REGRESSION DISCONTINUITY

Michal Kolesár*

April 10, 2024

1. IDENTIFICATION

We're interested in the effect of a treatment D_i on an outcome Y_i. In a regression discontinuity (RD) design, the treatment is determined, either fully or partially, by the value of some variable X_i, called a running variable, crossing a threshold. Without loss of generality, normalize the threshold to 0. In the sharp RD design,

$$D_i = \mathbb{1}\{X_i \geq 0\}, \tag{1}$$

so that all units with X_i exceeding 0 are treated. For example, D_i may be an indicator for being elected, and X_i may be the margin of victory, as in Lee (2008). In the fuzzy RD, the running variable induces a jump in the treatment probability,

$$\lim_{x \downarrow 0} P(D_i = 1 \mid X_i = x) - \lim_{x \uparrow 0} P(D_i = 1 \mid X_i = x) > 0. \tag{2}$$

For instance, van der Klaauw (2002) is interested in the effect of financial aid on attending college. Since students are put into "financial aid groups", having a numerical score X_i based on the objective part of the application (SAT scores, grades) over some cutoff 0 discontinuously increases the chances of receiving aid. D_i is an indicator for receiving aid. Another nice example comes from Bleemer and Mehta (2022). Here D_i is an indicator for majoring in economics, and Y_i is earnings. The paper exploits the fact that UC Santa Cruz students can't declare an econ major if their GPA in intro econ courses (X_i) is below 2.8.

We observe $\{(Y_i, X_i, D_i)\}_{i=1}^n$. Let us assume this triple is drawn i.i.d. from some well-defined population. For any variable A_i, let $\mu_A(x) = E[A_i \mid X_i = x]$.

*Email: mkolesar@princeton.edu.

1.1. *Sharp RD*

Here the parameter of interest is the discontinuity in the regression function $\mu_Y(x) := E[Y_i \mid X_i = x]$ at the cutoff:

$$\tau_Y = \lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x).$$

The variable X_i may itself be associated with the potential outcomes, but this association is assumed to be smooth, and so any discontinuity of the conditional distribution (or of a feature of this conditional distribution such as the conditional expectation) of the outcome as a function of this covariate at the cutoff value can be interpreted as evidence of a causal effect of the treatment. In particular, assume

Assumption 1 (Continuity). $\mu_{Y(0)}(x) := E[Y_i(0) \mid X_i = x]$ and $\mu_{Y(1)}(x) := E[Y_i(1) \mid X_i = x]$ are both continuous in x at 0.

Although in theory, we only need continuity at $x = 0$, it is rare that such assumption is reasonable without having continuity at all values of x. Indeed, examining continuity of the regression function away from the cutoff is a good way of checking whether Assumption 1 is reasonable in practice, as we discuss below.

Under eq. (1) and Assumption 1, τ_Y identifies the treatment effect for individuals at the cutoff:

$$\begin{aligned}\tau_Y &\overset{(i)}{=} \lim_{x \downarrow 0} E[Y_i(1) \mid X_i = x] - \lim_{x \uparrow 0} E[Y_i(0) \mid X_i = x] \\ &\overset{(ii)}{=} E[Y_i(1) - Y_i(0) \mid X_i = 0].\end{aligned}$$

where (i) follows from eq. (1) and (ii) follows from Assumption 1. See Figure 1.

Remark 1 (Practical implication). In practice, Assumption 1 requires:

1. No perfect manipulation: individuals cannot perfectly manipulate their running variable. Imperfect manipulation is allowed, although it may create problems with estimation and inference, as we'll discuss below. So in the Lee (2008) application, it is fine if Mark Harris hires McCready to tamper with votes, so long as McCready can't do it in a way that ensures victory.

2. Nothing else happens at the cutoff except for a change in treatment status. This is a strong assumption in geography-based or spatial RDs. Age-based cutoffs also need to be treated with care: if say the cutoff is retirement age (as in, e.g., Battistin et al. 2009), one needs to be careful when say estimating the effect of retirement on an outcome of interest, as other things may also change one an individual reaches retirement age (they become eligible for Medicare, they downsize and children move out etc.).

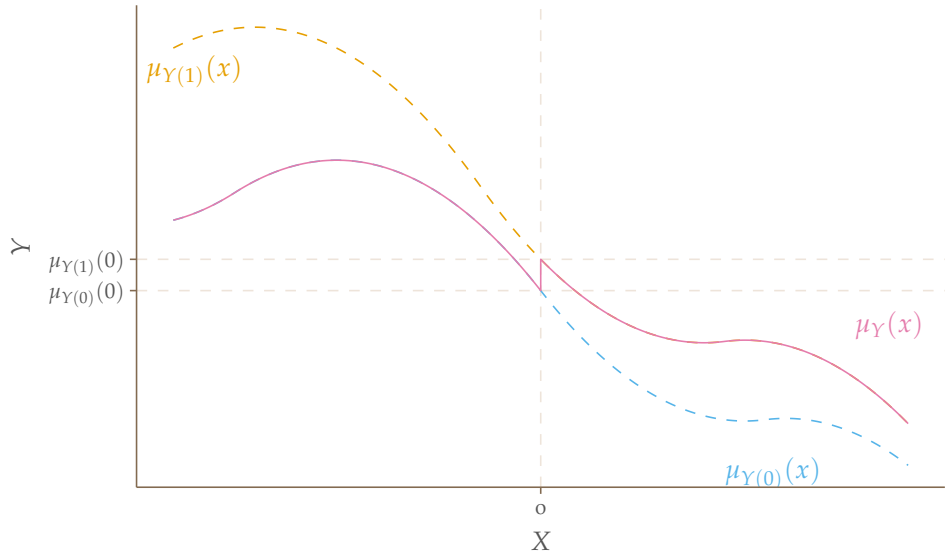

Figure 1: Regression functions for observed outcome (solid) and potential outcomes (dashed). The parameter of interest is the jump in the observed regression function at zero, which corresponds to $\tau_Y = \mu_{Y(1)}(0) - \mu_{Y(0)}(0)$.

1.2. *Fuzzy RD*

In many cases, "treatment eligibility" $\mathbb{1}\{X_i \geq 0\}$ does not perfectly determine the actual treatment: the jump in the treatment probability in eq. (2) is positive, but doesn't go all the way from zero to one. In such fuzzy cases, we scale the jump τ_Y by the size of the jump in treatment probability:

$$\theta = \frac{\lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x)}{\lim_{x \downarrow 0} \mu_D(x) - \lim_{x \uparrow 0} \mu_D(x)} =: \frac{\tau_Y}{\tau_D}.$$

where $\mu_D(x) := P(D_i = 1 \mid X_i = x)$ is the propensity score. To interpret this ratio, let $D_i(1)$ denote the potential treatment status of an individual if we were to make them eligible, and let $D_i(0)$ denote their treatment status if we were to make them ineligible. This may require the cutoff to be in principle manipulable, but that is often the case in administrative settings. The observed treatment corresponds to $D_i = D_i(\mathbb{1}\{X_i \geq 0\})$.

