Stat Papers (2017) 58:1217–1246
DOI 10.1007/s00362-016-0745-z

REGULAR ARTICLE

Regression discontinuity: review with extensions

Jin-young Choi[1] · Myoung-jae Lee[2]

Received: 30 June 2015 / Revised: 4 January 2016 / Published online: 5 February 2016

Abstract In treatment effect analysis, often the treatment takes a particular structure: 'on' if an underlying continuous variable crosses a threshold, and 'off' otherwise. Such a treatment occurs in various institutional settings such as a test score crossing a threshold to graduate, or income falling below a threshold to qualify for an aid. In this kind of cases, the study design is called 'regression discontinuity (RD)', which is popular in analyzing observational data, as long as the treatment takes the required form. This paper reviews RD to convey its essentials, and provides some extensions. First, the main RD idea based on *local randomization due to an institutional/legal break* is introduced. Second, treatment effects identified by RD are explored. Third, popular RD estimators are reviewed. Fourth, main specification tests are examined. Fifth, special RD topics are reviewed. Also, an empirical illustration is provided.

Keywords Regression discontinuity · Instrumental variable estimator · Nonparametrics

Mathematics Subject Classification 62G08 · 62G99

✉ Myoung-jae Lee
myoungjae@korea.ac.kr

Jin-young Choi
Choi@econ.uni-frankfurt.de

[1] Faculty of Economics and Business Administration, Goethe University, Frankfurt, Germany

[2] Department of Economics, Korea University, Seoul 02841, South Korea

1 Introduction

Randomization has been the golden rule of inference in statistics. In an experiment with a treatment group ($D = 1$) and a control group ($D = 0$), randomization of D assures that the two groups are different only in the treatment status and balanced in all covariates, observed or not. For instance, persons with different levels of ability or different types of genes are assigned to different groups, but the distribution of ability levels or gene types becomes almost the same across the two groups. Hence $E(Y|D = 1) - E(Y|D = 0)$ for a response variable Y reveals the mean effect of D on Y.

But randomization cannot be done if the treatment is possibly harmful as in smoking or radiation. Also, randomization is unthinkable in most observational studies, which has been accepted as a fact for a long time. Recently, it has been realized that regression discontinuity (RD), started long ago by Thistlethwaite and Campbell (1960), offers 'local randomization' using an institutional or legal break around a cutoff: $D = 1$ if an underlying continuous variable crosses a cutoff, and $D = 0$ otherwise. Many policy/program/treatment variables take this form; e.g., a test score crossing a cutoff to graduate from a school, a vote proportion crossing 0.5 to win an election, age crossing a cutoff to retire, etc.

RD has been steadily rising as a main vehicle of inference for observational data in social sciences as long as D takes the aforementioned form. The goal of this paper is to convey the essentials of RD in a nontechnical and concise fashion. Earlier reviews on RD can be seen in Van der Klaauw (2008), Imbens and Lemieux (2008) and Lee and Lemieux (2010), and this paper updates the reviews with emphasis on recent developments since the existing reviews. Although purported to be a review, this paper provides some extensions as well.

Despite the popularity of RD in social sciences, the RD literature in statistics has been nearly nonexistent as pointed out by Cook (2008), with exceptions being Berk and Rauma (1983), Berk and de Leeuw (1999) and Battistin and Rettore (2002). Robbins and Zhang (1991) examined the RD form of D, calling it a " biased allocation" of treatment, although they did not implement RD, whereas Hansen (2008) just mentioned RD in passing. Nevertheless, RD studies have been appearing slowly in statistics journals in recent years: Cattaneo et al. (2015) for randomized inference when large sample inference is inappropriate, Calonico et al. (2015) for an optimal way to draw RD data plots, and Angrist and Rokkanen (2015) for RD identification away from the cutoff. In addition to these theoretical contributions, several mostly applied works are available as well: Mealli and Rampichini (2012), Crawford et al. (2014), Dickens et al. (2014), Keele et al. (2015), and MacDonald et al. (2016). Undoubtedly, many more will come in future both on the theoretical and applied fronts in statistics, which makes the review in this paper timely for the statistics community.

Before proceeding further, some words on notation are needed. In RD, although there are exceptions, both identification and estimation are done only locally around a known cutoff c of a *'running/forcing/assignment variable'* S. As S can be recentered as $S - c$ always, we set $c = 0$ unless otherwise necessary, and denote a local neighborhood of 0 as $(-h, h)$ for a small positive *bandwidth* h; everything will be local in RD unless otherwise mentioned. Various functions of S will appear and their (dis-) continuity matters only at $S = 0$, and we will thus often omit the qualifier 'at $S = 0$'. Although the name running/forcing/assignment variable is well established in the RD literature, since it will appear very often in the remainder of this paper, we will call it just '*score*' (S for score). Throughout the paper, $1[A] = 1$ if A holds and 0 otherwise. A realized value of S will be denoted as s, and $E(\cdot|S = s)$ will be often written just as $E(\cdot|s)$.

The rest of this paper is organized as follows. Section 2 introduces RD main ideas and features, the details of which will be seen in the remaining sections. Section 3 discusses RD identification. Section 4 examines RD estimators, and Sect. 5 reviews specification tests. Section 6 collects RD topics; many recent theoretical developments can be found here. Section 7 provides an empirical illustration applying some of the introduced methods. Finally, Sect. 8 concludes.

2 RD main ideas and features

In RD, the main variables are (D, Y, S), and

$$Z \equiv 1[0 \leq S].$$

The following examples will help understanding what these variables are.

Example 1 Effect of entering a college on income, with the entrance determined solely by a normalized test score equal to or greater than 0; S is the test score, $D = Z$, and Y is income.

Example 2 Effect of schooling on health, where an educational law dictates that only persons with birth date $\geq c$ are subject to extra years of education in principle, but not everybody obeys the law; S is birth date minus c, D ($\neq Z$) is schooling years, and Y is health level.

To understand RD local randomization in Example 1, consider using $E(Y|D = 1) - E(Y|D = 0)$ to find the effect of D on Y. The problem is that the treatment group has higher test score (thus higher ability) individuals than the control group, so that $E(Y|D = 1) \neq E(Y|D = 0)$ may happen even if D has no true effect; D is not randomized and the two groups are systematically different (in ability). But the local treatment group with S at or just above 0 (i.e., $S \in [0, h)$) and the local control group with S just below 0 ($S \in (-h, 0)$) should have almost the same ability levels. For instance, with SAT score, $c = 1200$ (S is SAT score minus c) and $h = 1$, imagine the local treatment group with $S = 1$ and control group with $S = -1$. Since two point SAT difference is almost no difference (i.e., getting two points more/less is almost a pure luck), the two local groups would be different only in D and balanced in all covariates, which is a local randomization.

Example 1 has $Z = 1[0 \leq S]$, but depending on each RD case, Z may take the opposite form $1[S \leq 0]$; e.g., S is family income, and '$S \leq 0$' is the eligibility condition for an income aid program D. Also, as Example 2 shows, D may not be binary.

Although $D = Z$ in Example 1, college admission in reality would depend, not just on S, but also on other variables, say ε, so that $D \neq Z$. RD with D determined only by S (thus $E(D|S) = D$) is ‘ *sharp RD (SRD)*’ where $D = Z$ is a prime example; $D \neq Z$ is still possible in SRD as in $D = ZS$. RD with D determined by (S, ε) (thus $E(D|S) \neq D$) is ‘*fuzzy RD (FRD)*’; we may write $D = D(S, \varepsilon)$. In FRD, D is a “ fuzzy version” of Z. Although there are FRD’s with a non-binary D as in Example 2, there is little loss of generality in considering only binary D because there are essentially only two levels of S around 0: the treatment group with $S = h$ and the control group with $S = -h$.

In RD, D matters for Y through the ‘degree of sharpness’

$$E(D|0^+) - E(D|0^-) \quad \text{where} \quad E(D|0^+) \equiv \lim_{s\downarrow 0} E(D|s) \;\&\; E(D|0^-) \equiv \lim_{s\uparrow 0} E(D|s).$$

For FRD with a binary D, the degree of sharpness is less than one, whereas it is one in its SRD version. So far, we explained what RD looks like, and how fuzzy RD differs from sharp RD. In the rest of this section, we present the main features of RD.

RD local randomization requires a break of $E(D|S)$ at $S = 0$, differently from the usual randomization (e.g., flipping a coin) without S. In Example 1, test score S may affect income Y directly as well as indirectly through D because S has an ability component. Suppose

$$E(Y|S) = \beta_d E(D|S) + m(S) \tag{2.1}$$

where β_d is the college entry effect and $m(S)$ is the direct impact of S on Y. But $m(S)$ can be ignored because $m(S)$ should be continuous at $S = 0$ (to change little as S crosses 0) whereas $E(D|S)$ is not: there is no reason for the direct impact of ability on Y to be discontinuous at $S = 0$. *It is this contrast between the break* (i.e., *discontinuity*) *of $E(D|S)$ and no break* (i.e., *continuity*) *of $m(S)$ at $S = 0$ that identifies $\beta_d$in* (2.1).

For covariates W, a parameter β_w, an error U, and i indexing individuals, suppose

$$Y_i = \beta_d D_i + W_i'\beta_w + U_i, \quad i = 1, \ldots, N. \tag{2.2}$$

(2.1) holds with $m(S) = E(W'|S)\beta_w + E(U|S)$ if $E(W'|S)\beta_w + E(U|S)$ is continuous at $S = 0$. *Since $E(U|S)$ in $m(S)$ is allowed to be a non-trivial continuous function of S, RD is robust to the “ smooth” endogeneity of D through S*, and all smooth effects of covariates (observed or not) on Y can be ignored in RD as they can be buried in $m(S)$. Hence, *it is enough in RD to consider only (S, Z, D, Y) with no concern for the functional form issue on how covariates such as W enter the Y equation.*

In FRD with $D = D(S, \varepsilon)$, if we try to estimate β_d by the least squares estimator (LSE) of Y on (D, W) using (2.2), then D may be endogenous due to $COR(\varepsilon, U) \neq 0$; this endogeneity of D through ε is different from that through S. But *there is an “ automatic instrument” Z for D in FRD*, with which instrumental variable estimator (IVE) can be applied. If (2.1) instead of (2.2) is used for β_d estimation, however, then the endogeneity issue of D due to $COR(\varepsilon, U) \neq 0$ is mute because $E(D|S)$, not D itself, appears in (2.1).

3 Identification

This section addresses RD identification issues. But, before discussing identification, we have to make sure that D represents a treatment of interest. Whereas there is no problem in general to call D a treatment of interest in FRD, multiple treatments can occur together when $D = Z$ in SRD. For instance, if S is age and $c = 65$, then one may become eligible for several public assistance programs by turning 65. In this case, unless there is an extra variable such as ε to characterize D as in FRD, we are bound to find the effect of the combined treatment defined as the interaction of those multiple programs at $S = 65$. We will thus proceed from the premise that D represents a single treatment of interest.

3.1 Ratio identification

In an " abstract" formation, RD refers to D and Y related through (2.1). For SRD, (2.1) becomes $E(Y|S) = \beta_d D + m(S)$ which is a semi-linear model because $m(S)$ is an unknown (i.e., nonparametric) function. Take $\lim_{s\downarrow 0}$ and $\lim_{s\uparrow 0}$ on (2.1):

$$E(Y|0^+) = \beta_d E(D|0^+) + m(0^+) \text{ and } E(Y|0^-) = \beta_d E(D|0^-) + m(0^-).$$

Subtract the latter from the former, and then solve the difference for β_d:

$$\beta_d = \frac{E(Y|0^+) - E(Y|0^-)}{E(D|0^+) - E(D|0^-)}; \tag{3.1}$$

the unknown $m(\cdot)$ drops out due to $m(0^+) = m(0^-)$. The break of $E(D|S)$ at $S = 0$ ensures a non-zero denominator.

