Applied Causal Inference Powered by ML and AI

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

February 28, 2024

Publisher: Online
Version 0.1.1

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

Regression Discontinuity Design

17

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

– Gottfried Leibniz [1].

In this chapter we discuss 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, whether due to lack of observations near the RD or the lack of generalizability away from the RD. We show how modern machine learning methods can be utilized for estimation in RDDs with very many covariates.

17.1 Introduction

Like many other methods presented in the Advanced Materials – IV, proxy controls, and DiD – RDDs are also 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 the 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 averages and are comparable, but those just above have a scholar and those just below do not.

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 is a random event that has nothing to do with the academic preparedness of the student – it can happen to any one. This is an alternative approach to identification in RDDs based on local randomization [2].

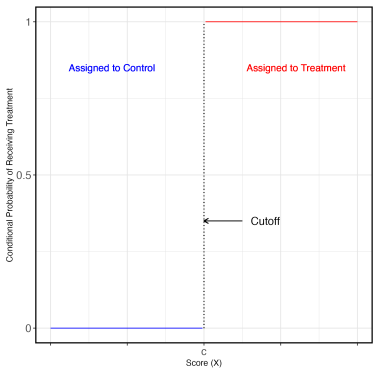

Figure 17.1: In the sharp RDD the assignment of treatment depends in a deterministic way on the underlying . Units with values of the running variable below a cutoff are not treated, while units above the threshold are treated.

Figure 17.2: In the sharp RDD the assignment of treatment depends in a deterministic way on the underlying running variable. Units with values of the running variable below a cutoff are not treated, while units above the threshold are treated.

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 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. An 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 $W = (Y, X, Z)$.

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

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

For identification of this causal effect it is required, that i) the conditional mean functions of the potential outcome $\mathbb{E}\left(Y_i(t) \mid X_i = x\right)$ are continuous at the cutoff level for $t \in \{0, 1\}$ and ii) that the density of the running variable near the cutoff is positive.

Under these conditions we have

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

$\lim_{x \downarrow c}$ and $\lim_{x \uparrow c}$ denote the right-sided and left-sided limit.

Hence, the jump in the conditional expectation functions $\mathbb{E}\left(Y_i \mid X_i = x\right)$ of the observed outcome at the threshold determines the causal effect of interest.

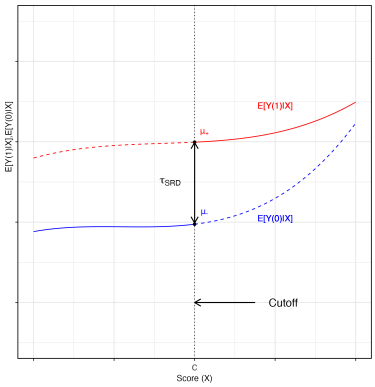

Figure 17.3: Identifictation and estimation in the sharp RDD.

Estimation

In the sharp RDD we are faced with the problem of estimating the jump in the conditional mean functions at the cutoff value which boils down to estimation of the conditional mean functions at the left and right of the cutoff value. For this nonparametric methods, like sieves, kernel, and local polynomials can be used. Local polynomial estimation has become the default method for this, and therefore we will focus on this method following the notation and exposition in [3].

Standard RD 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:

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

K denotes a kernel function, $h > 0$ the bandwidth, $K_h(x) = K(x/h)/h, V_i = (1, D_i, X_i/h, D_i X_i/h)^\top$ 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.

In a setting where standard conditions are met, such as the continuity of the running variable and the bandwidth h approaching zero at a suitable rate, the estimator $\widehat{\tau}_{\text{base}}(h)$ demonstrates an approximate normal distribution in large samples with a bias of the order h^2 and a variance of the order of $(nh)^{-1}$:

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

Bias and variance are given by

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

Here $\bar{v}$ and $\bar{\kappa}$ are kernel constants, defined as $\bar{v} = \left(\bar{v}_2^2 - \bar{v}_1\bar{v}_3\right)/\left(\bar{v}_2\bar{v}_0 - \bar{v}_1^2\right)$ for $\bar{v}_j = \int_0^\infty v^j K(v) dv$ and $\bar{\kappa} = \int_0^\infty \left(K(v)\left(\bar{v}_1 v - \bar{v}_2\right)\right)^2 dv / \left(\bar{v}_2\bar{v}_0 - \bar{v}_1^2\right)^2$, and f_X denotes the density of X_i.

