Journal of Economic Theory and Econometrics, Vol. 32, No. 3, Sep. 2021, 1–68

Basics and Recent Advances in Regression Discontinuity: Difference versus Regression Forms*

Jin-young Choi† Myoung-jae Lee‡

Abstract This paper reviews the basics of regression discontinuity (RD) design, whose hallmark is having a treatment determined by an underlying score (i.e., 'running variable') crossing a known cutoff or not. Following the basics, recent advances in RD are examined, where the topics are grouped into those related to score and those not. The former topics include multiple scores, measurement errors in score, integer score, and score-density continuity. The latter topics include regression kink (RK), high-order effects, and extending RD identification range (i.e., external validity). Detailed empirical examples are provided for the RD topics, but not for the RD basics which are fairly well-known these days. RD is simple, which can thus appeal even to lay audiences, and this review accordingly emphasizes the intuitive nature of RD and its applicability in practice. Practical and widely applicable techniques are given more coverage, whereas theoretically-motivated but less-practically-relevant ones are only briefly mentioned. The beauty of RD is in its simplicity, and temptation to make it too sophisticated should be resisted.

Keywords Regression Discontinuity, Score, Cutoff, Local Randomization, Instrument, Regression Kink, High-order Effects

JEL Classification C24, C31, C33, C36

*The authors are grateful to two anonymous reviewers for their helpful comments. Myoung-jae Lee's research has been supported by Korea University Research Grant. Regarding RD software, "https://rdpackages.github.io" should be helpful for RD programs written in `R`, Python or Stata. In Stata, the main RD package is `rdrobust`. Thoemmes *et al.* (2017) compares the `R` packages for RD: `rdd`, `rddtools`, `rdrobust` and `rddapp`. Also, simple RD programs written in GAUSS are available for Lee (2016) in his *Google* homepage.

†Chow Center, WISE, and School of Economics, Xiamen University, 422 Siming S Rd, Siming District, Xiamen, Fujian, China, 361005. E-mail: jychoi@xmu.edu.cn

‡Corresponding author. Department of Economics, Korea University, 145 Anam-ro, Anam-dong, Seongbuk-gu, Seoul, Republic of Korea. E-mail: myoungjae@korea.ac.kr

Received April 14, 2021, Revised September 9, 2021, Accepted September 9, 2021

1. BASICS: LOCAL RANDOMIZATION IN RD

In assessing effects of a binary treatment D on a response/outcome variable Y, randomization is the golden rule of inference in science. In a randomized experiment with treatment ($D = 1$) and control ($D = 0$) groups, randomization of D ensures that the two groups differ only in the treatment status, otherwise being "balanced" in all covariates, observed or not. For instance, individuals with different abilities or genes are assigned to different groups, but the distribution of ability levels or gene types should be almost the same across the two groups. Hence the mean difference $E(Y|D = 1) - E(Y|D = 0)$ reveals the mean effect of D on Y.

If the treatment group ("T group") and control group ("C group") systematically differ in some covariates, then we cannot attribute $E(Y|D = 1) \neq E(Y|D = 0)$ to the D difference, which is why covariate balance is critical. When individuals self-select $D = 0, 1$, the two groups are likely to be unbalanced in covariates, as individuals with certain traits tend to select $D = 0$ or 1; e.g., older persons tend to vote ($D = 1$) more than younger ones, and the more educated tend to exercise ($D = 1$) more than the less educated. Randomization avoids this kind of problems.

Randomization, however, 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. Despite this, regression discontinuity (RD), started long time ago by Thistlethwaite and Campbell (1960), offers 'local randomization' using an institutional or legal break/intervention.

The hallmark of RD is that D is fully (or partly) determined by an underlying 'running variable' or 'score' S crossing a known cutoff c or not, so that D (or $E(D|S)$) has a break at c. Such a break happens because accommodating a partial qualification/admittance (to a program/treatment) is too costly or infeasible. If we use only some local observations around c, then the local T group with S just above c and the local C group with S just below c should be similar in all aspects except in the treatment status, which is the local randomization with observational data. An example is entering a college based on a test score S being at least c—a "partial entry" is hard to imagine here—and those who barely enter and those who barely fail should be similar in all aspects.

RD local randomization differs from the usual non-local randomization as in flipping a coin. Because RD uses only local observations around c with observations far away from c discarded, RD is not efficient. If one does an experiment as in a medical trial, there is no reason to use RD; i.e., D there does not have to

be determined by S crossing c or not. The real value of RD is materialized in observational data.

After the birth of RD by Thistlethwaite and Campbell (1960), RD went dormant for decades until its revival around the year 2000. Review papers on RD have appeared in various disciplines: Imbens and Lemieux (2008), Lee and Lemieux (2010), Bloom (2012), Venkataramani *et al.* (2016) and Choi and Lee (2017), among others. Book chapters on RD can be found in Lee (2005, 2016) and Angrist and Pischke (2009). Cattaneo and Escanciano (2017) and Cattaneo *et al.* (2019) are books entirely on RD. See Choi and Lee (2017) for RD studies in statistics where the main outlet has been *Journal of the Royal Statistical Society (Series A)* and *Journal of the American Statistical Association*. See also Cook (2008) for a historical account on RD. The goal of this paper is to introduce recent advances in RD, for which the RD basics are covered as well.

Regarding notation and assumptions, we assume iid observations across $i = 1,...,N$. The subscript i as in D_i and Y_i is often omitted, as has been done already. We often write $E(Y|S=s)$ just as $E(Y|s)$ for a realized value s of S. $\int E(Y|s)ds$ is often written as $\int E(Y|s)\partial s$ to avoid the confusion with d in $D=d$. Estimates for a parameter β are denoted with '$\hat{}$', '$\tilde{}$', '–', etc. as in $\hat{\beta}$, $\tilde{\beta}$, $\bar{\beta}$, etc. The distribution or distribution function for $W|S$ is denoted as $F_{W|S}(\cdot|s)$, and its density as $f_{W|S}(w|s)$. *The continuity of various functions at c matters, and we often omit the qualification "at c". For simplification, redefine S as* $S-c$ *so that the cutoff becomes* 0 *unless otherwise mentioned, although c is still used when its specific value may be of interest.* The starred sections in this review are relatively more technical, and may be skipped.

2. BASICS: SHARP RD (SRD)

Suppose D is fully determined by a continuous variable S crossing the cutoff 0: $D = 1[0 \leq S]$, where $1[A] \equiv 1$ if A holds and 0 otherwise. This is 'sharp RD (SRD)', compared with 'fuzzy RD (FRD)' where D is only partly determined by S. S is often called a running/forcing/assignment variable, but we call S a 'score' (S from score).

For instance, to enter a competitive college ($D=1$), a test score S should be equal to or greater than a cutoff c. Local randomization is that those who barely pass the test (the local T group with '$S=c^+$', i.e., S just above c) and those who barely fail (the local C group with '$S=c^-$', i.e., S just below c) should be homogeneous in all covariates, observed or not, because a few point difference in a test (with the maximum score, say, 100) should be a matter of pure luck.

Another example is that S is age, $c = 65$ and D is a government-provided health care. The local T group who are a couple of weeks/months older than 65 should be almost the same as the local C group a couple of weeks/months younger than 65.

Treatments of the opposite direction also abound; e.g., an income aid program with $D = 1[S < c]$ where S is income and c is a poverty threshold. With little loss in generality, *we assume* $D = 1[c \leq S]$ *henceforth, neither* $1[S < c]$ *nor* $1[c < S]$*, unless otherwise necessary*. If $D = 1[S < c]$, then we just have to define the new treatment $D' \equiv 1 - 1[S < c] = 1[c \leq S]$ to switch the T and C group labels.

2.1. SRD IDENTIFICATION WITH INTERCEPT BREAK

Having seen how the treatment D is determined in RD, now we turn to how D relates to Y. Suppose D affects Y through

$$\begin{aligned} E(Y|S) &= \beta_d D + m(S) = \beta_d 1[0 \leq S] + m(S) \\ &= m(S) \text{ for } S < 0 \quad \text{and} \quad \beta_d + m(S) \text{ for } S \geq 0 \end{aligned}$$

for a parameter β_d and an unknown function $m(S)$ continuous at $S = 0$. As S changes from just below 0 (i.e., 0^-) to just above 0 (i.e., 0^+), $E(Y|S)$ changes from $m(0^-)$ to $\beta_d + m(0^+)$. The break of size β_d in $E(Y|S)$ at $S = 0$ occurs due to the break of size 1 in D at $S = 0$.

Suppose, as S changes from 0^- to 0^+, we see a break in $E(Y|S)$. Since only D had a break at $S = 0$ while everything else did not—this is what the continuity $m(S)$ at $S = 0$ essentially means—it must be the D break that caused the break in $E(Y|S)$. Finding a causal effect of D on Y using a break gives the name RD, and RD concludes no causal effect of D if $E(Y|S)$ has no break at $S = 0$ despite that D has a break. The well-known 'before-after (BA)' design is a special case of RD when S is time.

In $E(Y|S) = \beta_d D + m(S)$, S affects Y directly through $m(S)$ ('$S \to Y$') and indirectly through D ('$S \to D \to Y$'). In the college entrance example, S reflecting ability can affect Y directly, but there is no reason for this effect through $m(S)$ to be discontinuous at 0. This contrast between the treatment break in D and no break in $m(S)$ is the source to identify β_d.

Let the right and left limits of $E(\cdot|S)$ at c be

$$E(\cdot|c^+) \equiv \lim_{s \downarrow c} E(\cdot|S = s) \quad \text{and} \quad E(\cdot|c^-) \equiv \lim_{s \uparrow c} E(\cdot|S = s).$$

As Lee (2016, p. 102) shows (see also the appendix), *the "regression form"* $E(Y|S) = \beta_d D + m(S)$ *is not an assumption; rather, it is equivalent to the "difference form"* $\beta_d \equiv E(Y|0^+) - E(Y|0^-)$. The regression form is not a structural form (SF) (i.e., data generating process), but it is not a typical reduced form (RF) either because the SF parameter β_d appears there. See Lee (2018, 2021c) for "reduced structural forms" capturing a SF parameter with a RF.

Let (Y^0, Y^1) be the 'potential responses', so that $Y = (1-D)Y^0 + DY^1$; i.e., the untreated response Y^0 is observed when $D = 0$, and the treated response Y^1 when $D = 1$. Then

$$\begin{aligned} \beta_d &\equiv E(Y|0^+) - E(Y|0^-) = E(Y^1|0^+) - E(Y^0|0^-) \\ &= E(Y^1|0^+) - E(Y^0|0^+) = E(Y^1 - Y^0|0^+) \quad \text{if } E(Y^0|0^+) = E(Y^0|0^-), \end{aligned} \tag{2.1}$$

which is the *continuity of* $E(Y^0|S)$ *at* 0. This shows that *the local mean difference* $E(Y|0^+) - E(Y|0^-)$ *identifies the mean treatment effect on the just treated* (i.e., those with $S = 0^+$) which is $E(Y^1 - Y^0|0^+)$. If we further assume the continuity of $E(Y^1|S)$ at 0, then $\beta_d = E(Y^1 - Y^0|0^+) = E(Y^1 - Y^0|0)$ which is the *mean effect at the cutoff* 0.

Interpreting the difference $Y^1 - Y^0$ requires that differences of Y values be comparable, which allows continuous or counting Y. Binary Y is allowed too, because the differences $(-1, 0, 1)$ are comparable: Y changing 1 to 0, no change in Y, and Y changing 0 to 1, respectively. However, categorical/multinomial Y is not allowed which may be ordered or unordered; e.g., Y changing from category 1 to 3 is not comparable to Y changing from category 2 to 4, despite that both changes result in the same difference $\Delta Y = 2$. For categorical Y, define binary Y_j for category j to use $E(Y_j|0^+) - E(Y_j|0^-)$.

The critical identification (ID) condition $E(Y^0|0^+) = E(Y^0|0^-)$ in (2.1) is best understood by taking plastic surgery as a BA example. In BA, we compare Y^0 just before the treatment (i.e., the facial beauty before the surgery) to Y^1 just after (the facial beauty after) to take $E(Y|0^+) - E(Y|0^-) = E(Y^1|0^+) - E(Y^0|0^-)$ as the effect. For this comparison, it is essential to assume that Y^0 would have stayed the same, had it not been for the treatment, which is $E(Y^0|0^+) = E(Y^0|0^-)$; i.e., the facial beauty would have remained the same if no surgery.

The "opposite" to BA is 'spatial/geographical' RD where a boundary line is drawn abruptly in a region and S is the shortest distance to the boundary. An example is the African country boundaries drawn by Europeans in the 19th century, splitting more than 200 ethnicities into different countries. Also, the 38 parallel divided Korea into South and North after World War II, to be governed

by the U.S. and Russia. Individuals close to such a boundary are homogeneous to result in local randomization, because the boundary was drawn arbitrarily. For spatial RD examples, see Bayer *et al.* (2007), Dell (2010), Michalopoulos abd Papaioannou (2014), Turner *et al.* (2014), MacDonald *et al.* (2016), and Keele *et al.* (2017) among many others. Spatial local randomization, however, may not last long, because effects of different institutions take place over time (see Henderson *et al.* (2012) for drastic examples "from above"), and also because people selectively move around if moving is allowed.

2.2. SRD ESTIMATION WITH OLS

As for estimating $\beta_d \equiv E(Y|0^+) - E(Y|0^-)$, the simplest approach is using a sample analog of $E(Y|0^+) - E(Y|0^-)$: for a small constant $h > 0$,

$$\bar{\beta}_d \equiv \hat{E}(Y|0^+) - \hat{E}(Y|0^-) \quad \text{where}$$
$$\hat{E}(Y|0^+) \equiv \frac{1}{N_1}\sum_i Y_i 1[S_i \in (0,h)], \quad \hat{E}(Y|0^-) \equiv \frac{1}{N_0}\sum_i Y_i 1[S_i \in (-h,0)],$$

and N_1 and N_0 are the numbers of observations with $S \in (0,h)$ and $S \in (-h,0)$. Here, h is a 'bandwidth' or 'smoothing/tuning parameter'; how to choose h will be discussed later.

We can estimate β_d also using $E(Y|S) = \beta_d D + m(S)$. As $m(S)$ is continuous at 0, the simplest approach is replacing $m(S)$ with a constant β_0 to get $E(Y|S) = \beta_0 + \beta_d D$; $m(S) = \beta_0$ is a "trivially" continuous function of S. We can estimate (β_0, β_d) with the ordinary least squares estimator (OLS) of Y on $(1, D)$ using only the observations with $Q = 1$, where

$$Q_i \equiv 1[S_i \in (-h,h)].$$

As well-known, the slope estimand of the OLS of Y on $(1, D)$ is the mean difference $E(Y|D=1) - E(Y|D=0)$ (see, e.g., Lee 2016, p. 19), and when only the observations with $Q = 1$ are used, the slope estimand becomes, with $D = 1[0 \le S]$,

$$E(Y|D=1, Q=1) - E(Y|D=0, Q=1) = E(Y|0 \le S < h) - E(Y|-h < S < 0).$$

This reveals that the slope of the OLS is also $\bar{\beta}_d$. Setting $m(S) = \beta_0$ is a '*Local Constant regression (LCR)*', and the OLS is a 'local-constant OLS'.

Going one step further from $m(S) = \beta_0$, we can replace $m(S)$ with a linear function of S continuous at 0, which gives rise to '*Local Linear Regression*

(LLR)'. Specifically, let $m(S)$ be a 'linear spline' (or a piecewise linear function):

$$m(S) = \beta_0 + \beta_-(1-\delta)S + \beta_+\delta S = \beta_0 + \beta_- S + \Delta\beta \cdot \delta S,$$
$$\delta_i \equiv 1[0 \le S_i], \quad \Delta\beta \equiv \beta_+ - \beta_-.$$

Here, β_0 is the intercept, β_- is the left slope of S at 0, and β_+ is the right slope, or alternatively, β_- is the "base" slope and $\Delta\beta$ is the slope change at 0. Clearly, $m(S) = \beta_0 + \beta_-(1-\delta)S + \beta_+\delta S$ is continuous at 0, because of $m(0^-) = m(0^+) = \beta_0$ and the terms with S attached becoming 0 at $S = 0$. Although $\delta = D$ here, RD with $D \neq \delta$ will appear shortly.

'Local linear OLS' uses only the $Q = 1$ observations to estimate

$$E(Y|S) = \beta_0 + \beta_d D + \beta_-(1-\delta)S + \beta_+\delta S$$

which is almost "the industry standard" in RD practice. Despite the linear model and OLS, *RD is a nonparametric approach because the linear model is used only locally around* 0*, not globally. The asymptotic inference can be done, using the usual OLS asymptotic variance estimator.*

Biases in LCR can be much larger than those in LLR. Suppose $\beta_d = 0$ and $Y = \beta_0 + \beta_1 S$ with $\beta_1 > 0$. As Y is an increasing function of S, the left sample mean $\hat{E}(Y|0^-)$ over $S \in (-h, 0)$ is smaller than the right sample mean $\hat{E}(Y|0^+)$ over $S \in (0, h)$, resulting in $\bar{\beta}_d > \beta_d = 0$. Although the bias would disappear as $h \to 0$, it could be substantial in small samples to explain why LLR is more popular than LCR. LLR is "boundary-adaptive".

The word 'spline' in linear spline refers to the case where the right and left slopes at 0 are allowed to differ although the function is continuous at 0. Instead of the linear spline, we may use the linear $\beta_0 + \beta_s S$ with the restriction $\beta_s \equiv \beta_- = \beta_+$. In reality, $\beta_- = \beta_+$ may hold, but it is still good to allow $\beta_- \neq \beta_+$, because D may cause, not just an intercept break in $E(Y|S)$, but also a change in the slope of S. It can happen that $\beta_d = 0$ (no intercept break) but $\beta_- \neq \beta_+$ (slope break), which is called 'regression kink (RK)' to be examined in detail later.

Quadratic $\beta_2 S^2$ (same slope) or $\beta_{2-}(1-\delta)S^2 + \beta_{2+}\delta S^2$ (different slopes) may be used extra for $m(S)$. Gelman and Imbens (2019) recommend a linear or quadratic $m(S)$, but not higher orders for the following reasons. First, β_d can be written as the difference of weighted averages of (Y_i^0, Y_i^1)'s, and the weights for a high order $m(S)$ can be nonsensical. Second, estimates can be sensitive to the order of $m(S)$. Third, inference with a high order $m(S)$ is often poor. Hence, it is enough to use LLR or 'local quadratic regression (LQR)'.

Instead of OLS, weighed OLS (WLS) may be used with a weighting function assigning higher weights to observations closer to the cutoff than those far away. However, WLS hardly differs from the OLS in practice, and thus we do not consider WLS any further. Compared with WLS, OLS gives the same weight to all observations regardless how far off they are from the cutoff. That is, OLS uses the "uniform weight/kernel".

Using W as regressors in observational data amounts to controlling W in experiments. Although unnecessary in principle, covariates W may be controlled in RD; see, e.g., Kim (2013), Calonico *et al.* (2019) and Frölich and Huber (2019). First, controlling W pulls W out of the error term to reduce its variance. Second, if h is large, W may not be balanced to ruin the local randomization, which is avoided by controlling W. Third, $E(W|S)$ may have a break at 0 to bias the OLS (see e.g., Urquiola and Verhoogen (2006), which is also avoided by controlling W.

2.3. BANDWIDTH CHOICE AND COVARIATE BALANCE CHECK

Choosing the localizing bandwidth h matters greatly, as there is a trade-off in choosing h: 'too small' entails too few observations for low efficiency, and 'too large' entails a bias because local randomization breaks down and covariates get unbalanced across the two groups. Theoretically optimal or "robust" bandwidths have been proposed by Imbens and Kalyanaraman (2012), Calonico *et al.* (2014), 2020, and Arai and Ichimura (2018), but they do not necessarily work well in reality; see, e.g., Card *et al.* (2015) and Önder and Shamsuddin (2019) for RK and RD examples. It is hazardous simply to use a canned "default" bandwidth in an econometric/statistical software without checking the sensitivity of the effect estimates to h.

The basic approach is to think of a reasonable bound on h for randomization—how big h has to be to ruin local randomization—and presents estimates corresponding to different values of h within the bound to show the sensitivity of the effect estimate to h. With *SD* denoting standard deviation, *a sensible approach is thus starting from a rule-of-thumb bandwidth h such as* $SD(S)N^{-1/5}$, *and then shrinking/expanding h until the covariate balance is restored/maintained.*

If one still desires an automatic choice of h, then 'cross-validation (CV)' may be used: for a kernel $K(\cdot)$ such as the $N(0,1)$ density, minimize with respect to h,

$$\frac{1}{N}\sum_i \{Y_i - \hat{E}_{-i}(Y|S_i,h)\}^2 \quad \text{where} \quad \hat{E}_{-i}(Y|S_i,h) \equiv \frac{\sum_{j=1,j\neq i}^N K\{(S_j - S_i)/h\}Y_j}{\sum_{j=1,j\neq i}^N K\{(S_j - S_i)/h\}}.$$

That is, the value of h that gives the best predictor for Y_i without using Y_i per se is chosen by CV. The minimand is nearly convex, and the conventional CV bandwidth is asymptotically optimal. The reason why this is not often used in RD is that $E(Y|S)$ has a break, instead of being continuous in S, which makes $\hat{E}_{-i}(Y|S_i,h)$ biased for $E(Y|S_i)$ when $S_i \simeq 0$. However, since the goal is finding a reasonable value for h, not predicting Y per se well, the bias is hardly an issue, and we recommend CV. See Choi and Lee(2018b, pp. 263-4) on why 'one-sided kernel' to account for the break in $E(Y|S)$ fails in choosing h.

In checking out the balance of covariates W, the simplest way is testing for $E(W|0^+) = E(W|0^-)$ as in the LCR for Y. Instead of LCR, however, it is better to employ the same local model as used for Y; i.e., if LCR/LLR/LQR is used for Y, then the same should be used for W. To see why, suppose LLR is used for Y and a covariate W_k:

$$E(W_k|S) = \zeta_0 + \zeta_\delta \delta + \zeta_-(1-\delta)S + \zeta_+ \delta S.$$

Then we can take $\zeta_\delta \neq 0$ as evidence for W_k imbalance, and $\zeta_\delta \neq 0$ would bias $\hat{\beta}_d$ in the LLR for Y when W_k is not controlled for. In contrast, any non-zero ζ_0, ζ_- or ζ_+ is merged into β_0, β_- or β_+, not to cause any bias for $\hat{\beta}_d$. If we employ LLR for Y but LCR for W_k as in $E(W_k|S) = \xi_0 + \xi_\delta \delta$, then $\xi_\delta = 0$ may be rejected despite $\zeta_\delta = 0$ in the LLR for W_k.

