Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

March 5, 2025

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Regression Discontinuity Designs

17

"la nature ne fait jamais des sauts."
("nature never makes jumps.")

– Gottfried Leibniz [1].

In this chapter we discuss the Regression Discontinuity Design (RDD). First, we introduce the basic idea of Regression Discontinuity (RD). RDDs, when they exist, offer a highly credible way to identify causal effects. However, leveraging RDDs without covariates can fall short in practice. We show how modern machine learning methods can be utilized for estimation in RDDs with many covariates.

17.1 Introduction

Like many other methods presented in the Advanced Topics – instrumental variables, proxy controls, and difference-in-differences – Regression Discontinuity Designs (RDDs) are widely used in empirical work for measuring causal effects in non-experimental settings where we cannot reliably measure all confounders.

The basic RDD structure relies on a so-called running variable or score which determines treatment: units whose score is above a cutoff value are assigned to the treatment, while units with score below the cutoff are assigned to control. Examples are reward of a scholarship if a student's grade average exceeds a certain threshold, bestowing of license to practice (say, medicine or law) if one's exam score exceeds a threshold, assignment of a particular medical treatment if a biomarker is above a cutoff, or getting social benefits if income is below some income threshold.

We can always negate the running variable or rename the treatment if the relationship is the other way.

The intuition for identification is that units marginally above and below the threshold are comparable in terms of potential outcomes, since they are the same in all ways except the assignment to treatment, assuming of course that there are no other discontinuities at the cutoff that would also render them different in other ways. The latter continuity in potential outcomes is the identifying assumption in RDDs. For example, suppose we are interested in the causal effect of a student receiving a scholarship on their future academic success. While the future academic success of students with low grade averages is very different from those with high averages, with or without a scholarship, the students right at the cutoff essentially have the same grade averages and are comparable. However, those just above the cutoff have a scholarship and those just below do not. We can thus compare those just above the cutoff to those just below to learn the effect of having a scholarship on people at the grade average cutoff.

We can also conceive of being above or below as random "luck," i.e., exogenous variation. E.g., getting just one more question right on the exam might be viewed as a random event that has nothing to do with the academic preparedness of the student – anyone can happen to accidentally guess right on one question. Viewing falling just to one side or the other of a cutoff as a purely is an alternative approach to identification in RDDs based on *local randomization* [2].

17.2 The Basic RDD Framework

Setting

In the *sharp RDD* the binary treatment variable $D_i \in \{0, 1\}$ for individual i is assigned on the basis of a running variable X_i in a deterministic ("sharp") way: $D_i = 1(X_i \geq c)$, where 1 denotes the indicator function and c the cutoff value. That is, a unit is treated ($D_i = 1$) if the value of the running variable is above the threshold and in the control group ($D_i = 0$) otherwise. For each individual, we observe additionally the outcome Y_i and potentially some pre-treatment variables $Z_i \in \mathbb{R}^p$. The observed data $\{W_i\}_{i=1}^n = \{(Y_i, X_i, Z_i)\}_{i=1}^n$ are an i.i.d. sample of size n from the distribution of $W = (Y, X, Z)$. Note that we also observe $D_i = 1(X_i \geq c)$ for each individual because we know the cutoff value c and observe X_i.

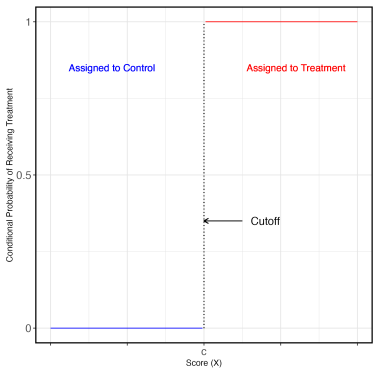

Figure 17.1: In the sharp RDD, the assignment of treatment depends in a deterministic way on the underlying running variable (or score variable). Units with values of the running variable below a cutoff are not treated, while units above the threshold are treated.

The parameter of interest in RDD is the ATE at the cutoff value c:

$$\tau_{\text{RD}} = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right].$$

This parameter can be identified under mild conditions in the RDD context. Learning treatment effects away from the RDD cutoff c generally requires stronger conditions that allow extrapolation.

We now state a simple sufficient condition under which τ_{RD} is identified.