RDD with Covariates: In empirical work, covariates (pretreatment variables) Z_i are often available that could also be included in the analysis. This is analogous to randomized control trials, where additional covariates can reduce the variance of the estimator and usually do not effect the point estimate. There are several ways how to adjust the RD estimator for covariates. [4] analyse in detail the use of additional regressors in RDD. The standard approach is simply to take up the regressors in the weighted least squares regression. The modified estimator is given by:

$$\hat{\tau}_{h,\text{adj}} = e_2^\top \underset{(\theta,\gamma)\in\mathbb{R}^{4+p}}{\operatorname{argmin}} \sum_i^n K_h\left(X_i\right)\left(Y_i - V_i^\top\theta - Z_i^\top\gamma\right)^2. \quad (17.2.1)$$

Z_i denotes the vector of covariates and γ the corresponding coefficient vector.

An important insight is, that the estimator can be equivalently written as a RD estimator without covariates, but with a

covariate-adjusted outcome, $Y_i - Z_i^\top \widehat{\gamma}_h$, where $\widehat{\gamma}_h$ is the vector of linear projection coefficients. The adjusted estimator is then given by:

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

with data-dependent weights $w_i(h)$ which depend only on the realizations of the running variable.

[4] show that $\widehat{\tau}_{\text{lin}}(h)$ is consistent for the RD parameter if the conditional distribution of the regressors given the running variable varies smoothly around the cutoff. The surprising part is that no functional form assumptions on the underlying conditional expectations are required.

Specifically, if $\mathbb{E}[Z_i \mid X_i = x]$ is twice continuously differentiable around the cutoff, then

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

under regularity conditions similar to those for the estimator without covariates, where the bias term B_{base} is as above and the new variance term is

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

with γ_0, a non-random vector of projection coefficients, the probability limit of $\widehat{\gamma}_h$ (see also [5]).

The linear adjustment estimator generally has smaller asymptotic variance than the estimator without covariates, i.e. $V_{\text{lin}} \leq V_{\text{base}}$ which was shown in [5]. See also the discussions in [4].

17.3 RDD with (Many) Covariates

Motivation for Using Covariates

For the identification and estimation of the average treatment effect at the cut-off value no covariate information is required except the running variable, but nevertheless in many applications additional covariates are collected, which might be exploited

for the analysis. Following [6], the use of covariates is beneficial for:

1. *Efficiency and power improvements:* Similar as in randomized control trials, using covariates can increase efficiency and improve power, as we discussed in the previous section. [7] show that the inclusion of covariates in a local polynomial analysis (additional to the score) can lead to asymptotic efficiency gains, if carefully implemented.
2. *Auxiliary information:* In RDD the score determines the assignment of the treatment and measurement errors in the running variable can distort the results. Additional covariates can be exploited to overcome these issues or deal with missing data problems.
3. *Treatment effect heterogeneity:* Covariates can be used to define subgroups in which the treatment effects differ.
4. *Other parameters of interest and extrapolation:* As the identified treatment effect in RDD is local at the cutoff, additional covariates might help for extrapolation of the treatment effects or identify other causal parameters.

For an extensive discussion of the use of covariates in RDDs we refer to [6].

High-Dimensional Covariates

RDD with LASSO estimation

In the case where many covariates are potentially included in the local polynomial regression of the RDD, Lasso can be used for variable selection. This has been analyzed by [8] and [5]. Here we follow [5] closely. The idea is that in a first step the relevant variables are selected with a localized / weighted Lasso regression. In the second step, the local linear RDD estimation with the selected covariates from the first step is conducted. In detail, the procedure is given by:

1. Using a preliminary bandwidth b and a penalty parameter λ, one solves the following Lasso version of the weighted least squares problem by adding a penalty term:

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

where

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

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

2. Using a final bandwidth h, one computes the restricted post-Lasso estimate of τ_{RD} as $\hat{\tau}_h\left(\hat{J}\right)$ as in 17.2.1, where $\hat{J} = \left\{k \in \{1,\ldots,p\} : \widetilde{\gamma}^{(k)} \neq 0\right\}$ is the set of the indices of those covariates selected in the first step.

