSE –bias and variance and accurate coverage probabilities across different choices of the sample size n and the parameter γ. In summary, our simulation studies confirm that one should apply the proposed framework to the RD study if there exists some potential discontinuity of the conditional distribution of the covariates given the running variable. More simulation results can be found in the appendix.

		bias		variance		Coverage		CI length	
n	γ	RD	WLL	RD	WLL	RD	WLL	RD	WLL
500	0.2	0.19	**0.08**	0.62	**0.43**	0.94	0.96	2.42	1.69
	0.4	0.45	**0.12**	0.55	**0.43**	0.87	0.95	2.16	1.70
	0.6	0.57	**0.18**	0.59	**0.49**	0.85	0.94	2.31	1.93
	0.8	0.76	**0.24**	0.57	**0.47**	0.71	0.93	2.23	1.84
	1	0.98	**0.36**	0.53	**0.44**	0.58	0.88	2.10	1.73
1000	0.2	0.19	**0.09**	**0.41**	0.49	0.92	0.97	1.62	1.92
	0.4	0.41	**0.13**	0.41	**0.34**	0.85	0.94	1.61	1.32
	0.6	0.55	**0.18**	0.40	**0.34**	0.72	0.92	1.58	1.34
	0.8	0.82	**0.25**	0.46	**0.39**	0.58	0.90	1.81	1.51
	1	1.01	**0.32**	**0.44**	0.51	0.40	0.94	1.74	2.02
2000	0.2	0.21	**0.06**	0.32	**0.27**	0.91	0.97	1.24	1.06
	0.4	0.38	**0.09**	0.31	**0.25**	0.76	0.94	1.20	0.98
	0.6	0.58	**0.17**	0.32	**0.26**	0.55	0.90	1.25	1.01
	0.8	0.84	**0.22**	0.32	**0.30**	0.25	0.91	1.24	1.17
	1	1.01	**0.22**	**0.29**	0.36	0.05	0.95	1.13	1.39
5000	0.2	0.21	**0.04**	0.22	**0.17**	0.83	0.95	0.88	0.69
	0.4	0.41	**0.07**	0.24	**0.19**	0.62	0.95	0.94	0.76
	0.6	0.61	**0.13**	0.23	**0.22**	0.30	0.94	0.91	0.87
	0.8	0.83	**0.18**	0.23	**0.27**	0.05	0.92	0.90	1.07
	1	1.03	**0.23**	0.23	**0.21**	0.01	0.98	0.88	0.84

Table 1: Comparison of the standard RD estimator and the proposed weighted local linear estimator (WLL) in the second setting.

5.2 Generalized Second Price Auction (GSP)

Next we apply our method to study the generalized second price auction (GSP) problem. GSP is an auction mechanism for multiple items and it has been used widely for the assignment of advertisement positions by internet search engine like Google and Bing. Let n be the number of bidders, and let $b^1 \geq b^2 \geq \cdots \geq b^n$ be the bids from high to low. Denote by $v_{(1)}, v_{(2)}, \cdots, v_{(n)}$ the bidders' valuation associated with the rank of bids and r^k the click through rate for the kth position. The kth bidder's payoff in a GSP is given as $(v_{(k)} - b^{k+1})r^k$.

An important metrics is the click through rate r^k for the kth position. Bidders are interested in the potential growth in their search traffic by winning the auction. And furthermore, in a Vickrey-Clarke-Groves (VCG) auction, r^k will determine the total cost for placing each bidder in the sponsored advertisement region. In real world, GSP is usually implemented through a reservation score. A search score is formed for every bidder based on their bid and other quality measures. When the search score is bigger than a pre-set reservation score, the bidder's link will be displayed in the sponsored area. Otherwise, they will be displayed after all the sponsored advertisements. The reservation score cut-off creates a natural regression discontinuity setting to evaluate r^k.