Assumption 17.2.1 (RDD Assumptions [3]) *Suppose that i) the conditional mean of the potential outcomes* $\mathrm{E}\left[Y(t) \mid X = x\right]$ *for* $t \in \{0, 1\}$ *are continuous at the cutoff level* c*, ii) that the density of the running variable,* f_X *near the cutoff is positive –* $f_X(c) > 0$*, and iii) there is no selection on gains local to the cutoff:* $\mathrm{E}[Y(1) - Y(0)|D, X = x] = \mathrm{E}[Y(1) - Y(0)|X = x]$ *for* $x \in (c - \epsilon, c + \epsilon)$ *for some small* $\epsilon > 0$*.*

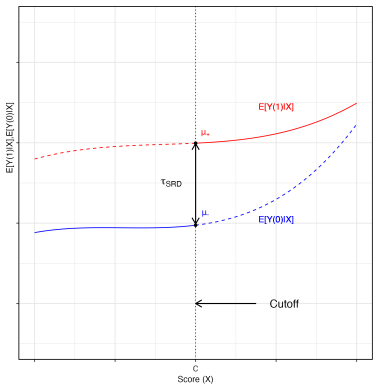

Figure 17.2: Identification and estimation in the sharp RDD.

Under Assumption 17.2.1, we have

$$\tau_{\text{RD}} = \lim_{x \downarrow c} \mathrm{E}\left(Y_i \mid X_i = x\right) - \lim_{x \uparrow c} \mathrm{E}\left(Y_i \mid X_i = x\right)$$

where $\lim_{x \downarrow c}$ and $\lim_{x \uparrow c}$ denote the right-sided and left-sided limit. Hence, the jump in $\mathrm{E}\left(Y_i \mid X_i = x\right)$, the conditional expectation function of the observed outcome, at the threshold determines the causal effect of interest.

Estimation

In sharp RDD, we are faced with the problem of estimating the jump in the conditional mean functions at the cutoff value. As we do not see treated and control observations exactly at the cutoff, this problem boils down to estimation of the conditional mean functions at points to the left and right of, but local to, the cutoff value. In practice, we can estimate the conditional mean at points local to the cutoff using conventional local nonparametric methods. Local polynomial estimation has become the default method for this local estimation task in RDD, and we therefore focus on this method following the notation and exposition in [4] closely.

Standard RDD Estimator: Without covariates, a weighted linear regression of Y_i on X_i is estimated locally around the cutoff to estimate the parameter of interest:

$$\hat{\tau}_{\text{base}}(h) = e_2^\top \underset{\theta \in \mathbb{R}^4}{\operatorname{argmin}} \sum_{i=1}^{n} K_h\left(X_i - c\right)\left(Y_i - \theta' V_i\right)^2 .$$

Here, K denotes a kernel function, $h > 0$ a bandwidth, $K_h(x) = K(x/h)/h$, $V_i = (1, D_i, (X_i - c), D_i(X_i - c))^\top$ a vector of appropriate transformations of the running variable, and $e_2 = (0, 1, 0, 0)^\top$ the unit vector to select the coefficient of D_i, which is the target parameter.

By kernel function, we mean a function that integrates to one and is symmetric around 0. Common examples are the uniform kernel, $K(x) = \frac{1}{2}1(-1 < x < 1)$, and the triangular or Bartlett kernel, $K(x) = (1 - |x|)1(-1 < x < 1)$.

Under standard conditions for local linear regression such as continuity of the running variable and having bandwidth h approach zero at a suitable rate, the estimator $\widehat{\tau}_{\text{base}}(h)$ follows an approximate normal distribution in large samples:

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

In the asymptotic distribution, the term $h^2 B_{\text{base}}$ with

$$B_{\text{base}} = \frac{\bar{\nu}}{2}\left(\partial_x^2 \mathbb{E}\left[Y_i \mid X_i = x\right]\Big|_{x=c^+} - \partial_x^2 \mathbb{E}\left[Y_i \mid X_i = x\right]\Big|_{x=c^-}\right)$$

represents bias, which is of the order h^2. The variance

$$\frac{1}{nh} V_{\text{base}} = \frac{1}{nh}\frac{\bar{\kappa}}{f_X(c)}\left(\mathbb{V}\left[Y_i \mid X_i = c^+\right] + \mathbb{V}\left[Y_i \mid X_i = c^-\right]\right)$$

is of the order of $(nh)^{-1}$. The terms $\bar{\nu}$ and $\bar{\kappa}$ in the bias and variance expressions are constants related to the kernel defined as

$$\bar{\nu} = \left(\bar{\nu}_2^2 - \bar{\nu}_1 \bar{\nu}_3\right) / \left(\bar{\nu}_2 \bar{\nu}_0 - \bar{\nu}_1^2\right)$$

for $\bar{\nu}_j = \int_0^\infty v^j K(v) dv$ and

$$\bar{\kappa} = \int_0^\infty \left(K(v)\left(\bar{\nu}_1 v - \bar{\nu}_2\right)\right)^2 dv \Big/ \left(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2\right)^2 .$$

The choice of the bandwidth h plays an important role in feasible implementation of RDD estimation. MSE optimal choice of h involves equating squared bias and variance which renders conventional confidence intervals invalid. Calonico et al. (2014) [5] considers the use of MSE optimal bandwidths along with bias correction and adjustment to standard errors to account for estimating the bias. Their proposed methods are widely used in practice and available in, e.g., R and python.