Differently from W, we cannot test for the balance of unobserved ε. However, there is an an indirect way to test for ε balance: test for the continuity of S-density $f_S(s)$ at c. Although $f_S(s)$ continuity tests are widely used, they can be misleading, not least because the continuity of $f_S(s)$ is neither necessary nor sufficient for ε balance. This issue will be examined later.

2.4. EFFECT HETEROGENEITY AND ITS WEIGHTED AVERAGE*

In reality, most treatment effects are heterogeneous, varying as S or W does. Since the treatment effect in RD is specific to the cutoff c, RD treatment effect can be heterogeneous also depending on the value of c. Here, we examine effect heterogeneity and related issues.

Suppose D and S interacts, so that the 'interaction' DS matters with slope $\beta_{ds} > 0$. In the college-entrance (D) and income (Y) example, as test score S reflects ability, $\beta_{ds} > 0$ is that the effect of D on Y is higher for those with a higher S. However, DS cannot be used as a separate regressor in RD, because $DS \simeq Dc$ locally around c, so that c in cD merges into β_d in $\beta_d D$.

Observe

$$E(Y|S) = \beta_d D + \beta_{ds} SD + m(S) = (\beta_d + \beta_{ds} c)D + \beta_{ds}(S-c)D + m(S)$$
$$= \dot{\beta}_d D + \dot{m}(S) \quad \text{where} \quad \dot{\beta}_d \equiv \beta_d + \beta_{ds} c \text{ and } \dot{m}(S) \equiv \beta_{ds}(S-c)\delta + m(S).$$

Here, $\dot{\beta}_d$ consists of the 'direct effect' β_d of D and the 'indirect effect' β_{ds} through $S = c$. Interestingly in $\dot{\beta}_d$, $\beta_{ds} = \partial \dot{\beta}_d / \partial c$ shows the effect of raising c. In the college-entrance and income example, better students are admitted into the college as c is raised, and consequently, the effect on income goes up by β_{ds}. A closely related example is a summer remedial class, where S is the spring-term GPA, $D = 1[S < c]$ and Y is a future GPA: here again, better students attend the summer class to increase Y as c is raised. Although β_d and β_{ds} are not separately identified, if another cutoff c_{new} is available in some data, we can find both $\beta_d + \beta_{ds} c$ and $\beta_d + \beta_{ds} c_{new}$, with which β_d and β_{ds} can be derived.

WD with slope β_{dw} may also matter for Y. With WD controlled, both β_d and β_{dw} are identified. With DW not controlled, however, $E(WD|S)$ becomes part of $m(S)$ that is likely to be discontinuous at c unless W takes the form such as $W = (S-c)M$ for a covariate M.

Similarly to Lee and Lemieux (2010, p. 298), suppose, for a function $\beta_d(\cdot)$,

$$Y = \beta_d(W)D + m(S) + error \text{ with } E\{\beta_d(W)|s\} \text{ and } E(error|s) \text{ continuous at } c$$
$$\Longrightarrow E(Y|c^+) - E(Y|c^-) = E\{\beta_d(W)D|c^+\} - E\{\beta_d(W)D|c^-\} = E\{\beta_d(W)|c\} - 0$$
$$= \int \beta_d(w)\partial F_{W|S}(w|c) = \int \beta_d(w) \frac{f_{S|W}(c|w)}{f_S(c)} \partial F_W(w) \text{ (using the Bayes' rule).} \tag{2.2}$$

This is a weighted average of $\beta_d(w)$, where the weight is $\{f_{S|W}(c|w)/f_S(c)\}$ $\partial F_W(w)$, not $\partial F_W(w)$, that is higher when $f_{S|W}(c|w)$ is higher. If D is a remedial class and W is IQ, then the weight is higher for w with $S = c$ more likely. For instance, $w = 120$ gets a higher weight than $w = 150$ because $f_{S|W}(c|150) < f_{S|W}(c|120)$ as smarter students are less likely to have $S \simeq c$. Hsu and Shen (2019, 2021) proposed tests for RD effect heterogeneity.

In two-party elections, we have $c = 0.5$, but c becomes random in multi-party elections, because which party wins depends on the vote shares of the other parties. In this case, the cutoff becomes a random variable C, but RD can still be done with the normalized score $S_C \equiv S - C$. The identified treatment effect is a weighted average of $E(Y^1 - Y^0|C = c, S = c^+)$ as follows. Suppose $f_{C|S}(c|s)$ is continuous in s for all c. Then, somewhat differently from Cattaneo *et al.*(2016,

p. 1236), the appendix proves

$$\lim_{h\downarrow 0}\{E(Y|S_C=h)-E(Y|S_C=-h)\}=\lim_{h\downarrow 0}\{E(Y|S=C+h)-E(Y|S=C-h)\}$$
$$=\int \lim_{h\downarrow 0} E(Y^1-Y^0|C=c,S=c+h)\cdot f_{C|S}(c|c)\partial c \quad (2.3)$$

under the continuity of $E(Y^0|C=c,S=s)$ in s for all c. *The integrand is 'the c-heterogeneous effect on the just treated', the $f_{C|S}(c|c)$-weighed average of which is the identified effect.* Multi-cutoffs arise also in other contexts, and the treatment effect may be estimated separately at each cutoff. See e.g. Önder and Shamsuddin (2019) and Bertanha (2020) for more on multi-cutoffs.

3. BASICS: FUZZY RD (FRD)

'$D=\delta\equiv 1[0\le S]$' occurs due to laws/regulations. However, often there are exceptions/loopholes in laws/regulations, which result in 'fuzzy RD (FRD)' with $D=D(S,\varepsilon)\neq\delta$; D in FRD is determined by S and a random variable ε. In contrast to FRD, $D=\delta$ that has been examined so far is called 'sharp RD (SRD)'. D may be fully determined by S, yet '$D\neq\delta$' can happen as in $D=\delta S$ to be seen in RK. FRD includes SRD as a special/limiting case.

3.1. SRD VERSUS FRD

FRD occurs often in college admission, which is hardly ever determined solely by a test score. Another FRD example can be seen in effects of schooling (Kan and Lee, 2018): a law stipulates that students with $c\le S$ be subject to more schooling, but not everybody obeys the law, where S is birth date and D ($\neq\delta$) is schooling years. Although there are RD's with non-binary D as in this example (see also Angrist and Lavy (1999), Urquiola (2006) and Urquiola and Verhoogen (2006) for non-binary class size as D, as well as Dong *et al.* (2021) for continuous D in general), we consider mostly binary D to deal with "clean" T and C groups unless otherwise mentioned. We present more SRD and FRD examples next.

Let S be the vote share in a last election, and Y the vote share in the current election. Being incumbent in the current election as a treatment takes the form $D=1[0.5\le Y_{t-1}]$ where $c=0.5$ and $S=Y_{t-1}$, as winning the previous election (i.e., being the incumbent in the current election) means $0.5\le Y_{t-1}$. Lee (2008) showed that incumbent advantage in the U.S. house elections is about 8%. A related example appeared in DiNardo and Lee (2004) for impacts of unionization

D on productivity, wage, etc., where small effects on most outcome variables were found, and near zero effect on wage; $D = 1$ if $S \geq 0.5$, with S being the vote share for unionization.

The above two election cases are SRD's. Two FRD examples for "double hurdle" treatment are, with the α's and γ's being parameters and $(\varepsilon, \varepsilon_0, \varepsilon_1)$ errors,

$$\begin{aligned} D &= 1[0 \leq S] \cdot 1[0 \leq \alpha_1 + \alpha_s S + \varepsilon], \\ D &= 1[S < 0] 1[0 \leq \gamma_1 + \gamma_2 S + \varepsilon_0] + 1[0 \leq S] 1[0 \leq \alpha_1 + \alpha_s S + \varepsilon_1]. \end{aligned}$$

The former appeared in Battistin and Rettore (2008): $1[0 \leq S]$ is the legal eligibility for a public treatment and the individual take-up decision is $1[0 \leq \alpha_1 + \alpha_s S + \varepsilon]$. The latter appeared in Battistin and Rettore (2002): the first part is a privately-taken treatment if the subject is rejected by (or does not take) the public treatment.

3.2. FRD IDENTIFICATION WITH LESSER BREAK

Generalizing the SRD regression form $E(Y|S) = \beta_d D + m(S)$ is the *FRD regression form*:

$$E(Y|S) = \beta_d E(D|S) + m(S). \tag{3.1}$$

As in SRD identification, the contrast between the break of $E(D|S)$ and no break of $m(S)$ in (3.1) is the ID source for β_d in FRD.

As $E(Y|S) = \beta_d D + m(S)$ for SRD is not a restriction because it is equivalent to $\beta_d \equiv E(Y|S = 0^+) - E(Y|S = 0^-)$, (3.1) for FRD is not a restriction either as it is equivalent to (Lee 2016, pp. 102-103) the "*break ratio*"

$$\beta_d \equiv \frac{E(Y|0^+) - E(Y|0^-)}{E(D|0^+) - E(D|0^-)} \quad \text{under} \quad E(D|0^+) \neq E(D|0^-). \tag{3.2}$$

This becomes $E(Y|S = 0^+) - E(Y|S = 0^-)$ for SRD due to $E(D|0^+) - E(D|0^-) = 1$. The fact that (3.1) implies (3.2) can be easily seen by taking $\lim_{s \downarrow 0}$ and $\lim_{s \uparrow 0}$ on (3.1) to obtain

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

and then solving this for β_d while invoking $m(0^+) = m(0^-)$. The break ratio reveals that FRD needs a break of $E(D|S)$ at 0, whose size is necessarily smaller than one.

For SRD, we showed $\beta_d = E(Y^1 - Y^0|0^+)$ under the continuity of $E(Y^0|S)$ at 0. In view of this, the natural question is what kind of treatment effect the break ratio identifies, and under what conditions. For this, define 'potential treatments' (D^0, D^1) corresponding to $\delta = 0, 1$; here δ is taken as the "deep/underlying" treatment, and D as an (intermediate) outcome. Classify individuals as '*never taker*' $(D^0, D^1) = (0,0)$, '*complier*' $(D^0, D^1) = (0,1)$, '*defier*' $(D^0, D^1) = (1,0)$ and '*always taker*' $(D^0, D^1) = (1,1)$, following Imbens and Angrist (1994). In words, compliers are those who would get treated iff $\delta = 1$. Under the '*monotonicity*' $D^0 \leq D^1$ to rule out defiers, Hahn *et al.* (2001) showed that

$$\beta_d \equiv \frac{E(Y|0^+) - E(Y|0^-)}{E(D|0^+) - E(D|0^-)} = E(Y^1 - Y^0|0^+, \text{complier})$$

under some conditions; see Choi and Lee (2018c) for the weakest conditions yet.

Figure 1: Break Ratio as Effect in FRD; for SRD, Left Panel is Step with Height 1

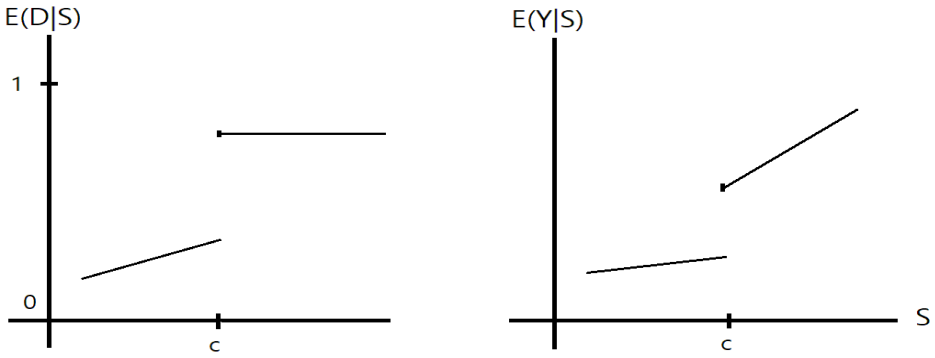


Figure 'Break Ratio as Effect in FRD' shows the FRD ratio ID graphically: the treatment effect is the break ratio at $S = c$ of $E(Y|S)$ and $E(D|S)$. Presenting the treatment effect graphically is a big advantage of RD, compared with other study designs. In (3.1), unless $\beta_d = 0$, the break in $E(D|S)$ causes a break in $E(Y|S)$.

3.3. FRD ESTIMATION WITH IVE

In estimating β_d using the FRD regression form $E(Y|S) = \beta_d E(D|S) + m(S)$, differently from $E(Y|S) = \beta_d D + m(S)$ where we replaced $m(S)$ with a function

of S, it seems that we need to replace $E(D|S)$ as well with a function of S. However, this is not the case as the following shows, where the effect of δ on D is found first within the SRD framework.

With D taken as an outcome of δ, the effect α_δ of δ on D can be found with

$$E(D|S) = \alpha_\delta \delta + m_D(S) \quad (m_D(S) \text{ that is analogous to } m(S) \text{ is continuous at 0}).$$

Substitute this into $E(Y|S) = \beta_d E(D|S) + m(S)$ in (3.1) to obtain

$$E(Y|S) = \gamma_\delta \delta + m_Y(S), \quad \gamma_\delta \equiv \beta_d \alpha_\delta$$
$$(m_Y(S) \equiv m(S) + \beta_d m_D(S) \text{ is continuous at 0}).$$

Specifying $m_D(S)$ and $m_Y(S)$, we can estimate α_δ and γ_δ. Then β_d can be found with $\gamma_\delta/\alpha_\delta$, where γ_δ is the numerator in the break ratio and α_δ is the denominator whose role is to remove α_δ lurking in $\gamma_\delta = \beta_d \alpha_\delta$.

This two-step estimation for β_d looks cumbersome. Fortunately, the two-step estimation with OLS is the same as the single-step instrumental variable estimator (IVE) where D is instrumented by δ; it goes without saying that only the $Q = 1$ observations should be used. The simplest IVE is the 'local constant IVE' for $Y = \beta_0 + \beta_d D + error$ with $(1, \delta)$ as the instrument, which appears when we replace the $m(\cdot)$ functions with constants. As can be seen in Lee (2016, p. 227), the slope of the local constant IVE is, with $\hat{E}(D|0^+)$ and $\hat{E}(D|0^-)$ defined analogously to $\hat{E}(Y|0^+)$ and $\hat{E}(Y|0^-)$,

$$\bar{\beta}_d \equiv \frac{\hat{E}(Y|0^+) - \hat{E}(Y|0^-)}{\hat{E}(D|0^+) - \hat{E}(D|0^-)} \quad \{= \hat{E}(Y|0^+) - \hat{E}(Y|0^-) \text{ in SRD}\}.$$

Better than the local constant IVE in terms of bias is the local linear IVE for

$$Y = \beta_0 + \beta_d D + \beta_-(1-\delta)S + \beta_+ \delta S + V$$

where D is instrumented by δ and V is an error term. As the local linear OLS's for δ on D and δ on Y are nonparametric, this local linear IVE is nonparametric as well.

Specifically, the local linear IVE is (the second element is $\hat{\beta}_d$)

$$\hat{\beta} = (\sum_i Q_i Z_i X_i')^{-1} \sum_i Q_i Z_i Y_i, \quad X \equiv \{1, D, (1-\delta)S, \delta S\}', \quad Z \equiv \{1, \delta, (1-\delta)S, \delta S\}'.$$

The asymptotic inference can be done with the usual IVE variance estimator

$$\hat{\Omega} \equiv (\sum_i Q_i Z_i X_i')^{-1} \cdot \sum_i Q_i Z_i Z_i' (Y_i - X_i'\hat{\beta})^2 \cdot (\sum_i Q_i X_i Z_i')^{-1};$$

$\hat{\beta}_d/\sqrt{\hat{\omega}_d}$ is the t-value for β_d, where $\hat{\omega}_d$ is the second diagonal element of $\hat{\Omega}$. In words, β_d is the intercept shift due to D, while the base intercept β_0 is picked up by 1 in X and the possibly different slopes around 0 are accounted for by the regressors $(1-\delta)S$ and δS. The IVE includes the OLS for SRD as a special case when $Z = X$, i.e., when $\delta = D$.

Consider OLS for D and Y with only the right-side observations of 0 minimizing

$$\sum_i (D_i - \tau_0 - \tau_1 S_i)^2 1[S_i \in (0,h)] \quad \text{and} \quad \sum_i (Y_i - \rho_0 - \rho_1 S_i)^2 1[S_i \in (0,h)]$$

for (τ_0, τ_1) and (ρ_0, ρ_1); let the OLS intercepts be $(\hat{\tau}_0^+, \hat{\rho}_0^+)$. We can do the same using only the left-side observations; let the OLS intercepts be $(\hat{\tau}_0^-, \hat{\rho}_0^-)$. Then we get

$$\hat{\beta}_d = \frac{\hat{\rho}_0^+ - \hat{\rho}_0^-}{\hat{\tau}_0^+ - \hat{\tau}_0^-}.$$

This equality is natural, because doing OLS separately on either side allows different intercepts and slopes as LLR with linear spline does above.

The equivalence between two-stage OLS and IVE means that dealing with SRD is enough in addressing an issue in RD. This is because, if we know how to estimate a treatment effect in SRD using OLS, then we can always estimate the treatment effect in the corresponding FRD using IVE with δ as an instrument for D. The effect found in the IVE is numerically the same as the ratio of the δ's effect on Y (the second OLS) to the δ's effect on D (the first OLS), as long as the same local functional form is used in the two OLS' for SRD and in the IVE for FRD.

For some observed covariates W including 1 and an error term U, suppose that the data are generated by a linear SF model:

$$Y = \beta_d D + W'\beta_w + U.$$

Taking $E(\cdot|S)$ on this equation, (3.1) holds with $m(S) = E(W'|S)\beta_w + E(U|S)$, which shows that it is enough to consider only (S, δ, D, Y) in FRD while ignoring W. Since $m(S)$ can include $E(U|S) \neq 0$ as long as $E(U|S)$ is continuous at 0, RD is robust to the endogeneity of D through S as long as $E(U|S)$ is continuous at 0. Other than through $E(U|S) \neq 0$, D can be endogenous to U also through $COR(\varepsilon, U) \neq 0$ when $D = D(S, \varepsilon)$, but RD always has an "automatic" instrument δ for D to apply IVE.

If D is exogenous to U, we can then apply OLS to $Y = \beta_d D + W'\beta_w + U$. However, this OLS requires the linear SF assumption, differently from the nonparametric IVE based on the equivalence of $E(Y|S) = \beta_d E(D|S) + m(S)$ to the break ratio β_d. Nevertheless, the linear SF assumption is useful in controlling W when $E(W|S)$ is discontinuous at 0. If not controlled, such a discontinuous $E(W|S)$ causes a bias similar to the omitted variable bias in OLS, which can be seen in Choi and Lee (2017) (2017, p. 1222).

In FRD, $E(D|S)$ should have a break at 0, which can be tested with the OLS to

$$E(D|S) = \lambda_0 + \lambda_\delta \delta + \lambda_-(1-\delta)S + \lambda_+ \delta S$$

using the local sample; $\lambda_\delta \neq 0$ indicates a break. If $|\lambda_\delta|$ is small, then even if '$H_0 : \lambda_\delta = 0$' is rejected, the inference for β_d can be tenuous. This is a 'weak ID' problem plaguing IVE, which is addressed by Feir *et al.* (2016). It is thus better to test for D exogeneity first as follows, and if no rejection, apply OLS instead of IVE.

One test for D exogeneity is based on 'control function approach': using only the local sample, obtain the OLS residual $\hat{\varepsilon}$ of D on Z to do the OLS of Y on $(X, \hat{\varepsilon})$: insignificance of $\hat{\varepsilon}$ means D exogeneity. Closely related to this is the OLS-IVE equality 'Hausman' test. Bertanha and Imbens (2020) also considered the Hausman test, but they recommended a related test based on an idea in Angrist (2004).

4. SCORE TOPICS: MULTIPLE-SCORE RD (MRD)

So far, we addressed single-score RD, but there are many RD's with multiple scores. For instance, to graduate from high school, a student may have to pass multiple exams. To be eligible for pension, one may have to be at least 60 years old with the pension contribution years of at least 10. Yet another example is spatial/geographical RD where longitude and latitude appear as two scores (Keele and Titiunik, 2015).

In examining *multiple-score RD (MRD)*, as there are SRD and FRD for single-score RD—we use 'SRD' and 'FRD' only for single-score RD in this section—there are also '*sharp multiple-score RD (SMRD)*' as in Lalive (2008) and Schmieder *et al.* (2012), and '*fuzzy multiple-score RD (FMRD)*' as in Jacob and Lefgren (2004) and Matsudaira (2008).

For simplicity, we examine only *two scores* $S \equiv (S_1, S_2)'$ with $c \equiv (c_1, c_2)'$. Extensions to more-than-two scores is straightforward—at least conceptually—although the details could be "messy". *Redefine* $S - c$ *as* S *so that the cutoff*

becomes $(0,0)'$. Let

$$\delta_j \equiv 1[0 \leq S_j] \quad \text{for} \quad j = 1,2.$$

Differently from single-score RD, SMRD has "AND cases" and "OR cases" as in

$$D = 1[0 \leq S_1,\ 0 \leq S_2] \quad \text{and} \quad D = 1[0 \leq S_1 \text{ or } 0 \leq S_2].$$

To simplify exposition, we consider only AND cases, because OR cases can be "flipped" to the AND cases; i.e., switch the labels of the T and C groups. Specifically, with $S_1' \equiv -S_1$ and $S_2' \equiv -S_2$, define for the OR case D:

$$D' \equiv 1 - 1[0 \leq S_1 \text{ or } 0 \leq S_2] = 1[S_1 < 0, S_2 < 0] = 1[0 < S_1', 0 < S_2']$$

so that $Y^{D=1} - Y^{D=0} = Y^{D'=0} - Y^{D'=1}$. We can find the effect with D' as the treatment and $S' \equiv (S_1', S_2')'$ as the score, and then multiply the effect by -1 to obtain the desired effect.