Results:

A key assumption, which is widely used for studying Lasso, is an approximate sparsity condition, which has been already discussed earlier in this book. To state and adapt this assumption more formally, the following population regression coefficients and corresponding residuals for any $J \subset \{1,\ldots,p\}$ and bandwidth h are defined:

$$(\theta_0(J,h), \gamma_0(J,h)) = \underset{(\theta,\gamma)}{\text{argmin}}\mathbb{E}\left[K_h(X_i)\left(Y_i - V_i^\top\theta - Z_i(J)^\top\gamma\right)^2\right],$$
$$r_i(J,h) = Y_i - V_i^\top\theta_0(J,h) - Z_i(J)^\top\gamma_0(J,h).$$

Approximate sparsity then means that there exist covariate sets $J \subset \{1,\ldots,p\}$ that contain a "small" number $s \equiv |J| \ll p$ of regressors, and are such that the local correlation between the corresponding regression errors $r_i(J,h)$ and each component of Z_i is small relative to the estimation error:

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

Moreover, this condition needs to be satisfied for an appropriate range of bandwidths, so that the sequence J does not depend on the exact choice of h.

Under this and other regularity conditions, [5] can show that the post-Lasso estimator $\hat{\tau}_h\left(\hat{J}\right)$ has the same first-order asymptotic properties as an infeasible estimator $\hat{\tau}_h(J)$ that uses the true target set, and then prove an asymptotic normality result for the latter. Taken together, this yields the main result of [5], which is that the post-Lasso estimator $\hat{\tau}_h\left(\hat{J}\right)$ of τ_{RD} satisfies

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

with asymptotic bias and variance, respectively, such that

$$\mathcal{B}_n \approx \frac{C_{\mathcal{B}}}{2}\left(\mu''_{\widetilde{Y}+} - \mu''_{\widetilde{Y}-}\right) \quad \text{and} \quad \mathcal{S}_n^2 \approx \frac{C_{\mathcal{S}}}{f_X(0)}\left(\sigma^2_{\widetilde{Y}+} + \sigma^2_{\widetilde{Y}-}\right).$$

Here $C_{\mathcal{B}}$ and $C_{\mathcal{S}}$ are constants that depend on the kernel function K only, and

$$\widetilde{Y}_i = Y_i - Z_i\left(J_n\right)^\top \gamma_n, \text{ with } \gamma_n = \left(\sigma^2_{Z(J_n)-} + \sigma^2_{Z(J_n)+}\right)^{-1}\left(\sigma^2_{YZ(J_n)-} + \sigma^2_{YZ(J_n)+}\right),$$

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 "baseline" sharp RD estimator with the covariate-adjusted outcome $\widetilde{Y}_i$ replacing the original outcome Y_i

Here we used the following notation: For generic random vectors A and B, we use the notation that $\mu_A(x) = \mathbb{E}(A \mid X = x)$, $\mu_{AB}(x) = \mathbb{E}\left(AB^\top \mid X = x\right)$, $\sigma^2_{AB}(x) = \mu_{AB}(x) - \mu_A(x)\mu_B(x)^\top$; and write $\sigma^2_A(x) = \sigma^2_{AA}(x)$ for simplicity. For a generic function f, we also write $f_+ = \lim_{x\downarrow 0} f(x)$ and $f_- = \lim_{x\uparrow 0} f(x)$ for its right and left limit at zero, respectively, so that $\tau_{\text{RD}} = \mu_{Y+} - \mu_{Y-}$.