17.3 RDD with (Many) Covariates

Motivation for Using Covariates

For identification and estimation of the average treatment effect at the cut-off value, no covariate information is required except the running variable. Of course, additional covariates are available in many applications. As outlined in Cattaneo et al. (2023) [6] provide an extensive discussion of the use of covariates in RDDs. They note there are several reasons that using covariates in RDD settings may be useful:

1. *Efficiency and power improvements:* As in randomized controlled trials, using covariates can increase efficiency and improve power; see, e.g., [7] and [8].
2. *Auxiliary information:* In RDD, the running variable determines the assignment of the treatment, and measurement error in the running variable can distort the results. Additional covariates can help overcome measurement issues and help deal with missing data problems.
3. *Treatment effect heterogeneity:* Covariates can be used to explore heterogeneity in treatment effects. For example, we can adapt methods from Chapter 14 and Chapter 15.
4. *Other parameters of interest and extrapolation:* Conventional RDD without covariates only identifies treatment effects at the cutoff value of the running variable. Making use of additional covariates may allow extrapolation of the treatment effects to other values of the runnign variable away from the cutoff point and may also be useful for identifying other causal parameters.

Low-Dimensional Covariates

There are several ways to adjust the RDD estimator to allow for the presence of covariates. Calonico et al. (2019) [7] provide a detailed analysis of the use of additional regressors in RDD in a setting with relatively few covariates. A transparent approach is to simply include the controls linearly within the objective function defining the baseline local linear RDD estimator. That is, we solve

$$\hat{\tau}_{\text{lin}}(h) = e_2^\top \underset{(\theta,\gamma)\in\mathbb{R}^{4+p}}{\text{argmin}} \sum_{i=1}^{n} K_h\left(X_i - c\right)\left(Y_i - \theta' V_i - \gamma' Z_i\right)^2 \quad (17.3.1)$$

where γ are the coefficients on the p dimensional vector of pretreatment variables, Z_i.

It is clear that the estimator for θ can be equivalently written as a RDD estimator with a covariate-adjusted outcome, $\tilde{Y}_i = Y_i - Z_i^\top \widehat{\gamma}_h$ and no covariates, where $\widehat{\gamma}_h$ is the vector of linear projection coefficients associated with Z_i obtained from solving (17.3.1). That is,

$$\widehat{\gamma}_h = \arg\min_{\gamma} \sum_{i=1}^{n} K_h\left(X_i - c\right)\left(\check{Y}_i - \gamma' \check{Z}_i\right)^2$$

$$\hat{\tau}_{\text{lin}}(h) = e_2^\top \underset{(\theta)\in\mathbb{R}^4}{\text{argmin}} \sum_{i=1}^{n} K_h\left(X_i - c\right)\left(\tilde{Y}_i - \theta' V_i\right)^2.$$

where $\check{W}_i$ is the residual from locally partialling V out from W (with partialling out interpreted elementwise if W is a vector). That is, $\check{W}_i = W_i - \widehat{\zeta}_h' V_i$ with

$$\widehat{\zeta}_i = \arg\min_{b} \sum_{i=1}^{n} K_h\left(X_i - c\right)\left(W_i - b' \check{V}_i\right)^2.$$

$\widehat{\tau}_{\text{lin}}(h)$ is consistent for the conditional average treatment effect at $X = c$, τ_{RD} if the conditional distribution of the regressors given the running variable varies smoothly around the cutoff. As the estimator is just a local linear regression involving the variables V and Z, this results essentially follows immediately from properties of local regression which will hold under mild smoothness conditions without requiring parametric functional form restrictions.Specifically, if $\mathbb{E}\left[Z_i \mid X_i = x\right]$ is twice continuously differentiable around the cutoff, we will have

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

under regularity conditions similar to those for the estimator without covariates. Interestingly, the bias term B_{base} is the same as in the case without including Z_i, but the variance term differs with

$$V_{\text{lin}} = \frac{\bar{\kappa}}{f_X(c)}\left(\mathbb{V}\left[Y_i - Z_i^\top \gamma_0 \mid X_i = c^+\right] + \mathbb{V}\left[Y_i - Z_i^\top \gamma_0 \mid X_i = c^-\right]\right)$$

where γ_0 is a non-random vector corresponding to the probability limit of $\widehat{\gamma}_h$ (see also [8]). Unsurprisingly, the linear adjustment estimator generally has smaller asymptotic variance than the estimator without covariates; i.e. $V_{\text{lin}} \leq V_{\text{base}}$. See [7] and [8] for formal statement of the properties of V_{lin} and discussion of additional potential avenues for inclusion of covariates.