There are at least two difficulties with MRD. First, we have to deal with two-, not one-, dimensional continuity in (S_1, S_2). Second, the "partial effect" of δ_1 or δ_2 may be present due to each score crossing its own cutoff, in addition to the effect due to D. 'MRD for a treatment' differs from one-score RD with multiple cutoffs as in Van der Klaauw (2002) and Angrist and Lavy (1999). MRD for a treatment also differs from 'MRD for multiple treatments' as in Leuven *et al.* (2007) and Papay *et al.* (2011) where each score dictates one treatment.

4.1. IDENTIFICATION FOR SHARP MULTIPLE-SCORE RD (SMRD)

Consider four potential responses $(Y^{00}, Y^{10}, Y^{01}, Y^{11})$ corresponding to $\delta_1, \delta_2 = 0, 1$. In AND-case two-score SMRD addressed by Choi and Lee (2018b), the treatment is $D = \delta_1 \delta_2$; e.g., a student graduates high school by passing two exams (and Y is lifetime income). Other than through $D = \delta_1 \delta_2$, δ_1 and δ_2 may separately affect Y. For instance, to graduate high school, one has to pass both math (δ_1) and English (δ_2) exams, but failing the math test may stigmatize the student ("I cannot do math") to affect Y. The separate effects of δ_1 and δ_2 are '*partial effects*'.

Before we get into the details on SMRD, we summarize the main results and compare them to SRD in Table 'SRD (1 Score) versus SMRD (2 Scores)'. The first line presents the form of difference used: single difference for SRD and double difference or 'difference in differences' (DD) for SMRD. The second line is for the identified treatment effects. The third line shows for which response the continuity at 0 is required; in SMRD, there appear three (partly) untreated

Table 1: SRD (1 Score) versus SMRD (2 Scores)

SRD (1 Score) versus SMRD (2 Scores)	
$E(Y\|0^+)-E(Y\|0^-)$	$E(Y\|0^+,0^+)-E(Y\|0^+,0^-)-E(Y\|0^-,0^+)+E(Y\|0^-,0^-)$
$E(Y^1-Y^0\|0^+)$	$E(Y^{\pm}\|0^+,0^+)$ with $Y^{\pm}\equiv Y^{11}-Y^{10}-Y^{01}+Y^{00}$
$E(Y^0\|S)$ continuous at 0	$E(Y^{00}\|S)$, $E(Y^{10}\|S)$, $E(Y^{01}\|S)$ continuous at 0
$\beta_d\delta+m(S)$	$\beta_1\delta_1+\beta_2\delta_2+\beta_d D+m(S)$ with $D=\delta_1\delta_2$

responses Y^{00},Y^{10},Y^{01}. The fourth line presents the regression form equivalent to the difference form in the first line. The main point is that we need DD, not a single difference, for two scores.

Turning to the details, the treatment effect of interest is

$$E(Y^{\pm}|0^+,0^+) \quad \text{where} \quad Y^{\pm}=Y^{11}-Y^{00}-(Y^{10}-Y^{00})-(Y^{01}-Y^{00}).$$

$Y^{11}-Y^{00}$ is the 'gross effect' of D, and $Y^{10}-Y^{00}$ and $Y^{01}-Y^{00}$ are the partial effects of δ_1 and δ_2. The *desired 'net effect' of D is obtained by subtracting the partial effects of δ_1 and δ_2 from the gross effect of D.* Those familiar with DD would recognize $Y^{\pm}\equiv(Y^{11}-Y^{10})-(Y^{01}-Y^{00})$ as a DD, which is known to isolate the interaction effect of two factors by removing their partial effects. See Lee and Sawada (2020), Lee (2021a) and references therein for DD in general.

Choi and Lee (2018b) showed that the first two lines in the table are the same:

$$\begin{aligned}
&E(Y^{\pm}|0^+,0^+)=\beta_d\equiv DD\{E(Y|S)\} \quad \text{where}\\
&DD(\cdot)\equiv E(\cdot|0^+,0^+)-E(\cdot|0^+,0^-)-E(\cdot|0^-,0^+)+E(\cdot|0^-,0^-),\\
&E(\cdot|0^+,0^+)\equiv\lim_{s_1\downarrow 0,s_2\downarrow 0}E(\cdot|s_1,s_2),\quad E(\cdot|0^+,0^-)\equiv\lim_{s_1\downarrow 0,s_2\uparrow 0}E(\cdot|s_1,s_2)\quad \text{and so on,}
\end{aligned}$$

under the continuity of $E(Y^{00}|S)$, $E(Y^{10}|S)$ and $E(Y^{01}|S)$ at 0. Choi and Lee (2021) also proved that, for some β_1 and β_2,

$$\beta_d\equiv DD\{E(Y|S)\}\Longleftrightarrow E(Y|S)=\beta_1\delta_1+\beta_2\delta_2+\beta_d\delta_1\delta_2+m_Y(S),\ \ DD\{m_Y(S)\}=0$$

where *the former is 'the difference form' for SMRD, and the latter is 'the regression form'.*

4.2. ESTIMATION FOR SMRD

Choi and Lee (2018b) estimate β_d by replacing $m_Y(S)$ with a function continuous at 0; note that if $m_Y(S)$ is not continuous at 0 while satisfying $DD\{m_Y(S)\} = 0$, then β_1 and β_2 are not identified. The simplest estimation approach is the LCR with $m_Y(S) = \beta_0$ to have

$$E(Y|S) = \beta_0 + \beta_1\delta_1 + \beta_2\delta_2 + \beta_d\delta_1\delta_2 :$$

estimate this by the OLS of Y on $(1, \delta_1, \delta_2, \delta_1\delta_2)$ with the local sample $|S_j| < h_j$ for the bandwidths h_j, $j = 1,2$. The LLR version replaces $m_Y(S)$ with $\beta_0 + \beta_{s1}S_1 + \beta_{s2}S_2$, or more generally,

$$\begin{aligned}\breve{m}_Y(S) \equiv \beta_0 &+ \beta_{11}\delta_1^-\delta_2^-S_1 + \beta_{12}\delta_1^-\delta_2^-S_2 + \beta_{21}\delta_1^-\delta_2^+S_1 + \beta_{22}\delta_1^-\delta_2^+S_2 \\ &+ \beta_{31}\delta_1^+\delta_2^-S_1 + \beta_{32}\delta_1^+\delta_2^-S_2 + \beta_{41}\delta_1^+\delta_2^+S_1 + \beta_{42}\delta_1^+\delta_2^+S_2, \\ &\delta_j^- \equiv 1[-h_j < S_j < 0], \;\; \delta_j^+ \equiv 1[0 \le S_j < h_j].\end{aligned}$$

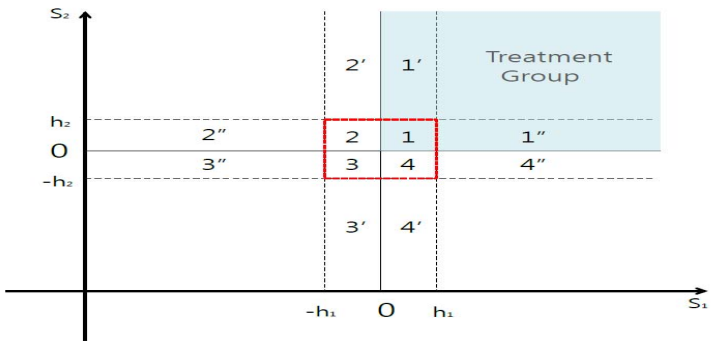
$\breve{m}_Y(S)$ is a piecewise-linear function continuous at 0, because $\breve{m}_Y(0) = 0$ and the limit of $E\{\breve{m}_Y(S)|S\}$ at 0 is β_0 regardless of from which direction 0 is approached, but $\breve{m}_Y(S)$ allows different slopes across the four quadrants determined by (δ_1^+, δ_2^+). In practice, $\beta_0 + \beta_{s1}S_1 + \beta_{s2}S_2$ may be preferred to $\breve{m}_Y(S)$, because there are too many terms in $\breve{m}_Y(S)$.

Partial effects are not new, as we almost always allow regressors δ_1 and δ_2 when $\delta_1\delta_2$ is used as a regressor. Nevertheless, most estimators in the MRD literature work only when the partial effects are zero. To understand those approaches, examine Figure 'AND-Case SMRD, where the continuity of $m(S)$ at 0 is $\lim_{s\to 0} m(s) = m(0)$ regardless of in which direction s approaches 0.

In the figure, the upper right colored area is treated, and the white area is the control. Two-dimensional localization around 0 gives T group (square 1) and C group (squares 2,3,4), which were used in our preceding estimation approach. This is inefficient, however, as there are boundary lines extending vertically or horizontally from $(0,0)$. Using one-dimensional localization with $S_1|0 \le S_2$ and $S_2|0 \le S_1$ can provide a more efficient estimator. Indeed, along this line, Choi and Lee (2018a) suggested a 'minimum distance estimator (MDE)', which is far more efficient than the preceding two-dimensional localization, although MDE is cumbersome to do.

Turning to estimators that are inconsistent when partial effects are present, suppose we try one-dimensional localization with S_1 for those with $\delta_2 = 1 \Longleftrightarrow 0 \le S_2$ to result in the T group (squares 1 and 1') and the C group (squares 2 and

Figure 2: AND-Case SMRD



2'), where a boundary line (not a boundary point) extends from $(0,0)$ upward. Analogously, we can also do localization with $S_2|\delta_1 = 1$. An empirical example below, however, will show that these approaches fail under partial effects.

Instead of conditioning on $\delta_2 = 1$ or $\delta_1 = 1$, Battistin *et al.* (2009) used $S_{\min} \equiv \min(S_1, S_2)$ based on $0 \leq S_{\min} \Longleftrightarrow 0 \leq S_1, S_2$. This approach, however, gets to approximate $m_Y(\cdot)$ with $m_Y(S_{\min}) = m_Y\{\min(S_1, S_2)\}$, not with the more flexible $m_Y(S) = m_Y(S_1, S_2)$. More importantly, denoting the bandwidth for $S_{\min}$ as $h_{\min}$, localization with $-h_{\min} < \min(S_1, S_2) < h_{\min}$ is insufficient because $-h_{\min} < S_1, S_2$ holds, but not necessarily $S_1, S_2 < h_{\min}$. Clark and Martorell (2014) used $S_{\min} = \min(S_1, S_2, S_3)$ for math, reading and writing scores. Although problematic, this min-based approach is convenient when there are many scores, because of $0 \leq S_{\min} \Longleftrightarrow 0 \leq S_1, S_2, S_3, ...$ Other than the min-based approach, 'boundary estimation' approaches estimate the treatment effects along the boundary lines to average them (Wong *et al.*, 2013; Keele and Titiunik, 2015). These approaches also fail when there are partial effects.

4.3. FUZZY MULTIPLE-SCORE RD (FMRD)*

For FMRD, define 4 potential treatments $(D^{00}, D^{10}, D^{01}, D^{11})$ corresponding to $\delta_1, \delta_2 = 0, 1$. We summarize the main results for FMRD and compare them to FRD in Table 'FRD (1 Score) versus FMRD (2 Scores)'. The first line presents the form of difference ratios used: single difference ratio for FRD, and DD ratio for FMRD under $DD\{E(D|S)\} \neq 0$ which reduces to $DD\{E(Y|S)\}$ when $DD\{E(D|S)\} = 1$. The second line is for the identified treatment effects. The

Table 2: FRD (1 Score) versus FMRD (2 Scores)

FRD (1 Score) versus FMRD (2 Scores)	
$\{E(Y\|0^+)-E(Y\|0^-)\}/\{E(D\|0^+)-E(D\|0^-)\}$	$DD\{E(Y\|S)\}/DD\{E(D\|S)\}$
$E(Y^1-Y^0\|0^+,\text{complier})$	$E(Y^1-Y^0\|0^+,0^+,\text{complier})$
$D^0 \le D^1$	$D^{ab} \le D^{cd}$; $D^{10}+D^{01} \le D^{11}+D^{00}$
$\beta_d D+m(S)$	$\beta_1\delta_1+\beta_2\delta_2+\beta_d D+m(S)$

third line shows the requisite monotonicity conditions, where $a \le b$ and $c \le d$. Various continuity conditions are needed, which are omitted in the table though. The fourth line presents the regression form equivalent to the difference form in the first line.

Turning to the details now, although the same word 'complier' appears for both FRD and FMRD, their definitions differ. Choi and Lee (2021) defined 'compliers' in FMRD as those with

$$D_2^{\pm} \equiv D^{11}-D^{10}-D^{01}+D^{00}=1 \quad \text{under the monotonicity condition in the table}$$

which needs some explanations as follows. The first inequality '$D^{ab} \le D^{cd}$' in the monotonicity is the natural extension of $D^0 \le D^1$ in FRD. The second inequality is to ensure that $D_2^{\pm}$ takes only 0 or 1, because of $D_2^{\pm}=0,1 \Longleftrightarrow 0 \le D_2^{\pm} \le 1$ that is

$$D^{10}+D^{01} \le D^{11}+D^{00} \le D^{10}+D^{01}+1.$$

The second inequality easily holds, but the first inequality that appeared in the table is restrictive. '$D^{10}+D^{01} \le D^{11}+D^{00}$' is to rule out $(D^{00},D^{10},D^{01},D^{11})=(0,1,1,1)$, because this would be an OR case (i.e., the treatment is taken if either $\delta_1=1$ or $\delta_2=1$), whereas our FMRD is a fuzzy version of AND-case SMRD.

Choi and Lee (2021) proved that

$$(i): DD\{E(D|S)\} = P(\text{complier}|0^+,0^+),$$
$$(ii): \frac{DD\{E(Y|S)\}}{DD\{E(D|S)\}} = E(Y^1-Y^0|0^+,0^+,\text{complier});$$

(i) characterizes the denominator in the DD ratio as the complier probability, which was also the case in the FRD ratio, and (ii) gives a causal meaning to the DD ratio.

The regression form for SMRD with D as the outcome and $\delta_1\delta_2$ as the treatment is, for some α parameters and a function $m_D(S)$,

$$E(D|S)=\alpha_1\delta_1+\alpha_2\delta_2+\alpha_\delta\delta_1\delta_2+m_D(S) \quad \text{with} \quad DD\{m_D(S)\}=0.$$

The DD ratio form $\beta_d = DD\{E(Y|S)\}/DD\{E(D|S)\}$ *for FMRD is equivalent to the regression form* (Choi and Lee, 2021):

$$E(Y|S) = \beta_d E(D|S) + m(S) \quad \text{with} \quad DD\{m_Y(S)\} = 0.$$

Substitute the $E(D|S)$ regression form into this $E(Y|S)$ regression form to obtain

$$E(Y|S) = \gamma_1\delta_1 + \gamma_2\delta_2 + \gamma_\delta\delta_1\delta_2 + m_Y(S) \quad \text{where}$$
$$\gamma_1 \equiv \beta_d\alpha_1, \quad \gamma_2 \equiv \beta_d\alpha_2, \quad \gamma_\delta \equiv \beta_d\alpha_\delta, \quad m_Y(S) \equiv m(S) + \beta_d m_D(S).$$

With the γ's found using this equation and the α's found using the $E(D|S)$ equation, we have

$$\beta_d = \frac{\gamma_1}{\alpha_1} = \frac{\gamma_2}{\alpha_2} = \frac{\gamma_\delta}{\alpha_\delta}.$$

Since α_1 and α_2 may be zero, we use the last ratio $\gamma_\delta/\alpha_\delta$ for β_d, not the first two.

As for estimation, the two-step local OLS estimating the α's and γ's with the observations $Q \equiv 1[|S_j| < h_j, j = 1,2] = 1$ to find $\beta_d = \gamma_\delta/\alpha_\delta$ is the same as the single-step local IVE to

$$Y = \beta_1\delta_1 + \beta_2\delta_2 + \beta_d D + m(S) + error \quad \text{with } \delta_1\delta_2 \text{ as an instrument for } D$$

when $m_D(S)$ and $m_Y(S)$ take the same form as $m(S)$ does; note a slight abuse of notation, because $m(S)$ here is not the same as $m(S)$ in $E(Y|S) = \beta_d E(D|S) + m(S)$. We can specify $m(S)$ as a constant, or as a (piecewise) linear function that appeared for SMRD. Essentially, $\alpha_\delta \neq 0$ is the 'inclusion restriction' for the the instrument $\delta_1\delta_2$, and the 'exclusion restriction' is $DD\{m_Y(S)\} = 0$ because $DD\{m_Y(S)\} = 0$ does not hold if $\delta_1\delta_2$ lurks in $m_Y(S)$.

4.4. EXAMPLE: SUMMER SCHOOL EFFECT ON TEST SCORE

Matsudaira (2008) examined the effect of remedial summer school on test scores, where a student has to attend summer school if he/she fails either math or reading test; all scores are standardized, and Y is a next year test score. This is an OR-case FMRD with $\delta \equiv 1[S_1 < 0 \text{ or } S_2 < 0]$. Matsudaira (2008) adopted one-dimensional localization approach, and used a third-order polynomial for $m(S)$.

Conditioning on passing the reading test ($0 \leq S_2$), Matsudaira (2008) applied FRD with math score S_1, and part of his Table 2 is in Table 'Effect of Summer School on Math Score Given Reading' for grades 3, 5 and 6. The first column

Table 3: Effect of Summer School on Math Score Given Reading: SE in (·)

Effect of Summer School on Math Score Given Reading: SE in (·)			
($\delta \equiv 1[S_1<0$ or $S_2<0]$)	$\alpha_\delta : \delta$ on D	$\gamma_\delta : \delta$ on Y	$\beta_d : D$ on Y
Grade 3: $N = 55931$	0.383 (0.016)	0.049 (0.020)	0.128 (0.055)
Grade 5: $N = 59258$	0.385 (0.006)	0.093 (0.015)	0.241 (0.039)
Grade 6: $N = 51810$	0.320 (0.011)	0.061 (0.014)	0.190 (0.047)

presents the D equation estimates, where α_δ is highly significant for the inclusion restriction. The second column is the effect γ_δ of δ on Y. The third column is obtained dividing the second column by the first; e.g., $0.128 = 0.049/0.383$ for Grade 3. The summer school effect on math score is $0.19 \sim 0.24$ of one SD.

Unfortunately, the one-score conditional approach gives biased estimates. To see the bias, let $\delta_j = 1$ *if failing test* j, and consider, for errors ε and U,

$$D = \alpha_0 + \alpha_1\delta_1 + \alpha_2\delta_2 + \alpha_\delta\delta_1\delta_2 + \varepsilon \quad \text{and} \quad Y = \beta_0 + \beta_1\delta_1 + \beta_2\delta_2 + \beta_d D + U;$$

the D model is "saturated" with four parameters and four cells due to $\delta_1, \delta_2 = 0, 1$. Rewrite the D equation to have $\delta \equiv 1[S_1 < 0 \text{ or } S_2 < 0] = \delta_1 + \delta_2 - \delta_1\delta_2$ explicit: due to $\delta_1\delta_2 = \delta_1 + \delta_2 - \delta$,

$$\begin{aligned} D &= \alpha_0 + \alpha_1\delta_1 + \alpha_2\delta_2 + \alpha_\delta(\delta_1 + \delta_2 - \delta) + \varepsilon \\ &= \alpha_0 + (\alpha_1 + \alpha_\delta)\delta_1 + (\alpha_2 + \alpha_\delta)\delta_2 - \alpha_\delta\delta + \varepsilon. \end{aligned}$$

Substitute this D equation into the Y equation to get

$$\begin{aligned} Y &= \beta_0 + \beta_1\delta_1 + \beta_2\delta_2 + \beta_d\{\alpha_0 + (\alpha_1 + \alpha_\delta)\delta_1 + (\alpha_2 + \alpha_\delta)\delta_2 - \alpha_\delta\delta + \varepsilon\} + U \\ &= \beta_0 + \beta_d\alpha_0 + \{\beta_1 + \beta_d(\alpha_1 + \alpha_\delta)\}\delta_1 + \{\beta_2 + \beta_d(\alpha_2 + \alpha_\delta)\}\delta_2 - \beta_d\alpha_\delta\delta + \beta_d\varepsilon + U. \end{aligned}$$

Given passing the reading exam ($\delta_2 = 0$), due to $\delta = \delta_1$ given $\delta_2 = 0$ from $\delta = \delta_1 + \delta_2 - \delta_1\delta_2$,

$$\begin{aligned} D &= \alpha_0 + (\alpha_1 + \alpha_\delta)\delta_1 - \alpha_\delta\delta_1 + \varepsilon = \alpha_0 + \alpha_1\delta_1 + \varepsilon, \\ Y &= \beta_0 + \beta_d\alpha_0 + \{\beta_1 + \beta_d(\alpha_1 + \alpha_\delta)\}\delta_1 - \beta_d\alpha_\delta\delta_1 + (\beta_d\varepsilon + U) \\ &= \beta_0 + \beta_d\alpha_0 + (\beta_1 + \beta_d\alpha_1)\delta_1 + (\beta_d\varepsilon + U). \end{aligned}$$

The ratio 'slope of δ_1 in Y equation over slope of δ_1 in D equation' is

$$\frac{\beta_1+\beta_d\alpha_1}{\alpha_1} = \frac{\beta_1}{\alpha_1}+\beta_d \neq \beta_d \quad \text{unless } \beta_1 = 0$$

$$\Longleftrightarrow \text{ no partial effect of failing math test.}$$

Since the next year math score would depend positively on the current year math score, β_1 is likely to be negative. Since α_1 should be positive, the effect estimates in the above table under-estimates the true effect by $|\beta_1/\alpha_1|$.

5. SCORE TOPICS: ERROR-RIDDEN SCORE AND INTEGER SCORE

So far, we assumed that a continuous S is available to determine D, fully or partly. Sometimes, however, S is not the true score, but an error-ridden version of the true score G. Also, S may be discrete, because D is determined by a discrete score crossing a cutoff, or because a discrete transformation of a true continuous score is observed. These issues are examined here.

5.1. CONTINUOUS SCORE MEASURED WITH ERROR

Despite that a break of $E(D|G)$ at 0 is expected, sometimes no or a lesser break is found. In this case, the most likely reason is a measurement error: an error-ridden score $S = G + error$ is observed, instead of the *genuine score* G. Note that, if S, not G, determines D, then there is no problem in doing RD with S despite the error in S—a point misunderstood for a while in the RD literature. That is, we tackle $D = 1[0 \leq G]$ with S observed, not $D = 1[0 \leq S]$ with S observed–the latter poses no problem. Under 'errors-in-variable' $S = G + error$, SRD becomes a FRD as the treatment gets "fuzzy", and thus we examine only FRD here.