RDD with generic ML Methods

As mentioned above, instead of using covariates in the weighted least squares regression (linear adjustment estimator), [4] show that it is asymptotically equivalent to run a local linear RD regression with a modified outcome variable $Y_i - Z_i'\gamma$ with a projection coefficient γ. [3] argue that this approach can be extended to allow for more general modifications of the form $Y_i - \eta_0(Z_i)$ for any function η_0. Different choices of η_0 give the same estimand, since treatment has no effect on Z, but may change the performance of an estimator based on such a modified centered outcome variable. The optimal choice of η_0 with regard to the asymptotic variance 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. In fact, that we get the same estimand for any η_0

also means that we must be insensitive to any errors in η_0, that is, we have Neyman-orthogonality. Thanks to this, by using a DML procedure, modern machine learning methods can then be used to estimate the function η_0 (especially, the optimal one) in a first step and then the modified outcome is used in a local RDD regression as second step, all with cross-fitting to ensure independence between the steps.

[3] extend the approach of [4] to allow for flexible covariate adjustment in high-dimensional settings using modern machine learning methods. The estimator they propose employs cross-fitting and consists of two steps:

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

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

$$\widehat{\tau}(h;\widehat{\eta}) = \sum_{i=1}^{n} w_i(h) M_i\left(\widehat{\eta}_{s(i)}\right).$$

[3] establish that the estimator $\widehat{\tau}(h;\widehat{\eta})$ is asymptotically equivalent to the infeasible estimator $\widehat{\tau}(h;\bar{\eta}) = \sum_{i=1}^{n} w_i(h) M_i(\bar{\eta})$ that uses the variable $M_i(\bar{\eta})$ as the outcome, where $\bar{\eta}$ is a deterministic approximation of $\widehat{\eta}$ whose error vanishes in large samples in some appropriate sense. It then holds that

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

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 the moment condition (3.2) holds for (essentially) all adjustment functions, and not just the optimal one. In that sense, the procedure allows for misspecification in the first stage. Moreover, even under misspecification $V(\bar{\eta})$ is typically smaller than V_{base}. 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 $M_i\left(\widehat{\eta}_{s(i)}\right)$, ignoring sampling uncertainty about the estimated adjustment function.

Heterogeneous Treatment Effects and Adjustments for Heterogeneity

So far we have used covariates in order to increase efficiency for the same estimand, τ_{RD} that was defined in their absence. Covariates, however, can also help us understand and and control for heterogeneity. In particular, at a conceptual level, we can repeat the setup in Section 17.2 for (almost) every stratum $Z = z$, leading to the *CATE at the cutoff*:

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

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

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

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

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

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

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

$$\tilde{\tau}_h = \int_{-\infty}^{\infty} (4\mathbb{I}[x > c] - 2) K_h(x - c) \text{E}[g_0(x, Z)] \text{d}x,$$

where $K_h(x) = K(x/h)/h$ for a kernel K. Note that under appropriate continuity of $g_0(x, W)$ near but not at $x = c$, for almost every W, we have that $\lim_{h \to 0} \tilde{\tau}_h = \tau_{\text{A-C-RD}}$. The quantity $\theta_0 = \tilde{\tau}_h$ is a simple linear summary of g_0, similar to those we studied in Chapter 10. We can then apply DML to estimate it

using the Neyman orthogonal-score

$$\psi(W;\theta,\eta) = \int_{-\infty}^{\infty} (4\mathbb{1}[x > c] - 2)K_h(x-c)g(x,Z)\mathrm{d}x + \frac{(4\mathbb{1}[x > c] - 2)K_h(X-c)}{f(X \mid Z)}(Y - g(X,Z)) - \theta,$$

where $\eta = (g, f)$ are the nuisances, with the true value for the latter nuisance being f_0, the conditional density of X given Z.

Note we can do this for every h, with our MSE to $\tau_{\text{A-C-RD}}$ (up to $o_p(1/n)$) consisting of the variance $\mathrm{E}[\psi(W;\tilde{\tau}_h,\eta_0)^2]/n$ and the squared bias $(\tau_{\text{A-C-RD}} - \tilde{\tau}_h)^2$. What remains is to choose h to balance the two. Depending on the smoothness of g_0, we can further reduce the bias by using higher-order kernels (see [9]) or leveraging higher-order local-polynomial regression (instead of the local-constant regression used to define $\tilde{\tau}_h$ above). Depending on how much we can drive the bias down, we can achieve a better MSE rate.