The *RD ratio identification of* β_d*in* (3.1) avoids the unknown $m(S)$ by invoking the continuity of $m(S)$ at $S = 0$. This identification result thus applies only to $S = 0$, and the identification of single-point effect is thought as a serious limitation of RD. But, if a model such as (2.2) holds, then β_d found at a single point is enough. We will examine this single-point identification issue further below. A big question for (3.1) is what happens if the continuity assumption of $m(S)$ is violated, which is to be discussed in the next subsection.

A better understanding of the ratio (3.1) would come from the following structural forms (SF): for an error term U', a parameter α_z and an unknown function $\mu(S)$ continuous at $S = 0$, suppose

$$Y = \beta_d D + m(S) + U' \text{ and } D = \alpha_z Z + \mu(S) + \varepsilon \text{ with } E(\varepsilon|S) \text{ continuous at } 0. \tag{3.2}$$

Substituting the D SF into the Y SF gives the Y reduced form (RF):

$$Y = \beta_d\{\alpha_z Z + \mu(S) + \varepsilon\} + m(S) + U' = \beta_d \alpha_z Z + \{\beta_d \mu(S) + m(S)\} + (\beta_d \varepsilon + U'). \tag{3.3}$$

The D SF in (3.2) gives $E(D|0^+) - E(D|0^-) = \alpha_z$, and the Y RF (3.3) gives $E(Y|0^+) - E(Y|0^-) = \beta_d \alpha_z$. Therefore (3.1) states nothing but $\beta_d = \beta_d \alpha_z / \alpha_z$, with α_z playing the role of a scaling factor in case $E(D|0^+) - E(D|0^-) \neq 1$. This shows that the ratio identification can be viewed as an *indirect identification*; in contrast, if β_d is identified in the Y SF as in (3.2), it would be a *direct identification*.

3.2 Identified effects when continuity fails

One might wonder what happens if $m(S)$ has a break at $S = 0$, contrary to the continuity assumption. In this case, β_d has to be redefined as $\mathring{\beta}_d$ to restore the continuity of $m(S)$:

$$\mathring{\beta}_d \equiv \beta_d + \frac{m(0^+) - m(0^-)}{E(D|0^+) - E(D|0^-)} \{= \beta_d + m(0^+) - m(0^-) \text{when} D = Z\}; \quad (3.4)$$

$\mathring{\beta}_d$ absorbs the $m(S)$ break magnitude. Call β_d the '*net effect*', and $\mathring{\beta}_d$ the '*gross effect*'. The proof for (3.4) in the appendix and much of this subsection are new in the RD literature.

Identifying the gross effect $\mathring{\beta}_d$ would be fine if $\mathring{\beta}_d$ per se is of interest; otherwise, it is a failure of identification. To see this point, consider SRD with $D = Z$, and imagine a covariate A lurking in $m(S)$ with $E(A|S)$ discontinuous at $S = 0$. In this case, $\mathring{\beta}_d = \beta_d + E(A|0^+) - E(A|0^-)$. If A is a post-treatment variable, then $\mathring{\beta}_d$ may be viewed as the total effect consisting of the 'direct effect' β_d and the 'indirect effect' $E(A|0^+) - E(A|0^-)$ of D through A; e.g., A may be an interaction with Z, say, $A = ZA'$ for a variable A' so that $E(A|0^+) - E(A|0^-) = E(A'|0^+)$. If A is a pre-treatment variable, however, then the indirect effect interpretation is inappropriate and the RD identification fails. In this case, we should separate β_d from $\mathring{\beta}_d$ by explicitly using A as a regressor as in $Y = \beta_d Z + \beta_a' A + error$.

For a covariate W *not* causing a break in $m(S)$ because $E(W|S)$ is continuous at 0, there is a trade-off in accounting for W or burying W in $m(S)$. Accounting for W has the advantage of reducing the model error term variance, but if the W part is misspecified, then the model error term variance may not go down because the misspecification error is added, although such a misspecification does not make the RD estimator inconsistent.

Differently from an observed covariate, an unobserved covariate (i.e., an error term) is allowed to influence Y only smoothly through $E(U|S)$ in $m(S)$. This may sound restrictive, but RD's ability to allow $E(U|S)$ to be a non-trivial function of S is a big advantage over other study designs that require $E(U|S)$ to be a constant. Not just this, RD allows D to be endogenous through ε in FRD; in (3.2), $D = D(S, \varepsilon)$ is allowed to be endogenous through $COR(\varepsilon, U') \neq 0$ as well as through $COR(S, U') \neq 0$ because Z can instrument for D.

4 Estimators

Rewrite (2.1) as

$$Y = \beta_d D + m(S) + e, \;\; e \equiv Y - E(Y|S) - \beta_d \{D - E(D|S)\} \left[\Longrightarrow E(e|S) = 0 \right] \quad (4.1)$$

where $m(S)$ is to be replaced by a (piecewise) polynomial function continuous at $S = 0$. Since $E(e|S) = 0$, LSE can be applied for SRD to (4.1). For FRD with $D = D(S, \varepsilon)$, if D is endogenous through ε, then IVE can be applied with Z instrumenting for D. Hence, there are LSE and IVE for RD estimation, and both LSE and IVE are equivalent to sample versions of (3.1), as will be seen shortly.

4.1 LSE for exogenous treatment

Since $E(e|S) = 0$ in (4.1), S and SZ are exogenous regressors in (4.1). Suppose D is exogenous, which holds always for SRD, and holds for FRD under $COR(\varepsilon, e|S) = 0$. Let $m(S) = \beta_0 + \beta_1 S + \beta_{1z} SZ$ so that

$$Y = \beta_d D \ + \beta_0 + \beta_1 S + \beta_{1z} SZ \ + e;$$

having $\beta_1 S + \beta_{1z} SZ$ in $m(S)$ is equivalent to having $\beta_- S(1 - Z) + \beta_+ SZ$ to allow different slopes around $S = 0$. With the LSE of Y on $(D, 1, S, SZ)$, we can estimate $(\beta_d, \beta_0, \beta_1, \beta_{1z})$.

In practice, it is better to use a piecewise cubic (or quartic) $m(S)$ continuous at $S = 0$ because, with $m(S)$ better approximated, the local identification at $S = 0$ can be expanded for a higher 'external validity' of RD. For instance, with $m'(S)$ being the derivative of $m(S)$,

$$\begin{aligned} Y &= \beta_d D \ + \ \beta_0 + \beta_1 S + \beta_2 S^2 + \beta_3 S^3 + \beta_{1z} SZ + \beta_{2z} S^2 Z + \beta_{3z} S^3 Z \ + e \\ &\Longrightarrow \ m'(S) = \beta_1 + 2\beta_2 S + 3\beta_3 S^2 + \beta_{1z} Z + 2\beta_{2z} SZ + 3\beta_{3z} S^2 Z \\ &\Longrightarrow \ m'(0^-) = \beta_1 \ \neq \ m'(0^+) = \beta_1 + \beta_{1z} \text{ whereas } m(0^-) = \beta_0 = m(0^+). \end{aligned} \tag{4.2}$$

In practice, LSE is done typically only for SRD, not for FRD. But, in FRD with $D = D(S, \varepsilon)$, so long as $COR(\varepsilon, e|S) = 0$, LSE can be applied; there is no reason to apply IVE in this case, because IVE is inefficient compared with LSE. The efficiency of the LSE can be much greater, because ε provides an extra variation to D beyond the " default" variation due to S. Hence it is recommended to do a ('Hausman') test of equality of the LSE and IVE; if not rejected, use the LSE. The test was also considered in Bertanha and Imbens (2014).

4.2 IVE for endogenous treatment

Suppose D is endogenous. To be an instrument for D in (4.1), Z should meet three conditions: (i) excluded from the Y equation, (ii) $COR(Z, e) = 0$, and (iii) included in the D equation. Condition (i) cannot be tested, but it is plausible because Z (i.e., the cutoff) should have no direct bearing on Y. Condition (ii) holds automatically due to $E(e|S) = 0$. Condition (iii) can be verified by the LSE of D on $(1, Z)$ (or on $(1, W, Z)$ if W is also used as a regressor): a significant slope of Z indicates (iii). Note that, differently from $COR(Z, e) = 0$ holding by construction in RD, for IVE in general, zero correlation between the instrument and the model error term is not verifiable (and thus only can be argued for).

Before discussing IVE, we start with a more intuitive 'nonparametric ratio estimator' for β_d in (3.1) that replaces the one-sided limits with local sample averages. Define $N^+ \equiv \sum_i 1[S_i \in (0,h)]$ and $N^- \equiv \sum_i 1[S_i \in (-h,0)]$, and

$$\hat{E}(Y|0^+) \equiv \frac{1}{N^+}\sum_i Y_i 1\big[S_i \in (0,h)\big] \; \& \; \hat{E}(Y|0^-) \equiv \frac{1}{N^-}\sum_i Y_i 1\big[S_i \in (-h,0)\big].$$

Define $\hat{E}(D|0^+)$ and $\hat{E}(D|0^-)$ analogously with Y replaced by D. The ratio estimator is

$$\hat{\beta}_d \equiv \frac{\hat{E}(Y|0^+) - \hat{E}(Y|0^-)}{\hat{E}(D|0^+) - \hat{E}(D|0^-)}.$$

Call $\hat{\beta}_d$ '*local-constant regression (LCR) estimator*' for a reason to become clear shortly.

The ratio estimator equals the slope IVE applied to the " artificial" linear model

$$Y = \beta_d D + \beta_0 + error \tag{4.3}$$

with Z instrumenting for D. The equality follows from the 'Wald estimator form' of IVE (e.g., Lee 2005, p.137), as was noted in (Hahn et al. 2001, p. 206). In this sense, the IVE is a sample version of (3.1). If we apply LSE to (4.3), then the slope LSE equals $\hat{E}(Y|0^+) - \hat{E}(Y|0^-)$; in this sense, the LSE is also a sample version of (3.1) when the denominator of (3.1) is one.

Unfortunately, LCR's finite-sample bias is large. To see this, suppose $\beta_d = 0$ and Y is generated by

$$Y = \beta_0 + \beta_1 S + U \text{ with } \beta_1 > 0.$$

Since Y is linearly increasing at $S = 0$, the 'left average' of Y's over $S \in (-h, 0)$ is smaller than the 'right average' of Y's over $S \in (0, h)$, resulting in $\hat{\beta}_d > 0 = \beta_d$. To overcome this problem, ' *local linear regression (LLR)*' was proposed by Hahn et al. (2001): fit a linear line over each local region $(-h, 0)$ and $(0, h)$ to obtain the two lines' heights at $S = 0$; the difference of the two heights is the estimated effect. To see the LLR idea better, consider

$$\begin{aligned}&Y = \beta_d D + \beta_0 + \beta_1 S + U \text{ with } \beta_1 > 0\\ &\quad \Longrightarrow Y = \beta_0 + \beta_1 S + U \text{ for } \; S \in (-h,0) \quad \& \, Y = \beta_d + \beta_0 + \beta_1 S + U \quad \text{for } S \in (0,h).\end{aligned}$$

The height at $S = 0$ from the left model is β_0 and that from the right model is $\beta_d + \beta_0$; the difference is thus β_d.

Although this may look complicated, there exists a simple IVE that gives the same numerical result, as was noted in Imbens and Lemieux (2008), p. 627: apply IVE to

$$Y = \beta_d D + \beta_0 + \beta_- S(1 - Z) + \beta_+ SZ + error \tag{4.4}$$

where D is instrumented by Z. The usual IVE standard errors can be used for inference.

The artificial linear model (4.4) for LLR is a refined version of (4.3) for LCR: $m(S)$ is " linear-splinely" approximated in the former while all but ignored with $m(S) \simeq \beta_0$ in the latter. If cubic terms are used in (4.4), then we get a 'local cubic regression' that is the same as the IVE for (4.2) with Z as an instrument for D. But local polynomial regressions other than LLR tend to be too variable. Hence, there are essentially *two practical RD estimators: LSE to* (4.2) (*or its lower/higher order version), and IVE to* (4.4).