With only S observed, what is available is

$$\text{"}available\ ratio\text{"}: \frac{E(Y|S=0^+)-E(Y|S=0^-)}{E(D|S=0^+)-E(D|S=0^-)}.$$

One question is whether $E(D|S)$ has a break at 0 when $E(D|G)$ does. Another is whether the available ratio equals the desired ratio with G in the conditioning events.

Specifically, consider the '*full errors-in-variable*' case:

$$S = G+V \quad \text{where } G \perp\!\!\!\perp V \text{ and the error } V \text{ has density } f_V(v) \text{ continuous at 0.} \tag{5.1}$$

Lee (2017) showed that RD fails in this case, because $E(D|S)$ has no break at 0. Also, often the continuity of score density f_S at 0 is tested to detect unobserved confounders, but Lee (2017, p. 3) showed that score density continuity test does not work, because f_S is continuous at 0 even when f_G is not as V smooths things out.

As an example, in Germany, it is possible to opt out of the public insurance to buy a private insurance instead, if the income exceeds a cutoff. Using this RD feature, Hullegie and Klein (2010) examined effects of private insurance. With S being a reported income and D private insurance, however, Hullegie and Klein found no break in $E(D|S)$; S must be an error-ridden version of the true income G. Hullegie and Klein thus used a 'selection-correction' approach under normality on V and $V \perp\!\!\!\perp S$. They found negative effects on the number of doctor visits, no effect on the number of hospital nights, and positive effects on health. The assumption '$V \perp\!\!\!\perp S$' is implausible though, when S is generated by adding V to G as in (5.1).

Davezies and Le Barbanchon (2017) addressed the full errors-in-variable setting under $P(D=1|G)>0$ for all G. They presumed availability of an auxiliary sample on treated subjects where both G and S are observed, and proposed a "sieve inverse-weighting GMM".

Now consider a '*part errors-in-variable*': for an unobserved binary B ('B' for binary),

$$S = BG + (1-B)H \quad \text{for a continuous "hazy score" } H \quad \text{and} \quad P(B=1|S=0)>0.$$

The true score G is observed when $B=1$, but an error-ridden score H (e.g., $H = G+V$) is observed when $B=0$. If the RD break is smaller than expected, then part errors-in-variable is a highly likely reason. Call those with $G=S \Longleftrightarrow B=1$ "*truth-tellers*". '$P(B=1|S=0)>0$' is that the proportion of the truth-tellers is not 0 when S equals the cutoff.

For the part errors-in-variable case, Lee (2017, p. 4) showed that the available ratio becomes

$$\frac{E(Y|G=S=0^+)-E(Y|G=S=0^-)}{E(D|G=S=0^+)-E(D|G=S=0^-)} = \frac{E(Y|G=0^+,B=1)-E(Y|G=0^-,B=1)}{E(D|G=0^+,B=1)-E(D|G=0^-,B=1)}:$$

the "effect on the truthful margin $G=S=0^+ \Longleftrightarrow G=0^+, B=1$" is identified by the available ratio. As compliers are not observed in FRD, "truth-tellers"

($G = S$) are not observed either, but at least we know for which group the effect is. The truth tellers are not compliers, because they are not necessarily treated iff $\delta = 1[0 \leq G] = 1$.

If we assume the 'no-selection problem assumption' $(D,Y) \perp\!\!\!\perp S|G$ as in Battistin *et al.* (2009), then we can remove S from the last display to obtain the usual RD ratio. The assumption is, however, unnecessarily strong, because the effect on the truthful margin is still meaningful without the assumption.

In the right side expression of the ratio in the last display, we have

$$E(Y|G = 0^-, B = 1) = E(Y^0|G = 0^-, B = 1)$$
$$E(D|G = 0^-, B = 1) = E(D^0|G = 0^-, B = 1)$$

and '$G = 0^-$' here can be replaced with '$G = 0^+$' under the continuity assumption for $E(Y^0|G = g, B = 1)$ and $E(D^0|G = g, B = 1)$ at $g = 0$. This would then makes the ratio equal to

$$\frac{E(Y^1 - Y^0|G = 0^+, B = 1)}{E(D^1 - D^0|G = 0^+, B = 1)} = E(Y^1 - Y^0|G = 0^+, B = 1, \text{complier}) :$$

the *"effect on the just-treated ($G = 0^+$), truth-telling ($B = 1$) compliers" is identified*.

With $J(g,s) \equiv E(D|G = g, S = s)$, the *break magnitude of $E(D|S)$ at $S = 0$ is proportional to the truth-teller proportion at $S = 0$* (Lee 2017, p. 4):

$$E(D|S = 0^+) - E(D|S = 0^-) = P(B = 1|S = 0) \cdot \{J(0^+, 0^+) - J(0^-, 0^-)\}.$$

For SRD, $J(g,s) = 1[0 \leq g] \Longrightarrow J(0^+, 0^+) - J(0^-, 0^-) = 1$: the break size of $E(D|S)$ at $S = 0$ is $P(B = 1|S = 0)$. See Schanzenbach (2009) for an empirical example.

5.2. INTEGER SCORE: GENUINE VERSUS NON-GENUINE INTEGERS

Although scores in RD are supposed to be continuous, often we face integer scores. Some integer scores are inherently integers, as in the number of students (Angrist and Lavy, 1999) or events in a given time interval ("counts"), which we call "genuine integers". In contrast, some are integer-transformed versions of continuous scores, which we call "non-genuine integers". With $\lfloor G \rfloor$ denoting 'the integer part of G not greater than G', let

$$S \equiv \lfloor G \rfloor.$$

Sometimes, $\lfloor G \rfloor$ is used to denote the rounded-down version of G as in $\lfloor 1.7 \rfloor = 1$, but "round-down" can be confusing for negative G (as in $\lfloor -1.7 \rfloor = -1$ or -2?). Hence, we use the expression 'integer part of G not greater than G' (so that $\lfloor -1.7 \rfloor = -2$, not -1). There may be non-integer discrete scores, but scores should be at least cardinal in RD, so that they are convertible to integers with an appropriate rescaling.

As long as the observations on the integer support points in a chosen local neighborhood around c are balanced in all covariates, observed or unobserved, the integer nature of S does not matter. A problem does occur, however, when there are not enough observations near c, and thus observations further away from c should be used than one feels comfortable with for local randomization. For example, birth date G may be observed only in years S for confidentiality, and persons years apart are compared. This can make RD biased, as the cohort and other unobserved differences may creep in. Birth date G may be observed also in quarters/months, but it can be converted to an integer with the appropriate rescaling as was just noted. Comparing persons a few quarters/months apart would not be a problem in RD.

Instead of the usual $S - c$ that may not be an integer, it is better to location normalize with

$$S - \lfloor c \rfloor = \lfloor G \rfloor - \lfloor c \rfloor.$$

so that $S - \lfloor c \rfloor$ is an integer; call those with $S = \lfloor c \rfloor$ the "*cutoff sample*". It helps to consider cases of integer c and non-integer c separately as follows, since $c = \lfloor c \rfloor$ only for integer c.

Suppose S is birth year, and $c = 1985.67$ for Sep. 1st, 1985, which is not an integer. Then

$$c = 1985.67, \quad \lfloor c \rfloor = 1985, \quad S - \lfloor c \rfloor = S - 1985, \quad c - \lfloor c \rfloor = 0.67;$$

0.67 shows how far off the actual cutoff is from the cutoff integer $\lfloor c \rfloor = 1985$. Here, we cannot tell whether an individual in the cutoff sample was born before/after c, which makes the cutoff sample non-informative for the treatment $D = 1[c \leq G]$. If c is an integer as in $c = 1985$, then

$$c = 1985, \quad \lfloor c \rfloor = 1985, \quad S - \lfloor c \rfloor = S - 1985, \quad c - \lfloor c \rfloor = 0.$$

S is enough to tell whether the person was born before/after c (i.e., whether $D = 0$ or 1). If the cutoff sample is dropped, then there is no real difference between integer and non-integer c.

5.3. ESTIMATION WITH GENUINE INTEGER SCORE

For genuine integer score $S = G$ and $D = 1[0 \leq G]$ with the normalization $G - c$, an interpolation of $E(Y^0|S = -1)$ toward $S = 0$ is needed to find the effect at 0:

$$\begin{aligned} & E(Y|S=0) - \text{`interpolated version of } E(Y|S=-1) \text{ toward } S=0\text{'} \\ = \;& E(Y^1|S=0) - E(Y^0|S=0) = E(Y^1 - Y^0|S=0). \end{aligned}$$

For instance, we may find a linear line $\alpha_0 + \alpha_1 S$ going through $E(Y|S=-1)$ and $E(Y|S=-2)$ to use $\alpha_0 = \alpha_0 + \alpha_1 \times 0$ as $E(Y^0|S=0)$. More generally, we may find a quadratic line $\alpha_0 + \alpha_1 S + \alpha_2 S^2$ going through three points $E(Y|S=s)$, $s = -1, -2, -3$ and use α_0 as $E(Y^0|S=0)$.

Analogously to $E(Y|G) = \beta_d \delta + m(G)$, Lee and Card (2008; "LC") considered $E(Y|S) = \beta_d \delta + m(S)$ where S takes on J integer values. Then LC replaced $m(S)$ with a parametric function $m(S;\beta_m)$ with $C-1$ parameters. Recall that we also use parametric functions for $m(\cdot)$ with a continuous score, but the difference is that the parametric functional form needs to hold only locally around 0 for the continuous score, whereas it should hold broadly across the J integers for S.

Let $\hat{U}_R \equiv Y - \hat{\beta}_d \delta - m(S;\hat{\beta}_m)$ be the parametric residual, and $\tilde{U}_{UR}$ the (nonparametric) residual with the full set of dummies for J integers. Under H_0 : $m(S) = m(S;\beta_m)$, LC proposed a specification test with

$$(\sum_i \hat{U}_{i,R}^2 - \sum_i \tilde{U}_{i,UR}^2)/\{\sum_i \tilde{U}_{i,UR}^2/(N-J)\} \rightsquigarrow \chi_{J-C}^2$$

where $J - C$ is the difference between the numbers of parameters in the full dummy model and the parametric model.

LC further suggested to use a clustered variance estimator because the observations with the same $S = s$ are clustered, which has been widely used. However, Kolesár and Rothee (2018) showed that the resulting confidence intervals (CI's) do not make sense and should not be used.

Kolesár and Rothee (2018) recommended using either the usual RD inference if the number of the support points is not so small, or one of the two different CI constructions in their paper. The two CI's require a bound on the second derivative $|m''(S)|$, or the assumption of the smallest specification error at the cutoff given h. Since these bring in arbitrariness as h does, sticking to the usual RD inference seems better.

5.4. NO BIAS CONDITION DESPITE NON-GENUINE INTEGER SCORE

For SRD, consider a linear spline model for Y:

$$E(Y|G) = \beta_0 + \beta_d\delta + \beta_1(G-c) + \beta_{1\delta}\delta(G-c), \quad \delta \equiv 1[c \le G]. \tag{5.2}$$

When S is observed, not G, the question is whether we can replace $G-c$ in (5.2) with $S-c$ (or $S-\lfloor c \rfloor$) to apply OLS to the resulting model. The answer is "no" in general due to the next two reasons, but "yes" under some condition.

First, depending on whether c is an integer or not, $1[c \le G] \ne 1[c \le S]$ can happen:

$$\begin{aligned} \text{integer } c \quad &: \quad 1[c \le G] = 1[c \le S], \\ \text{non-integer } c \quad &: \quad 1[c \le G] = 1[c \le S] \text{ if } S \ne \lfloor c \rfloor \quad \text{and} \quad \text{unclear if } S = \lfloor c \rfloor. \end{aligned}$$

For instance, when $c = 1985$, $1[1985 \le G] = 1[1985 \le S]$ holds, so that we have $\delta \equiv 1[c \le G] = 1[c \le S]$. When $c = 1985.67$, $S = 1984$ implies that the subject is untreated, $S = 1986$ implies that the subject is treated, but the treatment status is unclear when $S = 1985 = \lfloor c \rfloor$.

Second, we cannot obtain the $E(Y|S)$ model by replacing G with S in (5.2); rather, $E(Y|S)$ should be derived from $E(Y|G)$ using $E(Y|S) = E\{E(Y|G)|S\}$, as "G is finer than S". Let $e \equiv G - S \iff G = S + e$, and denote the uniform distribution on $[0,1]$ as '$Uni[0,1]$'. Observe

$$e \sim Uni[0,1] \perp\!\!\!\perp S \implies E(G|S) = E(S+e|S) = S + E(e|S) = S + E(e) = S + 0.5.$$

This shows that only $E(e)$ is to be used, not the full distribution assumption of e, and that a distribution other than $e \sim Uni[0,1]$ can be used, as long as $E(e)$ is known.

When c is an integer, $\delta \equiv 1[c \le G] = 1[c \le S]$ holds, as was just noted. We can then take $E(\cdot|S)$ on (5.2) to obtain, using $E(G|S) = S + 0.5$,

$$\begin{aligned} E(Y|S) &= \beta_0 + \beta_d\delta + \beta_1(S + 0.5 - c) + \beta_{1\delta}\delta(S + 0.5 - c) \\ &= \{\beta_0 + \beta_1(0.5-c)\} + \{\beta_d + \beta_{1\delta}(0.5-c)\}\delta + \beta_1 S + \beta_{1\delta}\delta S. \end{aligned} \tag{5.3}$$

When c is not an integer, dropping the cutoff sample restores $\delta \equiv 1[c \le G] = 1[c \le S]$, and thus (5.3) still holds with the cutoff sample dropped.

The $E(Y|S)$ model in (5.3) reveals that, if we do the OLS of Y on $(1, \delta, S, \delta S)$ ignoring the non-genuine integer score problem, then the slope estimator for δ

is consistent for $\beta_d + \beta_{1\delta}(0.5-c)$, not for β_d. Hence, this finding gives the 'no bias condition' (Lee and *et al.*, 2021):

$$\beta_{1\delta}(0.5-c) = 0. \tag{5.4}$$

In words, *if the slope is symmetric ($\beta_{1\delta} = 0$) or the cutoff falls in the middle ($c = 0.5$), then the non-genuine integer score problem can be ignored to do the OLS of Y on $(1, \delta, S, \delta S)$*—keep in mind the caveat that the cutoff sample should be dropped if c is not an integer. Even if $\beta_{1\delta}(0.5-c) \neq 0$, still the product can be small if $\beta_{1\delta} \simeq 0$ or $c \simeq 0.5$.

5.5. ESTIMATION WITH NON-GENUINE INTEGER SCORE AVOIDING BIAS*

The idea of assuming $e \sim Uni[0,1] \perp\!\!\!\perp S$ was initiated by Dong (2015). Dealing essentially only with integer c, Dong noted that the symmetry $\beta_{1\delta} = 0$ in (5.4) makes the OLS consistent, and examined polynomial models of a general order making use of moments other than $E(e)$, going beyond the linear spline model in (5.2). Based on (5.3), Dong proposed an "indirect" approach to find β_d, followed by bootstrap inference: regardless of (5.4) holding or not, dropping the cutoff sample if c is not an integer,

$$\beta_d = \{\delta \text{ slope in } E(Y|S)\} - \{\delta S \text{ slope in } E(Y|S)\}(0.5-c). \tag{5.5}$$

Bartalotti *et al.* (2021) extended the Dong's (2015) approach to multiple clusters with cluster-specific measurement errors, when either the error distribution is known or an auxiliary sample is available to provide information on moments of e. Such a case arises for spatial RD, where G is the individual distance to a borderline/boundary but what is available is only the distance S to the borderline from the centroid of a larger region, say, the county of residence for each individual. Clustering for RD in general was addressed by Bartalotti and Brummet (2017).

Lee and *et al.* (2021) noted that (5.5) is unnecessary, because we can modify the OLS regressors such that β_d can be found as a slope of a modified regressors. For this, note $S + 0.5 - c = S - \lfloor c \rfloor + 0.5 - (c - \lfloor c \rfloor)$ to redefine $S - \lfloor c \rfloor$ as S and $c - \lfloor c \rfloor$ as c, and define

$$\delta_- \equiv 1[S \leq -1], \quad \delta_0 \equiv 1[S = 0], \quad \delta_+ \equiv 1[1 \leq S], \quad S_{0.5c} \equiv S + 0.5 - c.$$

Regardless of (5.4) holding or not, for any S including $S=0$ for the cutoff sample,

$$E(Y|S) = \beta_0 + \beta_d\{(1-c)\delta_0 + \delta_+\} + \beta_-(-0.5c^2\delta_0 + \delta_- S_{0.5c}) + \beta_+\{0.5(1-c)^2\delta_0 + \delta_+ S_{0.5c}\}$$

where $\beta_- \equiv \beta_1$ and $\beta_+ \equiv \beta_1 + \beta_{1\delta}$ are the left and right slopes around $S=0$. If the cutoff sample with $\delta_0 = 1$ is dropped, then this becomes the usual RD linear spline model with score $S_{0.5c}$: $E(Y|S) = \beta_0 + \beta_d\delta_+ + \beta_-\delta_- S_{0.5c} + \beta_+\delta_+ S_{0.5c}$.

For estimation, do the OLS of Y on

$$W_c \equiv \{1,\ (1-c)\delta_0+\delta_+,\ -0.5c^2\delta_0+\delta_- S_{0.5c},\ 0.5(1-c)^2\delta_0+\delta_+ S_{0.5c}\}' \quad (5.6)$$

to estimate β_d as the slope of the regressor $(1-c)\delta_0 + \delta_+$. Lee and *et al.* (2021) also presented estimators for the quadratic model with $(G-c)^2$ and $\delta(G-c)^2$ appearing extra in (5.2).

For FRD, instead of (5.2), consider

$$E(D|G) = \alpha_0 + \alpha_\delta\delta + \alpha_1(G-c) + \alpha_{1\delta}\delta(G-c),$$
$$E(Y|G) = \gamma_0 + \gamma_\delta\delta + \gamma_1(G-c) + \gamma_{1\delta}\delta(G-c).$$

Then, recalling (5.3) for SRD with $\beta_d\delta$ replaced by $\beta_d D$, IVE can be applied to

$$Y = \beta_0 + \beta_d D + \beta_1 S + \beta_{1\delta}\delta S + error$$

with δ as an instrument for D, under the no bias condition $\alpha_{1\delta}(0.5-c) = \gamma_{1\delta}(0.5-c) = 0$.

Regardless of the no-bias condition holding or not, the generalization of (5.5) for FRD in Dong (2015) is, again dropping the cutoff sample if c is not an integer,

$$\beta_d = \frac{\{\delta \text{ slope in } E(Y|S)\} - \{\delta S \text{ slope in } E(Y|S)\}(0.5-c)}{\{\delta \text{ slope in } E(D|S)\} - \{\delta S \text{ slope in } E(D|S)\}(0.5-c)}. \quad (5.7)$$

Instead of this, however, it is simpler and more efficient (as the cutoff sample is always used) to do IVE, using W_c as an instrument for (Lee and *et al.*, 2021)

$$X_c \equiv \{1,\ D,\ -0.5c^2\delta_0+\delta_- S_{0.5c},\ 0.5(1-c)^2\delta_0+\delta_+ S_{0.5c}\}'.$$

As an empirical example for FRD, Dong (2015) estimated the effects of retirement on food consumption, using China Urban Household Survey, 1997-2006. Age G is observed as yearly age S, and $c=60$. In Table 'Effect of Retirement on Consumption', "Naive" means ignoring the non-genuine integer score

Table 4: Effect of Retirement on Consumption $Y = \ln(\text{Food Consumption})$

Effect of Retirement on Consumption $Y = \ln(\text{Food Consumption})$				
	h years (N)	δ on Y	δ on D	Ratio
Naive	6 (12,866)	-0.041 (0.016)**	0.19 (0.024)***	-0.21 (0.085)**
	10 (22,296)	-0.054 (0.013)***	0.19 (0.022)***	-0.29 (0.078)***
Correct	6 (12,866)	-0.029 (0.016)**	0.21 (0.022)***	-0.13 (0.074)*
	10 (22,296)	-0.045 (0.013)***	0.21 (0.020)***	-0.22 (0.061)***

problem to use S as if $S = G$, and "Correct" means not ignoring it; *, ** and *** denote significance at 10%, 5% and 1% levels. The column 'Ratio' is the desired effect, which is (5.7) for "Correct": the naive and correct effects differ much. There are two shortcomings in this analysis. First, covariates may have breaks at $c = 60$; e.g., various discounts/benefits may start for seniors at age 60, which can result in a break in Y just as retirement does. Second, $h = 6$ and 10 years seem too large, as persons as much as 20 years apart in age are compared.

5.6. HEAPING PROBLEM

When the observed score S has both genuinely continuous score G and a discrete component (i.e., a discrete transformation $\tau(G)$), there is a "heaping" problem. Heaping in S can occur due to rounding/grouping, low precision in the measurement tool, custom/practice (working 40 hours per week,...), etc.; heaps can occur at any value of S including the cutoff. Also, heaps can occur when respondents do not know S well. For example, respondents to a survey may not know the family members' birthdays, in which case birthdays may heap at the first day, the 15th day or the last day of each month. Barreca *et al.* (2016) showed heaping in birth weight in relation to Almond *et al.* (2010), (2011) in birthday/age as in Edmonds *et al.* (2005) and in income as in Saez (2010); these variables are often used as S.

The fact that either G or $\tau(G)$ is observed for each subject is a 'selection problem': e.g., if heaping is due to low precision in the measurement tool, those reporting $\tau(G)$ may be poorer or less careful. If these covariates (income level and the degree of carefulness) affect Y, then the heaping is non-random to introduce biases into treatment effect estimates, because those at heaps are systematically different from those around the heaps. Attention should be paid not just to the RD cutoff, but also to heap points. That is, whether covariates and S have a

break or not at heap points as well as at the cutoff should be checked out.