High-Dimensional Covariates

RDD with Lasso

In the case where many covariates are available, one straightforward option is to adopt the procedure from the low-dimensional setting by including all variables linearly inside the local linear regression and then using (local) Lasso regression to estimate the parameters. This procedure has been analyzed by [9] and [8]. Here, we follow [8] closely. The idea is that in a first step the relevant variables are selected with a localized / weighted Lasso regression. In the second step, we then run local linear RDD estimation with the selected covariates from the first step.

Specifically, estimation proceeds as follows:

(Post)-Lasso Estimation in RDD

1. Using a preliminary bandwidth b and a penalty parameter λ, one solves a Lasso version of the local linear regression defining the RDD estimator by adding a penalty term to obtain preliminary estimates

$$\begin{aligned}(\widetilde{\theta},\widetilde{\gamma}) &= \underset{(\theta,\gamma)\in\mathbb{R}^{4+p}}{\operatorname{argmin}} \sum_{i=1}^{n} K_b\left(X_i - c\right)\left(Y_i - V_i^{\top}\theta - \left(Z_i - \hat{\mu}_Z\right)^{\top}\gamma\right)^2 \\ &\qquad + \lambda \sum_{k=1}^{p} \hat{w}_k \left|\gamma_k\right|,\end{aligned}$$

where

$$\hat{\mu}_Z = \frac{1}{n}\sum_{i=1}^{n} Z_i K_b\left(X_i - c\right)$$

and

$$\hat{w}_k^2 = \frac{b}{n}\sum_{i=1}^{n}\left(K_b\left(X_i - c\right) Z_i^{(k)} - \mu_Z^{(k)}\right)^2$$

are the local sample mean and variance, respectively, of the covariates.

2. Let $\hat{J} = \left\{k \in \{1, \ldots, p\} : \widetilde{\gamma}^{(k)} \neq 0\right\}$ denote the set of indices of those covariates whose first step Lasso estimates are non-zero. Using the variables in $\hat{J}$ and a final bandwidth h, we compute our estimate of the treatment effect, τ_{RD}, as $\hat{\tau}_{\hat{J}}(h)$ exactly as in (17.3.1) where the set of covariates is restricted to $\hat{J}$.

[8] provide formal asymptotic properties post-Lasso estimator $\hat{\tau}_{\hat{J}}(h)$. As the estimator fundamentally relies on properties of the Lasso, a key assumption is approximate sparsity – Definition 3.1.1 – of the coefficients on the controls Z_i.

To adapt approximate sparsity to the RDD setting, we define population regression coefficients, $(\theta_0(J, h), \gamma_0(J, h))$, and corresponding residuals, $r_i(J, h)$, for any $J \subset \{1, \ldots, p\}$ and bandwidth h:

$$\begin{aligned}
&(\theta_0(J, h), \gamma_0(J, h)) \\
&\qquad = \underset{(\theta,\gamma)}{\operatorname{argmin}} \, \mathrm{E}\left[K_h\left(X_i - c\right)\left(Y_i - V_i^\top \theta - Z_i(J)^\top \gamma\right)^2\right], \\
&r_i(J, h) = Y_i - V_i^\top \theta_0(J, h) - Z_i(J)^\top \gamma_0(J, h).
\end{aligned}$$

Approximate sparsity then means that there exists a covariate set $J^* \subset \{1, \ldots, p\}$ that contains a "small" number $s \equiv |J^*| \ll p$ of regressors that captures the majority of the local predictive power of the control variables Z_i. We formalize the requirement that variables in J^* capture the majority of the explanatory power by requiring that the local correlation between the regression errors $r_i\left(J^*, h\right)$ and each component of Z_i is small relative to the estimation error:

$$\max_{j=1,\ldots,p} \left|\mathrm{E}\left[K_h\left(X_i - c\right) Z_i^{(j)} r_i\left(J^*, h\right)\right]\right| = O\left(\sqrt{\frac{\log p}{nh}}\right). \quad (17.3.2)$$

To ensure that properties of the estimator are not heavily influenced by the exact bandwidth choice, we further require that (17.3.2) holds uniformly across an appropriate range of bandwidths.

Under this local approximate sparsity condition and other regularity conditions, [8] show that the post-Lasso estimator $\hat{\tau}_{\hat{J}}(h)$ has the same first-order asymptotic properties as an infeasible estimator $\hat{\tau}_{J^*}(h)$ that uses only the subset of variables

indexed by J^*. They then prove an asymptotic normality result for $\hat{\tau}_{J^*}(h)$. Taken together, we then obtain that $\hat{\tau}_{\hat{J}}(h)$ of τ_{RD} satisfies

$$\frac{\sqrt{nh}\left(\hat{\tau}_{\hat{J}}(h) - \tau_{\text{RD}} - h^2\mathcal{B}_n\right)}{\mathcal{S}_n} \xrightarrow{d} \mathcal{N}(0,1),$$

with asymptotic bias and variance, respectively, given by

$$\mathcal{B}_n \approx \frac{C_{\mathcal{B}}}{2}\left(\mu''_{\widetilde{Y}+} - \mu''_{\widetilde{Y}-}\right) \quad \text{and} \quad \mathcal{S}_n^2 \approx \frac{C_{\mathcal{S}}}{f_X(c)}\left(\sigma^2_{\widetilde{Y}+} + \sigma^2_{\widetilde{Y}-}\right).$$