As in Imbens and Angrist (1994), assume a version of monotonicity:

Assumption 2 (Monotonicity). $P(D_i(1) \geq D_i(0) \mid X_i) = 1$

This says that if I were to make the individual eligible (by, say, moving the cutoff), it either has no effect on their treatment, or else induces them to take the treatment; nobody selects out of treatment.

Assumption 3 (Continuity). $\mu_{Y(d)}(x)$, $\mu_{D(d)}(x)$, and $\mu_{D(d)Y(d')}(x)$ are continuous at $x = 0$ for $d, d' \in \{0, 1\}$.

This reduces to Assumption 1 if $D_i(d) = d$.

Under eq. (2) and Assumptions 2 and 3,

$$\theta = E[Y_i(1) - Y_i(0) \mid X_i = 0, D_i(1) > D_i(0)],$$

the local average treatment effect (LATE) for individuals who are at the cutoff, and who are compliers: they select into treatment if I move the cutoff so they clear it, out of treatment if I move the cutoff so they don't clear it. This is a very local treatment effect!

Proof. Letting $\tau_i = Y_i(1) - Y_i(0)$, we have

$$\begin{aligned}
\lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x) &\overset{(1)}{=} \lim_{x \downarrow 0} E[Y_i(0) + D_i(1)\tau_i \mid X_i = x] - \lim_{x \uparrow 0} E[Y_i(0) + D_i(0)\tau_i \mid X_i = x] \\
&\overset{(2)}{=} E[(D_i(1) - D_i(0))\tau_i \mid X_i = 0] \\
&\overset{(3)}{=} E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0), X_i = 0] P(D_i(1) > D_i(0) \mid X_i = 0),
\end{aligned}$$

where (1) follows since $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$, (2) follows from Assumption 3, and (3) follows by Assumption 2. Finally, by Assumption 3, $\lim_{x \downarrow 0} \mu_D(x) - \lim_{x \uparrow 0} \mu_D(x) = P(D(1) = 1 \mid X = 0) - P(D(0) = 1 \mid X = 0)$, which equals $P(D(1) > D(0) \mid X = 0)$ by Assumption 2. Equation (2) ensures that the denominator is non-zero. □

Note that this setup is a bit different from Hahn, Todd, and van der Klaauw (2001), who were the first to give a LATE interpretation to θ. In particular, they define potential treatments $D_i(x)$ as treatment status that would obtain if $X_i = x$, which requires the running variable to be manipulable. That is typically more restrictive than requiring the cutoff to be manipulable. They also assume that $(Y_i(1) - Y_i(0))$ and $D_i(x)$ are jointly independent of X_i near 0, which effectively requires that the running variable is as good as randomly assigned near the cutoff. In contrast, the setup here only requires continuity of the conditional distribution in X_i, not full independence.

Fuzzy RD thus is like a local instrumental variables (IV) model. Implicit in Assumption 3 is an exclusion restriction that eligibility itself doesn't affect potential potential outcomes. In fact, as we'll see, estimation (though not inference) will be exactly the same as in an IV model.

1.3. *Alternative frameworks*

The framework we describe above was based on sampling form a large population, and assuming continuity of potential outcome and treatment distribution at the cutoff. There are two other frameworks that have been proposed based on randomization.

LOCAL RANDOMIZATION FRAMEWORK The first is based on the idea that close enough to the cutoff, the treatment in sharp RD can be thought of "as good as randomly assigned": that is, not only do the potential outcomes change smoothly as a function of X, they are independent of X in a small neighborhood of the cutoff. The heuristic

justification is that if units either have imperfect knowledge of the cutoff or have no ability to precisely manipulate their own score, units with scores close enough to the cutoff will have the same chance of being barely above the cutoff as barely below it.

Cattaneo, Frandsen, and Titiunik (2015) formalize this idea by letting $Y_i(D_i, X_i)$ denote potential outcomes, and assuming that for some small window $\mathcal{W}$ around the cutoff, $Y_i(D_i, X_i)$ doesn't depend on X_i (exclusion restriction), and that $Y_i(D_i) \perp\!\!\!\perp X_i \mid X_i \in \mathcal{W}$ (random assignment). Note that this sort of exclusion restriction is not necessarily satisfied even if we assumed random assignment of treatment near the cutoff. In contrast, our framework in Section 1.1 doesn't require an exclusion restriction on X_i: $Y_i(D_i)$ may depend on X_i, as long as it does so smoothly. We do not require that $\mu_{Y(0)}$ and $\mu_{Y(1)}$ are exactly flat near the cutoff.

Under this setup, one can treat the potential outcomes as fixed, and conduct randomization inference based on repeated sampling of X_i: either by using randomization tests, or by other methods for analyzing randomized experiments that may have an asymptotic justification. The key issue is how to choose the window $\mathcal{W}$: choosing it too small loses a lot of power, and if we choose it larger, the exclusion restriction will be questionable. Cattaneo, Frandsen, and Titiunik (2015) propose a heuristic window selection mechanism based on the idea that in a randomized experiment, the distribution of observed covariates has to be equal between the treated and the controls. However, formalizing this trade-off is harder than the analogous bias-variance trade-off that arises when selecting bandwidths under the framework in Section 1.1.

RANDOM CUTOFF ASSIGNMENT Another approach is to think of the cutoff c as being as good as randomly assigned, drawn from a known distribution P, with the realized cutoff equal to $c = 0$. This approach has been proposed in Ganong and Jäger (2018). We can then use finite-sample randomization tests to test the sharp null that the policy has no effect on the outcomes.

In particular, let $\hat{\tau}_c$ denote the statistic of interest, computed under the assumption that the cutoff is at c. For example, it may correspond to the estimate of a treatment effect. The actual estimate is given by $\hat{\tau}_0$. Now, under the sharp null that the treatment has no effect, the distribution of $\hat{\tau}_0$ is given, under repeated sampling of the cutoff (holding everything else fixed) by the distribution of $\hat{\tau}_C$ with $C \sim P$. Therefore, can reject the null of no effect, if, say $|\tau_0|$ exceeds the 95% quantile of this distribution, which we can simulate if we know P. The key issue is how to pick P. Ganong and Jäger (2018) propose using a uniform distribution within a small window of the cutoff (which as the issue that the realized value of c is always in the middle), a uniform distribution over $\{X_i\}_{i=1}^n$, or using institutional knowledge of how c was determined.

2. FALSIFICATION TESTS

Assumption 1 (or Assumption 3 in fuzzy RD) is not testable directly. We can, however, test the two conditions in Remark 1 that imply it.

MANIPULATION OF THE RUNNING VARIABLE To test for presence of manipulation, we can check the continuity of the density of the running variable, that is, whether $\lim_{x\downarrow 0} f(x) = \lim_{x\uparrow 0} f(x)$. One often accompanies a formal test with a plot of the density on either side of the cutoff. While continuity of the density is, strictly speaking, neither necessary nor sufficient for manipulation, it makes intuitive sense.