To see if heaping matters in the sense that covariates have breaks at some heap points as they might do at the cutoff 0, we can do OLS: with heaps at $g_1,...,g_J$, apply OLS to a covariate X_k model such as (the former is local around g_j, and the latter is global)

$$\begin{aligned} X_k &= \gamma_0+\gamma_1 1[S=g_j]+\gamma_2 1[S<g_j](S-g_j)+\gamma_3 1[g_j\le S](S-g_j)+error; \\ X_k &= \gamma_0+\sum_{j=1}^{J}\gamma_j 1[S=g_j]+\gamma_\delta 1[0\le S]+\gamma_-(1-\delta)S+\gamma_+\delta S+error. \end{aligned}$$

Recall the part errors-in-variable setting with $S=BG+(1-B)H$, where the "truth-teller dummy" B is not observed and H is continuous. Differently from this, $S=BG+(1-B)\tau(G)$ occurs in heaping, where typically B is observed and $\tau(G)$ is discrete. With B observed, we may apply the Heckman selection correction approach by explaining $B=1$ with G and then using $E(Y|G,B=1)$ as follows.

With *Cov* and *Var* being covariance and variance, for SRD with $D=1[0\le G]$, suppose

$$B=1[0<\alpha_0+\alpha_1 G+\varepsilon],\ \ Y=\beta_d D+m(G)+U,\ \ E(U|\varepsilon,G)=E(U|\varepsilon)=\frac{Cov(\varepsilon,U)}{Var(\varepsilon)}\varepsilon.$$

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the $N(0,1)$ density and distribution function. With *Cor* standing for correlation, using the well-known selection correction term $E(U|G,B=1)=\rho\sigma_u\phi\{(\alpha_0+\alpha_1 G)/\sigma_\varepsilon\}/\Phi\{(\alpha_0+\alpha_1 G)/\sigma_\varepsilon\}$ where $\rho\equiv Cor(\varepsilon,U)$, $\sigma_u\equiv SD(u)$ and $\varepsilon\sim N(0,\sigma_\varepsilon^2)\perp\!\!\!\perp G$ gives

$$E(Y|G,B=1)=\beta_d D+m(G)+\rho\sigma_u\frac{\phi\{(\alpha_0+\alpha_1 G)/\sigma_\varepsilon\}}{\Phi\{(\alpha_0+\alpha_1 G)/\sigma_\varepsilon\}}\equiv\beta_d D+\tilde{m}(G),$$
$$\tilde{m}(G)\text{ continuous}$$

because $\phi(\cdot)/\Phi(\cdot)$ is continuous. Hence we may just use the $B=1$ subsample to do the usual RD analysis around the cutoff.

Barreca *et al.* (2011) proposed "donut RD" using only the $B=1$ group around heaping points to check out the sensitivity of the treatment effect estimator as more and more observations around the heaping points are removed. Barreca *et al.* (2016), however, noted that heaping can bias the effect estimator even if a heap does not fall near the cutoff 0; e.g., all males with G near 0 may move out of the local neighborhood to a heap far away. This implies that the

donut RD idea may not work well, because far away heaps may affect the local estimation around 0. Related to this problem is that small h's reduce the bias in RD while increasing the variance, but when heaping is present, too small a h may make the bias problem worse because one or two heaps nearby may unduly influence the effect estimator.

The opposite to using only the $B = 1$ observations is collapsing G into $\tau(G)$. In trying to use birthday as score, Shigeoka (2014) noticed heaps on the first day of the month, as well as at the multiples of the fifth and tenth days. Shigeoka collapsed birthday into birth month to use age-in-months $\tau(G)$, instead of age-in-days G, with $c = 70$. Differently from using age 60 as c at which concomitant breaks in Y can happen due to other factors such as discount schemes or lower tax rates, Shigeoka noted no such problems in Japan at age 70. Shigeoka found no effect of lower copayment (at age 70) on health measures despite more health care utilization. Using $\tau(G)$ instead of G works if $\tau(G)$ is not too coarse so that $D = 1[0 \leq G] = 1[0 \leq \tau(G)]$.

Instead of using only G or $\tau(G)$, we may estimate the effect using the two groups ($B = 1$ and $B = 0$) separately—the usual estimation with the $B = 1$ group and the estimation with integer score for the $B = 0$ group—to test if the two estimates are the same or not. If yes, combine the two estimates. If not, something went awry.

6. SCORE TOPICS: SCORE DENSITY BREAK

The unknown function $m(S)$ in $E(Y|S) = \beta_d D + m(S)$ for SRD consists of $E(W|S)$ and $E(U|S)$ where W is (observed) covariates and U is an error term (i.e., an unobserved covariate). Since W can be controlled, there is little loss of generality in regarding $m(S)$ as $E(U|S)$. Alternatively, since $Y = (Y^1 - Y^0)D + Y^0$ with $D = 1[0 \leq S]$ gives

$$E(Y|S) = E(Y^1 - Y^0|S)D + E(Y^0|S) \simeq \beta_d D + E(Y^0|S) \quad \text{locally around } S = 0,$$

we may regard $m(S)$ also as $E(Y^0|S)$ that consists of $E(W|S)$ and $E(U|S)$.

In the following, first, we consider $m(S) = E(U|S)$ to examine the issue of 'S manipulation' and f_S continuity. Then, to provide counter-examples against the misguided belief that the continuity of f_S is equivalent to the continuity of $m(S)$, we consider $m(S) = E(Y^0|S)$, which is convenient in constructing counter-examples.

6.1. PERFECT MANIPULATION AND INFORMATIVE DENSITY BREAK

For $m(S) = E(U|S)$, observe, using the Bayes' rule,

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

This shows that the continuity of $f_{S|U}(s|u)/f_S(s)$ at 0 is necessary for the continuity of $E(U|S)$. Since $f_{S|U}(s|u)$ involves the unobserved U, only the continuity of $f_S(s)$ is to be checked out.

Suppose S is a test score, $D = 1[0 \le S]$ is entry to college, Y is lifetime income after age 30, U is binary with $U = 1$ for socializing well and $U = 0$ otherwise, $p \equiv P(U = 1)$ with $0 < p < 1$, and persons with $U = 0$ try extra hard to get $D = 1$ as a way to make up for their low lifetime income due to $U = 0$. Assume first (Kim and Lee, 2016)

$$f_{S|U}(s|0) = 1[0 \le s < 1] \quad \text{and} \quad f_{S|U}(s|1) = \phi(s): \tag{6.1}$$

those with $U = 0$ have their S in $[0,1]$ whereas those with $U = 1$ have their S well spread around 0 with $\phi(\cdot)$. Then

$$\begin{aligned} &f_S(s) = f_{S|U}(s|0)(1-p) + f_{S|U}(s|1)p = 1[0 \le s < 1](1-p) + \phi(s)p \\ &\Longrightarrow\ f_S(0^+) - f_S(0^-) = \{(1-p) + \phi(0)p\} - \phi(0)p = 1-p. \end{aligned}$$

This break in f_S occurs because those with $U = 0$ manipulated their S to perfection, and the break size reveals the proportion $1-p = P(U = 0)$ of the "manipulators".

Since U is binary, using the Bayes' rule again,

$$\begin{aligned} &E(U|s) = P(U = 1|s) = \frac{f_{S|U}(s|1)P(U=1)}{f_S(s)} = \frac{\phi(s)p}{1[0 \le s < 1](1-p) + \phi(s)p} \\ &\Longrightarrow E(U|0^+) - E(U|0^-) = \frac{\phi(0)p}{1-p+\phi(0)p} - \frac{\phi(0)p}{\phi(0)p} \ne 0 \quad \text{as long as } 1-p \ne 0. \end{aligned}$$

Hence the break in f_S ($f_S(0^+) - f_S(0^-) = 1-p$) is informative for the break in $E(U|S)$.

Urquiola and Verhoogen (2006) showed an example of f_S break at c with class size D, the number of enrolled students S, and a test score Y. As S crosses c, D drops because a law dictates that the maximum number of students in a class be c, but household income and mother schooling jump at c to cause a break in f_S at c.

To check out the f_S continuity at c, we may simply draw a histogram around c with c as a histogram boundary point to see if the histogram has a break at c or not. A more advanced way is to draw f_S using a 'one-sided LCR kernel estimator' as in Choi and Lee (2017, p. 1229), or using an 'one-sided LLR kernel estimator' as will be seen shortly.

6.2. IMPERFECT MANIPULATION AND INTENT-TO-TREAT EFFECT

Suppose that, differently from (6.1), $f_{S|U}(s|0)$ is a continuous density, say $\psi(s)$, that is tilted heavily to the right of 0, which means that those with $U=0$ could not perfectly manipulate their S, although they could to a large extent. In this case,

$$f_S(s) = f_{S|U}(s|0)(1-p) + f_{S|U}(s|1)p = \psi(s)(1-p) + \phi(s)p \quad \text{(continuous at 0);}$$
$$E(U|S=s) = \frac{f_{S|U}(s|1)P(U=1)}{f_S(s)} = \frac{\phi(s)p}{\psi(s)(1-p)+\phi(s)p} \quad \text{(also continuous at 0).}$$

Here, the continuity of f_S at 0 is informative for that of $E(U|S)$. This example illustrates that, *even if individuals can manipulate S, this will not make* f_S *discontinuous at* 0*, so long as they cannot do it perfectly.*

Score S is not subject to manipulation if S is given as in age, or if c is unknown. For instance, in Van der Klaauw (2002) for college scholarship amount D, S and c are known to admission officers, but never to students to prevent manipulation. In Malamud and Pop-Eleches (2011), a fixed proportion of applicants get a voucher for PC purchase, depending on income rank S which in turn depends on who applied: S and c unknown to everybody beforehand.

A related case, which is not a genuine RD though, appeared in Card *et al.* (2008) for "tipping point": whites' exodus out of city center may suddenly increase at a certain point, where S is the neighborhood minority share. Here, S is known to everybody but c is unknown; c is estimated to be $5 \sim 20\%$. Hansen (2017) examined effects of debt-GDP ratio on economic growth, because economic growth was hypothesized to falter after the debt-GDP ratio crosses an unknown cutoff; this example pertains to RK though, as the slope break is at stake, not an intercept break. Porter and Yu (2015) estimated c along with the treatment effect to show that the estimated c is as good as the true c.

Suppose that a medicine D in a randomized clinical trial is hard to take, as it causes nausea. Then some subjects would not comply to the group assignment, say $\delta = 0, 1$. One effect of interest is $E(Y|D=1) - E(Y|D=0)$ where D is the

actual treatment taken, but as important as this is the '*intent-to-treat (ITT) effect*' $E(Y|\delta=1)-E(Y|\delta=0)$ which includes the 'noncompliance effect'. ITT effect is informative because many patients also will not comply when D is prescribed in the real world.

An example analogous to ITT can be seen in a summer remedial course D newly starting, based on a pre-summer score. Let G be the score before D started. With D in place newly, some teachers manipulate G so that a new score S emerges: the score of students who would benefit from the summer course is lowered if it is above c, and the score of those who would be harmed is raised if below c. Observing S instead of G, three choices appear: (i) trying to identify the no-manipulation effect $E(Y^1-Y^0|G=c^+)$, (ii) identifying the manipulated effect $E(Y^1-Y^0|S=c^+)$ analogous to ITT, and (iii) redesigning D to weaken manipulation as the following illustrates (Lee and Choi, 2021).

For an income-supporting program D, instead of $D=1[\text{income}-c<0]$, we may set $D=1[\text{'income rank among the applicants'}-c<0]$: as it is unclear who would apply to the program, there is a lesser room for manipulation. For a school graduation $D\equiv 1[c\leq S]$ with a test score S, we may use the class standing as S to make manipulation hard.

Gerard *et al.* (2020) proposed a "manipulation-robust" bounding approach for treatment effect, when manipulation is done by a proportion of subjects locating themselves deliberately on one side of c. Although these manipulators are not identified individually, their proportion $1-p$ can be found with the break magnitude of f_S as was illustrated above.

6.3. INDECISIVE DENSITY BREAK

The continuity of f_S is informative, but contrary to the common perception, *the f_S continuity is neither necessary nor sufficient for the continuity of $m(S)=E(Y^0|S)$*—the basic ID assumption in (2.1) for SRD. Indeed, McCrary (2008) provided two counter-examples to make this point, which seem to be ignored by most practitioners though. The two examples are just verbal, and we provide two formal counter-examples next, drawing on Lee and Choi (2021).

Suppose $Y=1$ is working in a tech sector at age 30, S is a test score, $D=1[0\leq S]$ is high school graduation, and $p\equiv P(Y^0=1)$; $Y^0=1$ means working in a tech sector without high school graduation, and $Y^1=1$ with high school

graduation. Observe

$$f_S(s) = f_{S|Y^0}(s|0)P(Y^0=0) + f_{S|Y^0}(s|1)P(Y^0=1),$$
$$E(Y^0|s) = P(Y^0=1|s) = \frac{f_{S|Y^0}(s|1)}{f_S(s)}P(Y^0=1) \quad \text{(using the Bayes' rule).}$$

First, suppose

$$p=0.5, \quad f_{S|Y^0}(s|0) = 2\times 1[-0.5\le s<0], \quad f_{S|Y^0}(s|1) = 2\times 1[0\le s<0.5].$$

In the last term $f_{S|Y^0}(s|1)$, those 'working in a tech sector without a high school diploma' (i.e., $Y^0=1$) have the test scores on $[0,0.5]$, which is higher than $[-0.5,0]$ for those 'not working in any tech sector without a high school diploma' (i.e., $Y^0=0$) in $f_{S|Y^0}(s|0)$, as the $Y^0=1$ group tend to be smarter than the $Y^0=0$ group. In this example, there is no break in f_S but there is a break in $E(Y^0|S)$, illustrating that a break in f_S is not necessary for the $E(Y^0|S)$ break:

$$f_S(s) = f_{S|Y^0}(s|0)0.5 + f_{S|Y^0}(s|1)0.5 = 1[-0.5\le s<0.5]$$
$$\implies f_S(0^-) = 1 = f_S(0^+);$$
$$E(Y^0|S=s) = \frac{f_{S|Y^0}(s|1)}{f_S(s)}P(Y^0=1) = \frac{2\times 1[0\le s<0.5]}{1[-0.5\le s<0.5]}0.5 = 1[0\le s<0.5]$$
$$\implies E(Y^0|0^+) - E(Y^0|0^-) = 1-0 = 1.$$

Since $E(Y^0|S)$ is discontinuous at $S=0$, the RD identification fails despite the f_s continuity.

Second, turning to 'f_S break but no break $E(Y^0|S)$ at 0', an example in McCrary (2008) is: for a summer school attendance D determined by S falling below 0, "Teachers give bonus points to some of those who just barely fail the exam (perhaps to reduce the size of summer school classes), and subtract points from no student. Then the density test would suggest a failure of ID. However, if teachers select at random which students receive bonus points, then an ATE (average treatment effect) would be identified."

To mathematically formalize this example, the bonus point should be specified, which is not straightforward however, because questions such as "is the bonus point infinitesimally small?" and "is it the same or different across students" arise. Lee and Choi (2021) thus replaced the bonus point with 'sign reversal': randomly selected subjects ($A=1$) who would be treated have their score sign-reversed not to be treated. Then Lee and Choi (2021) proved that both $E(Y^1-Y^0|G=0^+)$ and $E(Y^1-Y^0|S=0^+)$ are identified, despite a break in f_S of size $-2f_G(0)P(A=1)\neq 0$.

6.4. SCORE DENSITY CONTINUITY TESTS*

So far, we showed that f_S continuity is informative, but neither necessary nor sufficient for RD validity. Hence, it seems adequate to simply compare histograms around c, with c as a histogram boundary point, without doing anything further for f_S continuity. Despite this, f_S continuity tests have been applied routinely in practice, and RD studies with f_S continuity rejected are hard to find. There can be two reasons for this. First, the tests have low power, and the RD local sample size is small. Second, there is a "selection problem" that RD studies with the f_S continuity rejected are driven out of the academic community.

In this section, we review the popular McCrary's (2008) LLR test for f_S continuity, and then a recent test in Cattaneo *et al.* (2020). Although the two tests look differ much, a modified, possibly improved, version of the latter becomes similar to the former. Other tests are available in the literature as well, although hardly used: Otsu *et al.* (2013) based on empirical likelihood, Frandsen (2017) for non-manipulation of integer S, and Bugni and Canay (2021) for a sign test using order statistics.

The idea of the McCrary (2008) test is applying the local linear density estimator of Cheng *et al.* (1997) separately to the negative and positive sides. First, estimate a histogram for f_S with a bandwidth. Then the midpoint of each histogram interval is taken as an independent variable, and a normalized histogram height at the midpoint is taken as the dependent variable, to which LLR is applied with another bandwidth. Then the test uses the difference between the two density estimators at the cutoff.

Specifically, let h_1 be the interval size for the first-stage histogram. Let G_j, $j = 1...n$, be the midpoints of n intervals in the histogram, and R_j the histogram height at G_j divided by Nh_1 (i.e., the relative frequency divided by h_1). McCrary (2008, p. 705) suggested $h_1 = 2SD(S)N^{-1/2}$. The second stage is a test with the logarithm of the intercept ratio $\hat{\theta} \equiv \ln(\hat{\varphi}_0/\hat{\psi}_0) = \ln\hat{\varphi}_0 - \ln\hat{\psi}_0$, where (φ_0, φ_1) and (ψ_0, ψ_1) minimize the positive and negative side minimands, respectively:

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

$K(t) = (1-|t|)1[|t| \leq 1]$ is the triangular kernel and h_2 is a second bandwidth. Then,

$$\sqrt{Nh_2}(\hat{\theta} - \theta) \rightsquigarrow N\{0, \ \frac{24}{5}(\frac{1}{\varphi_0} + \frac{1}{\psi_0})\}. \tag{6.2}$$

Using $\ln(\cdot)$ in the test is, however, undesirable for two reasons. First, if $\hat{\varphi}_0 = 0$ or $\hat{\psi}_0 = 0$, then the test statistic does not exist. Second, even if $\hat{\varphi}_0 \neq 0$ and

$\hat{\psi}_0 \neq 0$, still small values of φ_0 or ψ_0 would blow up the asymptotic variance, as they appear in the denominators. The latter can diminish the test power much, which nonetheless may "please" the researcher, as the research would pass the McCrary test. The following test without $\ln(\cdot)$ should work better:

$$\sqrt{Nh_2}\{\hat{\varphi}_0 - \hat{\psi}_0 - (\varphi_0 - \psi_0)\} \rightsquigarrow N\{0,\ \frac{24}{5}(\varphi_0 + \psi_0)\}. \tag{6.3}$$

The asymptotic variance follows from $\ln\hat{\varphi}_0 - \ln\hat{\psi}_0 - (\ln\varphi_0 - \ln\psi_0) \simeq \varphi_0^{-1}(\hat{\varphi}_0 - \varphi_0) - \psi_0^{-1}(\hat{\psi}_0 - \psi_0)$; $\varphi_0 + \psi_0$ in (6.3) is due to $\hat{\varphi}_0 \perp\!\!\!\perp \hat{\psi}_0$ and multiplying φ_0^{-1} in (6.2) by φ_0^2 and ψ_0^{-1} by ψ_0^2.

A local-linear approximation of $f(s)$ implies a quadratic approximation of $F(s)$:

$$f(s) \simeq b_1 + b_2 s \implies F(s) \simeq b_0 + b_1 s + \frac{b_2}{2}s^2 \quad \text{(dropping the subscript } S\text{)}. \tag{6.4}$$

The quadratic form allows estimating $F(s)$ with the empirical distribution function $\hat{F}(s) \equiv N^{-1}\sum_i 1[S_i \leq s]$. Using this idea, Cattaneo *et al.* (2020) propose to minimize

$$\sum_i \{\hat{F}(S_i) - b_0 - b_1(S_i - s) - b_2(S_i - s)^2\}K(\frac{S_i - s}{h})$$

with respect to (wrt) (b_0, b_1, b_2), and take the slope as the density estimator: $\hat{f}(s) = \hat{b}_1(s)$. The solution $\hat{b}(s) \equiv \{\hat{b}_0(s), \hat{b}_1(s), \hat{b}_2(s)\}'$ is a WLS: with $r(t) \equiv (1, t, t^2)'$,

$$\hat{b}(s) = \{\sum_i r(S_i - s)r(S_i - s)'K(\frac{S_i - s}{h})\}^{-1} \cdot \sum_i r(S_i - s)\hat{F}(S_i)K(\frac{S_i - s}{h}).$$

Cattaneo *et al.* (2020) showed that, for a bias term *Bias* and constant $V_s > 0$,

$$\sqrt{Nh}\{\hat{f}(s) - f(s) - Bias\} \rightsquigarrow N(0, V_s), \quad \hat{V}_s \to^p V_s \quad \text{where} \quad S_i^{sh} \equiv \frac{S_i - s}{h},$$

$$\hat{V}_s \equiv (0,1,0)\cdot\hat{A}(s)^{-1}\hat{G}(s)\hat{A}(s)^{-1}\cdot(0,1,0)', \quad \hat{A}(s) \equiv \frac{1}{Nh}\sum_i r(S_i^{sh})r(S_i^{sh})'K(S_i^{sh}),$$

$$\hat{G}(s) \equiv \frac{1}{N^3h^3}\sum_{i,j,k=1}^{N} r(S_j^{sh})r(S_k^{sh})'K(S_j^{sh})K(S_k^{sh})\{1[S_i \leq S_j] - \hat{F}(S_j)\}\{1[S_i \leq S_k] - \hat{F}(S_k)\};$$

$\hat{G}(s)$ requires triple sums. Based on this, a test statistic for the f continuity at the cutoff 0 is

$$\{\frac{N_1}{N}\hat{f}_+(0) - \frac{N_0}{N}\hat{f}_-(0)\}/(\frac{N_1}{N}\frac{1}{Nh}\hat{V}_+ + \frac{N_0}{N}\frac{1}{Nh}\hat{V}_-)$$

where $N_0 \equiv \sum_i 1[S_i < 0]$, $N_1 \equiv N - N_0$, and $(\hat{f}_+, \hat{f}_-)$ and $(\hat{V}_+, \hat{V}_-)$ are the density and asymptotic variance estimators using the positive and negative side observations only.

Cattaneo *et al.* (2020) suggest an elaborate choice of h, noting that an ad-hoc choice would not work. Relative to the McCrary test, their test is better in using only one h, but seems less robust to the choice of h. Cattaneo *et al.* (2020) allow different bandwidths on the negative and positive sides, but this would negate the advantage of using only one bandwidth; in their empirical analysis, their data-driven bandwidths on the two sides differ much. Also, $\hat{G}(s)$ in the above asymptotic variance estimator involves a time-consuming triple sum.