4.3 Relationship between LLR and IVE and remarks

For a bandwidth h and a kernel K (e.g., $K(\cdot)$ is the $N(0, 1)$ density), minimize

$$\sum_{i=1}^{N}(D_i - \tau_0 - \tau_1 S_i)^2 K\left(\frac{S_i}{h}\right)1[0 < S_i] \;\&\; \sum_{i=1}^{N}(Y_i - \rho_0 - \rho_1 S_i)^2 K\left(\frac{S_i}{h}\right)1[0 < S_i]$$

for (τ_0, τ_1) and (ρ_0, ρ_1). Let the minimizers for the intercepts τ_0 and ρ_0 be $\hat{\tau}_0^+$ and $\hat{\rho}_0^+$. Define $\hat{\tau}_0^-$ and $\hat{\rho}_0^-$ analogously with $1[0 < S_i]$ replaced by $1[S_i < 0]$. Then LLR is

$$\hat{\beta}_{d,LLR} \equiv \frac{\hat{\rho}_0^+ - \hat{\rho}_0^-}{\hat{\tau}_0^+ - \hat{\tau}_0^-}.$$

If the uniform kernel $K(t) = 1[|t| < 1]/2$ *is used, then* $\hat{\beta}_{d,LLR}$ *is the same as the IVE for* (4.4). Otherwise (i.e., if another kernel is used), $\hat{\beta}_{d,LLR}$ differs from the IVE for (4.4). The asymptotic variance of $\hat{\beta}_{d,LLR}$ with a general kernel K is involved, as can be seen in Hahn et al. (2001) and Imbens and Lemieux (2008). To avoid this problem, Otsu et al. (2015) proposed empirical-likelihood-based RD estimators, where confidence intervals for β_d are drawn using pivotal asymptotic χ^2 distributions of likelihood ratios.

While practitioners use almost exclusively the above LSE and IVE/LLR estimators for RD, there are other estimators as well in the literature (Porter 2003; Calonico et al. 2014; Yu 2015). Also, Calonico et al. (2014) and Feir et al. (2015) addressed better RD estimator inference using bias correction and weak instrument approach, respectively.

If observed covariates W are to be controlled, $\beta_w' W$ or $\gamma_-' W(1-Z)+\gamma_+' WZ$ can be added to (4.2) and (4.4), respectively, where β_w and γ's are parameters. Alternatively, residuals such as $Y - \hat{\beta}_W' W$ may be used instead of Y, which essentially nullifies the presence of W. Using the residuals keeps the data dimension low so that we can focus on $(residual, Z, D, S)$. Graphical tools to be explained below can be then easily applied, as W no more appears explicitly.

In nonparametrics, series approximation as in (4.2) is dubbed a 'global approach' where the smoothing degree depends on the order of the approximating function. In contrast, a kernel estimator with a smoothing parameter h is dubbed 'local'. Such a distinction, however, becomes blurry in RD, because RD estimation is done using only local observations around the cutoff. The equivalence of the IVE for (4.4) using a series approximation (of the first order) to the LLR when the uniform kernel is used corroborates this point.

4.4 Bandwidth choice and summary

So far, we have not discussed how to choose the bandwidth h. For this, note that a local-constant nonparametric kernel estimator for $E(Y|S = S_i)$ without using Y_i is

$$\hat{E}_{-i}(Y|S_i, h) \equiv \frac{\sum_{j \neq i} K\big((S_j - S_i)/h\big)Y_j}{\sum_{j \neq i} K\big((S_j - S_i)/h\big)}$$

where $\sum_{j \neq i}$ is the sum over $j = 1, \ldots, N$ except the ith observation. There is no single best way to choose h, but a good rule of thumb in practice is $h = SD(S)N^{-1/5}$, and a more systematic way is to use ‘cross-validation (CV)’ as follows.

The usual CV chooses h by minimizing

$$\sum_i \big\{Y_i - \hat{E}_{-i}(Y|S_i, h)\big\}^2.$$

Y_i can be used to estimate $E(Y|S_i)$, but CV is based on the idea of predicting Y_i with an estimator as this minimand shows, and it would be “ silly” to use Y_i in predicting Y_i; this is why $\hat{E}_{-i}(Y|S_i, h)$ is used in CV. When there might be a break in $E(Y|S)$ at $S = 0$, $\hat{E}_{-i}(Y|S_i, h)$ can be replaced by

$$\frac{\sum_{j \neq i} K\big((S_j - S_i)/h\big)1[S_j < S_i < 0]Y_j}{\sum_{j \neq i} K\big((S_j - S_i)/h\big)1[S_j < S_i < 0]} \text{ or } \frac{\sum_{j \neq i} K\big((S_j - S_i)/h\big)1[0 < S_i < S_j]Y_j}{\sum_{j \neq i} K\big((S_j - S_i)/h\big)1[0 < S_i < S_j]} \tag{4.5}$$

depending on whether $S_i < 0$ or $0 < S_i$. The idea is simple: if $S_i < 0$, then only the left observations with $S_j < S_i$ are used; if $0 < S_i$, only the right observations with $S_i < S_j$.

In (4.5), a local-constant version is used, and the LLR version for the first term of (4.5) is the intercept estimator in minimizing with respect to (ρ_0, ρ_1)

$$\sum_{j \neq i} \big\{Y_j - \rho_0 - \rho_1(S_j - S_i)\big\}^2 \cdot K\big(\frac{S_j - S_i}{h}\big)1[S_j < S_i < 0].$$

This CV scheme was used by Ludwig and Miller (2007). Imbens and Kalyanaraman (2012) considered a variation of the CV scheme (p. 944), and more importantly, they suggested a theoretically optimal choice of h as reviewed in the appendix. Calonico et al. (2014) proposed a bias-correction approach for RD inference, where the extra variability induced by the bias-correcting term is taken into account; they also discussed choosing optimal bandwidths.

In summary, for SRD, do LSE to (4.2) *or its lower/higher order version; use the LSE standard errors for inference. For FRD, do LSE to* (4.2) *or its lower/higher order version if* ε *in* D *is unlikely to make* D *endogenous, and otherwise do LLR with the uniform kernel that equals the IVE to* (4.4); *for inference, use the LSE/IVE standard errors. For bandwidth, use* $h = SD(S)N^{-1/5}$, *or the h chosen by the above CV method; do all estimation locally using only the observations with* $S \in (-h, h)$, *although we write "* $\sum_{i=1}^{N}$*" for simplicity.*

5 Specification tests

There are various specification tests for RD, but the most important ones are those for breaks of $E(Y|S)$ and $E(D|S)$, and those for the continuity of $m(S)$.

5.1 Conditional mean breaks

Most RD studies present a graph plotting $E(Y|S)$ vs. S to demonstrate the break of $E(Y|S)$; no break means $\beta_d = 0$, or Z having no explanatory power for D ($\alpha_z = 0$ in (3.2)). For FRD, $E(D|S)$ versus S is also shown because $E(D|0^+) - E(D|0^-) \neq 0$ is necessary. Informal graphical presentations can be formalized into the following LSE-based tests.

Consider an artificial linear model analogous to (4.4):

$$D = \zeta_z Z + \zeta_0 + \zeta_- S(1 - Z) + \zeta_+ SZ + error \tag{5.1}$$

where ζ's are parameters. *Applying LSE to this, a non-zero slope of* Z *indicates a break of* $E(D|S)$ *at* $S = 0$*, for which the LSE standard errors can be used*. The logic is that ζ_z equals the intercept break $E(D|0^+) - E(D|0^-)$ due to Z, whereas the base intercept is picked up by 1 and the slopes are accounted for by the last two regressors.

As for a break in $E(Y|S)$ at $S = 0$, consider an artificial linear model analogous to (5.1):

$$Y = \xi_z Z + \xi_0 + \xi_- S(1 - Z) + \xi_+ SZ + error \tag{5.2}$$

where ξ's are parameters. *A non-zero slope LSE of* Z *indicates a break of* $E(Y|S)$ *at* $S = 0$. The LSE-based tests with (5.1) and (5.2) seem unknown in the literature despite their simplicity.

Although LSE to (5.1) and (5.2) can be used to test for a break at the known point 0, it is possible that a break may occur somewhere else. Seeing a break where it is not supposed to be suggests a misspecification, and checking out breaks over a range for S around 0 can be done with the difference of one-sided kernel regression estimators. For $E(Y|s)$, plot

$$\tilde{L}_Y(s) \equiv \frac{\sum_i K\big((S_i - s)/h\big)1[s < S_i]Y_i}{\sum_i K\big((S_i - s)/h\big)1[s < S_i]} - \frac{\sum_i K\big((S_i - s)/h\big)1[S_i < s]Y_i}{\sum_i K\big((S_i - s)/h\big)1[S_i < s]} \tag{5.3}$$

versus s. $\tilde{L}_Y(s)$ has been used to find structural breaks in statistics [e.g., Qiu (2005)], but hardly so in the RD literature; for structural breaks in general, see, e.g., Breitung and Kruse (2013), Ciuperca (2014), and the references therein.

Porter and Yu (2015) estimated the unknown cutoff by maximizing $\tilde{L}_Y(s)^2$ with respect to s for SRD where the local-constant estimators are replaced by local polynomial estimators; for FRD, $\tilde{L}_Y(s)^2$ plus the analogous expression with Y replaced by D is used. They found that the cutoff estimator is super-consistent, converging at the rate N instead of the usual $\sqrt{N}$. Hence the estimated cutoff is as good as the true cutoff, and the estimation does not affect the asymptotic distribution of the treatment effect estimator.

Since $\tilde{L}_Y(s)$ is for different values of s, the appropriate bandwidth may differ from the h chosen for $s = 0$ only. The most practical way to choose h for $\tilde{L}_Y(s)$ is " eye-balling" : choose h so that the $\tilde{L}_Y(s)$ graph is not too smooth nor too jagged. As for inference, the asymptotic distribution may be derived for $\tilde{L}_Y(s)$, but confidence bands based on nonparametric bootstrap resampling from the original sample with replacement would be adequate in practice. To detect breaks in $E(D|s)$, replace Y in $\tilde{L}_Y(s)$ with D.

Turning to the continuity of $m(S)$, since $m(S)$ comes from the conditional means of the ignored covariates W and U—we will use only U to denote errors in the rest of this paper—the continuity of $E(W|S)$ and $E(U|S)$ should be checked out. We can test for breaks in $E(W|S)$ by the LSE to (5.1) with D replaced by W; Urquiola and Verhoogen (2009) showed an example where $E(W|S)$ is not continuous at $S = 0$. As for the continuity of $E(U|S)$, it is discussed next.

5.2 Score-density continuity at cutoff

Since U is not observed, the continuity of $E(U|s)$ cannot be seen. Instead, necessary conditions can be tested. Observe

$$E(U|s) = \int u f_{U|S}(u|s)\mathrm{d}u = \int u \frac{f_{S|U}(s|u)}{f_S(s)} f_U(u)\mathrm{d}u$$

where $f_{U|S}$ denotes $U|S$ density and f_S denotes S density; $f_{S|U}$ and f_U are defined analogously. This shows that *the continuity of* $f_{S|U}(s|u)/f_S(s)$ *at* $s = 0$ *is necessary for the continuity of* $E(U|s)$. *Since the continuity of* $f_{S|U}(s|U)$ *at* $s = 0$ *cannot be tested, only the continuity of* $f_S(s)$ *is to be checked out.*

To enhance understanding, recall Example 1 with a discrete U taking on -1 and 1. Suppose that U stands for 'socializing well' and persons with $U = -1$ try extra hard to get $D = 1$ to make up for their lower income due to $U = -1$; let $\pi \equiv P(U = -1)$. Also suppose

$$f_{S|U}(s|-1) = 1[0 \le s < 1],\ f_{S|U}(s|1) = \phi(s) \text{ and } \phi \text{ is a density continuous at } 0 :$$

those with $U = -1$ attain the scores in [0, 1] whereas those with $U = 1$ have scores well spread around 0 with ϕ. Then

$$f_S(s) = f_{S|U}(s|-1)\pi + f_{S|U}(s|1)(1-\pi) = \pi 1[0 \le s < 1] + (1-\pi)\phi(s)$$
$$\Longrightarrow f_S(0^-) = (1-\pi)\phi(0),\ f_S(0^+) = \pi + (1-\pi)\phi(0) \Longrightarrow f_S(0^+) - f_S(0^-) = \pi.$$

The break of f_S occurs because those with $U = -1$ manipulated their S to " perfection" . Now suppose that $f_{S|U}(s|-1)$ is continuous at 0 but tilted heavily to the right of 0: those with $U = -1$ could not perfectly manipulate S although they could to a large extent. Then, $f_S(0^+) - f_S(0^-) = 0$. In principle, RD does not require f_S to be continuous at 0, but it is this concern for manipulating S that prompts testing for the continuity.