The original test, developed by McCrary (2008) used local polynomial estimators to estimate the density to the right and to the left of the cutoff, based on pre-binned data. Cattaneo, Jansson, and Ma (2020) modify the test to reduce the number of tuning parameters: instead of pre-binning the density, they propose running a local polynomial regression of the empirical cumulative distribution function (CDF) onto X_i. The difference in estimated slopes at the cutoff is then the estimate of the jump in the density. Otsu, Xu, and Matsushita (2013) develop a test using a local likelihood approach.

An alternative approach, studied in Bugni and Canay (2021), is to formalize the idea that under no manipulation, the running variable should be effectively randomized close to the cutoff. In particular, Bugni and Canay (2021) suggest picking q values of the running variable closest (in absolute value) to the cutoff, and count the number of positive values. Under the null, the density should be locally constant near the cutoff, so that the number of positive values should be distributed $\text{Binom}(q, 1/2)$. Under Lispchitz smoothness of the density f, if $q = o(n^{2/3})$, the bias from assuming that the density is exactly locally constant will be asymptotically negligible, and the test will control size.

Example 1. In one of the first applications of fuzzy RD in economics, Angrist and Lavy (1999) exploit the fact that following the advice of a rabbinic scholar Maimonides, class sizes in Israel are capped at 40 students, so that a grade cohort with 41 students is supposed to be split into 2 classes, while a cohort with 39 students remains in one large class. In a follow-up analysis using a more recent sample (2002–2011 vs 1991 in the original paper), Angrist et al. (2019) document clear enrollment manipulation at the Maimonides cutoffs—see Figure 2. This leads them to use predicted enrollment, rather than actual enrollment, as a running variable in their analysis. The original sample also displays similar, albeit slightly less striking evidence of sorting, as reported by Otsu, Xu, and Matsushita (2013). ⊠

COVARIATE BALANCE CHECKS This idea goes back to Lee (2008): the treatment should have no effect on pre-determined covariates. Therefore, if one estimates the effect of the treatment on some pre-determined variable Z_i (using the same estimator and confidence interval as used for estimating the effect of the treatment on the outcome of interest Y_i), the estimated treatment effect should not be significantly different from

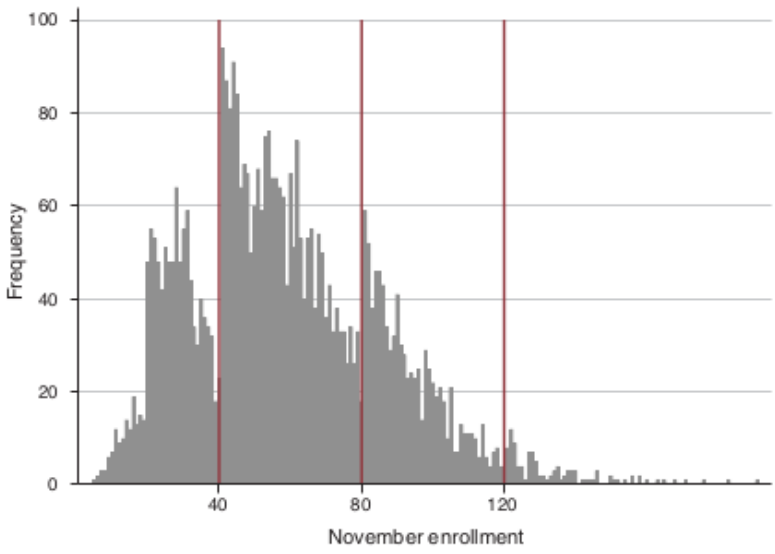

Figure 2: Figure 1 from Angrist et al. (2019). The fifth-grade enrollment distribution, as reported by school headmasters in November. Red reference lines indicate Maimonides' rule cutoffs at which an additional class is added.

zero. This test is an analog of a covariance balance check in randomized experiments.

Again, one could alternatively use the local randomization approach. In particular, Canay and Kamat (2018) propose a permutation test based on q closest observations to the cutoff. Order the covariates W_i according to the running variable, obtaining $S = (W^-_{(q)}, \ldots, W^-_{(1)}, W^+_{(1)}, \ldots, W^+_{(q)})$. Compare their empirical CDFs $\hat{F}^+(w) = \frac{1}{q}\sum_j \mathbb{1}\{W^+_{(q)} \leq w\}$ using the Cramér-von Mises test statistic

$$T(S) = \frac{1}{2q}\sum_{j=1}^{2q}\left[\hat{F}^-(S_j) - \hat{F}^+(S_j)\right]^2,$$

Compute the critical value using a permutation test by permuting the elements of S. Note that the rule of thumb for picking q proposed in the published version of the paper is not correct, one needs to pick q that grows more slowly to control the bias. Note also that although the intuition behind this test is based on a local randomization framework, the justification is asymptotic, and under the usual framework.

DISCONTINUITY AWAY FROM CUTOFF Similar to testing for discontinuity in the density, it's also a good idea to at least visually inspect that we don't observe jumps in μ_Y away from the cutoff.

3. ESTIMATION AND INFERENCE

Let us focus on sharp RD for concreteness. Statistically, the problem of estimating τ_Y just amounts to estimating a conditional mean at 0 separately for the treated and untreated subpopulations, and taking a difference.

The key issue is that since 0 is a boundary point in both regression problems, there is an unavoidable need for extrapolation, because by design there are no units with $X_i = 0$ for which we observe $Y_i(0)$, and also because typically there are too few units that are very close to the cutoff.

This is why using parametric methods, such as specifying that $\mu_{Y(1)}$ and $\mu_{Y(0)}$ are exactly polynomial of order p, or using global nonparametric methods, such as using series estimators (where, if the basis is polynomial, the only difference is that p grows with sample size), is unattractive. Global estimators in general, and global polynomial estimators in particular, may place large weight on observations far away from the cutoff in constructing the estimate $\hat{\mu}_{Y(1)}(0)$ ($\hat{\mu}_{Y(0)}(0)$), which corresponds to the intercept in the regression of Y_i on powers of X_i for the treated (untreated) units. That is, polynomial estimators can be written as

$$\hat{\tau}_Y = \sum_i w(X_i) Y_i, \tag{3}$$

where the weight $w(X_i)$ on Y_i depends on the value of the running variable X_i. The weights sum to one for observations above the cutoff, and sum to minus one for observations below the cutoff. Some of these weights are negative, unless the order of the polynomial is 0. Furthermore, the average magnitude of the weight tends to increase with the order of the polynomial, so that if p is large, some observations will receive very large weights. As a result, in such cases, small amounts of misspecification (true regression function that's not exactly polynomial, small measurement error in the outcomes) may generate large biases in the intercept estimates. This leads to estimators with large mean squared error, and confidence intervals with poor coverage. See Gelman and Imbens (2019) for a thorough discussion of this point.