Here $C_{\mathcal{B}}$ and $C_{\mathcal{S}}$ are constants that depend on the kernel function K only, and $\widetilde{Y}_i = Y_i - \gamma_n' Z_i(J_n)$ with

$$\gamma_n = \left(\sigma^2_{Z(J_n)-} + \sigma^2_{Z(J_n)+}\right)^{-1}\left(\sigma^2_{YZ(J_n)-} + \sigma^2_{YZ(J_n)+}\right),$$

is a covariate-adjusted version of the outcome variable that uses a vector γ_n that can be thought of as an approximation of $\gamma_0(J^*, h)$ that is independent of the bandwidth. The estimator is thus first-order asymptotically equivalent to a basic sharp RDD estimator with the covariate-adjusted outcome $\tilde{Y}_i$ replacing the original outcome Y_i

For generic random vectors A and B, $\mu_A(x) = \text{E}(A \mid X = x)$, $\mu_{AB}(x) = \text{E}(AB' \mid X = x)$, $\sigma^2_{AB}(x) = \mu_{AB}(x) - \mu_A(x)\mu_B(x)'$. Denote $\sigma^2_A(x) = \sigma^2_{AA}(x)$ for simplicity. For a generic function g, we also write $g_+ = \lim_{x \downarrow c} g(x)$ and $g_- = \lim_{x \uparrow c} g(x)$ for its right and left limit at the RDD cutoff c, respectively. Thus, $\tau_{\text{RD}} = \mu_{Y+} - \mu_{Y-}$.

RDD with generic ML Methods

As discussed above in the section on using low-dimensional control variables, adjusting for a small number of controls by including them linearly within the local linear RDD estimator is asymptotically equivalent to running a local linear RDD regression with a modified outcome variable $Y_i - Z_i'\gamma$ for appropriately defined γ. Noack et al. (2023) [4] consider flexible inclusion of control variables, allowing for high-dimensional settings, by considering more general outcome variable adjustments of the form $Y_i - \eta_0(Z_i)$ for potentially nonlinear function η_0. That is, they consider employing the conventional, covariate free RDD estimator using the adjusted outcome $Y_i - \eta_0(Z_i)$ in place of Y_i.

Under the assumption that $\text{E}[\eta(Z)|X = x]$ is twice continuously differentiable at points local to $X = c$ for a large class of functions η, different choices of η_0 result in the same estimand for the

adjusted RDD estimator because

$$\begin{aligned}\tau_{RD} = &\lim_{x \downarrow c} \mathrm{E}[Y - \eta_0(Z)|X = x] \\ &- \lim_{x \uparrow c} \mathrm{E}[Y - \eta_0(Z)|X = x]\end{aligned} \tag{17.3.3}$$

for any choice of η_0 withing this class of functions under this assumption. This assumption seems reasonable in settings where treatment can be safely assumed to have no effect on Z, such as scenarios where Z are pretreatment characteristics. However, the choice of η_0 impacts the performance of the RDD estimator by altering the estimator's asymptotic variance. [4] show that the optimal choice of η_0 with regard to the asymptotic variance of the RDD estimator of the treatment effect is the average of the conditional expectation functions of the outcome given the running variables and covariates just to the right and left of the cutoff value.

Given these observations,[4] provide an approach to allow for flexible covariate adjustment in high-dimensional settings using modern machine learning methods to estimate the optimal η_0. They consider a DML style estimator that employs cross-fitting and consists of two steps.

General ML Estimation in RDD [4]

1. Randomly split the data $\{W_i\}_{i \in 1,\ldots,n}$ into S folds of (approximately) equal size, collecting the corresponding indices in the sets I_s, for $s \in 1, \ldots, S$. Let $\widehat{\eta}_s(z) = \widehat{\eta}\left(z; \{W_i\}_{i \in I_s^c}\right)$, for $s \in 1, \ldots, S$ and I_s^c the set of indices of observations not included in fold s, be the researcher's preferred estimator of η_0 calculated using only data from outside the s^{th} fold.