The above example, that is a generalized version of an example in Kim and Lee (2016), shows that we can identify π with $f_S(0^+) - f_S(0^-)$. Gerard et al. (2015) called the individuals with $U = -1$ " manipulators', and assuming that all manipulators have $S \ge 0$, they indeed identified π using $f_S(0^+)$ and $f_S(0^-)$. Declaring 'the treatment effect on the non-manipulators' as the main parameter of interest, they went on to bound the effect, and proposed how to estimate the bound and conduct inference.

The easiest way to see the break of f_S at 0 is constructing a histogram for f_S such that 0 becomes a boundary point. A smoothed version of the histogram is

$$\tilde{f}_S(s) \equiv \frac{1}{Nh}\sum_i 2K\Big(\frac{S_i - s}{h}\Big)\big\{1[S_i < s < 0] + 1[0 < s < S_i]\big\};$$

the part $\{\cdot\}$ is to use only the left observations ($S_i < s$) when $s < 0$, and only the right observations ($s < S_i$) when $0 < s$. With the uniform kernel $K((S_i - s)/h) = 1[|S_i - s| < h]/2$, $\tilde{f}_S(s)$ becomes

$$\hat{f}_S(s) \equiv \frac{1}{Nh}\sum_i 1[s-h < S_i < s] \ \ \text{if}\, s < 0 \ \& \ \frac{1}{Nh}\sum_i 1[s < S_i < s+h] \ \ \text{if}\, 0 < s.$$

whereas $\tilde{f}_S(s)$ and $\hat{f}_S(s)$ are constructed with a break at 0 in mind, we may want to explore unknown break locations. This can be done by plotting

$$\tilde{J}(s) \equiv \frac{1}{Nh}\sum_i 2K\Big(\frac{S_i - s}{h}\Big)\big\{1[s < S_i] - 1[S_i < s]\big\} \tag{5.4}$$

versus s, which is a " density analog" of $\tilde{L}_Y(s)$; $\tilde{J}(s)$ is little known in the literature as $\tilde{L}_Y(s)$ is. Choosing h and doing inference can be done analogously to what was done for $\tilde{L}_Y(s)$.

Whereas the test with $\tilde{f}_S(s)$ is a " local-constant" variety, McCrary (2008) suggested a two-stage LLR-type test for the break of f_S at 0, as reviewed in the appendix. In the first stage, estimate f_S around $s = 0$ over n intervals using $\hat{f}_S$; let $s_1, \ldots, s_n$ be the centers of the histogram intervals. In the second-stage, taking $\{s_j, \hat{f}_S(s_j)\}$, $j = 1, \ldots, n$ as data (analogous to (X_j, Y_j), $j = 1, \ldots, n$), do LLR on the negative and positive sides separately to see if the two estimated lines meet at $s = 0$ or not. Despite the popularity, however, the test requires two bandwidths (one for $\hat{f}_S$ and the other for LLR), which is a disadvantage. Also the asymptotic distribution was derived only for the 'triangular kernel' $K(t) = (1 - |t|)1[|t| < 1]$, although it would be certainly possible to use other kernels for the test.

Otsu et al. (2013) proposed empirical-likelihood-based methods to construct confidence intervals (CI) for $f_S(0^+) - f_S(0^-)$. The methods allow popular kernels including the triangular kernel, and by constructing CI's using the likelihood ratio that follows a χ^2 distribution asymptotically, the methods obviate the need to estimate the asymptotic variance.

6 RD topics

So far, we examined RD essentials. In practice, there also arise other issues going beyond the " basic" RD. This section reviews those RD topics.

6.1 Multiple scores

In a typical RD, a single score crossing a cutoff affects D non-trivially. There are, however, many RD cases where multiple scores are involved to determine a single treatment. For instance, multiple test scores crossing cutoffs may be required for school graduation, and age and pension-contribution-years should cross cutoffs to receive pension.

'Multiple-score RD for a single treatment' that is examined in this subsection differs from 'single-score RD with multiple cutoffs' as in Van der Klaauw (2002) and Angrist and Lavy (1999), which is easily handled by looking at each cutoff one at a time. The treatments at the multiple cutoffs may be ordered, and as such, they fall in the domain of 'multi-valued (or multiple) treatments' (see Imbens (2000), Lechner (2001), Uysal (2015), and the references therein), and are either to be taken as such, or weight-averaged to come up with a single representative effect (Bertanha 2015). Multiple-score RD for a single treatment differs also from 'multiple-score RD for multiple treatments' (Leuven et al. 2007; Papay et al. 2011) where each score dictates one treatment.

When there are two scores S_1 and S_2—the dominating case in mulitple-score RD—two cases arise: '*OR case*' where any score can cross a cutoff to get treated (Jacob and Lefgren 2004; Matsudaira 2008), and '*AND case*' where both scores should cross cutoffs to get treated (Schmieder et al. 2012; Caliendo et al. 2013). But an OR case can be converted to the AND case simply by " flipping" the treatment, i.e., by relabeling the treatment and control groups. Hence, dealing with AND case is enough in two-score RD. For more than two scores, mixed cases can arise; e.g., given three, only two may cross the cutoffs to get treated.

There appeared three ways to reduce multiple-score RD to single-score RD. First, multiple scores are combined to form a single score/index as in Van der Klaauw (2002). Second, when $D = 1[0 \le S_1, 0 \le S_2]$, D becomes $1[0 \le S_1]$ on the subpopulation with $0 \le S_2$, in which case the familiar toolkit for single-score RD can be used; this has been the dominant approach (Jacob and Lefgren 2004; Lalive 2008; Schmieder et al. 2012; Caliendo et al. 2013, etc.). This dimension-reduction strategy also works analogously for more than two scores. Third, defining $S_m \equiv \min(S_1/\sigma_1, S_2/\sigma_2)$ for some scale-normalizing constants σ_1 and σ_2 (Battistin et al. 2009; Clark and Martorell 2014), D becomes $1[0 \leq S_m]$ or a fuzzy version of $1[0 \leq S_m]$, and we can set up the following to apply LSE or IVE:

$$E(Y|S_m) = \beta_m D + \beta_0 + \beta_- S_m(1-Z) + \beta_+ S_m Z.$$

This can be easily generalized to J scores with $S_m = \min(S_j/\sigma_j, j = 1, \ldots, J)$.

Instead of these, Imbens and Zajonc (2009) dealt with both multiple-score SRD and FRD head-on: they discussed identification and estimation with multidimensional LLR. The main complication with multiple scores is the appearance of a boundary instead of a cutoff. For instance, with B denoting the treatment and control boundary, the treatment effect at $s \in B$ for FRD is

$$\beta_d(s) \equiv \frac{\lim_{\nu\to 0} E\{Y|S \in N_\nu^+(s)\} - \lim_{\nu\to 0} E\{Y|S \in N_\nu^-(s)\}}{\lim_{\nu\to 0} E\{D|S \in N_\nu^+(s)\} - \lim_{\nu\to 0} E\{D|S \in N_\nu^-(s)\}}$$

where $N_\nu^+(s)$ denotes the 'ν-treated-neighborhood' of s and $N_\nu^-(s)$ denotes the ν-control-neighborhood of s. Imbens and Zajonc proposed also an integrated version of $\beta_d(s)$:

$$\beta_d \equiv \int_{s\in B} \beta_d(s) f_S(s|S \in B)\partial s = \frac{\int_{s\in B} \beta_d(s) f_S(s)\partial s}{\int_{s\in B} f_S(s)\partial s}.$$

Tests for the effect heterogeneity along B, the asymptotic distribution of the LLR, and the optimal bandwidth choice are also shown in Imbens and Zajonc (2009).

Wong et al. (2013) examined two-score OR-case SRD, and Keele and Titiunik (2015) two-score AND-case SRD where the two scores are latitude and longitude. These studies were done independently of Imbens and Zajonc (2009), but essentially, they are special cases of Imbens and Zajonc (2009). Choi and Lee (2015) also examined two-score SRD, but their study differs from the other studies because they allowed 'partial effects' as follows that were ruled out by Imbens and Zajonc (2009), Wong et al. (2013) and Keele and Titiunik (2015).

Suppose that a student has to pass both math exam ($Z_1 = 1$) and English exams ($Z_2 = 1$) to graduate high school ($D = Z_1 Z_2 = 1$), then it is possible that Z_1 and Z_2 may affect Y (say, lifetime income) separately from D. In a case like this, one may postulate

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_d D + m(S_1, S_2) + U$$

and estimate the partial effects β_1 and β_2 along with β_d. A simple estimator is obtained by approximating $m(S_1, S_2)$ polynomially (or with splines). Choi and Lee (2015) also proposed a nonparametric estimators that takes a form of local 'difference in differences', which is natural because D is the interaction of Z_1 and Z_2.

6.2 Incompletely observed or discrete score

Sometimes we observe only an error-ridden score S, instead of the genuine score G determining D; e.g., $S = G + V$ for an error V, where S is a reported income in a survey and G is the true income. This is a measurement error (or errors in variable) problem in RD score. In terms of (3.1), we have the desired ratio (DR) and the available ratio (AR):

$$DR \equiv \frac{E(Y|G=0^+) - E(Y|G=0^-)}{E(D|G=0^+) - E(D|G=0^-)} \quad \& \quad AR \equiv \frac{E(Y|S=0^+) - E(Y|S=0^-)}{E(D|S=0^+) - E(D|S=0^-)}.$$

Bear in mind that only observing G is problematic while D and Y are fully observed, and that D is determined by G, not by S ; if S itself determines D, then there is no identification problem—a point misunderstood in the literature for a while (see Cappelleri et al. (1991) and the references therein). Yu (2012) examined identification and local polynomial estimation for various measurement-error SRD and FRD cases depending on D determined by G or S, and the measurement error V shrinking toward zero or not.

First, for a continuously distributed error V, suppose ('⨿' stands for independence)

$$S = G + V \text{ where } V \amalg G. \tag{6.1}$$

This is a classical "*full errors-in-variable* " , as S is a smooth-error-ridden version of G and '$V \amalg G$' holds. Lee (2016) showed that $E(D|S)$ has no break at $S=0$ under (6.1) as V smooths G out, and thus (3.1) fails. This is a " hopeless" case, unless one is willing to impose strong assumptions as in Hullegie and Klein (2010) where D is private health insurance, Y is health care utilization, G is the true income and S is a self-reported income. In contrast to (6.1) with continuous G and V, Pei (2011) assumed that G and V are discrete with bounded supports to nonparametrically identify the distribution of G and the treatment effect.

Yanagi (2014) considered SRD after rewriting (6.1) as $S = G + \sigma V'$ with $SD(V') = 1$. Yanagi (2014) (as well as Yu 2012) considered

$$\tau_{DS} \equiv E(Y|D=1, S=0^+) - E(Y|D=0, S=0^-)$$

that is more informative than $E(Y|S=0^+) - E(Y|S=0^-)$ because G is used (although not localized) through D. Yanagi (2014) found, with $\tau_G \equiv E(Y|G=0^+) - E(Y|G=0^-)$,

$$\tau_{DS} = \tau_G + \sigma^2 \times (\text{entity depending on } S, D, Y)$$

which characterizes the identification error $\tau_{DS} - \tau_G$. It is possible to fully identify τ_G if auxiliary data are available for σ. For example, a survey on income gives S whereas σ is announced by the government based on a census.