A better alternative is to use local methods, which by design only place non-zero weight on observations that are near the cutoff. This can be seen as the key distinction between "parametric" and "nonparametric" thinking: while in "parametric" models, we don't worry about extrapolation bias, in "nonparametric" models, we both (i) take into account the potential extrapolation bias when choosing between different estimators; we don't just minimize variance, and (ii) we should try to account for the potential bias when conducting inference.

3.1. Standard approach to estimation

We now discuss the standard approach to estimation, based on local polynomial regression. In local polynomial regression, one picks a bandwidth h and a polynomial order

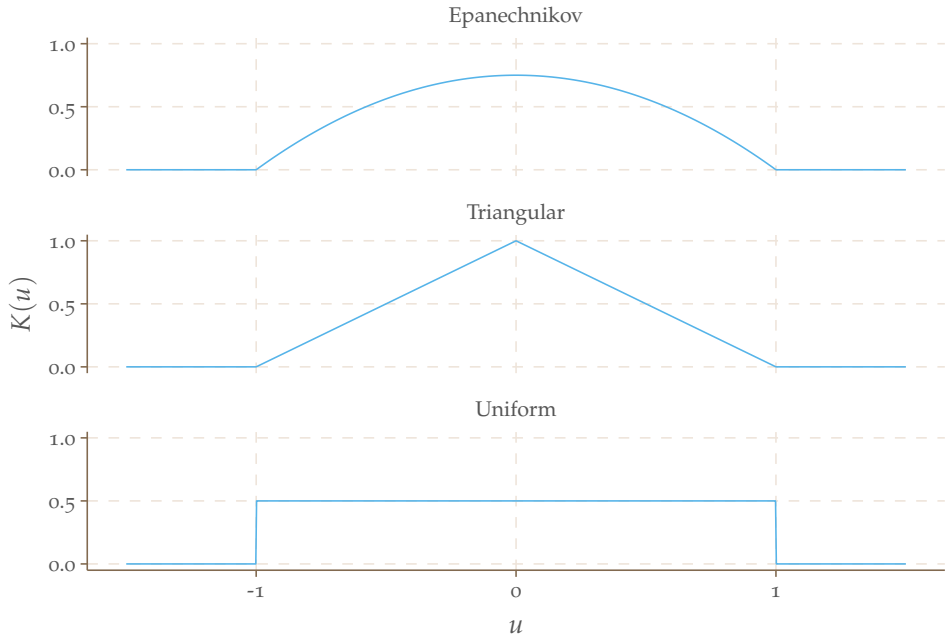

Figure 3: Uniform, Triangular, and Epanechnikov kernels

p. All observations with distance further away from the cutoff than h are discarded, and the estimates $\hat{\mu}_{Y(1)}(0)$ and $\hat{\mu}_{Y(0)}(0)$ correspond to the intercept in the regression of Y_i on powers of X_i for the treated (untreated) units, using only observations within distance h of the cutoff. If we set $h = \infty$, we get back to global polynomial regression. More generally, one may want to downweight observations that are relatively further away from the cutoff, putting weight $K(x/h)$ on observation i if $X_i = x$. Here K is a kernel function. If we use the uniform kernel $K(u) = \mathbb{1}\{|u| \leq 1\}$, then we get back to the unweighted case, placing equal weight on observations within h of the cutoff, and 0 weight on observations farther away than h. Other popular choices of kernel include the triangular kernel $K(u) = (1 - |u|)_+$, or the Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)_+$. See Figure 3.

The intercept estimate is then obtained by a weighted least squares regression of Y onto powers of X:

$$\hat{\mu}_{Y(1)}(0) = e_1' \left(\sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) m(X_i)' \right)^{-1} \sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) Y_i$$

where $m(X) = (1, x, \ldots, x^p)$, and $e_1 = (1, 0, \ldots, 0)'$ is the first unit vector. With $\hat{\mu}_{Y(0)}(0)$ defined analogously, the estimate is given by

$$\hat{\tau}_{Y,h} = \hat{\mu}_{Y(1)}(0) - \hat{\mu}_{Y(0)}(0).$$

We can compute $\hat{\tau}_{Y,h}$ all in one step as the coefficient on $D_i = 1\{X_i \geq 0\}$ in a weighted

regression of Y_i onto D_i interacted with $m(X_i)$, with weights $K(X_i/h)$.

To implement this method, one needs to pick K, p, and h. The first two are relatively easy:

1. The order of the polynomial depends on the amount of smoothness we assume. In particular, suppose that $\mu_{Y(1)}$ and $\mu_{Y(0)}$ both belong to the Hölder class of order $p+1$, that is, they are (almost everywhere) $p+1$ times differentiable, with a bounded $(p+1)$th derivative. Then it is optimal to use a polynomial of order p. In practice, $p=1$ or $p=2$ are typically used.

2. The kernel choice typically matters less. The triangular and Epanechnikov kernel are both slightly more efficient choices than the uniform (see Cheng, Fan, and Marron (1997) and Armstrong and Kolesár (2020) for formal results).

BANDWIDTH SELECTION The choice of bandwidth, in contrast, is more complicated, and more consequential. The key tradeoff is between bias and variance: while larger h leads to a lower variance of the estimate, since we're using more data, it also leads to a larger bias: unless the true regression function is exactly polynomial of order p inside the estimation window, putting weight on observations away from the cutoff will bias the estimate.

Observe that the local polynomial estimate has the form of a linear estimator, as in eq. (3), with the weights given by

$$w(x;h) = e_1' \left(\sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) m(X_i)' \right)^{-1} \mathbb{1}\{x \geq 0\} K(x/h) m(x)$$
$$- e_1' \left(\sum_i \mathbb{1}\{X_i < 0\} K(X_i/h) m(X_i) m(X_i)' \right)^{-1} \mathbb{1}\{x < 0\} K(x/h) m(x)$$

The finite-sample conditional bias is therefore given by

$$\text{bias}(\hat{\tau}_{Y,h}) = \sum_i E[w(X_i;h) Y_i \mid X] - (\mu_{Y(1)}(0) - \mu_{Y(0)}(0))$$
$$= \sum_i \mathbb{1}\{X_i \geq 0\} w(X_i)(\mu_{Y(1)}(X_i) - \mu_{Y(0)}(0)) + \sum_i \mathbb{1}\{X_i < 0\} w(X_i)(\mu_{Y(1)}(X_i) - \mu_{Y(0)}(0)),$$

where the second equality follows since for the local polynomial estimator, $\sum_i \mathbb{1}\{X_i \geq 0\} w(X_i) = -\sum_i \mathbb{1}\{X_i < 0\} w(X_i) = 1$. The variance is given by

$$\text{var}(\hat{\tau}_{Y,h} \mid X) = \sum_i w(X_i;h)^2 \sigma^2(X_i),$$

where $\sigma^2(x) = \text{var}(Y_i \mid X_i = x)$. One could get a consistent estimate of the variance (discussed below) that's consistent uniformly over all bandwidth choices considered. Therefore, if one could also get an estimate of the bias, one could estimate the mean squared error (MSE) for each bandwidth choice and pick the bandwidth that minimizes