For (6.4), using a smooth estimator such as $\tilde{F}(s) \equiv N^{-1}\sum_i \Phi\{(s-S_i)/h_d\}$ for a bandwidth h_d seems better, because the eventual goal is estimating f, not F; $\partial\tilde{F}(s)/\partial s$ is the usual kernel density estimator. However, $\tilde{F}(s)$ requires choosing the extra bandwidth h_d, which makes the Cattaneo *et al.*'s test similar to the McCrary test. A detailed study is called for, comparing the McCrary and Cattaneo *et al.* tests in their performance and sensitivity to bandwidth choice.

7. OTHER TOPICS: REGRESSION KINK (RK)

In RD, we identify the effect of a treatment D, using an intercept break. Sometimes, however, we may identify the effect using a slope break, not an intercept break. This is 'RK', which is examined here. A slope break may occur with or without any intercept break, and D can be binary or continuous as a function of S. For continuous D, RK is the natural way of effect identification, as there is no intercept break.

7.1. RK IDENTIFICATION

In nonparametrics, estimating $\nabla E(\cdot|s) \equiv \partial E(\cdot|s)/\partial s$ is more difficult than estimating $E(\cdot|s)$, because the estimation error $\widehat{\nabla E(\cdot|s)} - \nabla E(\cdot|s)$ converges in probability to zero as $N \to \infty$ at the rate $\sqrt{Nh^2}$ that is slower than $\sqrt{Nh}$ for $\widehat{E(\cdot|s)} - E(\cdot|s)$. Derivative estimation requires "finer" information than the mean

estimation. Define the right and left derivatives at c as

$$\nabla_+ E(\cdot|c) \equiv \lim_{\xi \downarrow 0} \frac{E(\cdot|c+\xi) - E(\cdot|c)}{\xi} \quad \text{and} \quad \nabla_- E(\cdot|c) \equiv \lim_{\xi \downarrow 0} \frac{E(\cdot|c) - E(\cdot|c-\xi)}{\xi}.$$

Now recall for FRD:

$$E(Y|S) = \beta_d E(D|S) + m(S) \Longleftrightarrow \beta_d = \frac{E(Y|0^+) - E(Y|0^-)}{E(D|0^+) - E(D|0^-)}.$$

Suppose the break in $E(D|S)$ at the cutoff 0 is small: the "ID power" is weak in RD. Then one may identify β_d using the difference of derivatives instead of the difference of means. This may provide a better ID power despite the slower convergence rate, if the derivative difference is greater enough than the mean difference to overcome the convergence rate disadvantage.

Specifically, suppose $\nabla E(D|s)$ *has a break at* 0*, but* $\nabla m(s)$ *including* $\nabla E(W|s)$ *for covariates* W *is continuous.* Then we can identify β_d with the 'derivative difference ratio' instead of the mean difference ratio, because differentiating $E(Y|S) = \beta_d E(D|S) + m(S)$ from right and left at the cutoff and solving it for β_d renders the derivative difference ratio:

$$\begin{aligned} &\nabla_+ E(Y|0) = \beta_d \nabla_+ E(D|0) + \nabla_+ m(0) \ \text{ and } \ \nabla_- E(Y|0) = \beta_d \nabla_- E(D|0) + \nabla_- m(0) \\ &\qquad \Longrightarrow \ \beta_d = \frac{\nabla_+ E(Y|0) - \nabla_- E(Y|0)}{\nabla_+ E(D|0) - \nabla_- E(D|0)} \quad \{\text{as } \nabla_+ m(0) = \nabla_- m(0)\}. \end{aligned}$$

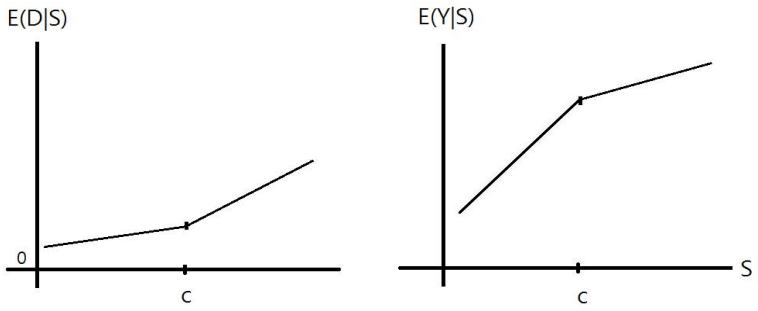
Bear in mind that β_d is the effect of D on Y, not the derivative of the effect.

Using a slope break instead of an intercept break entails restrictive conditions, however, such as the continuity of $\partial f_{S|W}(s|w)/\partial s$ at 0. Card *et al.* (2015) proposed to test for the continuous differentiability (i.e., no kink) of f_S, which is stronger than the continuity (i.e., no break) of f_S for RD. They further suggested to test for the continuous differentiability of $E(W|s)$, which is stronger than the continuity of $E(W|s)$.

In Figure 'RK Effect', the treatment effect at c can be seen visibly as the slope difference in the right panel divided by the slope difference in the left panel. If D exerts a causal effect on Y and there is a "kink relation" between D and S at 0, then there should be a kink relation between Y and S at 0. As there are SRD and FRD, there are Sharp RK (SRK) with D determined solely by S, and fuzzy RK (FRK) with D determined by S and a random variable ε. An example of SRK is a non-binary treatment $D = \alpha_{\delta s} \delta S = \alpha_{\delta s} 1[0 \le S] S$ for a constant $\alpha_{\delta s}$, which has $E(D|0^+) = E(D|0^-) = 0$ but $\nabla_+ E(D|0) - \nabla_- E(D|0) = \alpha_{\delta s}$ (Nielsen *et al.*, 2010), because

$$\nabla_+ E(D|0) = \lim_{s \downarrow 0} \alpha_{\delta s} 1[0 \le s] = \alpha_{\alpha s} \quad \text{and} \quad \nabla_+ E(D|0) = \lim_{s \uparrow 0} \alpha_{\delta s} 1[0 \le s] = 0.$$

Figure 3: RK Effect: Right Slope Difference Divided by Left Slope Difference



7.2. CONTINUOUS VERSUS NON-CONTINUOUS TREATMENT*

One controversy in RK is whether D can be binary or not; whereas Card *et al.* (2015), (2017) examined continuous treatments, Dong (2018) looked at binary treatments. Recall, for FRD,

$$E(Y|S) = \beta_d E(D|S) + m(S)$$
$$\iff \beta_d = \frac{E(Y|0^+) - E(Y|0^-)}{E(D|0^+) - E(D|0^-)} = E(Y^1 - Y^0|0^+, \text{complier}).$$

As far as the equivalence between the regression and difference forms goes here, D does not have to be binary, but the rightmost causal interpretation with (Y^0, Y^1) requires D to be binary.

When D is binary in SRD, looking at the difference $E(Y|0^+) - E(Y|0^-) = E(Y^1|0^+) - E(Y^0|0^-)$ is natural as it leads to the familiar $E(Y^1 - Y^0|0^+)$ under the continuity of $E(Y^0|s)$ at 0. When $D = d(S)$ in SRK is continuous (e.g., $\alpha_{\delta s}\delta S$) with its realized version $d(s)$, there are infinitely many potential responses, and we may look at the 'per d-unit change' in Y measured by $E(Y^d - Y^0|0^+)/d$. Letting $d \to 0^+$ provides a "d-free" effect, taking the form of a derivative. That is, for continuous D, RK is a natural starting point.

More formally, suppose $D = d(S)$ and $Y = y(D,S,U)$ for an error term U, which includes separable models such as $Y = \beta_d D + m(S) + U$ as special cases. Then a natural effect to look at in SRK with (S,U) fixed at $(0,u)$ is

$$\nabla_1 y\{d(0),0,u\} \quad \text{where} \quad \nabla_1 y(d^*,s,u) \equiv \frac{\partial y(d,s,u)}{\partial d}|_{d=d^*}.$$

Card *et al.* (2015) proved that, using the Bayes' rule,

$$\frac{\nabla_+ E(Y|0) - \nabla_- E(Y|0)}{\nabla_+ E(D|0) - \nabla_- E(D|0)} = \tau\{d(0),0\} \equiv E[\nabla_1 y\{d(0),0,u\}|S=0]$$
$$= \int \nabla_1 y\{d(0),0,u\}\partial F_{U|S}(u|0) = \int \nabla_1 y\{d(0),0,u\}\frac{f_{S|U}(0|u)}{f_S(0)}\partial F_U(u):$$

the causal interpretation of SRK effect is a weighted average of the effect $\nabla_1 y\{d(0), 0, u\}$ at $S = 0$ and $U = u$*, where values of u with a high* $f_{S|U}(0|u)$*—that is, close to the cutoff—receive a relatively higher weight.* An analogous weighted average holds for FRK. Recall that we also saw analogous weighted averages when RD effect heterogeneity was discussed in (2.2) and (2.3).

So far we considered binary or continuous D in SRK, but there can be mixed types such as

$$D = \alpha_\delta \delta + \alpha_{\delta S}\delta S = \alpha_\delta 1[0 \le S] + \alpha_{\delta s} 1[0 \le S]S = (\alpha_\delta + \alpha_{\delta s}S)1[0 \le S].$$

There is a constant effect α_δ of δ and a S-interacting effect $\alpha_{\delta s}$. For instance, D may be tax amount where α_δ is the base tax rate as the normalized income $S-c$ exceeds 0 and $\alpha_{\delta s}$ is the tax proportional to the income exceeding c (i.e., $S-c$).

In $D = \alpha_\delta \delta + \alpha_{\delta S}\delta S$, there are breaks in both $E(D|S)$ and $\nabla E(D|S)$. If the effect β_d of D on Y in $E(Y|S) = \beta_d E(D|S) + m(S)$ is the target, asymptotically there is no gain in using both RD and RK to estimate β_d, because RD estimators are asymptotically more efficient than RK estimators. However, as was mentioned above, the RK break can be greater than the RD break, greater enough to overcome the disadvantage in the $\sqrt{Nh^2}$ versus $\sqrt{Nh}$ convergence rates. This means that, in small samples, we may estimate β_d with either RD or RK, and combine them if desired. Indeed, Dong (2018) suggested IVE with both δ and δS as instruments for D.

7.3. RK ESTIMATION AND EXAMPLES

A simple RK estimator for FRK is IVE (Card *et al.* (2012) to

$$Y = \eta_0 + \eta_1 S + \eta_1^\Delta D + error \tag{7.1}$$

with δS as an instrument for D; for SRK with $D = \delta S$, this IVE reduces to OLS. In this Y equation, there is no intercept break allowed as S crosses 0.

An alternative estimator allowing an intercept break is a LQR minimizing wrt (τ_0,τ_1,τ_2) and (ρ_0,ρ_1,ρ_2):

$$\sum_i (D_i-\tau_0-\tau_1 S_i-\tau_2 S_i^2)^2 1[S_i\in(0,h)],\ \ \sum_i (Y_i-\rho_0-\rho_1 S_i-\rho_2 S_i^2)^2 1[S_i\in(0,h)].$$

The minimizers are $(\hat\tau_0^+,\hat\tau_1^+,\hat\tau_2^+)$ and $(\hat\rho_0^+,\hat\rho_1^+,\hat\rho_2^+)$; define $(\hat\tau_0^-,\hat\tau_1^-,\hat\tau_2^-)$ and $(\hat\rho_0^-,\hat\rho_1^-,\hat\rho_2^-)$ analogously with $S\in(-h,0)$. The LQR estimator for RK is the slope difference ratio:

$$\hat\beta_d \equiv \frac{\hat\rho_1^+-\hat\rho_1^-}{\hat\tau_1^+-\hat\tau_1^-}$$

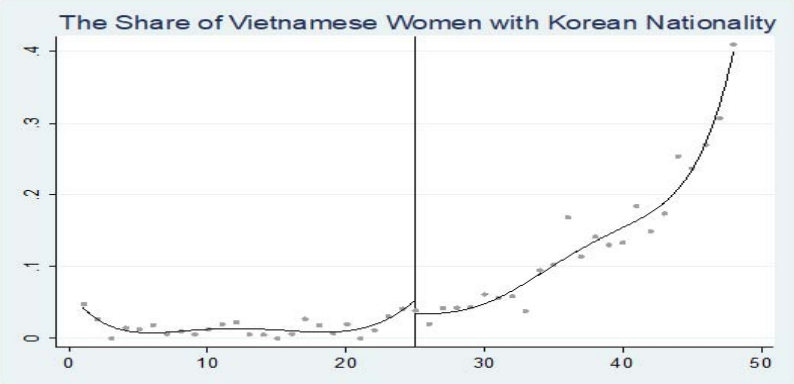
If $\tau_2=\rho_2=0$, then the LLR estimator obtains, which differs from the IVE to (7.1), because the intercept continuity ($\tau_0^+=\tau_0^-$) is not imposed in LLR. LQR tends to be numerically unstable, and thus LLR is recommended.

In Dahlberg *et al.* (2008) with local government expenditure/tax Y, the central government grant amount D is proportional to $S-2$ when emigration share S is at least 2%, and D is a fuzzy version of $\alpha_\delta 1[2\le S](S-2)$. In Kim and Lee (2016), Y is work hours, S is the previous year income, and D is the previous year average tax rate. As the marginal tax rate has jumps, its "integral" average rate D is kink-continuous. See also Simonsen *et al.* (2016) for a RK example on price sensitivity of demand for prescription drugs in Denmark.

Lee and Wie (2020) examined the effect of legal entitlement on women's empowerment in South Korea, where foreign brides have to wait two years before they become South Korean citizens. S is the duration of residence in months with $c=24$, D is the dummy for obtaining South Korean citizenship, and several variables are used for Y, such as labor market participation, living with mother-in-law, etc. This is a FRK, because there were exceptions to the residence requirement. Figure 'Proportion of Vietnamese Brides with Korean Citizenship' plots $E(D|S)$ versus S to show that there is no intercept break at c but the slope becomes steeper after c. Using only the Vietnamese brides in the National Survey of Multicultural Family in 2009 with $N=8903$, Lee and Wie found a significantly negative effect of D on experiencing discrimination (about $-35\sim-55\%$); also the relationship with the parents-in-law significantly improved.

Card *et al.* (2015) examined the RK effect of unemployment insurance benefit D on unemployment duration Y in Austria, where S is the average earnings in a base yea and $N\simeq 275{,}000$. Card *et al.* (2015) tried various bandwidth-choice and bias-correction schemes. They obtained better results with a relatively large h without bias correction. Card *et al.* concluded that, first, optimal bandwidth

Figure 4: Proportion of Vietnames Brides with Korean Citizenship



selectors choose a relatively small h, which leads to imprecise estimates. Second, LQR tends to be quite noisy. Third, the bias-corrected estimates from LLR do not differ much from the uncorrected estimates, but the correction adds much imprecision. *"Optimal bandwidths" and "bias corrections" in RK (and RD) should be taken with a grain of salt*.

With RK, Iwasaki *et al.* (2019) verified 'prospect theory' with natural experimental data from the Fukushima Nuclear Disaster in 2011, Japan. The two main tenets of prospect theory are reference dependence and loss aversion: utility functions due to income, health, family size, house size, etc. change the slope around a reference point where the left side (i.e. loss from the reference point) is steeper than the right side (i.e., gain). The disaster provides a natural experiment, where losses/gains are randomized, not self-selected. Based on survey data on displaced residents, Iwasaki *et al.* (2019) found a kink in the utility due to health (and income).

8. OTHER TOPICS: HIGH-ORDER EFFECTS

We have seen a break in intercept for RD and a break in slope for RK. This raises the question: are there higher-order breaks? We address this question here with a focus on second-order effect, drawing on Lee (2020). Sharp cases are dealt with first, followed by fuzzy cases.

8.1. HIGH-ORDER EFFECT IDENTIFICATION

For SRD with $D = \delta \equiv 1[0 \leq S]$, take $E(\cdot|S)$ on the observed $Y = (Y^1 - Y^0)\delta + Y^0$ to get

$$E(Y|S) = \beta(S)\delta + m(S) \quad \text{where} \quad m(S) \equiv E(Y^0|S) \quad \text{and}$$
$$\beta(s) \equiv E(Y^1 - Y^0|S = s) \text{ are assumed to be continuous at 0.}$$

As any continuous function can be approximated well around 0 with a power function, let

$$\beta(S) = \sum_{r=0}^{R} \beta_{r\delta} S^r = \beta_{0\delta} + \beta_{1\delta} S + , ..., + \beta_{R\delta} S^R.$$

The S-conditional effect of the binary treatment δ is $\beta(S)$, which is decomposed into effects $\beta_{0\delta}$, $\beta_{1\delta}$, , ..., $\beta_{R\delta}$ of various orders. Call $\beta_{r\delta}$ the 'RD$_r$ effect' of the binary treatment δ.

For example, suppose S is time, a lower speed limit law δ goes into effect on day $S = 0$, and Y is traffic fatality. The treatment δ may result in an immediate intercept break $\beta_{0\delta} < 0$, or its effect may be gradual with $\beta_{0\delta} = 0$ and $\beta_{1\delta} < 0$. The effect may be even more gradual with $\beta_{0\delta} = \beta_{1\delta} = 0$ and $\beta_{2\delta} < 0$, which is a "deceleration" compared with the decrease $\beta_{1\delta} < 0$.

Substituting the $\beta(S)$ power function into $E(Y|S) = \beta(S)\delta + E(Y^0|S)$ gives

$$E(Y|S) = \sum_{r=0}^{R} \beta_{r\delta} S^r \delta + m(S) = \beta_{0\delta}\delta + \beta_{1\delta} S\delta + , ..., + \beta_{R\delta} S^R \delta + m(S). \quad (8.1)$$

If $m(s)$ is continuously differentiable up to order $R \geq 2$, then $\beta_{0\delta}, \beta_{1\delta}, \beta_{2\delta}, ...$ are the breaks of order-0, order-1, order-2, etc.. *The effects of various orders emerge because δ gets attached to each term in $\beta(S)$.* Using second-order power functions for both $\beta(S)$ and $m(S)$ renders $E(Y|S) = \sum_{r=0}^{2} \beta_{r\delta} S^r \delta + \sum_{r=0}^{2} \beta_r S^r$; $R = 0, 1, 2$ gives LCR, LLR and LQR, respectively.

For a function $G(s)$, denote its right and left limits at c as

$$\nabla_+^0 G(c) \equiv G(c^+) \equiv \lim_{\lambda \downarrow 0} G(c + \lambda) \quad \text{and} \quad \nabla_-^0 G(c) \equiv G(c^-) \equiv \lim_{\lambda \downarrow 0} G(c - \lambda).$$

Then define the order-r right and left derivatives at c for $r = 1, 2, ..., R$ iteratively as

$$\nabla_+^r G(c) \equiv \lim_{\xi \downarrow 0} \frac{\nabla_+^{r-1} G(c + \xi) - \nabla_+^{r-1} G(c)}{\xi}$$
$$\text{and} \quad \nabla_-^r G(c) \equiv \lim_{\xi \downarrow 0} \frac{\nabla_-^{r-1} G(c) - \nabla_-^{r-1} G(c - \xi)}{\xi}.$$

The ID assumption for RD_r with $r \leq R$ is that $\beta(S) = \sum_{r=0}^{R} \beta_{r\delta} S^r$ holds and $m(s)$ is at least r-times continuously differentiable at 0 so that $m(s)$ drops out when we take the difference between the right and left order-r derivatives of $E(Y|s)$ at 0. *Due to $\nabla_+^r E(Y|0) - \nabla_-^r E(Y|0) = r!\beta_{r\delta}$ then, the RD_r identification finding is*

$$\beta_{r\delta} = \frac{1}{r!}\{\nabla_+^r E(Y|0) - \nabla_-^r E(Y|0)\}, \qquad r = 0, 1, 2, ...$$

For $r = 0$, this is the RD identification $\beta_{0\delta} = \nabla_+^0 E(Y|0) - \nabla_-^0 E(Y|0) = E(Y|0^+) - E(Y|0^-)$.

In SRD, the treatment $D = \delta \equiv 1[0 \leq S]$ jumps from 0 to 1 at $S = 0$; the break in δ results in a break in $E(Y|S)$ as long as δ affects Y. In SRK, the treatment $D = D(S)$ changes its slope at $S = 0$; the kink in D results in a kink in $E(Y|S)$ as long as D affects Y. *In $E(Y|S) = \beta_{0\delta}\delta + \beta_{1\delta} S\delta + \beta_{2\delta} S^2\delta + \sum_{r=0}^{2} \beta_r S^r$, if we regard the non-binary $S\delta$ as the RK treatment, then the "order-0" RK effect (say, "RK_0") is $\beta_{1\delta}$, whereas $\beta_{1\delta}$ is the order-1 RD effect (RD_1) when we consider the binary δ as the treatment. As RD_1 with treatment δ is closely related to RK with treatment $S\delta$, RD_2 with treatment δ is closely related to a causal framework with treatment $S^2\delta$ ("Regression Acceleration"?).*

8.2. HIGHER-ORDER EFFECT ESTIMATION

Consider a local polynomial regression (LPR) of order R minimizing, wrt $(\rho_0^+, \rho_1^+, ..., \rho_R^+)$,

$$\sum_{i=1}^{N} (Y_i - \rho_0^+ - \rho_1^+ S_i -, \ ..., \ -\rho_R^+ S_i^R)^2 1[0 < S_i];$$

a kernel $K(S_i/h)$ may be attached, if desired. Denote the minimizer as $(\hat{\rho}_0^+, \hat{\rho}_1^+, ..., \hat{\rho}_R^+)$. Define $(\hat{\rho}_0^-, \hat{\rho}_1^-, ..., \hat{\rho}_R^-)$ analogously with $1[0 < S_i]$ replaced by $1[S_i < 0]$. Due to $\nabla_+^r E(Y|0) - \nabla_-^r E(Y|0) = r!(\rho_r^+ - \rho_r^-)$, an estimator for $\beta_{r\delta} = (1/r!)\{\nabla_+^r E(Y|0) - \nabla_-^r E(Y|0)\}$ is

$$\hat{\beta}_{r\delta} \equiv \hat{\rho}_r^+ - \hat{\rho}_r^- .$$

The LPR applied separately to the positive and negative sides is equal to the OLS to

$$Y = \beta_{0\delta}\delta + \beta_{1\delta} S\delta +, ..., +\beta_{R\delta} S^R\delta + \sum_{r=0}^{R} \beta_r S^r + U \tag{8.2}$$

using only the local subsample with $Q = 1$, because different slopes are allowed in this OLS. Once the OLS is obtained, do the inference with the usual OLS asymptotic variance estimator:

$$\Omega_N \equiv (\sum_i X_i' X_i Q_i)^{-1} \cdot \sum_i (X_i X_i' \hat{U}_i^2 Q_i) \cdot (\sum_i X_i X_i' Q_i)^{-1}$$
$$\text{where} \quad X_i \equiv (1, \delta_i,\ S_i, S_i\delta_i,\ ...,\ S_i^R, S_i^R\delta_i)' \quad \text{and} \quad \hat{U}_i \equiv Y_i - X_i'\hat{\beta}.$$

Since the nonparametric dimension is one, the most practical way to choose h is by nonparametrically estimating a graph for $\nabla^r E(Y|s)$ and then "eye-balling": choose h such that that the graph is neither too jagged nor too smooth. In estimating such a graph, the often-used rule-of-thumb bandwidth $h = SD(S)N^{-1/5}$ can serve as a lower bound for h, because order-r derivative estimation for $r \geq 1$ requires a bandwidth larger than the bandwidth for $E(Y|s)$. Of course, various estimates corresponding to different h's should be presented in practice.