Second, for an unobserved binary variable R, suppose

$$S = RG + (1-R)(G+V) : \tag{6.2}$$

G is observed when $R=1$, and an error-ridden score $G+V$ is observed otherwise; "*part errors-in-variable* " (6.2) occurred in Battistin et al. (2009) and Schanzenbach

(2009). Differently from (6.1), G is observed when $R = 1$: there are " truth-tellers $(G = S)$" .

Lee (2016) showed

$$AR = \frac{E(Y|G = S = 0^+) - E(Y|G = S = 0^-)}{E(D|G = S = 0^+) - E(D|G = S = 0^-)} :$$

although DR is not identified, the effect on the " truthful margin" $G = S = 0$ is identified. The condition '$G = S$' is reminiscent of 'compliers' in Imbens and Angrist (1994). If $(D, Y) \amalg (R, V)|G$ holds additionally as assumed by Battistin et al. (2009), then $(D, Y) \amalg S|G$ holds and S thus drops out of the last display to render $AR = DR$.

Third, suppose $S = 0, 1, 2, \ldots$ is a *grouped transformation* of G:

$$S = \sum_j 1[G \geq \gamma_j] \iff S = j \ \text{ iff } \ G \in [\gamma_j, \gamma_{j+1}) \text{ for some known } \gamma_j \text{ 's.} \quad (6.3)$$

This occurs to yearly rounded-down age ($S = \sum_{j=1} 1[G \geq j]$), or income recorded in groups due to confidentiality. Dong (2015) listed many applied papers for (6.3) while proposing a LSE-based estimator under the uniform distribution for the rounding error.

Related to (6.3), Lee and Card (2008) addressed the *pure discrete score* case $S = G$ with no measurement error. For this, unless one " settles" with nearby support points of the cutoff, extension/interpolation toward the cutoff from those points is inevitable, which requires a parametric assumption on $m(\cdot)$ in (2.1). Lee and Card (2008) suggested to test for the parametric model specifications against the saturated nonparametric model, and use a cluster variance estimator for inference because the observations at each support point would share the same specification error.

Fourth, a general score model for both continuous and discrete components is

$$S = RG \ + \ (1 - R) \sum_j 1[G \geq \gamma_j]. \quad (6.4)$$

'*Heaping*' (i.e., probability masses despite that the score is supposed to be continuous) in Almond et al. (2010), Almond (2011) and Barreca et al. (2011, 2016) is an example. Heaping can occur for many reasons so that the part next to $1 - R$ can take a form other than $\sum_j 1[G \geq \gamma_j]$; e.g., a matter of practice (retiring at 60, working 40 h per week, ...), limited precision in measurement, top coding, etc.

One identification strategy facing (6.4) is conditioning on $R = 1$; this works under no 'selection-problem' $R \amalg (G, D, Y)$. The opposite strategy is conditioning on $R = 0$ to turn to (6.3), and then the estimates obtained under $R = 1$ and $R = 0$ may be weight-averaged to come up with a single effect. Barreca et al. (2016) recommended plotting disaggregated data not to miss heaping features and estimating a model for a covariate W such as

$$W = \alpha_0 + \alpha_1 1[S = \gamma] + \alpha_2(S - \gamma) + error \ \text{ where } \ \gamma \ \text{ is a heaping point}$$

to see $\alpha_1 = 0$; if $\alpha_1 \neq 0$, then W differs systematically at the heaping point γ. For instance, with S being birth weight and W income, poor district hospitals may use scales of poor precision to result in heaping, in which case $\alpha_1 < 0$. A systematic difference in W may suggest the same for unobserved variables to result in the aforementioned selection problem.

6.3 Regression kink (RK)

Recalling (2.1), in ‘regression kink (RK)’ design, $\nabla E(D|S)$ is assumed to be discontinuous at 0 whereas $\nabla m(S)$ is continuous. For instance, Kim and Lee (2016) made use of the fact that marginal income tax rate has breaks/jumps at income (S) cutoffs in income tax schedule, which implies the average tax rate having slope breaks at those cutoffs because it is based on the integral of the marginal tax rate. Using this, Kim and Lee (2016) estimated labor supply elasticity with respect to after-tax income.

Simonsen et al. (2015) estimated the price elasticity of prescription drug demand using RK, where the price that each individual faces differs depending on the accumulated prescription drug purchase amount S: if $S + \bar{P}$ crosses a threshold c where $\bar{P}$ is the shelf price for the current purchase, then the out-of-pocket price D changes from $\bar{P}$ to a subsidized price (e.g., to $0.5\bar{P}$). This results in D decreasing linearly as a function of S over the range $c - \bar{P} \leq S < c$, and slope breaks occur, going in and out of this range.

Define the right and left derivative at 0:

$$\nabla E(Y|0^+) \equiv \lim_{\nu \to 0^+} \frac{E(Y|S=\nu) - E(Y|S=0)}{\nu},$$
$$\nabla E(Y|0^-) \equiv \lim_{\nu \to 0^+} \frac{E(Y|S=0) - E(Y|S=-\nu)}{\nu}.$$

The difference of the two one-sided derivatives of (2.1) at 0 is, as $\nabla m(0^+) = \nabla m(0^-)$,

$$\nabla E(Y|0^+) - \nabla E(Y|0^-) = \beta_d \cdot \left\{\nabla E(D|0^+) - \nabla E(D|0^-)\right\}$$
$$\Longrightarrow \beta_d = \frac{\nabla E(Y|0^+) - \nabla E(Y|0^-)}{\nabla E(D|0^+) - \nabla E(D|0^-)}. \tag{6.5}$$

As in RD, sharp RK (SRK) refers to D determined only by S, and fuzzy RK (FRK) refers to $D = D(S, \varepsilon)$. The derivation leading to (6.5) for SRK was shown in Nielsen et al. (2010), p. 214, where the denominator in (6.5) becomes a known constant. For instance, $D = \alpha_s ZS$ for a parameter $\alpha_s \neq 0$ makes the denominator of (6.5) $\alpha_s - 0 = \alpha_s$, although $E(D|S)$ is continuous at 0.

For estimation of FRK, Card et al. (2012) showed that β_d can be estimated by the slope $\hat{\eta}_1^{\Delta}$ of D in the IVE applied to (η's are parameters)

$$Y = \eta_0 + \eta_1 S + \eta_1^{\Delta} D + error \tag{6.6}$$

with D instrumented by ZS (not by Z); the published version (Card et al. 2015) of Card et al. (2012) discussed only local polynomial estimators. For sharp RK, apply LSE to (6.6).

In RK, β_d represents the effect of D on Y, not the derivative of the effect; it is just that we identify β_d using the derivatives in RK. For instance, $Y = \beta_d D + m(S) + U$ may hold where $E(D|S)$ and $m(S)$ are subject to the above RK conditions. In contrast, letting the cutoff be c and writing the RD effect at c as $\beta_d(c)$, Dong and Lewbel (2015) looked at the derivative $\beta_d'(c)$ for a number of reasons. First, we may be interested in the effect constancy to test $H_0 : \beta_d'(c) = 0$. Second, the sign of $\beta_d'(c)$ will tell whether the effect will become smaller or larger for those with S a little smaller or larger than c. Third, $\beta_d'(c)$ shows the effect of changing the cutoff c. Fourth, knowing derivatives expands the RD external validity by extrapolating the estimated effect away from the cutoff.

An estimator for $\beta_d(c)$ in SRK can be an estimator for $\beta_d'(c)$ in SRD, because both are sample versions of $\nabla E(Y|c^+) - \nabla E(Y|c^-)$. An estimator for $\beta_d(c)$ in FRK, however, cannot be an estimator for $\beta_d'(c)$ in FRD in general, because the former is a sample version of (6.5) whereas the latter is a sample version of the derivative of (3.1) with cutoff c. A better understanding on this can be gained by the following.

The equation (6.5) and the ensuing estimation in (6.6) are based on the premise that there is no break in $E(D|S)$ at c. What if there is a break in $E(D|S)$ at c and yet we use (6.5)? This question is addressed for binary D by Dong (2014) who showed

$$\frac{\nabla E(Y|c^+) - \nabla E(Y|c^-)}{\nabla E(D|c^+) - \nabla E(D|c^-)} = \beta_d(c) + \beta_d'(c)\frac{E(D|c^+) - E(D|c^-)}{\nabla E(D|c^+) - \nabla E(D|c^-)}.$$

This reveals how the RK estimand (6.5) with cutoff c is related to $\beta_d(c)$ and $\beta_d'(c)$. If $\beta_d'(c) = 0$, then both (6.5) and RD estimate the same parameter $\beta_d(c)$. This opens up the possibility to use both RD and RK in estimating $\beta_d(c)$. Since RD estimators are more efficient than the RK estimators, asymptotically, there will be no gain in using both and combining them, although there would be some gain in finite samples. Dong (2014) suggested IVE with all of S, Z and SZ as instruments for D under $\beta_d'(c) = 0$. The resulting IVE works if there is a break in either $E(D|S)$ or $\nabla E(D|S)$.

6.4 Quantile RD and external validity

So far we examined mean-based RD, and one may wonder if there exists a quantile-based RD. Indeed, Frandsen et al. (2012) proposed a quantile RD. Although we have not used 'potential treatments/responses' up to now because RD can be explained without them, we now introduce them to explicate Frandsen et al. (2012). Consider a fuzzy D, and imagine potential treatments (D^0, D^1) corresponding to $Z = (0, 1)$ and potential responses (Y^0, Y^1) corresponding to $D = 0, 1$; the relationship between (D^0, D^1) and Z is analogous to that between (Y^0, Y^1) and D . Given these, Frandsen et al. (2012) proposed 'complier quantile effects' as follows.

Observe

$$E(Y^1|\text{complier}) = \frac{E(YD|Z=1) - E(YD|Z=0)}{E(D|Z=1) - E(D|Z=0)} \tag{6.7}$$

which was proven by Abadie (2002), where 'compliers' are those with ($D^0 = 0, D^1 = 1$). Replacing Y in (6.7) with $1[Y \leq y]$ and then localizing at $S = 0$ gives $F_{Y^1|Complier,S=0}(y)$ where $F_{Y^1|Complier,S=0}(\cdot)$ denotes the distribution function of Y^1 given ($complier$, $S = 0$), from which quantiles can be found. Replacing D with $1 - D$ in (6.7) and proceeding analogously gives $F_{Y^0|Complier,S=0}(y)$, from which quantiles can be found. Then the corresponding quantile differences give complier quantile effects.

For SRD with $D = Z$, under the continuity of $E(Y^0|S)$ at 0, (3.1) becomes 'the effect on the just treated':

$$\beta_d = E(Y|0^+) - E(Y|0^-) = E(Y^1|0^+) - E(Y^0|0^-) = E(Y^1 - Y^0|0^+). \quad (6.8)$$

Denote a quantile of $Y|S$ as $q(Y|S)$. In quantile RD, what can be identified is not $q(Y^1-Y^0|0^+)$ as in (6.8), but only $q(Y^1|0^+)-q(Y^0|0^+)$ at best. This restriction should be borne in mind; see Lee (2000), however, for a limited scope to find $q(Y^1 - Y^0|0^+)$.

In contrast to (6.8), due to Imbens and Angrist (1994), (3.1) for FRD becomes the effect on 'the just treated compliers':

$$\beta_d = E(Y^1 - Y^0|0^+, \text{complier}) \quad (6.9)$$

Hence, the RD identification only on the cutoff/margin 0 in (6.8) becomes further restricted to the " marginal compliers" .

In an effort to increase the external validity of RD, Bertanha and Imbens (2014) explored if the identified RD effect is independent of the individual type such as complier, 'always taker' ($D^0 = D^1 = 1$) and 'never taker' ($D^0 = D^1 = 0$). They recommended plotting $E(Y|D = 0, S)$ around $S = 0$ to see if the untreated compliers ($D = 0, S = 0^-$) differ from the never takers ($D = 0, S = 0^+$); also, an analogous plot of $E(Y|D = 1, S)$ will reveal the difference between the always takers ($D = 1, S = 0^-$) and the treated compliers ($D = 1, S = 0^+$) around $S = 0$. If no difference, then the qualifier 'complier' may be dropped from (6.9) for the external validity across the individual types.