it. Estimating the bias directly is difficult, however, so the classic approach is to use a Taylor approximation to the bias. In particular, if we Taylor-expand $\mu_{Y(d)}(x)$ around 0, then as $h \to 0$ and $n \to \infty$ (see Theorem 3.2 in Fan and Gijbels 1996):

$$\begin{aligned}\text{bias}(\hat{\tau}_{Y,h}) &= \left[C_B(p,K)\mu_{Y(1)}^{(p+1)}(0)h^{p+1} - C_B(p,K)\mu_{Y(0)}^{(p+1)}(0)h^{p+1}\right](1+o(1)),\\ &= C_B(p,K)h^{p+1}\left[\mu_{Y(1)}^{(p+1)}(0) - \mu_{Y(0)}^{(p+1)}(0)\right](1+o(1)),\end{aligned}$$

where $C_B(p,K)$ is a constant that depends only on the order of the polynomial and the kernel. One could similarly approximate the variance as $nh \to \infty$ (again, see Theorem 3.2 in Fan and Gijbels 1996):

$$\begin{aligned}\text{var}(\hat{\tau}_{Y,h} \mid X) &= \left[C_V(p,K)\frac{\sigma^2(0_+)}{f_X(0_+)nh} + C_V(p,K)\frac{\sigma^2(0_+)}{f_X(0_+)nh}\right](1+o(1))\\ &= C_V(p,K)\left[\frac{\sigma^2(0_+)+\sigma^2(0_-)}{2f_X(0)nh}\right](1+o(1)),\end{aligned}$$

where $C_V(p,K)$ is a constant that depends only on the order of the polynomial and the kernel, $\sigma^2(0_+) = \lim_{x \downarrow 0}\text{var}(Y_i \mid X_i = x)$, $\sigma^2(0_-)$ is defined similarly, $f_X(0_+)$ and $f_X(0_-)$ are the densities at 0 of the running variable for the treated and untreated, respectively, and the second line assumes that there is no jump in the density $f_X(x)$ of the running variable at 0, so that $f_X(0_+) = f_X(0_-) = 2f_X(0)$.

Then the asymptotic approximation to the MSE is given by

$$\text{AMSE}(h) = C_B(p,K)^2 h^{2(p+1)}(\mu_{Y(1)}^{(p+1)}(0) - \mu_{Y(0)}^{(p+1)}(0))^2 + C_V(p,K)\left[\frac{\sigma^2(0_+)+\sigma^2(0_-)}{2f_X(0)nh}\right],$$

which we can minimize analytically over h to yield the (pointwise) optimal bandwidth

$$h_{\text{PT}}^* = \left(\frac{C_V(p,K)}{2(p+1)C_B(p,K)^2}\frac{\sigma^2(0_+)+\sigma^2(0_-)}{2f_X(0)(\mu_{Y(1)}^{(p+1)}(0) - \mu_{Y(0)}^{(p+1)}(0))^2 \cdot n}\right)^{\frac{1}{2p+3}}. \tag{4}$$

Of course, this bandwidth is not feasible, because we do not know the variances $\sigma^2(0_+)$, $\sigma^2(0_-)$, the derivatives $\mu_{Y(1)}^{(p+1)}(0)$, $\mu_{Y(0)}^{(p+1)}(0)$, or the density $f_X(0)$. Imbens and Kalyanaraman (2012) propose a feasible version of this bandwidth based on plugging in estimates of these unknown quantities.

Notice that, so long as $\mu_{Y(1)}^{(p+1)}(0) \neq \mu_{Y(0)}^{(p+1)}(0)$, the optimal bandwidth shrinks at the rate $O(n^{-\frac{1}{2p+3}})$. This is optimal rate: it ensures that the squared bias is of the same order $O(n^{-\frac{2p+2}{2p+3}})$ as the variance. As long as we choose the bandwidth to be of this order, the resulting convergence rate of $\hat{\tau}_Y$ will be $O_p(n^{-\frac{p+1}{2p+3}})$. So, for example, if $p = 1$ (local linear regression), the estimator converges at the rate $n^{-2/5}$, slower than the parametric rate $n^{-1/2}$. But one could get closer to the optimal rate by assuming more smoothness: if one assumes a third-order Hölder class, then one can run a local quadratic regression ($p = 2$), with a faster convergence rate equal to $n^{-3/7}$ if the bandwidth is picked optimally.

3.2. *Problems with the standard approach*

There are two issues with this approach. First, its performance can be arbitrarily bad even if we use the infeasible bandwidth choice h^*_{PT}. This is because the Taylor-expansion method effectively assumes that we can approximate $\mu_{Y(d)}$ locally around zero by a polynomial of order $p+1$. This is fine as long as the bandwidth h^*_{PT} we end up choosing is not too large. But if in this approximation the $p+1$th derivative to the right and to the left of the cutoff are similar, so that $\mu^{(p+1)}_{Y(1)}(0) \approx \mu^{(p+1)}_{Y(0)}(0)$, the implied optimal bandwidth choice h^*_{PT} will be large, at which point the local polynomial approximation may become very misleading. In this case, the Taylor approximation effectively decides that close to zero, the bias of $\hat{\mu}_{Y(1)}(0)$ is similar to that of $\hat{\mu}_{Y(0)}(0)$, so that the biases cancel out, implying that we should use a large bandwidth. But while such conclusion may be accurate for bandwidths close zero, it may be quite inaccurate for large bandwidths.

To illustrate this point, consider the following example from Armstrong and Kolesár (2020). Suppose that $-\mu_{Y(0)}(x) = \mu_{Y(1)}(x) = x^{p+2}$ if p is even, or $-\mu_{Y(0)}(x) = \mu_{Y(1)}(x) = x^{p+3}$ if p is odd. Then the $(p+1)$th derivative of both $\mu_{Y(0)}$ and $\mu_{Y(1)}$ at zero is zero, implying that the optimal bandwidth is infinite. The resulting estimator is therefore a global pth order polynomial least squares estimator. Its mean squared error will be large, since this estimator is not even consistent.[1]

To address this problem, plug-in bandwidths such as the Imbens and Kalyanaraman (2012) bandwidth selector that estimate h^*_{PT} include tuning parameters to prevent the bandwidth from getting too large. However, the MSE of the resulting estimator at such functions is then determined almost entirely by these tuning parameters.

The second problem with the standard approach is that, in order to implement the plug-in method, one needs to estimate the $(p+1)$th order derivatives $\mu^{(p+1)}_{Y(1)}(0)$ and $\mu^{(p+1)}_{Y(0)}$. This is a harder problem than the original problem of estimating the intercepts $\mu_{Y(1)}(0)$ and $\mu_{Y(0)}(0)$. Formally, this shows up in the smoothness requirement on $\mu_d(x)$: we need these functions to be in the Hölder class of order $p+2$ or higher. But if we are willing to assume this higher order of smoothness, it is no longer optimal to use local polynomial regression of order p: we should be using local polynomial regression of order $p+1$! So the resulting estimator is optimal in the class of estimators (local polynomial estimators of order p), that is itself suboptimal.