As is discussed in the next section, Dong and Lewbel (2015) proposed a linear approximation approach to extend the SRD identification range at c to $c_{new} \neq c$ using

$$E(Y^1 - Y^0|c_{new}) \simeq E(Y^1 - Y^0|c) + \beta_{1\delta}(c_{new} - c) = \beta_{0\delta} + \beta_{1\delta}(c_{new} - c). \quad (8.3)$$

Our second-order effect provides a generalization of this:

$$E(Y^1 - Y^0|c_{new}) \simeq \beta_{0\delta} + \beta_{1\delta}(c_{new} - c) + \beta_{2\delta}(c_{new} - c)^2. \quad (8.4)$$

For FRD, we can imagine the $E(D|S)$ equation with the β's replaced by α's in (8.1), and the $E(Y|S)$ equation with the β's replaced by γ's. Then Lee (2020) showed that the second-order approximation to $E(Y^1 - Y^0|S,\text{complier})$ is

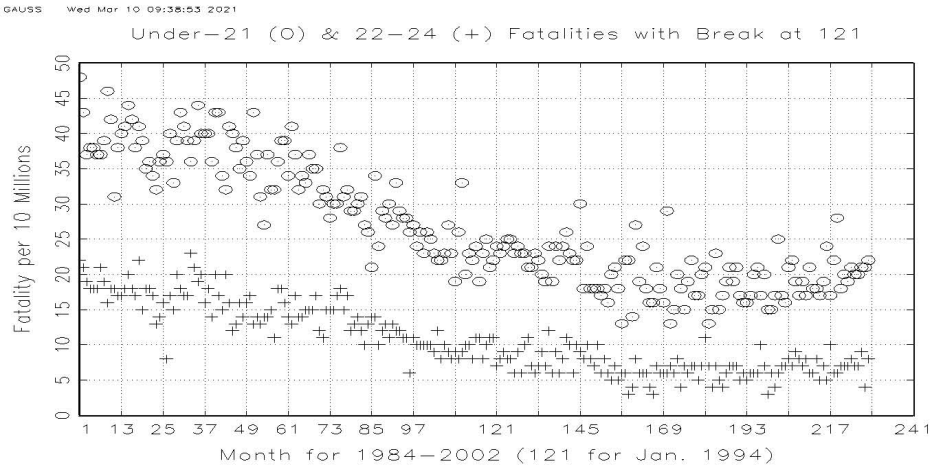
$$\frac{\gamma_{0\delta}}{\alpha_{0\delta}} + (\frac{\gamma_{1\delta}}{\alpha_{0\delta}} - \frac{\gamma_{0\delta}}{\alpha_{0\delta}}\frac{\alpha_{1\delta}}{\alpha_{0\delta}})S + (\frac{\gamma_{2\delta}}{\alpha_{0\delta}} - \frac{\gamma_{0\delta}}{\alpha_{0\delta}}\frac{\alpha_{2\delta}}{\alpha_{0\delta}} - \frac{\gamma_{1\delta}}{\alpha_{0\delta}}\frac{\alpha_{1\delta}}{\alpha_{0\delta}} + \frac{\gamma_{0\delta}}{\alpha_{0\delta}}\frac{\alpha_{1\delta}^2}{\alpha_{0\delta}^2})S^2. \quad (8.5)$$

The slope of S is the same as "$\hat{\pi}_f'(c)$" in Dong and Lewbel (2015, equation 19). That is, (8.5) includes the first-order approximation for FRD in Dong and Lewbel (2015) as a special case. Both α and γ parameters can be estimated as in the OLS to (8.2), and the asymptotic inference for the slopes in (8.5) can be done with bootstrap.

8.3. EMPIRICAL EXAMPLE: DRUNK-DRIVING LAW ON FATALITY

An under-age drunk drinking law for people aged below 21 went into effect in January 1994 in California. Figure 'Traffic Fatalities in California' plots traffic fatality in California for people below age 21 (circles) and people of age 22-24 (plus signs) per 10 million persons, based on the monthly data used in Kuo (2012) over 1984-2002. Persons of age 22-24 serve as a control group, because they are not subject to the law. Although the cutoff (January 1994) is $c = 121$, there is no apparent break at $c = 121$; rather, the fatality seems to have decreased steadily first and then picked up in 1999 (181∼192). Kuo (2012) attributed this lack of break to the lack of 'awareness of the law', which must have spread to affect traffic fatality gradually.

Figure 5: Traffic Fatalities in California (for Ages below 21 and 22-24)



Let Y be the "proportional treatment effect" Lee and Kobayashi (2001):

$$Y \equiv \frac{(\text{treatment group traffic fatality}) - (\text{control group traffic fatality})}{(\text{control group traffic fatality})}.$$

This is to control for the common multiplicative time trend (e.g., $\exp(\beta_0 S)$) between the treatment and control groups, because $\exp(\beta_0 S)$ gets cancelled out in the ratio form of Y. Let S be a month in 1984-2002 divided by 100, so that $c = 1.21$.

Figure 6: RD_0 (dotted), RD_1 (dashed) and RD_2 (solid) Breaks at 1.21

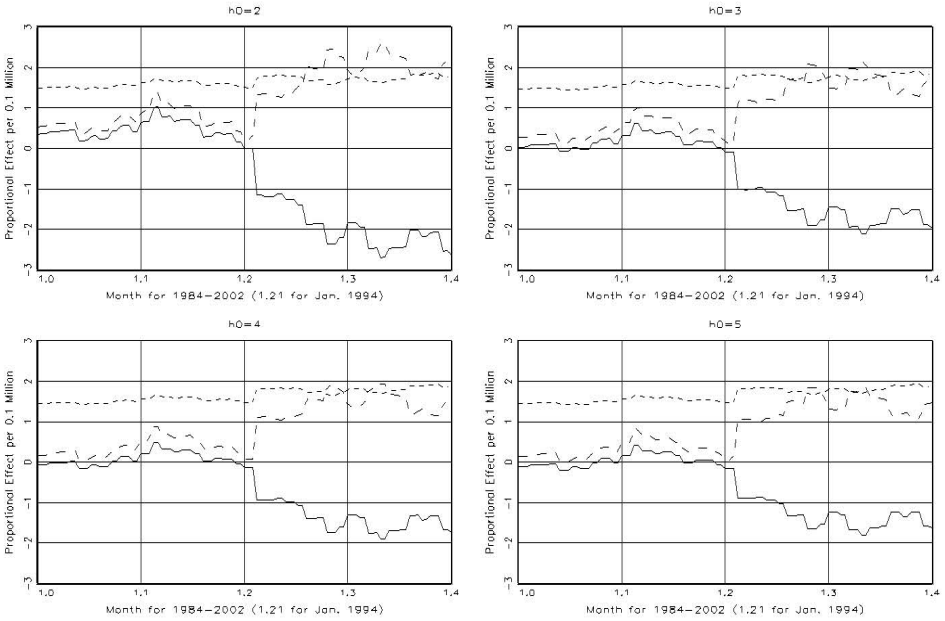


Figure 'RD$_0$ (dotted), RD$_1$ (dashed) and RD$_2$ (solid) Breaks at 1.21' shows the RD$_0$, RD$_1$ and RD$_2$ estimated with one-sided LQR. Four bandwidths $h_0 = 2,3,4,5$ are used for $h = h_0 SD(S)N^{-1/6}$. The figures reveals almost no RD$_0$ (dotted line), but a positive RD$_1$ (dashed) and a negative RD$_2$ (solid). As it is hard to imagine the law increasing traffic fatality (i.e., a positive RD$_0$ or RD$_1$), the figure demonstrates the importance of considering RD$_2$.

Using $h = 4$, OLS with only $(\beta_{0\delta}, \beta_{1\delta})$ gives the order 0 and 1 effects, which leads to the long-run effects at c_{new} (2, 4, 6, 8 years from c) using (8.3) as in Dong and Lewbel (2015). Also, OLS with $(\beta_{0\delta}, \beta_{1\delta}, \beta_{2\delta})$ gives the order 0, 1 and 2 effects, which leads to the long-run effects using (8.4). The actual long-run effect estimates are (Lee, 2020):

Effects after 2, 4, 6, 8 years based on order 0,1 effects : 0.92, 1.37, 1.83, 2.28;

Effects after 2, 4, 6, 8 years based on order 0,1,2 effects : 0.30, -0.29, -1.59, -3.60.

The first-order approximation finding based on (8.3) is that the under-age drinking law increases traffic fatality, which does not make sense, whereas the second-

order approximation finding based on (8.4) is the opposite, which is more sensible. Essentially, this is because of the negative RD_2 effect in Figure 'RD_0 (dotted), RD_1 (dashed) and RD_2 (solid) Breaks at 1.21'.

9. OTHER TOPICS: EXTENDING IDENTIFICATION RANGE

One critical shortcoming in RD is that the effect ID is local, only at the cutoff c, and the effect is only on compliers for FRD. Hence there appeared a couple of ways to extend the effect ID range away from c or from compliers, which are reviewed next; extending the ID range of RD effect also goes by the expression"increasing external validity". Other than the two approaches reviewed below, Angrist (2015) proposed a way to identify effects away from c as briefly reviewed in Choi and Lee (2017, p. 1236), but their approach is a matching, not a RD. Also, Wing and Bello-Gomez (2018) provided a review on extending RD identification range, adding one more approach with a control group. This additional approach is, however, just a DD which identifies effects far away from c, using a control group.

9.1. APPROACHES BASED ON DERIVATIVES

As was mentioned already, Dong and Lewbel (2015) showed that the SRD effect at $c_{new} \neq c$ can be found with the effect at c using (8.3), and Lee (2020) generalized the linear approximation in (8.3) using high-order terms, a special case of which is the second order approximation in (8.4). For FRD, (8.5) is the second order approximation for $E(Y^1 - Y^0|S$, complier), which includes the first-order approximation in Dong and Lewbel (2015, equation 19) as a special case.

Dong and Lewbel presented an empirical example for SRD, where S is a test score in Massachusetts (MA) and D is a scholarship program which waives tuition at in-state public colleges if $c < S$. Part of their results for Y being 'attending 4-year public college in MA' or 'attending 4-year private college' with $h = 10$ is in Table 'Effects of Scholarship Program'. The table shows that D increased the probability of attending 4-year public college in MA by 8.1%. The effect -8.0% on 4-year private college is the "mirror image" of that on 4-year public college; i.e., D just switched students from private to public colleges.

Turning to finding the effect at c_{new}, Dong and Lewbel set $c_{new} = c - 2$ (1% decrease in c), and the last row of the table shows the effect at c_{new}:

$$E(Y^1 - Y^0|S = c_{new}) \simeq E(Y^1 - Y^0|S = c) + \hat{\beta}_{1\delta} \times (-2) \Longrightarrow 0.12 = 0.081 + 0.019 \times 2.$$

Table 5: Effects of Scholarship Program

Effects of Scholarship Program		
$N = 18456$	4 yr. public college	4 yr. private college
effect $\hat{\beta}_{0\delta}$ at c (SD)	0.081 (0.015)	-0.080 (0.015)
$\hat{\beta}_{1\delta}$ (SD)	-0.019 (0.003)	0.018 (0.002)
effect at c_{new} (SD)	0.12 (0.015)	-0.12 (0.015)

9.2. APPROACHES BASED ON SUBJECT-TYPE INDEPENDENCE IN FRD*

The subject types in FRD are $\{a, comp, n\}$ which stands for {always taker, complier, never taker}, along with the usually ruled-out defier. The identified FRD effect is $E(Y^1 - Y^0|S = c^+,$ complier). This ID range can be extended across the subject types and across S values, which is explored here, drawing on Bertanha and Imbens (2020).

Examine $E(Y|S,\ D = 0)$ around $S = c$ to see if the 'untreated compliers who constitute part of $(S = c^-, D = 0) = (\delta = 0, D = 0)$' differ from the never takers $(S = c^+, D = 0) = (\delta = 1, D = 0)$, which was first proposed by Battistin and Rettore (2008). An analogous examination of $E(Y|S,\ D = 1)$ around $S = c$ reveals the difference between the always takers $(S = c^-, D = 1) = (\delta = 0, D = 1)$ and the 'treated compliers who constitute part of $(S = c^+, D = 1) = (\delta = 1, D = 1)$'.

If no difference found around $S = c$ in the $D = d$ group, $d = 0, 1$, then the 'complier' in $E(Y^1 - Y^0|S = c^+,$ complier) may be dropped to establish the external validity across the subject types. Formalizing this idea in the following, first, we show that $E(Y^1 - Y^0|s)$ for $s \neq c$ is identified under the "subject-type independence" assumption $T \perp\!\!\!\perp (Y^0, Y^1)|S$ with $T \in \{a, comp, n\}$. Second, we explain how this assumption can be verified.

Observe

$$T \perp\!\!\!\perp (Y^0, Y^1)|S \implies E(Y^1 - Y^0|S = c^+, T) = E(Y^1 - Y^0|S = c^+). \qquad (9.1)$$

Because (Y^0, Y^1) are indexed by $D = 0, 1$, we have

$$\begin{aligned}
(i)\ s &< c \Longleftrightarrow \delta = 0 : E(Y|D=1,s) = E(Y^1|T=a,\ s) = E(Y^1|s); \\
(ii)\ s &< c \Longleftrightarrow \delta = 0 : E(Y|D=0,s) = E(Y^0|T=n \text{ or } comp,\ s) = E(Y^0|s); \\
(iii)\ c &\leq s \Longleftrightarrow \delta = 1 : E(Y|D=0,s) = E(Y^0|T=n,\ s) = E(Y^0|s); \qquad (9.2) \\
(iv)\ c &\leq s \Longleftrightarrow \delta = 1 : E(Y|D=1,s) = E(Y^1|T=a \text{ or } comp,\ s) = E(Y^1|s).
\end{aligned}$$

Hence $E(Y^0|s)$ *and* $E(Y^1|s)$ *are identified for all* s *using* $E(Y|D=0,s)$ *and* $E(Y|D=1,s)$ *under (9.1), and thus we can identify* $E(Y^1 - Y^0|s)$ *for all* s. Since $D = (D^1 - D^0)\delta + D^0$ consists of (δ, D^0, D^1), under the usual $\delta \perp\!\!\!\perp (Y^0, Y^1)|S$, '$T \perp\!\!\!\perp (Y^0, Y^1)|S \Longleftrightarrow (D^0, D^1) \perp\!\!\!\perp (Y^0, Y^1)|S$' implies $D \perp\!\!\!\perp (Y^0, Y^1)|S$. Then $E(Y^d|D,S) = E(Y^d|S)$ follows, which is essentially (9.2).

Turning to verifying $T \perp\!\!\!\perp (Y^0, Y^1)|S$, take $\lim_{s \uparrow c}$ on $E(Y|D=0,s) = E(Y^0|s)$ in (9.2)(ii), and take $\lim_{s \downarrow c}$ on $E(Y|D=0,s) = E(Y^0|s)$ in (9.2)(iii) to obtain, respectively,

$$E(Y|D=0,c^-) = E(Y^0|c^-) \quad \text{and} \quad E(Y|D=0,c^+) = E(Y^0|c^+). \qquad (9.3)$$

This implies the continuity of $E(Y|D=0,s)$ at c because $E(Y^0|s)$ is continuous at c. Therefore, $T \perp\!\!\!\perp (Y^0, Y^1)|S$ *in (9.1) can be verified with the continuity of* $E(Y|D=0,s)$ *at* c.

Specifically, do the OLS of Y on $(1, \delta, S, \delta S)$ with the subsample $D = 0$, and test for zero slope of δ as δ captures the discontinuity of $E(Y|D=0,s)$. If not rejected, then we may adopt (9.1) and extend $E(Y^1 - Y^0|S = c^+$, complier) to $E(Y^1 - Y^0|S = c^+)$, which further leads to $E(Y^1 - Y^0|s)$ for all s as in (9.2). Bertanha and Imbens also allowed for covariates in the test, so that the test can avoid being misled due to covariate differences across the subject types.

Take $\lim_{s \uparrow c}$ on $E(Y|D=1,s) = E(Y^1|s)$ in (9.2)(i), and take $\lim_{s \downarrow c}$ on $E(Y|D=1,s) = E(Y^1|s)$ in (9.2)(iv) to obtain, analogously to (9.3),

$$E(Y|D=1,c^-) = E(Y^1|c^-) \quad \text{and} \quad E(Y|D=1,c^+) = E(Y^1|c^+).$$

This implies the continuity of $E(Y|D=1,s)$ at c under the extra assumption that $E(Y^1|s)$ is continuous at c. Therefore, $T \perp\!\!\!\perp (Y^0, Y^1)|S$ can be verified also with the continuity of $E(Y|D=1,s)$, which can be done by the OLS of Y on $(1, \delta, S, \delta S)$ with the subsample $D = 1$.

10. REMAINING TOPICS

We have covered many topics for RD up to this point. Yet there still remain some topics. Here we briefly examine them in no particular order.

First, instead of looking at the mean effect, Frandsen *et al.* (2012) and Qu and Yoon (2019) examined quantile RD effects with binary treatment. Chiang and Sasaki (2019) looked at quantile RK effects with continuous treatment, and Chen *et al.* (2020) with binary treatment. Quantile effects would enrich RD and RK studies, but as pointed out in Lee (2021b) among others, there is a fundamental difficulty in quantile treatment effect: $Q_\alpha(Y^1 - Y^0) \neq Q_\alpha(Y^1) - Q_\alpha(Y^0)$ unlike $E(Y^1 - Y^0) = E(Y^1) - E(Y^0)$, where $Q_\alpha(Y)$ denotes the α quantile of Y. This problem has been addressed by Lee (2000) to an extent, but it remains difficult to overcome, thus limiting the use of quantile effects in general.

Second, we typically do asymptotic inference, i.e., the inference under $N \to \infty$, which may not be, however, appropriate, as RD uses only a local sample around the cutoff. When the local sample size is too small, instead of asymptotic inference, we may do the following 'randomization/permutation' inference. Assess how unlikely zero effect is, based on the p-value computed by comparing the actual effect estimate to the "pseudo effect estimates" obtained by reassigning randomly each subject to the local C or T groups because all subjects are 'exchangeable' under the no effect hypothesis. See Cattaneo *et al.* (2015, 2017) and Canay and Kamat (2018); see also Ganong and Jäger (2018) for RK permutation test.

Third, a break in $E(D|s)$, $E(Y|s)$ or $f_S(s)$ at a point other than c suggests something wrong. Equations (5.3) and (5.4) in Choi and Lee (2017) show simple LCR type one-sided kernel estimators for $E(Y|s)$ and $f_S(s)$ to visually locate breaks, with an empirical example in Choi and Lee (2017, pp. 1240-1241). Of course, a LLR version can be used as well. For instance, to find breaks in $E(Y|s)$ at $s \neq 0$, obtain $\hat{\rho}_0^+(s)$ along with $\hat{\rho}_1^+(s)$ for $s > 0$ minimizing

$$\sum_i \{Y_i - \rho_0^+ - \rho_1^+(S_i - s)\}^2 1[S_i \in (s, s+h)].$$

Analogously, obtain the intercept estimator $\hat{\rho}_0^-(s)$ along with $\hat{\rho}_1^-(s)$ for $s < 0$ and $1[S_i \in (s-h, s)]$. Then plot $\hat{\rho}_0^-(s)$ for $s < 0$, and $\hat{\rho}_0^+(s)$ for $s > 0$ to find breaks in $E(Y|s)$.

Fourth, RD can be applied to limited dependent variables, using a local sample. For instance, censored MLE can be applied to $Y = \max(0, Y^*)$ for a latent continuous response Y^*, using the regressors $(1, \delta, S, \delta S)$ for LLR. Berk and de Leuuw (1999) applied logit to binary Y with regressors $(1, \delta, S)$. When

Y is categorical/multinomial, convert the multinomial Y into binary responses (one binary response for each category) to apply the usual RD approach. Instead of this "multi-dimensional" approach, we may apply multinomial logit with $(1, \delta, S, \delta S)$ as the regressors, although nonparametric approaches can be devised; see Koch and Racine (2016) and Xu (2017) for more. Xu (2018) addressed discrete duration Y while allowing for censoring.

APPENDIX

Proof for $E(Y|S) = \beta_d D + m(S) \iff \beta_d = E(Y|S=0^+) - E(Y|S=0^-)$ in SRD

First, take $E(\cdot|0^+)$ and $E(\cdot|0^-)$ on $E(Y|S) = \beta_d D + m(S)$ to get, as $m(0^+) = m(0^-)$,

$$E(Y|0^+) = \beta_d + m(0^+), \quad E(Y|0^-) = m(0^-) \implies \beta_d = E(Y|0^+) - E(Y|0^-).$$

Hence '$E(Y|S) = \beta_d D + m(S)$' implies '$\beta_d = E(Y|0^+) - E(Y|0^-)$'. Second, for the reverse, define $m(S) \equiv E(Y|S) - \beta_d D$ using the local mean difference β_d, and take $E(\cdot|0^+)$ and $E(\cdot|0^-)$:

$$\begin{aligned} & m(0^+) \equiv E(Y|0^+) - \beta_d, \;\; m(0^-) \equiv E(Y|0^-) \\ \implies & m(0^+) - m(0^-) = E(Y|0^+) - E(Y|0^-) - \beta_d. \end{aligned}$$

'$\beta_d \equiv E(Y|0^+) - E(Y|0^-)$' implies $m(0^+) - m(0^-) = 0$, which is the continuity of $m(S)$ at 0, and thus $E(Y|S) = \beta_d D + m(S)$ with $m(S)$ continuous at 0 follows from the definition of $m(S)$.