2. Estimate τ_{RD} by computing a local linear "no covariates" RDD estimator that uses the adjusted outcome $\tilde{Y}_i\left(\widehat{\eta}_{s(i)}\right) = Y_i - \widehat{\eta}_{s(i)}(Z_i)$ as the dependent variable, where $s(i)$ denotes the fold that contains observation i. I.e. estimate τ_{RD} as

$$\hat{\tau}_{\hat{\eta}}(h) = e_2^\top \underset{\theta \in \mathbb{R}^4}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i - c)\left(\tilde{Y}_i\left(\widehat{\eta}_{s(i)}\right) - \theta' V_i\right)^2.$$

[4] establish that the estimator $\hat{\tau}_{\hat{\eta}}(h)$ is asymptotically equivalent to the infeasible estimator

$$\widehat{\tau}_{\bar{\eta}}(h) = e_2^\top \underset{\theta \in \mathbb{R}^4}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i - c)\left(\tilde{Y}_i\left(\bar{\eta}_{s(i)}\right) - \theta' V_i\right)^2,$$

where $\bar{\eta}$ is a deterministic approximation of $\widehat{\eta}$ whose error vanishes in large samples in some appropriate sense. It then holds that

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

The asymptotic variance in the above expression is minimized if $\widehat{\eta}$ is consistent for η_0, in the sense that $\bar{\eta} = \eta_0$. However, the distributional approximation is valid even if $\bar{\eta} \neq \eta_0$ because (17.3.3) holds for (essentially) all adjustment functions, not just the optimal one. In that sense, the procedure allows for misspecification in the choice of model for $\hat{\eta}$. Moreover, $V(\bar{\eta})$ is typically smaller than V_{base} even under misspecification. Valid confidence intervals can easily be constructed for τ by applying standard methods developed for settings without covariates to a data set with running variable X_i and outcome $\tilde{Y}_i\left(\widehat{\eta}_{s(i)}\right)$ ignoring sampling uncertainty about the estimated adjustment function.

Heterogeneous Treatment Effects and Adjustments for Heterogeneity

Our treatment of covariates thus far has been focused on using covariates to increase efficiency of the average treatment effect at the cutoff value τ_{RD}. Covariates can also help us understand heterogeneity of treatment effects.

At a conceptual level, it is clear that we can repeat the setup in Section 17.2 conditional on $Z = z$, leading to the *CATE at the cutoff*:

$$\begin{aligned}\tau_{\text{C–RD}}(Z) &= \mathrm{E}[Y(1) - Y(0) \mid Z, X = c] \\ &= \lim_{x \downarrow c} g(x, Z) - \lim_{x \uparrow c} g_0(x, Z),\end{aligned}$$

We can also define GATEs at the cutoff exactly as in Chapter **??**. More generally, one may adapt the material on heterogeneous treatment effects from Chapter 14 and Chapter 15 to the RDD setting.

where $g_0(X, Z) = \mathrm{E}[Y \mid X, Z]$.

A potentially policy-relevant summary of $\tau_{\text{C–RD}}(Z)$ is its average:

$$\tau_{\text{A–C–RD}} = \mathrm{E}[\tau_{\text{C–RD}}(Z)] = \mathrm{E}[\mathrm{E}[Y(1) - Y(0) \mid Z, X = c]].$$

For example, if we were to assume that Z accounts for all treatment effect heterogeneity across values of the running variable, that is, $Y(1) - Y(0) \perp\!\!\!\perp X \mid Z$,[1] then we would conclude that $\tau_{\text{A–C–RD}} = \mathrm{E}[Y(1) - Y(0)]$ is the marginal ATE in the population, not just at the cutoff. More generally, we can say

1: The weaker conditional mean-independence of $Y(1) - Y(0)$ and $1[X = c]$, given Z, suffices, but is perhaps harder to reason about.

that $\tau_{\text{A-C-RD}}$ controls for the heterogeneity modulated by Z, without requiring that Z accounts for all effect heterogeneity.

We can leverage DML to estimate $\tau_{\text{A-C-RD}}$. For $h > 0$, consider a smoothed version of the same parameter:

$$\tilde{\tau}_h = \int_{-\infty}^{\infty} (41[x > c] - 2) K_h(x - c) \mathrm{E}[g_0(x, Z)] \mathrm{d}x.$$

Note that $\lim_{h \to 0} \tilde{\tau}_h = \tau_{\text{A-C-RD}}$ under appropriate continuity of $g_0(x, W)$ near $x = c$ for almost every W. The quantity $\theta_0 = \tilde{\tau}_h$ is a simple linear summary of g_0, similar to those we studied in Chapter 14. We can apply DML to estimate $\theta_0 = \tilde{\tau}_h$ using the Neyman orthogonal score

$$\psi(W; \theta, \eta) = \int_{-\infty}^{\infty} (41[x > c] - 2) K_h(x - c) g(x, Z) \mathrm{d}x + \frac{(41[x > c] - 2) K_h(X - c)}{f(X \mid Z)} (Y - g(X, Z)) - \theta,$$

where $\eta = (g, f)$ are nuisance functions where the population value of f, f_0, is the conditional density of X given Z. Implementing DML in this setting then requires estimation of the conditional expectation function g_0 and the conditional density function f_0.[2]

2: There are a variety of approaches for estimating conditional density functions using machine learning methods. See, for example, [10], [11], and [12].