3.3. *Bias-aware approach*

To prevent these issues, one can instead adapt a minimax approach: choose the bandwidth to minimize the *worst-case* mean squared error of $\hat{\tau}_Y$: this is the estimation ap-

1. To ensure consistency and finiteness of h^*_{PT}, one therefore needs to assume that $\mu^{(p+1)}_{Y(1)}(0) \neq \mu^{(p+1)}_{Y(0)}(0)$. However, the MSE at h^*_{PT} can still be arbitrarily poor whenever $\mu^{(p+1)}_{Y(1)}(x)$, and $\mu^{(p+1)}_{Y(0)}(x)$ are similar near zero, but not so globally.

proach proposed in Armstrong and Kolesár (2018). That is, minimize

$$\sup_{\mu_{Y(1)},\mu_{Y(0)}\in\mathcal{F}_{\mathrm{H},p+1}(M)} (\mathrm{bias}(\hat{\tau}_{Y,h})^2 + \mathrm{var}(\hat{\tau}_{Y,h}))$$
$$= \mathrm{var}(\hat{\tau}_{Y,h}) + \sup_{\mu_{Y(1)},\mu_{Y(0)}\in\mathcal{F}_{\mathrm{H},p+1}(M)} \mathrm{bias}(\hat{\tau}_{Y,h})^2 = \sum_i w(X_i;h)\sigma^2(X_i) +$$
$$\sup_{\mu_{Y(1)},\mu_{Y(0)}\in\mathcal{F}_{\mathrm{H},p+1}(M)} \Bigg[\sum_{i:\, X_i\geq 0} w(X_i;h)(\mu_{Y(1)}(X_i) - \mu_{Y(1)}(0))$$
$$+ \sum_{i:\, X_i<0} w(X_i;h)(\mu_{Y(0)}(X_i) - \mu_{Y(0)}(0)) \Bigg]^2,$$

where the equality follows since the variance of the estimator doesn't depend on μ, and $\mathcal{F}_{\mathrm{H},p+1}(M)$ is the Hölder class, the class of $p+1$ times differentiable functions, with the $(p+1)$th derivative bounded by a constant M (we'll discuss the choice of the curvature parameter M below). While the sup in the above display is an infinite-dimensional optimization problem, it turns out that one can solve it in closed form: see Armstrong and Kolesár (2020, Theorem B.3).

Example 2 ($p = 0$). For local constant regression ($p = 0$), the bias-maximizing function takes the form Mx. This is easy to see: the weights are simply given by $w(X_i;h) = K(X_i/h)$ for $X_i \geq 0$; since the weights are positive, we need to make $\mu_{Y(1)}(X_i) - \mu_{Y(1)}(0)$ as large as possible. Now, $\mu_{Y(1)} \in \mathcal{F}_{\mathrm{H},0}(M)$ if and only if $|\mu_{Y(1)}(x) - \mu_{Y(1)}(x')| \leq M|x - x'|$. Therefore, $\mu_{Y(1)}(X_i) - \mu_{Y(1)}(0) \leq MX_i$. The equality is sharp for all X_i if and only if $\mu_{Y(1)}(x) = a + Mx$, for an arbitrary intercept a, showing that $\mu_{Y(1)}(x) = Mx$ is indeed least favorable. The proof for $\mu_{Y(0)}(x)$ is similar. ⊠

If $p = 1$ (local linear regression), the bias-maximizing function takes the form $\mu_{Y(1)}(x) = -Mx^2/2$ and $\mu_{Y(0)}(x) = Mx^2/2$. So for $p = 1$, the worst-case MSE is given by

$$\sum_i w(X_i;h)\sigma^2(X_i) + B_{M,h}^2, \qquad B_{M,h} = -\frac{M}{2}\sum_{i=1}^n w(X_i;h)X_i^2 \,\mathrm{sign}(X_i). \tag{5}$$

We denote the minimizer by h^*_{MSE}. Computing it is not feasible, since we do not know the variance function σ^2. In practice, one can assume homoskedastic errors (in analogy to ordinary least squares (OLS)), and use a preliminary variance estimator $\hat{\sigma}^2$ to obtain a feasible version of this bandwidth.

- In contrast with the classic approach of minimizing the asymptotic approximation to the MSE, this approach doesn't rely on any asymptotic approximation (apart from the variance estimation), and doesn't require any regularization to prevent the bandwidth from getting too large.
- The approach doesn't require any assumptions on the distribution of X_i: in particular, nothing changes if the distribution of the running variable is discrete, as is often the case in practice.

To compare this resulting bandwidth to that based on the classic approach, it is useful to consider a large-sample version of the worst-case MSE criterion. If the density of the running variable f is well-behaved (which rules out discrete running variables), and $\sigma^2(X_i)$ is continuous, then (see Equation (19) in Armstrong and Kolesár 2020)

$$h^*_{\text{MSE}} = \left(\frac{C_V(p,K)}{2(p+1)\tilde{C}_B(p,K)^2} \cdot \frac{\sigma^2_+(0) + \sigma^2_-(0)}{2f_X(0) \cdot 4M^2 \cdot n} \right)^{\frac{1}{2p+3}} (1 + o_p(1)),$$

where $\tilde{C}_B(p,K)$ is a kernel constant that's slightly larger than the constant $C_B(p,K)$ in eq. (4). Comparing this expression with eq. (4), the key difference is in the term $4M^2$: rather than plugging in an estimate of the difference between the $(p+1)$th derivatives of $\mu_{Y(1)}$ and $\mu_{Y(0)}$ at 0, the bandwidth uses the priori worst-case difference, equal to $2M$ under $\mathcal{F}_{\text{H},p}(M)$. This ensures good performance of the resulting estimator simultaneously for all possible conditional means $\mu_{Y(1)}$ and $\mu_{Y(0)}$ in the parameter space $\mathcal{F}_{\text{H},p}(M)$ (i.e. uniformly over $\mathcal{F}_{\text{H},p}(M)$). Thus, unlike h^*_{PT}, this bandwidth doesn't yield poor performance in cases where the $(p+1)$th derivatives of $\mu_{Y(0)}$ and $\mu_{Y(1)}$ are similar locally to 0, but not so globally.

CHOICE OF M The key question is how to pick the curvature parameter M, as the optimal bandwidth depends on this choice. The issue is that if we pick M too conservatively, in the sense that, say, in the $p = 1$ case, the second derivatives of the functions $\mu_{Y(1)}$ and $\mu_{Y(0)}$ are much smaller than M over the support of X_i, the resulting bandwidth will be too conservative: it would be nice to be able to estimate the bound on the second derivative from the data, and use this estimate $\hat{M}$.

If one wants to conduct inference based on the estimator $\hat{\tau}_{Y,h^*_{\text{MSE}}}$, it turns out that such a data-driven rule for choosing M (or any data-driven rule) is not possible without further restrictions (see Armstrong and Kolesár 2018).