Angrist and Rokkanen (2015) found another way to enhance the external validity of RD. To see the idea, consider SRD with $D = Z$. They assumed the existence of covariates X such that (we use ⨿ for simplicity, although Angrist and Rokkanen (2015) used mean independence)

$$Y^d \amalg S|X \iff Y^d \amalg \eta|X, \quad with \quad S \text{ (hence } D\text{) determined by } (X, \eta) \text{ for an error } \eta:$$

the potential responses are independent of the unobserved part of S given X, which implies $Y^d \amalg D|X$. Then

$$E(Y|S, X, D = d) = E(Y|X, D = d), \quad d = 0, 1.$$

This implication can be tested, e.g., by parametrizing $E(Y|S, X, D = d) = \beta_{ds} S + \beta'_{dx} X$ and checking if $\beta_{ds} = 0$. More important is that

$$
\begin{aligned}
&E\big\{E(Y|X, D=1) - E(Y|X, D=0)\ |S=s\big\} \quad (\text{under } 0 < E(D|X) < 1)\\
&= E\big\{E(Y^1|X) - E(Y^0|X)\ |S=s\big\} \quad (\text{due to } Y^d \amalg D|X)\\
&= E\big\{E(Y^1|S,X) - E(Y^0|S,X)\ |S=s\big\} = E(Y^1 - Y^0|S=s). \qquad (6.10)
\end{aligned}
$$

Hence $E(Y^1 - Y^0|S = s)$, not just the effect at the cutoff, is identified by (6.10). Angrist and Rokkanen (2015) also considered FRD after strengthening $Y^d \amalg S|X$ to $(Y^0, Y^1, D^0, D^1) \amalg S|X$; for FRD, in essence, 'complier with $S = s$' replaces $S = s$ in the conditioning set.

7 Empirical illustration

In this section, we apply some of the RD estimators and tests introduced so far to a data set used in Angrist and Lavy (1999) to find effects of class size on students' achievement; S is the number of the enrolled students. This example may not be ideal as S is discrete (integer-valued), but it also reveals problems in RD, which is not necessarily a bad thing.

The 'Maimonides rule' in Israel public schools limits class size to 40; if 41 students, then there should be two classes; if 81, then there should be three classes, and so on. Hence 40 or 41, 80 or 81, 120 or 121,... are break points. But the actual class sizes did not exactly follow the rule. Figure 1 shows the actual and predicted class sizes by the Maimonides rule around the cutoff 40/41. The actual class size tends to be smaller than the predicted size (thus FRD), as some schools could afford to add classes before reaching the limit.

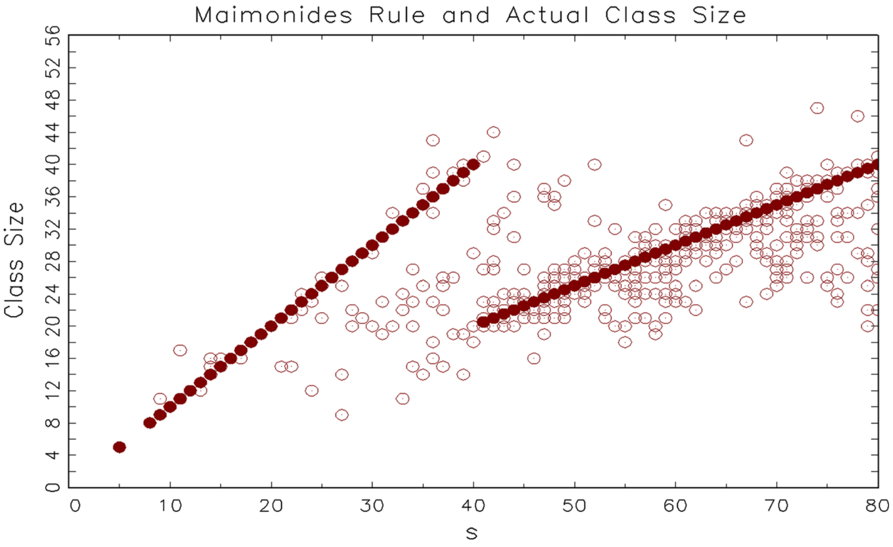

Fig. 1 Class size versus enrolment

Table 1 Data description

Variable	Mean (SD)	Min, max
Math score (math) Y	65.5 (10.0)	27.7, 93.7
Reading score (read) Y	73.2 (8.45)	34.8, 91.6
Class size D	27.3 (6.47)	5, 47
Enrolment S	51.2 (17.7)	5, 79
Proportion disadvantaged W	0.18 (0.16)	0.00, 0.76

Table 2 Class size effect

	Without W		With W	
S	LSE (SE)	IVE (SE)	LSE (SE)	IVE (SE)
Math				
39	0.89(0.18)**	−0.30(0.32)	0.49(0.16)**	−0.21(0.21)
40	0.88(0.18)**	−0.51(0.28)+	0.48(0.16)**	−0.32(0.18)+
41	0.84(0.18)**	−0.68(0.46)	0.49(0.16)**	−0.28(0.24)
42	0.78(0.19)**	−2.45(1.11)*	0.49(0.16)**	−0.64(0.37)+
Read				
39	0.71(0.19)**	−0.45(0.33)	0.30(0.14)*	−0.40(0.17)*
40	0.68(0.18)**	−0.62(0.29)*	0.25(0.14)+	−0.49(0.16)**
41	0.61(0.18)**	−0.62(0.46)	0.26(0.14)+	−0.45(0.21)*
42	0.50(0.18)**	−2.50(1.14)*	0.15(0.14)	−0.69(0.29)*

*, **, + Significance at 99, 95, 90 %

The data were collected in June 1991 on enrolment S, class size D , proportion of disadvantaged students W and average test score Y for mathematics and reading. Angrist and Lavy (1999) used third to fifth grade samples and several cutoffs, but we will focus only on $c = 40/41$ for fifth grade. The unit of observations is a class, and $N = 1127$; the descriptive statistics are in Table 1. The single covariate 'proportion disadvantaged' W reflects family income level. See Angrist and Lavy (1999) for details on the data.

The effects of class size on test scores are in Table 2. The left half is for LSE to (4.2) and IVE to (4.4), whereas the right half adds $\beta_w W$ to (4.2) and $\gamma_{w-} W(1-Z)+\gamma_{w+} WZ$ to (4.4) to explicitly account for W; this is done because $E(W|S)$ has a break at 42 as will be seen shortly. We evaluate the effects over $S \in [39, 42]$ because many breaks occur over these points, which can be either due to the discrete nature of S or $E(\cdot|S)$ having genuine breaks. The bandwidth was chosen by CV using (4.5).

In Table 2, LSE is significantly positive for all S with the effect ranging over 0.50–0.89 before W is used; once W is used, however, the effect drops to 0.15–0.49. These positive effects that are likely due to the D endogeneity through ε are, however, counter-intuitive. IVE is immune to this endogeneity problem, and it gives negative effects ranging over −0.30 to −2.5 before W is used; about a half of them are significant. But once W is used, IVE drops to −0.21 to −0.69, and the effects on reading are significant for all S. These changes due to W have two reasons: omitted

Table 3 LSE-based break tests at known points: estimate (SE)

S	$E(D\|S)$	$E(\text{Math}\|S)$	$E(\text{Read}\|S)$	$E(W\|S)$
37	−3.44(1.10)**	−3.00(2.01)	−0.62(1.48)	0.06(0.03)**
38	−6.74(0.95)**	−0.26(1.89)	1.40(1.39)	0.03(0.03)
39	−7.76(1.04)**	2.32(1.60)	3.36(1.32)**	0.05(0.03)*
40	−8.52(1.13)**	4.36(1.47)**	5.21(1.28)**	0.02(0.02)
41	−4.62(1.10)**	3.15(1.26)*	2.88(1.10)**	0.01(0.02)
42	−2.78(1.09)**	6.81(1.81)**	6.58(1.31)**	−0.08(0.02)**
43	−0.73(1.03)	6.96(1.19)**	3.48(1.14)*	−0.06(0.02)**
44	−0.15(1.03)	5.02(1.08)**	2.42(0.99)	−0.05(0.02)**
45	0.85(1.02)	3.03(1.01)**	−0.25(0.91)	−0.003(0.02)

W may make D endogenous (this affects only LSE), and W has a break near c (this affects both LSE and IVE). Notice that the SE of IVE is 2–4 times greater than the SE of LSE, and that the SE's of LSE and IVE drop by controlling W.

To test breaks of $E(\cdot|S)$, we apply LSE to (5.2) and (5.1) for Y and D to estimate the slope of Z with the same bandwidth as used for Table 2; we also test breaks for $E(W|S)$. Due to the problem of the S discreteness and the fuzziness in D as Fig. 1 shows, the LSE-based tests indicate many breaks, for which the test results over $S \in [37, 45]$ are presented in Table 3. There, $E(D|S)$ have breaks in many points with the largest at 40; $E(\text{Math}|S)$ also have many breaks with the largest at 42 and 43; $E(\text{Read}|S)$ have a couple of breaks with the largest break at 40 and 42; finally, $E(W|S)$ have breaks at 42, 43 and 44, which may make both LSE and IVE inconsistent if W is unaccounted for. It is not clear yet how this multiple break problem should be handled; at least, breaks at nearby points around c should be checked out.

Figure 2 presents $\tilde{L}_{\text{Math}}(s)$, $\tilde{L}_{\text{Read}}\ (s)$, $\tilde{L}_D(s)$ and $\tilde{L}_W(s)$ in (5.3), where the $N(0, 1)$ density kernel is used and $h = SD(S)N^{-1/5}$ is used as a rule of thumb; in each graph, the two dashed lines form (point-wise) 95 % confidence intervals obtained by nonparametric bootstrap. Although the nonparametric bootstrap may look ad-hoc, judging by s at which the confidence intervals exclude 0, Fig. 2 agrees well with Table 3 in terms of the break locations.

For the continuity of f_S, we obtained $\tilde{J}(s)$ in (5.4). Since the observation unit is a class, not a school, the schools with enrollment larger than 40 appear multiple times in the data, which could generate an artificial break in f_S. Hence we picked only one observation randomly from each school for $\tilde{J}(s)$. Figure 3 shows that f_S has multiple breaks around 40, which may be viewed as breaks of $E(U|S)$ to violate the RD premise.