17.4 Empirical Example

In this section, we examine the effect of the antipoverty program PROGRESA/Oportunidades on the consumption behavior of families in Mexico in the early 2000s using RDD. Data for this application are provided by [5]. We follow [4] in the presentation of the results.

The Notebooks 17.6.1 implement the empirical exercise.

The program was intended for families in extreme poverty and included financial incentives for participation in measures targeted at improving the family's health, nutrition, and children's education. The effect of this program has been widely studied in economics; see, e.g. Parker and Todd (2017) [13].

Eligibility for the program was determined based on a pre-intervention household poverty-index. Individuals above a certain threshold were eligible to receive a cash transfer through the program(and are coded as the treatment group), while individuals below the threshold were excluded and recorded

Table 17.1: RDD Estimates of Effects of Progresa

	Food_1	Non-Food_1	Food_2	Non-Food_2
No Controls	-22.17	-9.13	54.89	43.76
	(27.45)	(21.90)	(48.12)	(32.34)
Linear	-29.36	-6.49	55.59	45.44
	(21.86)	(20.62)	(44.36)	(29.76)
Lasso (Baseline)	-25.65	-2.20	58.32	46.85
	(21.86)	(20.84)	(45.69)	(31.83)
Lasso (Flexible)	-22.58	-1.98	52.32	38.07
	(21.79)	(20.76)	(45.78)	(32.42)
Random Forest	-20.28	-1.44	55.38	36.98
	(22.44)	(21.03)	(45.95)	(31.02)
Boosted Trees	-18.38	-3.74	53.37	43.09
	(21.76)	(20.94)	(46.00)	(31.78)

Note: RDD estimates of treatment effects of Progresa on food and non-food consumption. Columns "Food_1" and "Non-Food_1" provide estimates for the effects one year after implementation of the program, and columns "Food_2" and "Non-Food_2" provide estimates for the effects two years after implementation of the program. Row labels denote the method used to estimate the conditional expectation of the outcomes near the cutoff given the pretreatment variables. Standard errors are provided in parentheses.

as a control group. All observations above the threshold participated in the program, which makes the analysis fall into the standard (sharp) RDD framework.

Because our treatment variable is eligiblity for the program, we are technically estimating the average *intention to treat* effect at the eligibility cutoff.

We look at the outcome variables food and non-food consumption, both one and two years after the implementation of the program. The data set contains 1,944 observations and 27 measured variables including outcomes and pre-treatment demographic and socioeconomic variables. We summarize RDD estimates of the treatment effect on the four outcome variables when we do not include controls, adjust for controls linearly, and adjust for controls using a handful of machine learning methods in Table 17.1.

As the theory predicts, the point estimates obtained across the different methods are quite similar relative to standard errors. We do see that the methods that control for pretreatment variables are somewhat more precise than the baseline estimates that do not use controls, though the method used to adjust for the controls does not seem to have a major impact.

17.5 Notes

The ideas behind RDDs and IVs come together in *fuzzy RDDs*. Whereas in sharp RDDs the treatment assignment is determin-

istic depending on being above or below the cutoff, in fuzzy RDDs the assignment mechanism is assigned at random with a assignment probability that need not be 0 or 1. Nonetheless, as in the sharp case, there is a discontinuity at the cutoff level. Then, for the units in an infinitesimal neighborhood of the cutoff, being just above or just below can be understood as an *instrument* for the treatment, with the assignment probability reflecting the compliance and the size of the discontinuity therein being the strength of the instrument. Almost the same tools for IV can be used once we localize to the cutoff.

Excellent introductions and surveys for RDD are the "classics" [14] and [15]. Updates including recent results are [16], [17], [18] and the monographs [19] and [2].

17.6 Notebooks

Notebook 17.6.1 (RDD) R Notebook for RDD and Python Notebook for RDD provide an analysis of the effect of the antipoverty program Progresa/Opportunidades on the consumption behavior of families in Mexico in the early 2000s.

17.7 Exercises

Exercise 17.7.1 ((Theoretical). RDD) Derive the moment conditions which identify the target parameter in RDD and show that it is orthogonal with regard to covariates.

Exercise 17.7.2 (RDD in Practice) In Israel, there is a strict restriction on the maximum size of public-school classrooms. For several decades in the previous century, the maximum was 40, such that, say, having 81 enrolled in a single grade meant a school has to open three parallel classrooms for that grade so that no one classroom has more than 40 students. Discuss why this structure induces an RDD for the study of the impact of class size on academic performance? Assuming we have the school id, class id, and test scores of each individual student in, say the 5th grade in 1991, how would you construct an RDD: what would be the unit of analysis, the running variable, and the cutoff? How should we interpret the ATE and to what kind of student population might it not be relevant for and why? (Once you have thought about this study question, you can read about the study that famously leveraged this RDD

in [20].)