Research Question. It remains an open question how one could construct a data-driven rule if one was only interested in estimation. ⊠

This result is essentially an instance of the general issue with using pre-testing or using model selection rules, such as using cross-validation or information criteria like AIC or BIC to pick which controls to include in a regression: doing so leads to distorted confidence intervals. Here the curvature parameter M indexes the size of the model: a large M is the analog of saying that all available covariates need to be included in the model to purge omitted variables bias; a small M is the analog of saying that a small subset of them will do. Just like one needs to use institutional knowledge of the problem at hand to decide which covariates to include in a regression, ideally one uses problem-specific knowledge to select M. Analogous to reporting results based on different subsets of controls in columns of a table with regression results, one can vary the choice of M by way of sensitivity analysis.

Depending on the problem at hand, it may be difficult to translate problem-specific intuition about how close we think the regression function is to a linear function into

a statement about the curvature parameter M. In such cases, it is convenient to have a rule of thumb for selecting M using the data. Armstrong and Kolesár (2020) suggest the following rule of thumb for calibrating M, based on formalizing the heuristic that the local smoothness of μ_d is no smaller than its smoothness at large scales:

- Fit a global polynomial on either side of the cutoff, and calculate the largest second derivative of the fitted polynomial. Set M to this value.

Armstrong and Kolesár (2020) show that if the second derivative of $\mu_{Y(1)}$ and $\mu_{Y(0)}$ near zero is indeed bounded by the largest second derivative of a global polynomial approximation to $\mu_{Y(0)}$ and $\mu_{Y(1)}$, the rule of thumb will indeed consistently estimate an upper bound on the true M (other reasonable rules of thumb have been proposed—see, for instance, Imbens and Wager (2019)).

Research Question. The additional restriction justifying the above rule of thumb is a little hard to interpret—is it possible to come up with a better data-driven rule for the choice of M, based on a more interpretable restriction on μ? ⊠

Since these methods for choosing M are meant just as a rule of thumb for start of the analysis, it is a good idea to consider alternative choices by way of sensitivity analysis. Also, plotting an approximation of μ that imposes this value of M (by, say, fitting a spline inside the estimation window) can help visually assess whether this choice is reasonable.

3.4. *Inference*

Since $\hat{\tau}_Y$ is a weighted average of the outcomes (in the sense of eq. (3)), by the Lindeberg central limit theorem, the t-statistic will satisfy

$$\frac{\hat{\tau}_{Y,h} - \tau_Y}{\operatorname{var}(\hat{\tau}_{Y,h})^{1/2}} \approx \mathcal{N}\left(\frac{\operatorname{bias}(\hat{\tau}_{Y,h})}{\operatorname{var}(\hat{\tau}_{Y.h})^{1/2}}, 1\right) + o_p(1), \tag{6}$$

so long as the weights $w(X_i)$ are not too large for any given observation.

The key issue in inference is that if the bandwidth h was chosen to optimally trade of bias against variance, the bias-standard deviation ratio $b = \operatorname{bias}(\hat{\tau}_{Y,h}) / \operatorname{var}(\hat{\tau}_{Y.h})^{1/2}$ will not be approximately zero. There are two standard approaches to handle this issue.

The first is undersmoothing: calculate the optimal bandwidth (either the bandwidth h^*_{PT} with regularization to prevent the bandwidth from getting too large, or else the bandwidth h^*_{MSE}). Then use a bandwidth that is smaller than this optimal bandwidth in the sense that it shrinks to zero at a faster rate. Of course, the issue in finite samples is how to define "smaller": one can in practice justify any bandwidth choice.

The second is bias correction: try to estimate the bias of $\hat{\tau}_{Y,h}$ and subtract it off. For this to be feasible, one again needs to assume more smoothness than was initially assumed to make the local polynomial estimator of order p optimal. Furthermore, even if one had enough smoothness, the bias estimate is often noisy, and the coverage of

the resulting confidence intervals is often poor (Hall 1992). To ameliorate this issue, Calonico, Cattaneo, and Titiunik (2014) propose adjusting the variance estimator to take into account the variability of the bias estimate, which they call robust bias correction (RBC). In an important special case, when the pilot bandwidth used to estimate the bias is the same as the main bandwidth h for the local linear estimator that estimates τ_Y, this procedure amounts to running a local *quadratic* regression, but with the original bandwidth h (that was picked as to be optimal for local linear regression). As a result, one can think of this procedure as a particular way of implementing undersmoothing, with a more principled stance on the amount of undersmoothing.

Both of these approaches have the disadvantage that the resulting confidence intervals (CIs) will not be optimal under the smoothness assumptions originally used to justify the choice of the initial estimator $\hat{\tau}_{Y,h}$. Armstrong and Kolesár (2018, 2020) suggest an alternative approach that doesn't have this problem based on bounding the bias in eq. (6). In particular, although we do not know the ratio of the bias to the standard deviation, we can bound it by the ratio of the worst-case bias over $\mathcal{F}_{\mathrm{H},p}(M)$ to the standard deviation—we already calculated this worst-case bias in eq. (5): it is given by $B_{M,h}$, so that $b \leq B_{M,h}/\operatorname{var}(\hat{\tau}_{Y,h})^{1/2} =: \bar{B}$.

Thus, the t-statistic is asymptotically $\mathcal{N}(b,1)$, with $|b| \leq \bar{B}$. Since the quantiles of the absolute value $\mathcal{N}(b,1)^2$ are increasing in b, we can use the 95% percent quantile of the $|\mathcal{N}(\bar{B},1)|$ distribution as our critical value, which leads to the CI

$$\hat{\tau}_{Y,h} \pm \mathrm{cv}_\alpha(\bar{B}) \operatorname{var}(\hat{\tau}_{Y,h})^{1/2},$$

where $\mathrm{cv}_\alpha(b)$ is the α quantile of the $|\mathcal{N}(b,1)|$ distribution (equivalently, the square root of the $1-\alpha$ quantile of a χ^2 distribution with 1 degree of freedom, and non-centrality parameter b^2, which is readily available in statistical software). This CI is honest in the sense that its validity doesn't rely on undersmoothing, or any other asymptotic promises about how the bandwidth would shrink with the sample size, and it is valid uniformly over the whole parameter space $\mathcal{F}_{\mathrm{H},p}(M)$. It is also *bias-aware* in the sense that its length reflects the potential finite-sample bias of the estimator.

Remark 2 (Variance estimation). Since the estimator $\hat{\tau}_{Y,h}$ is just a weighted least squares estimator, one can use the Eicker-Huber-White (EHW) asymptotic variance estimator to estimate $\operatorname{var}(\tau_{Y,h})$: $\hat{V}_{EHW,h} = \sum_i w(X_i)^2(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i = m(X_i)'\hat{\beta}$ is the fitted value based on the local polynomial regression. This variance estimate is conservative in finite samples, for the same reasons that the EHW estimator is conservative for inference on the conditional best linear predictor (as we discussed before). Alternatively, one could use a nearest-neighbor variance estimator (Abadie and Imbens 2006; Abadie, Imbens, and Zheng 2014), replacing $(Y_i - \hat{Y}_i)^2$ with $\frac{J}{J+1}(Y_i - J^{-1}\sum_{j=1}^J Y_{j(i)})^2$, where $j(i)$ is the jth closest observation to Y_i (in terms of the distance $|X_{j(i)} - X_i|$) on the same side of the cutoff as i. Here J is a fixed small number, such as $J = 3$.