Weight-Averaged Effect for Random Cutoff

Due to the continuity of $f_{C|S}(c|s)$ in s for all c and of $E(Y^0|C=c, S=s)$ in s for all c,

$$\begin{aligned} \lim_{h\downarrow 0} E(Y|S=C+h) &= \lim_{h\downarrow 0} \int E(Y^1|C=c, S=c+h) \cdot f_{C|S}(c|c+h)\partial c \\ &= \int \{\lim_{h\downarrow 0} E(Y^1|C=c, S=c+h\} \cdot f_{C|S}(c|c)\partial c; \\ \lim_{h\downarrow 0} E(Y|S=C-h) &= \int \{\lim_{h\downarrow 0} E(Y^0|C=c, S=c-h)\} \cdot f_{C|S}(c|c-h)\partial c \\ &= \int \{\lim_{h\downarrow 0} E(Y^0|C=c, S=c-h)\} f_{C|S}(c|c)\partial c \\ &= \int \{\lim_{h\downarrow 0} E(Y^0|C=c, S=c+h)\} f_{C|S}(c|c)\partial c. \end{aligned}$$

Then $\lim_{h\downarrow 0} E(Y|S=C+h) - \lim_{h\downarrow 0} E(Y|S=C-h)$ gives (2.2).

REFERENCES

Almond, D., Doyle Jr., J.J., Kowalski, A.E., and H. Williams (2010). "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns," *Quarterly Journal of Economics* 125(2), 591-634.

Almond, D., Doyle Jr., J.J., Kowalski, A.E., and H. Williams (2011). "The Role of Hospital Heterogeneity in Measuring Marginal Returns to Medical Care: a Reply to Barreca, Guldi, Lindo, and Waddell," *Quarterly Journal of Economics* 126(4), 2125-2131.

Angrist, J. (2004). "Treatment Effect Heterogeneity in Theory and Practice," *Economic Journal* 114(494), C52-C83.

Angrist, J.D. and V. Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics* 114(2), 533-575.

Angrist, J.D. and J.S. Pischke (2009). *Mostly Harmless Econometrics*, Princeton University Press

Angrist, J.D. and M. Rokkanen (2015). "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association* 110(512), 1331-1344.

Arai, Y. and H. Ichimura (2004). "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator," *Quantitative Economics* 9(1), 441-482.

Barreca, A.I., Guldi, M., Lindo, J.M., and G.R. Waddell (2011). "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification," *Quarterly Journal of Economics* 126(4), 2117-2123.

Barreca, A.I., Lindo, J.M., and G.R. Waddell (2016). "Heaping-induced Bias in Regression-discontinuity Designs," *Economic Inquiry* 54(1), 268-293.

Bartalotti, O. and Q. Brummet (2017). "Regression Discontinuity Designs with Clustered Data, in Regression Discontinuity Designs: Theory and Applications," in *Advances in Econometrics* 38, eds. M.D. Cattaneo and J.C. Escanciano, Emerald Publishing, 383-420.

Bartalotti, O., Brummet, Q., and S. Dieterle (2021). " A Correction for Regression Discontinuity Designs with Group-specific Mismeasurement of the Running Variable," *Journal of Business and Economic Statistics* 39(3), 833-848.

Battistin, E. and E. Rettore (2002). " Testing for Programme Effects in a Regression Discontinuity Design with Imperfect Compliance," *Journal of the Royal Statistical Society (Series A)* 165(1), 39-57.

Battistin, E. and E. Rettore (2008). " Ineligibles and Eligible Non-participants as a Double Comparison Group in Regression-discontinuity Designs," *Journal of Econometrics* 142(2), 715-730.

Battistin, E., Brugiavini, A., Rettore, E., and G. Weber (2009). " The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach," *American Economic Review* 99(5), 2209-2226.

Bayer, P., Ferreira, F., and R. McMillan (2007). " A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy* 115(4), 588-638.

Berk, R.A. and J. de Leuuw (1999). " A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of the American Statistical Association* 94(448), 1045-1052.

Bertanha, M. (2020). "Regression Discontinuity Design with Many Thresholds," *Journal of Econometrics* 218(1), 216-241.

Bertanha, M. and G.W. Imbens (2020). "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics,* 38(3), 593-612.

Bloom, H.S. (2012). "Modern Regression Discontinuity Analysis," *Journal of Research on Educational Effectiveness,* 5(1), 43-82.

Bugni, F.A. and I.A. Canay (2021). "Testing Continuity of a Density via G-order Statistics in the Regression Discontinuity Design," *Journal of Econometrics,* 221(1), 138-159.

Calonico, S., Cattaneo, M.D., and M.H. Farrell (2020). " Optimal Bandwidth Choice for Robust Bias-corrected Inference in Regression Discontinuity Designs," *Econometrics Journal,* 23(2), 192-210.

Calonico, S., Cattaneo, M.D., Farrell, M.H., and R. Titiunik (2019). "Regression Discontinuity Designs Using Covariates," *Review of Economics and Statistics,* 101(3), 442-451.

Calonico, S., Cattaneo, M.D., and R. Titiunik (2014). "Robust Nonparametric Confidence Intervals for Regression-discontinuity Designs," *Econometrica,* 82(6), 2295-2326.

Canay, I.A. and V. Kamat (2018). "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," *The Review of Economic Studies,* 85(3), 1577-1608.

Card, D., Mas, A., and J. Rothstein (2008). "Tipping and the Dynamics of Segregation," *Quarterly Journal of Economics,* 123(1), 177-218.

Card, D., Lee, D., Pei, Z., and A. Weber (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, D., Pei, Z., and A. Weber (2015). "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica,* 83(6), 2453-2483.

Card, D., Lee, D., Pei, Z., and A. Weber (2017). "Regression Kink Design: Theory and Practice," in *Regression Discontinuity Designs: Theory and Applications* 38, eds. M.D. Cattaneo and J.C. Escanciano, Emerald Publishing, 2453-2483.

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

Cattaneo, M.D., Frandsen, B.R., and R. Titiunik (2015). "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the US Senate," *Journal of Causal Inference* 3(1), 1-24.

Cattaneo, M.D., Idrobo, N., and R. Titiunik (2019). *A Practical Introduction to Regression Discontinuity Designs,* Cambridge University Press.

Cattaneo, M.D., Jansson, M., and X. Ma (2020). "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association* 115(531), 1449-1455.

Cattaneo, M.D., Keele, L., Titiunik, R., and G. Vazquez-Bare (2016). "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *Journal of Politics* 78(4), 1229-1248.

Cattaneo, M.D., Titiunik, R., and G. Vazquez-Bare (2017). "Comparing Inference Approaches for RD Designs: a Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management* 36(3), 643-681.

Chen, H., Chiang, H.D., and Y. Sasaki (2020). "Quantile Treatment Effects in Regression Kink Designs," *Econometric Theory* 36(6), 1167-1191.

Cheng, M.Y., Fan, J., and J.S. Marron (1997). "On Automatic Boundary Corrections," *The Annals of Statistics* 25(4), 1691-1708.

Chiang, H.D. and Y. Sasaki (2019). "Causal Inference by Quantile Regression Kink Designs," *Journal of Econometrics* 210(2), 405-433.

Choi, J.Y. and M.J. Lee (2017). "Regression Discontinuity: Review with Extensions," *Statistical Papers* 58(4), 1217-1246.

Choi, J.Y. and M.J. Lee (2018a). "Minimum Distance Estimator for Sharp Regression Discontinuity with Multiple Running Variables," *Economics Letters* 162, 10-14.

Choi, J.Y. and M.J. Lee (2018b). "Regression Discontinuity with Multiple Running Variables Allowing Partial Effects," *Political Analysis* 26(3), 258-274.

Choi, J.Y. and M.J. Lee (2018c). "Relaxing Conditions for Local Average Treatment Effect in Fuzzy Regression Discontinuity," *Economics Letters* 173, 47-50.

Choi, J.Y. and M.J. Lee (2021). "Fuzzy Multiple-score Regression Discontinuity," unpublished paper

Clark, D. and P. Martorell (2014). "The Signaling Value of a High School Diploma," *Journal of Political Economy* 122(2), 282-318.

Cook, T.D. (2008). "Waiting for Life to Arrive: a History of the Regression-discontinuity Design in Psychology, Statistics and Economics," *Journal of Econometrics* 142(2), 636-654.

Dahlberg, M., Mörk, E., Rattsø, J. and H. Ågren (2008). "Using a Discontinuous Grant Rule to Identify the Effect of Grants on Local Taxes and Spending," *Journal of Public Economics* 92(12), 2320-2335.

Davezies, L. and T. Le Barbanchon (2017). "Regression Discontinuity Design with Continuous Measurement Error in the Running Variable," *Journal of Econometrics* 200(2), 260-281.

Dell, M. (2010). "The Persistent Effects of Peru's Mining Mita," *Econometrica* 78(6), 1863-1903.

DiNardo, J.E. and D.S. Lee (2004). "Economic Impacts of New Unionization on Private Sector Employers: 1984–2001," *Quarterly Journal of Economics* 119(4), 1383-1441.

Dong, Y. (2015). "Regression Discontinuity Applications with Rounding Errors in the Running Variable," *Journal of Applied Econometrics* 30(3), 422-446.

Dong, Y. (2016). "Jump or Kink? Regression Probability Jump and Kink Design for Treatment Effect Evaluation," unpublished paper.

Dong, Y., Lee, Y., and M. Gou (2021). "Regression Discontinuity Designs with a Continuous Treatment," *Journal of the American Statistical Association* forthcoming.

Dong, Y. and A. Lewbel (2015). "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Review of Economics and Statistics* 97(5), 1081-1092.

Edmonds, E., K. Mammen, and D.L. Miller (2005). "Rearranging the Family? Income Support and Elderly Living Arrangements in a Low-income Country," *Journal of Human Resources* 40(1), 186-207.

Feir, D., T. Lemieux, and V. Marmer (2016). "Weak Identification in Fuzzy Regression Discontinuity Designs," *Journal of Business and Economic Statistics* 34(2), 185-196.

Frandsen, B.R. (2017). "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design when the Running Variable is Discrete," in *Regression Discontinuity Designs: Theory and Applications* eds. M.D. Cattaneo and J.C. Escanciano, Emerald Publishing, 281-315.

Frandsen, B.R., Frölich, M. and B. Melly (2012). "Quantile Treatment Effects in the Regression Discontinuity Design," *Journal of Econometrics* 168(2), 382-395.

Frölich, M. and N. Huber (2019). "Including Covariates in the Regression Discontinuity Design," *Journal of Business and Economic Statistics* 37(4), 736-748.

Ganong, P. and S. Jäger (2018). "A Permutation Test for the Regression Kink Design," *Journal of the American Statistical Association* 113(522), 494-504.

Gelman, A. and G. Imbens (2019). "Why High-order Polynomials Should Not Be Used in Regression Discontinuity Designs," *Journal of Business and Economic Statistics* 37(3), 447-456.

Gerard, F., M. Rokkanen, and C. Rothe (2020). "Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable," *Quantitative Economics* 11(3), 839-870.

Hahn, J., Todd, P., and W. Van der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression-discontinuity Design," *Econometrica* 69(1), 201-209.

Hansen, B. (2017). "Regression Kink with an Unknown Threshold," *Journal of Business and Economic Statistics* 35(2), 228-240.

Henderson, V.J., A. Storeygard, and D. N. Weil (2012). "Measuring Economic Growth from Outer Space," *American Economic Review* 102(2), 994-1028.

Hsu, Y.C. and S. Shen (2019). "Testing Treatment Effect Heterogeneity in Regression Discontinuity Designs," *Journal of Econometrics* 208(2), 468-486.

Hsu, Y.C. and S. Shen (2021). "Testing Monotonicity of Conditional Treatment Effects under Regression Discontinuity Designs," *Journal of Applied Econonmetrics* 36(3), 346-366.

Hullegie, P. and T.J. Klein (2010). "The Effect of Private Health Insurance on Medical Care Utilization and Self-assessed Health in Germany," *Health Economics* 19(9), 1048-1062.

Imbens, G.W. and J.D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62(2), 467-475.

Imbens, G.W. and K. Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies* 79(3), 933-959.

Imbens, G.W. and T. Lemieux (2008). "Regression Discontinuity Designs: a Guide to Practice," *Journal of Econometrics* 142(2), 615-635.

Iwasaki, K., M.J. Lee, and Y. Sawada (2019). "Verifying Reference-dependent Utility and Loss Aversion with Fukushima Nuclear-disaster Natural Experiment," *Journal of the Japanese and International Economies* 52, 78-89.

Jacob, B.A. and L. Lefgren (2004). "Remedial Education and Student Achievement: a Regression Discontinuity Analysis," *Review of Economics and Statistics* 86(1), 226-244.

Kan, K. and M.J. Lee (2018). "The Effects of Education on Fertility: Evidence from Taiwan," *Economic Inquiry* 56(1), 343-357.

Keele, L.J., S. Lorch, M. Passarella, D. Small, and R. Titiunik (2017). "An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance," in *Regression Discontinuity Designs: Theory and Applications* eds. M.D. Cattaneo and J.C. Escanciano, Emerald Publishing, 147-194

Keele, L.J. and R. Titiunik (2015). "Geographic Boundaries as Regression Discontinuities," *Political Analysis* 23(1), 127-155.

Kim, K.I. (2013). "Regression Discontinuity Design with Endogenous Covariates," *Journal of Economic Theory and Econometrics* 24(4), 320-337.

Kim, Y.S. and M.J. Lee (2016). "Regression-kink Approach for Wage Effect on Male Work Hours," *Oxford Bulletin of Economics and Statistics* 78(3), 424-442.

Koch, S. and J.S. Racine (2016). "Health Care Facility Choice and User Fee Abolition: Regression Discontinuity in a Multinomial Choice Setting," *Journal of the Royal Statistical Society (Series A)* 179(4), 927-950.

Kolesár, M. and C. Rothe (2018). "Inference in regression discontinuity designs with a discrete running variable," *American Economic Review* 108(8), 2277-2304.

Kuo, T.C. (2012). "Evaluating California Under-age Drunk Driving Laws: Endogenous Policy Lags," *Journal of Applied Econometrics* 27(7), 1100-1115.

Lalive, R. (2008). "How Do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach," *Journal of Econometrics* 142(2), 785-806.

Lee, D.S. (2008). "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics* 142(2), 675-697.

Lee, D.S. and D. Card (2008). "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics* 142(2), 655-674.

Lee, D.S. and T. Lemieux (2010). "Regression Discontinuity Designs in Economics," *Journal of Economic Literature* 48(2), 281–355.

Lee, H. and D. Wie (2020). "Legal Entitlement and Empowerment of Marriage Immigrants in Korea," *Feminist Economics* 26(3), 90-118.

Lee, M.J. (2000). "Median Treatment Effect in Randomized Trials," *Journal of the Royal Statistical Society (Series B)* 62(3), 595-604.

Lee, M.J. (2005). *Micro-econometrics for Policy, Program, and Treatment Effects,* Oxford University Press.

Lee, M.J. (2016). *Matching, Regression Discontinuity, Difference in Differences, and Beyond*, Oxford University Press.

Lee, M.J. (2017). "Regression Discontinuity with Errors in the Running Variable: Effect on Truthful Margin," *Journal of Econometric Methods* 6(1), 1-8.

Lee, M.J. (2018). "Simple Least Squares Estimator for Treatment Effects Using Propensity Score Residuals," *Biometrika* 105(1), 149-164.

Lee, M.J. (2020). "High-order Effects and External Validity in Regression Discontinuity for Policy Analysis," unpublished paper.

Lee, M.J. (2021a). *Difference in Differences and Beyond,* Parkyoung Publishing Company.

Lee, M.J. (2021b). "Finding Correct Elasticities in Log-linear and Exponential Models Allowing Heteroskedasticity," *Studies in Nonlinear Dynamics & Econometrics* 25(3), 81-91.

Lee, M.J. (2021c). "Instrument Residual Estimator for Any Response Variable with Endogenous Binary Treatment," *Journal of the Royal Statistical Society (Series B)* 83(3), 612-635.

Lee, M.J. and J.Y. Choi (2021). "Score Manipulation, Density Continuity and Intent-to-treat Effect for Regression Discontinuity," unpublished paper.

Lee, M.J. and S. Kobayashi (2001). "Proportional Treatment Effects for Count Response Panel Data: Effects of Binary Exercise on Health Care Demand," *Health Economics* 10(5), 411-428.

Lee, M.J., Shim, H.C., and S.S. Park (2021). "Regression Discontinuity with Integer Score and Non-integer Cutoff," *Korean Economic Review*, forthcoming.

Lee, M.J. and Y. Sawada (2020). "Review on Difference in Differences," *Korean Economic Review* 36, 135-173.

Leuven, E., Lindahl, M., Oosterbeek, H., and D. Webbink (2007). "The Effect of Extra Funding for Disadvantaged Pupils on Achievement," *Review of Economics and Statistics* 89(4), 721-736.

MacDonald, J.M., Klick, J., and B. Grunwald (2016). "The Effect of Private Police on Crime: Evidence from a Geographic Regression Discontinuity Design," *Journal of the Royal Statistical Society (Series A)* 179(3), 831-846.

Malamud, O. and C. Pop-Eleches (2011). "Home Computer Use and The Development of Human Capital," *Quarterly Journal of Economics* 126(2), 987-1027.

Matsudaira, J.D. (2008). "Mandatory Summer School and Student Achievement," *Journal of Econometrics* 142(2), 829-850.

McCrary, J. (2008). "Manipulation of the Running Variable in the Regression Discontinuity Design: a Density Test," *Journal of Econometrics* 142(2), 698-714.

Michalopoulos, S. and E. Papaioannou (2014). "National Institutions and Subnational Development in Africa," *Quarterly Journal of Economics* 129(1), 151-213.

Nielsen, H.S., Sorensen, T.,and C.R. Taber (2010). "Estimating the Effect of Student Aid on College Enrollment: Evidence from a Government Grant Policy Reform," *American Economic Journal: Economic Policy* 2(2), 185-215.

Önder, Y.K. and M. Shamsuddin (2019). "Heterogeneous Treatment Under Regression Discontinuity Design: Application to Female High School Enrolment," *Oxford Bulletin of Economics and Statistics* 81(4), 744-767.

Otsu, T., Xu, K.L., and Y. Matsushita (2013). "Estimation and Inference of Discontinuity in Density," *Journal of Business and Economic Statistics* 31(4), 507-524.

Papay, J.P., Murnane, R.J., and J.B. Willett (2011). "Extending the Regression Discontinuity Approach to Multiple Assignment Variables," *Journal of Econometrics* 161(2), 203-207.

Porter, J. and P. Yu (2015). "Regression Discontinuity Designs with Unknown Discontinuity Points: Testing and Estimation," *Journal of Econometrics* 189(1), 132-147.

Qu, Z. and J. Yoon (2019). "Uniform Inference on Quantile Effects Under Sharp Regression Discontinuity Designs," *Journal of Business and Economic Statistics* 37(4), 625-647.

Saez, E. (2010). "Do Taxpayers Bunch at Kink Points?," *American Economic Journal: Economic Policy* 2(3), 180-212.

Schanzenbach, D.W. (2009). "Do School Lunches Contribute to Childhood Obesity?" *Journal of Human Resources* 44(3), 684-709.

Schmieder, J.F., Wachter, T.V., and S. Bender (2012). "The Effects of Extended Unemployment Insurance over the Business Cycle: Evidence from Regression Discontinuity Estimates over 20 Years," *Quarterly Journal of Economics* 127(2), 701-752.

Shigeoka, H. (2014). "The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection," *American Economic Review* 104(7), 2152-2184.

Simonsen, M., Skipper, L., and N. Skipper (2016). "Price Sensitivity of Demand for Prescription Drugs: Exploiting a Regression Kink Design," *Quarterly Journal of Economics* 31(2), 320-337.

Thistlethwaite, D. and D. Campbell (1960). "Regression-discontinuity Analysis: an Alternative to the ex post facto Experiment," *Journal of Educational Psychology* 51(6), 309-317.

Thoemmes, F., Liao, W., and Z. Jin (2017). "The Analysis of the Regression-discontinuity Design in R," *Journal of Educational and Behavioral Statistics* 42(3), 341-360.

Turner, M.A., Haughwout, A., and W. Van der Klaauw (2014). "Land Use Regulation and Welfare," *Journal of Educational and Behavioral Statistics* 82(4), 1341-1403.

Urquiola, M. (2006). "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia," *Review of Economics and Statistics* 88(1), 171-177.

Urquiola, M. and E. Verhoogen (2009). "Class-size Caps, Sorting, and the Regression-discontinuity Design," *American Economic Review* 99(1), 179-215.

Van der Klaauw, V. (2002). "Estimating the Effect of Financial Aid Offers on College Enrollment: a Regression-discontinuity Approach," *International Economic Review* 43(4), 1249-1287.

Venkataramani, A.S., Bor, J., and A.B. Jena (2016). "Regression Discontinuity Designs in Healthcare Research," *Journal of Educational and Behavioral Statistics BMJ* 352:i1216.

Wing, C. and R.A. Bello-Gomez (2018). "Regression Discontinuity and Beyond: Options for Studying External Validity in an Internally Valid Design," *American Journal of Evaluation* 39(1), 91-108.

Wong, V.C., Steiner, P.M., and T.D. Cook (2013). "Analyzing Regression Discontinuity Designs with Multiple Assignment Variables: a Comparative Study of Four Estimation Methods," *Journal of Educational and Behavioral Statistics* 38(2), 107-141.

Xu, K.L. (2017). "Regression Discontinuity with Categorical Outcomes," *Journal of Econometrics* 201(1), 1-18.

Xu, K.L. (2018). "A Semi-nonparametric Estimator of Regression Discontinuity Design with Discrete Duration Outcomes," *Journal of Econometrics* 206(1), 258-278.