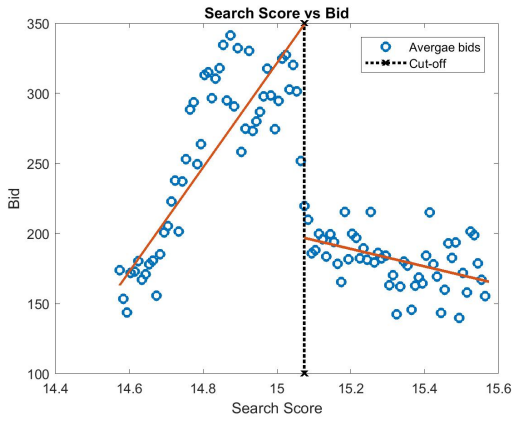

Figure 1: Search Score v.s. Bids

We study the Microsoft Bing search data from Oct 2nd to Oct 22nd in 2015 and estimate the effect of advertisement positions. We focus on a set of searches with first advertisement positions displayed but without third advertisement position displayed.[1] This allows us to analyze the effect of second advertisement positions by comparing click-abilities for bidders near the search score cutoff. Once the score passed the cut-off at 15.17, the customers' links will be placed at the sponsored advertisement area and a significant increase in the click traffic can be observed. Consider the bid as a covariate. Figure 1 plots the mean bidding price before and after the search score cut-off. The mean bids before the cut-off is higher than the mean bids after the cut-off, implying the discontinuity of the conditional distribution of the covariate. The conditional density of the covariate before and after the cut-off are also different. The plot is given in the Appendix. Although a local envy free equilibrium exists when bidders are all bidding their true valuation (Edelman et al., 2007), information asymmetry or bidder inertia may still lead to bidder selections. For example, active bidders may have the incentive to bid more aggressively to take advantage of the bidders with high inertia.

Table 2 presents the results of our estimator and a

[1] Microsoft Bing allows a maximum of 4 advertisement to be displayed at the time of study.

polynomial RD estimator. The classic RD estimator may not be valid in this case due to the discontinuity in the covariates. It estimates that placing the advertisement on the second position can increase the click-ability by 1.91% and it is statistically significant. On the other hand, the proposed estimator delivers only 1.20%, 57% less than the RD estimator and it is not statistically significant at 5% level. The difference is mainly because low quality bidders with high willingness to pay have the incentive to bid higher to move their search scores pass the threshold. But when we match bidders with similar bids before and after the cut-off, the effect goes away. Thus, the conclusion based on our proposed estimator is more reliable.

	RD	WLL
Estimates	1.91%***	1.20%
Standard Error	0.0033	0.0090

Table 2: Estimated effect of adversing at second position using the standard RD estimator and the proposed method. The bandwidth parameters are selected using cross-validation and standard errors are obtained via bootstrap. *** represents significance at $p < 0.05$

Acknowledgements

We would like to thank Zhuan Pei, Yanqing Fan, Peter Hull, the three reviewers and participants in seminars and conferences at which this paper was presented. All remaining errors are ours.

References

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84** 781–807.

Angrist, J. and Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association* **110** 1331–1344.

Armstrong, T. and Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics* **11** 1–39.

Calonico, S., Cattaneo, M., Farrellx, M. and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics* **101** 442–451.

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

Canay, I. A. and Kamat, V. (2018). Approximate permutation tests and induced order statistics in the regression discontinuity design. *The Review of Economic Studies* .

Cattaneo, M. D. and Escanciano, J. C. (eds.) (2017). *Regression Discontinuity Designs: Theory and Applications*, vol. 38. Emerald Publishing Limited.

Cattaneo, M. D., Frandsen, B. R. and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference* **3** 1–24.

D'Amour, A., Ding, P., Feller, A., Lei, L. and Sekhon, J. (2020). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics* .

Edelman, B., Ostrovsky, M. and Schwarz, M. (2007). Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review* **97** 242–259.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, vol. 66. CRC Press.

Frölich, M. and Huber, M. (2018). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics* **o** 0–0.

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

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663–685.

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

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics* **142** 675–697.

Rubin, D. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.