8 Conclusions

We reviewed regression discontinuity (RD) to show its essentials. We did the review following the logical order: identification, estimation, and specification tests; also

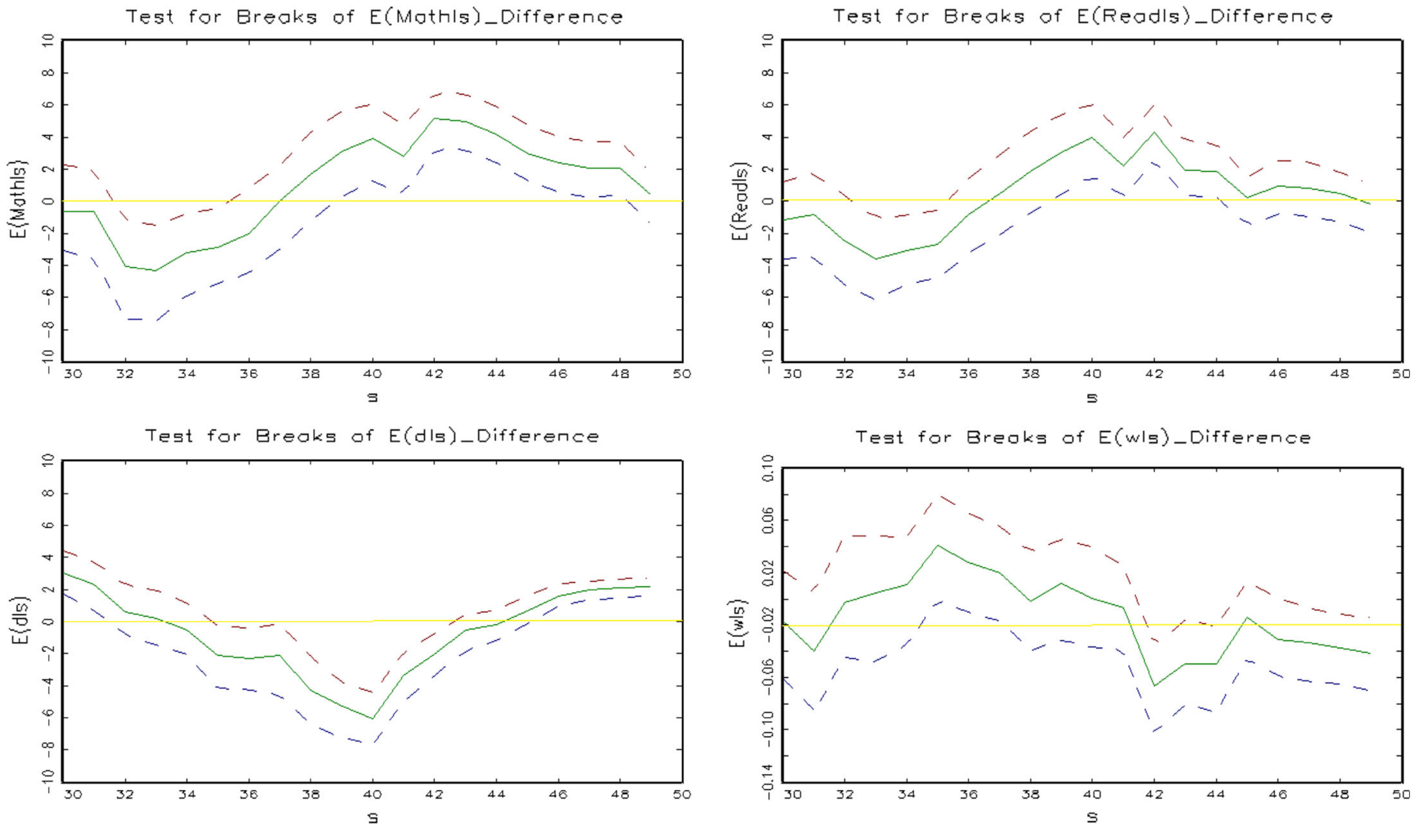

Fig. 2 Conditional mean break versus enrolment

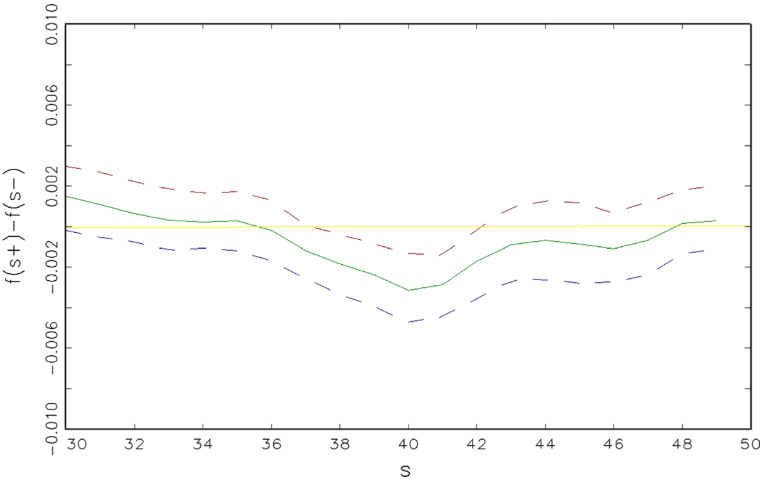

Fig. 3 Density break versus enrolment

some RD topics were examined and an empirical illustration was provided. Since RD provides local randomization which is not easily available in other study designs, the applicability of RD is high for observational data so long as the treatment of interest meets (part of) the RD requirement: an underlying continuous variable crosses an institutional/legal cutoff to get treated. This bodes well for RD, as there are many such examples in laws, policies and programs in the world.

Acknowledgements The authors are grateful to the Editor and two reviewers for their helpful comments. The research of Myoung-jae Lee has been supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5A2A01009718).

Appendix

Identified RD effect when continuity fails

Rewrite (2.1) as

$$E(Y|S) = \mathring{\beta}_d E(D|S) + \mathring{m}(S) \text{ where } \mathring{m}(S) \equiv m(S) - \frac{m(0^+) - m(0^-)}{E(D|0^+) - E(D|0^-)} E(D|S).$$

It holds that

$$E\{\mathring{m}(S)|0^+\} = m(0^+) - \frac{m(0^+) - m(0^-)}{E(D|0^+) - E(D|0^-)} E(D|0^+),$$
$$E\{\mathring{m}(S)|0^-\} = m(0^-) - \frac{m(0^+) - m(0^-)}{E(D|0^+) - E(D|0^-)} E(D|0^-).$$

Then $E\{\mathring{m}(S)|S\}$ is continuous at $S = 0$ because

$$E\{\mathring{m}(S)|0^+\} - E\{\mathring{m}(S)|0^-\} = m(0^+) - m(0^-) - \{m(0^+) - m(0^-)\} = 0.$$

Optimal plug-in bandwidth

The optimal bandwidth h_{opt} proposed by Imbens and Kalyanaraman (2012) for SRD is

$$h_{opt} = C_k \left[\frac{\hat{\sigma}^2(0^-) + \hat{\sigma}^2(0^+)}{\hat{f}_S(0)[\{\hat{m}^{(2)}(0^+) - \hat{m}^{(2)}(0^-)\}^2 + \hat{r}_- + \hat{r}_+]} \right]^{1/5} N^{-1/5}$$

where C_k is a kernel-dependent constant (5.40 for the uniform kernel and 3.44 for the triangular kernel $K(t) = \max(0, 1-|t|) = (1-|t|)1[|t| < 1]$), $\hat{f}_S(0)$ is an estimator for $f_S(0)$, $\hat{\sigma}^2(s)$ is an estimator for $V(Y|s)$, $\hat{m}^{(2)}(s)$ is an estimator for the second derivative of $E(Y|s)$, and $\hat{r}_-$ and $\hat{r}_+$ are functions of $\hat{\sigma}^2(0^-)$ and $\hat{\sigma}^2(0^+)$. Obtain h_{opt} with the following three steps.

First, using a pilot rule-of-thumb bandwidth $h_1 = 1.84 \cdot SD(S) \cdot N^{-1/5}$ for the uniform kernel, let $Q^-_{1i} \equiv 1[-h_1 < S_i < 0]$ and $Q^+_{1i} \equiv 1[0 < S_i < h_1]$ and

$$\hat{\sigma}^2(0^-) \equiv \frac{1}{N_{1-} - 1} \sum_i Q^-_{1i}(Y_i - \overline{Y}^-_1)^2 \quad \text{where } N_{1-} \equiv \sum Q^-_{1i} \text{ and } \overline{Y}^-_1 \equiv \frac{1}{N_{1-}} \sum_i Q^-_{1i} Y_i,$$

$$\hat{\sigma}^2(0^+) \equiv \frac{1}{N_{1+} - 1} \sum_i Q^+_{1i}(Y_i - \overline{Y}^+_1)^2 \quad \text{where } N_{1+} \equiv \sum Q^+_{1i} \text{ and } \overline{Y}^+_1 \equiv \frac{1}{N_{1+}} \sum_i Q^+_{1i} Y_i,$$

$$\hat{f}_S(0) = \frac{N_{1-} + N_{1+}}{2Nh_1}.$$

Second, set the second pilot bandwidths:

$$h_2^+ = 3.56 \left[\frac{\hat{\sigma}^2(0^+)}{\hat{f}_S(0)\{\hat{m}^{(3)}(0)\}^2} \right]^{1/7} N_{1+}^{-1/7} \text{ and } h_2^- = 3.56 \left[\frac{\hat{\sigma}^2(0^-)}{\hat{f}_S(0)\{\hat{m}^{(3)}(0)\}^2} \right]^{1/7} N_{1-}^{-1/7}$$

where $\hat{m}^{(3)}(0)$ for the third derivative of $E(Y|0)$ is to be obtained as $6\hat{\gamma}_4$ from the LSE to

$$Y = \gamma_0 + \gamma_1 Z + \gamma_2 S + \gamma_3 S^2 + \gamma_4 S^3 + error.$$

Let $Q^-_{2i} \equiv 1[-h_2^- < S_i < 0]$ and $Q^+_{2i} \equiv 1[0 < S_i < h_2^+]$. Apply LSE to

$$Q_2^- Y = Q_2^-(\lambda_0^- + \lambda_1^- S + \lambda_2^- S^2 + U^-) \text{ and } Q_2^+ Y = Q_2^+(\lambda_0^+ + \lambda_1^+ S + \lambda_2^+ S^2 + U^+)$$

to obtain $\hat{m}^{(2)}(0^-) = 2\hat{\lambda}_2^-$ and $\hat{m}^{(2)}(0^+) = 2\hat{\lambda}_2^+$.

Finally, obtain

$$\hat{r}_- = \frac{2160 \cdot \hat{\sigma}^2(0^-)}{N_{2-}(h_2^-)^4} \;\&\; \hat{r}_+ = \frac{2160 \cdot \hat{\sigma}^2(0^+)}{N_{2+}(h_2^+)^4} \quad \text{where } N_{2-} \equiv \sum_i Q^-_{2i} \;\&\; N_{2+} \equiv \sum_i Q^+_{2i}$$

and then h_{opt}. Imbens and Kalyanaraman (2012) proposed an optimal bandwidth for FRD in their Eq. (2.4), but then noted that the above h_{opt} for SRD would be adequate even for FRD, as it is based on the numerator estimation in FRD.

LLR-type score-density continuity test

Let h_1 be the interval size for the first-stage histogram of the McCrary (2008) test, and for an even number n, there are $n/2$ intervals to be constructed on either side of 0. Consider left and right intervals of width h_1 around 0:

$$[-jh_1,\ -(j-1)h_1),\ \ j = \frac{n}{2}, \frac{n}{2}-1, \ldots, 1 \quad \& \quad [(j-1)h_1,\ jh_1),\ \ j = 1, 2, \ldots, \frac{n}{2}.$$

The midpoints of these intervals are

$$G_j \equiv -(j-0.5)h_1,\quad j = \frac{n}{2}, \frac{n}{2} - 1, \ldots, 1 \quad \& \quad (j-0.5)h_1,\quad j = 1, 2, \ldots, \frac{n}{2};$$

'G' stands for grid points and there are n midpoints. Let R_j denote the histogram height at G_j; (R_j, G_j), $j = 1, \ldots, n$ are to be taken as " observations" in the second stage.