Bibliography

[1] Gottfried Wilhelm Leibniz. *Nouveaux Essais sur l'entendement humain*. 1765 (cited on page 470).

[2] Matias D. Cattaneo, Nicolas Idrobo, and Rocio Titiunik. 'A Practical Introduction to Regression Discontinuity Designs: Extensions'. In: 2023 (cited on pages 471, 483).

[3] Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. 'Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design'. In: *Econometrica* 69.1 (2001), pp. 201–209. (Visited on 04/19/2024) (cited on page 472).

[4] Claudia Noack, Tomasz Olma, and Christoph Rothe. 'Flexible Covariate Adjustments in Regression Discontinuity Designs'. In: (2023) (cited on pages 473, 478, 479, 481).

[5] Sebastian Calonico, Matias D. Cattaneo, and Rocio Titiunik. 'Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs'. In: *Econometrica* 82.6 (2014), pp. 2295–2326. (Visited on 02/08/2024) (cited on pages 474, 481).

[6] Matias D. Cattaneo, Luke Keele, and Rocio Titiunik. 'Covariate Adjustment in Regression Discontinuity Designs'. In: 2023 (cited on page 474).

[7] Sebastian Calonico, Matias D. Cattaneo, Max Farrell, and Rocío Titiunik. 'Regression Discontinuity Designs Using Covariates'. In: *The Review of Economics and Statistics* 101.3 (2019), pp. 442–451. DOI: 10.1162/rest_a_00760 (cited on pages 474–476).

[8] Alexander Kreiss and Christoph Rothe. 'Inference in regression discontinuity designs with high-dimensional covariates'. In: *The Econometrics Journal* 26.2 (Dec. 2022), pp. 105–123. DOI: 10.1093/ectj/utac029 (cited on pages 474, 476, 477).

[9] Yoichi Arai, Taisuke Otsu, and Myung Hwan Seo. 'Regression Discontinuity Design with Potentially Many Covariates'. In: 2022 (cited on page 476).

[10] Rafael Izbicki and Ann B. Lee. 'Converting high-dimensional regression to high-dimensional conditional density estimation'. In: *Electronic Journal of Statistics* 11.2 (2017), pp. 2800 –2831. DOI: 10.1214/17-EJS1302 (cited on page 481).

[11] Zijun Gao and Trevor Hastie. 'LinCDE: Conditional Density Estimation via Lindsey's Method'. In: *Journal of Machine Learning Research* 23 (2022), pp. 1 –55 (cited on page 481).

[12] Jonas Rothfuss, Fabio Ferreira, Simon Walther, and Maxim Ulrich. *Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks*. 2019 (cited on page 481).

[13] Susan W. Parker and Petra E. Todd. 'Conditional Cash Transfers: The Case of Progresa/Oportunidades'. In: *Journal of Economic Literature* 55.3 (2017), pp. 866–915. DOI: 10.1257/jel.20151233 (cited on page 481).

[14] Guido W. Imbens and Thomas Lemieux. 'Regression discontinuity designs: A guide to practice'. In: *Journal of Econometrics* 142.2 (2008). The regression discontinuity design: Theory and applications, pp. 615–635. DOI: https://doi.org/10.1016/j.jeconom.2007.05.001 (cited on page 483).

[15] David S. Lee and Thomas Lemieux. 'Regression Discontinuity Designs in Economics'. In: *Journal of Economic Literature* 48.2 (2010), pp. 281–355. DOI: 10.1257/jel.48.2.281 (cited on page 483).

[16] Blaise Melly and Rafael Lalive. *Estimation, inference, and interpretation in the regression discontinuity design*. eng. Discussion Papers. Bern, 2020. URL: http://hdl.handle.net/10419/228904 (cited on page 483).

[17] Matias D. Cattaneo and Rocío Titiunik. 'Regression Discontinuity Designs'. In: *Annual Review of Economics* 14.1 (2022), pp. 821–851. DOI: 10.1146/annurev-economics-051520-021409 (cited on page 483).

[18] Matias D. Cattaneo, Luke Keele, and Rocío Titiunik. 'A guide to regression discontinuity designs in medical applications'. In: *Statistics in Medicine* n/a.n/a (2023), pp. 1–31. DOI: https://doi.org/10.1002/sim.9861 (cited on page 483).

[19] Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, 2020 (cited on page 483).

[20] Joshua D Angrist and Victor Lavy. 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement'. In: *The Quarterly Journal of Economics* 114.2 (1999), pp. 533–575 (cited on page 484).

Index