17.4 Empirical Example

In this section, the effect of the antipoverty program Progresa/Opportunidades on the consumption behavior of families in Mexico in the early 2000s is analyzed. The analysis is accompanied by two notebooks.

The program was intended for families in extreme poverty and included financial incentives for participation in measures that improved the family's health, nutrition and children's education. The effect of this program is a widely studied problem in social and economic sciences and, according to the WHO, was a very successful measure in terms of reducing extreme poverty in Mexico.

Eligibility for the program was determined based on a pre-intervention household poverty-index. Individuals above a certain threshold received treatment (participation in the program), while individuals below the threshold were excluded and recorded as a control group. All observations above the threshold participated in the program, which makes the analysis fall into the standard (sharp) regression discontinuity design.

Data for this application are provided by [10] and in the presentation of the results we follow [3].*

* Links to notebooks for a replication are provided in the Notebook section.

Outcome variables are food and non-food consumption, one year and two years after the implementation of the program. The treatment variables is defined as eligibility for the cash transfer (intention-to-treat analysis). The data set contains 1,944 observations and 85 socio-economic pre-treatment variables like household size, gender, years of education and information on the house. Without considering pre-treatment variables participation in the program reduced food consumption by 22.1 units in the year following the intervention. With including additional pre-treatment variables and using ML methods for estimation, the point estimates for the effect of the program remain almost unchanged, but the confidence intervals are different. For an in-depth discussion of the results, we refer to the notebooks.

Without any covariate adjustments the effect of the cash transfer on food consumption one year after the program was introduced is estimated with −18.6 (s.e. 16.6). Utilizing linear adjustments for the covariates leads to an estimate of −14.8 and a reduced variance of 13.7. Using machine learning methods for the adjustment leads to to estimates of the effect between −16.0 and −21.5 and to a reduction of the standard errors compared to the baseline model (standard errors between 14 and 16). Notably, zero is contained in all confidence intervals (95% confidence level).

Notebooks

- Python notebook for RDD provides an analysis of the effect of the antipoverty program Progresa/ Oportunidades on the consumption behavior of families in Mexico in the early 2000s.

- R notebook version for RDD

Notes

The ideas behind RDDs and IVs come together in *fuzzy RDDs*. Whereas in sharp RDDs the treatment assignment is deterministic 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" [11] and [12]. Updates including recent results are [13], [14], [15] and the monographs [16] and [2].

Study Problems

1. Derive the moment conditions which identify the target parameter in RDD and show that it is orthogonal with regard to covariates.

2. 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 does this 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 [17].)

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, 482).

[3] Claudia Noack, Tomasz Olma, and Christoph Rothe. *Flexible Covariate Adjustments in Regression Discontinuity Designs*. 2023 (cited on pages 472, 477, 478, 480).

[4] Sebastian Calonico, Matias D. Cattaneo, and Max Farrell. '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 473, 474, 477, 478).

[5] 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).

[6] Matias D. Cattaneo, Luke Keele, and Rocio Titiunik. *Covariate Adjustment in Regression Discontinuity Designs*. 2023 (cited on page 475).

[7] M. D. Cattaneo, M. H. Farrell, and Y. Feng. 'Large Sample Properties of Partitioning-Based Series Estimators'. In: *ArXiv e-prints* (Apr. 2018) (cited on page 475).

[8] Yoichi Arai, Taisuke Otsu, and Myung Hwan Seo. 'Regression Discontinuity Design with Potentially Many Covariates'. In: 2022 (cited on page 475).

[9] Bruce E Hansen. 'Exact mean integrated squared error of higher order kernel estimators'. In: *Econometric Theory* 21.6 (2005), pp. 1031–1057 (cited on page 480).

[10] 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 page 480).

[11] 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 482).

[12] 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 482).

[13] 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 482).

[14] 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 482).

[15] 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 482).

[16] 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 482).

[17] 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 482).