The test is based on $\ln \hat{\varphi}_0 - \ln \hat{\psi}_0$ that come from the LLR intercept estimates in (h_2 is a second-stage bandwidth)

$$\min_{\varphi_0, \varphi_1} \sum_{j=1}^{n} (R_j - \varphi_0 - \varphi_1 G_j)^2 K(\frac{G_j}{h_2}) 1[0 < G_j],$$
$$\min_{\psi_0, \psi_1} \sum_{j=1}^{n} (R_j - \psi_0 - \psi_1 G_j)^2 K(\frac{G_j}{h_2}) 1[G_j < 0].$$

Specifically, the test statistic is

$$\hat{\theta} \equiv \ln \hat{f}^+ - \ln \hat{f}^- \text{ where } \hat{f}^+ \equiv \sum_j K\Big(\frac{G_j}{h_2}\Big) 1[0 < G_j] \frac{H_2^+ - H_1^+ G_j}{H_2^+ H_0^+ - (H_1^+)^2} R_j,$$
$$\hat{f}^- \equiv \sum_j K\Big(\frac{G_j}{h_2}\Big) 1[G_j < 0] \frac{H_2^- - H_1^- G_j}{H_2^- H_0^- - (H_1^-)^2} R_j,$$
$$H_k^+ \equiv \sum_j K\Big(\frac{G_j}{h_2}\Big) 1[0 < G_j] {G_j}^k \text{ and } H_k^- \equiv \sum_j K\Big(\frac{G_j}{h_2}\Big) 1[G_j < 0] {G_j}^k.$$

As for the asymptotic distribution, the proposition in McCrary (2008), p.702) is that, for the triangular kernel, under $h_2 \to 0$, $Nh_2 \to \infty$, $h_1/h_2 \to 0$, and $h_2^2\sqrt{Nh_2} \to$

$B \in [0, \infty)$,

$$\sqrt{Nh_2}(\hat{\theta} - \theta) \rightsquigarrow N\left\{\frac{B}{20}\left(\frac{-f^{+''}}{f^{+}} - \frac{-f^{-''}}{f^{-}}\right), \frac{24}{5}\left(\frac{1}{f^{+}} + \frac{1}{f^{-}}\right)\right\}$$

where $f^{+} \equiv f_S(0^{+})$, $f^{-} \equiv f_S(0^{-})$, $\theta \equiv \ln f^{+} - \ln f^{-}$ and '$''$' denotes the second derivative. With under-smoothing $h_2^2\sqrt{Nh_2} \to 0$, the asymptotic bias can be ignored. In practice, adopt the under-smoothing to ignore the asymptotic bias. Choosing two bandwidths h_1 and h_2 is a problem. McCrary (2008) stated that the choice of h_1 does not matter much whereas the choice of h_2 does, which is questionable though; (McCrary 2008, p. 705) suggested $h_1 = 2SD(S)N^{-1/2}$ (then h_2 may be chosen by visual inspection).

References

Abadie A (2002) Bootstrap tests for distributional treatment effects in instrumental variable models. J Am Stat Assoc 97:284–292

Almond D, Doyle JJ Jr, Kowalski AE, Willimans H (2010) Estimating marginal returns to medical care: evidence from at-risk newborns. Q J Econ 125:591–634

Almond D, Doyld JJ Jr, Kowalski AE, Williams H (2011) The role of hospital heterogeneity in measuring marginal returns to medical care: a reply to Barreca, Guldi, Lindo, and Waddell. Q J Econ 126:2125–2131

Angrist JD, Lavy V (1999) Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q J Econ 114:533–575

Angrist JD, Rokkanen M (2015) Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. J Am Stat Assoc 10:1331–1344

Barreca AI, Guldi M, Lindo JM, Waddell GR (2011) Saving babies? Revisiting the effect of very low birth weight classification. Q J Econ 126:1–7

Barreca AI, Lindo JM, Waddell GR (2016) Heaping-induced bias in regression-discontinuity designs. Econ Inq 54:268–293

Battistin E, Brugiavini A, Rettore E, Weber G (2009) The retirement consumption puzzle: evidence from a regression discontinuity approach. Am Econ Rev 99:2209–2226

Battistin E, Rettore E (2002) Testing for programme effects in a regression discontinuity design with imperfect compliance. J R Stat Soc (Ser A) 165:39–57

Berk RA, de Leeuw J (1999) An evaluation of California's inmate classification system using a generalized regression discontinuity design. J Am Stat Assoc 94:1045–1052

Berk RA, Rauma D (1983) Capitalizing on nonrandom assignment to treatments: a regression-discontinuity evaluation of a crime control program. J Am Stat Assoc 78:21–27

Bertanha M (2015) Regression discontinuity design with many thresholds **unpublished paper**

Bertanha M, Imbens GW (2014) External validity in fuzzy regression discontinuity designs. NBER working paper 20773

Breitung J, Kruse R (2013) When bubbles burst: econometric tests based on structural breaks. Stat Pap 54:911–930

Caliendo M, Tatsiramos K, Uhlendorff A (2013) Benefit duration, unemployment duration and job match quality: a regression-discontinuity approach. J Appl Econ 28:604–627

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

Calonico S, Cattaneo MD, Titiunik R (2015) Optimal data-driven regression discontinuity plots. J Am Stat Assoc 10:1753–1769

Cappelleri JC, Trochim WMK, Stanley TD, Reichardt CS (1991) Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. the case of no interaction. Eval Rev 15:395–419

Card D, Lee DS, Pei Z, Weber A (2012) Nonlinear policy rules and the identification and estimation of causal effects in a generalized regression kink design, NBER Working Paper 18564

Card D, Lee DS, Pei Z, Weber A (2015) Inference on causal effects in a generalized regression kink design. Econometrica 83:2453–2483

Cattaneo MD, Frandsen B, Titiunik R (2015) Randomization Inference in the regression discontinuity design: an application to party advantages in the U.S. Senate. J Causal Inference 3(1):1–24

Choi JY, Lee MJ (2015) Regression discontinuity with multiple running variables allowing partial effects, presented at the World Congress Meeting of the Econometric Society at Montreal

Ciuperca G (2014) Model selection by LASSO methods in a change-point model. Stat Pap 55:349–374

Clark D, Martorell P (2014) The signaling value of a high school diploma. J Polit Econ 122:282–318

Cook TD (2008) " Waiting for life to arrive" : a history of the regression-discontinuity design in psychology, statistics and economics. J Econ 142:636–654

Crawford C, Dearden L, Greaves E (2014) The drivers of month-of-birth differences in children's cognitive and non-cognitive skills. J R Stat Soc (Ser A) 177:829–860

Dickens R, Riley R, Wilkinson D (2014) The UK minimum wage at 22 years of age: a regression discontinuity approach. J R Stat Soc (Ser A) 177:95–114

Dong Y (2014) Jump or kink? Identification of binary treatment regression discontinuity design without the discontinuity. R&R J Political Econ **forthcoming**

Dong Y (2015) Regression discontinuity applications with rounding errors in the running variable. J Appl Econ 30:422–446

Dong Y, Lewbel A (2015) Identifying the effect of changing the policy threshold in regression discontinuity models. Rev Econ Stat 97:1081–1092

Feir D, Lemieux T, Marmer V (2015) Weak identification in fuzzy regression discontinuity designs. J Bus Econ Stat **forthcoming**

Frandsen BR, Frölich M, Melly B (2012) Quantile treatment effects in the regression discontinuity design. J Econ 168:382–395

Gerard F, Rokkanen M, Rothe M (2015) Partial identification in regression discontinuity designs with manipulated running variables **unpublished paper**

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

Hansen BB (2008) The prognostic analogue of the propensity score. Biometrika 95:481–488

Hullegie P, Klein TJ (2010) The effect of private health insurance on medical care utilization and self-assessed health in Germany. Health Econ 19:1048–1062

Imbens GW (2000) The role of the propensity score in estimating dose-response functions. Biometrika 87:706–710

Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 62:467–475

Imbens GW, Kalyanaraman K (2012) Optimal bandwidth choice for the regression discontinuity estimator. Rev Econ Stud 79:933–959

Imbens GW, Lemieux T (2008) Regression discontinuity designs: a guide to practice. J Econ 142:615–635

Imbens GW, Zajonc T (2009) Regression discontinuity design with vector-argument assignment rules **unpublished paper**

Jacob BA, Lefgren L (2004) Remedial education and student achievement: a regression discontinuity analysis. Rev Econ Stat 86:226–244

Keele LJ, Titiunik R (2015) Geographic boundaries as regression discontinuities. Polit Anal 23:127–155

Keele LJ, Titiunik R, Zubizarreta JR (2015) Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. J R Stat Soc (Ser A) 178:223–239

Kim YS, Lee MJ (2016) Regression-kink approach for wage effect on male work hours. Oxf Bull Econ Stat **forthcoming**

Lalive R (2008) How do extended benefits affect unemployment duration? Regression discontinuity approach. J Econ 142:785–806

Lechner M (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner M, Pfeiffer F (eds) Econometric evaluation of labor market policies. Physica-Verlag, New York, pp 43–58

Lee DS, Card D (2008) Regression discontinuity inference with specification error. J Econ 142:655–674

Lee DS, Lemieux T (2010) Regression discontinuity designs in economics. J Econ Lit 48:281–355

Lee MJ (2000) Median treatment effect in randomized trials. J R Stat Soc (Ser B) 62:595–604

Lee MJ (2005) Micro-econometrics for policy, program, and treatment effects. Oxford University Press, Oxford

Lee MJ (2016) Regression discontinuity with errors in the running variable: effect on truthful margin. J Econ Methods **forthcoming**

Leuven E, Lindahl M, Oosterbeek H, Webbink D (2007) The effect of extra funding for disadvantaged pupils on achievement. Rev Econ Stat 89:721–736

Ludwig J, Miller D (2007) Does head start improve children's life chances? evidence from a regression discontinuity design. Q J Econ 122:159–208

Matsudaira JD (2008) Mandatory summer school and student achievement. J Econ 142:829–850

McCrary J (2008) Manipulation of the running variable in the regression discontinuity design: a density test. J Econ 142:698–714

MacDonald JM, Klick J, Grunwald B (2016) The effect of private police on crime: evidence from a geographic regression discontinuity design. J R Stat Soc (Ser A) **forthcoming**

Mealli F, Rampichini C (2012) Evaluating the effects of university grants by using regression discontinuity designs. J R Stat Soc (Ser A) 175:775–798

Nielsen HS, Sorensen T, Taber CR (2010) Estimating the effect of student aid on college enrollment: evidence from a government grant policy reform. Am Econ J Econ Policy 2(2):185–215

Otsu T, Xu KL, Matsushita Y (2013) Estimation and inference of discontinuity in density. J Bus Econ Stat 31:507–524

Otsu T, Xu KL, Matsushita Y (2015) Empirical likelihood for regression discontinuity design. J Econ 186:94–112

Papay JP, Murnane RJ, Willett JB (2011) Extending the regression discontinuity approach to multiple assignment variables. J Econ 161:203–207

Pei Z (2011) Regression discontinuity design with measurement error in the assignment variable **unpublished paper**

Porter J (2003) Estimation in the regression discontinuity model. Department of Economics, University of Wisconsin, Madison **unpublished paper**

Porter J, Yu P (2015) Regression discontinuity designs with unknown discontinuity points: testing and estimation. J Econ 189:132–147

Qiu P (2005) Image processing and jump regression analysis. Wiley, Hoboken

Robbins H, Zhang CH (1991) Estimating a multiplicative treatment effect under biased allocation. Biometrika 78:349–354

Schanzenbach DW (2009) Do school lunches contribute to childhood obesity? J Hum Resour 44:684–709

Schmieder JF, Wachter TV, Bender S (2012) The effects of extended unemployment insurance over the business cycle: evidence from regression discontinuity estimates over 20 Years. Q J Econ 127:701–752

Simonsen M, Skipper L, Skipper N (2015) Price sensitivity of demand for prescription drugs: exploiting a regression kink design. J Appl Econ **forthcoming**

Thistlethwaite D, Campbell D (1960) Regression-discontinuity analysis: an alternative to the ex post facto experiment. J Educ Psychol 51:309–317

Urquiola M, Verhoogen E (2009) Class-size caps, sorting, and the regression-discontinuity design. Am Econ Rev 99:179–215

Uysal SD (2015) Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. J Appl Econ 30:763–786

Van der Klaauw W (2002) Estimating the effect of financial aid offers on college enrollment: a regression discontinuity approach. Int Econ Rev 43:1249–1287

Van der Klaauw W (2008) Regression-discontinuity analysis: a survey of recent developments in economics. Labour 22:219–245

Wong VC, Steiner PM, Cook TD (2013) Analyzing regression discontinuity designs with multiple assignment variables: a comparative study of four estimation methods. J Educ Behav Stat 38:107–141

Yanagi T (2014) The effect of measurement error in the sharp regression discontinuity design. KIER Discussion Paper, No 910

Yu P (2012) Identification in regression discontinuity designs with measurement error **unpublished paper**

Yu P (2015) Understanding estimators of treatment effects in regression discontinuity designs. Econ Rev **forthcoming**