Remark 3 (Discrete running variable). This bias-aware CIs also have the advantage that they allow the running variable to be discrete, which is formally ruled out in the under-

M	Estimate	Bias	SE	95% bias-aware CI	Effective obs.	h	$\bar{L}$
0.14	5.85	0.89	1.37	$(2.69, 9.01)$	764	7.7	0.01
0.04	6.24	0.71	1.12	$(3.66, 8.81)$	1250	12.8	0.01

Table 1: Lee (2008) RD example: estimation results. Bias refers to the worst-case bias under the assumed value of M. h refers to the estimate of the optimal bandwidth that minimizes the worst-case MSE given in eq. (5). $\bar{L}$ refers to maximum leverage.

smoothing and RBC approaches. In contrast, the other popular proposal for handling discrete covariates, to cluster the errors by the running variable (Lee and Card 2008), has a serious deficiency: it may lead to confidence intervals that are *shorter* than unclustered CIs. See Kolesár and Rothe (2018) for a detailed discussion of this point.

Of course, if the *sampling design* is clustered, then it is appropriate to use clustered standard errors to estimate $\text{var}(\hat{\tau}_{Y,h})$, as discussed the lecture on OLS.

Remark 4 (Leverage). The main condition to deliver the asymptotic normality result in (6) is that the leverage of any individual observation is not too large, in the sense that $\max_i w(X_i)^2 / \sum_{j=1}^n w(X_j) \to 0$. Since our estimator is just a weighted least squares estimator, this is the same partial leverage condition we encountered when discussing asymptotic normality in the OLS lecture. Computing this leverage as a routine diagnostic is a good idea.

4. EMPIRICAL ILLUSTRATION

We use the dataset from Lee (2008). The dataset contains 6,558 observations on elections to the US House of Representatives between 1946 and 1998. The running variable $X_i \in [-100, 100]$ is the Democratic margin of victory (in percentages) in election i. The outcome variable $Y_i \in [0, 100]$ is the Democratic vote share (in percentages) in the next election. Given the inherent uncertainty in final vote counts, the party that wins is essentially randomized in elections that are decided by a narrow margin, so that the RD parameter τ_Y measures the incumbency advantage for Democrats for elections decided by a narrow margin—the impact of being the current incumbent party in a congressional district on the vote share in the next election. Figure 4 plots the averages of the raw data.

For estimation, we use $p = 1$ (local linear regression), and the triangular kernel. To determine the bandwidth, we use the Armstrong and Kolesár (2020) rule of thumb, which yields $M = 0.14$, which is driven by observatinons with $X \leq -50$ (You can see from Figure 4 that there is a lot of curvature when $X \leq -05$). If (somewhat arbitrarily) we restrict attention to the 4,900 observations within distance 50 of the cutoff, we obtain $M = 0.04$. Table 1 shows the estimation results. Note the fairly small value of the bandwidth, in spite of the low value of the second derivative, M, selected. In contrast, the Imbens and Kalyanaraman (2012) bandwidth selector picks $h = 29.4$, due to small estimates of the second derivatives near the cutoff.

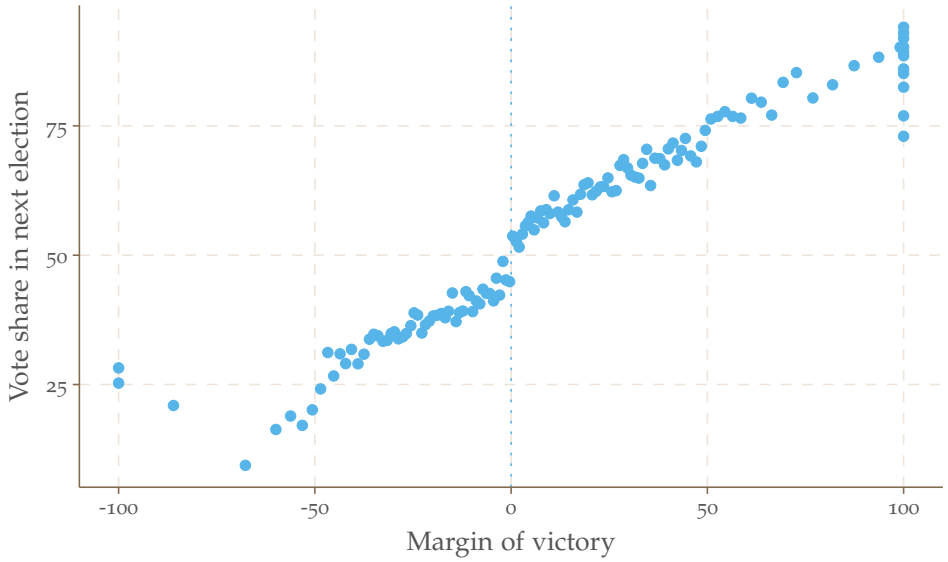

Figure 4: Lee (2008) RD example. Points correspond to 20-observation averages.

- Graphical analysis is both very useful and popular, and potentially misleading. Useful for detecting outliers or potential issues, and as a sanity check.

5. EXTENSIONS

5.1. *Covariate adjustments*

In practice, we commonly have available a vector of L pre-treatment covariates W_i. We can use such covariates in a balance check as discussed in Section 2, but are there any advantages to incorporating them directly into the local polynomial regression? In standard linear regression there are two reasons to include covariates: (i) we think that it helps to make the assumption that the treatment is as good as randomly assigned more plausible, or (ii) including covariates may help precision if they soak up variation in the regression residual. What about in RD?

Under the standard RD framework, where identification is achieved via Assumption 1, including covariates is hard to justify on "identification" grounds. That is, if we think that $E[Y_i(d) \mid X_i, W_i = w]$ is continuous in X_i, then, by iterated expectations, $\mu_{Y(d)}(x) = E[Y_i(d) \mid X_i = x] = \int E[Y_i(d) \mid X_i = x, W_i = w] f(w \mid x) dw$ will be continuous so long as the conditional density f of W_i is continuous. So, in contrast to worries about omitted variable bias in linear regression, the reason to include covariates in RD cannot be that we are worried about failure of Assumption 1, and think that it only holds conditional on covariates. In fact, since conditional expectations "smooth" non-linearities, we expect $\mu_{Y(d)}$ to be smoother than if we also condition on the covariates.

This leaves us with precision as the main reason to include covariates (actually, perhaps one could also argue that covariates help reduce the bias of the estimator). One option, explored in Frölich and Huber (2019) is to estimate the conditional treatment effects $E[Y_i(1) \mid X_i = 0, W_i] - E[Y_i(0) \mid X_i = 0, W_i]$, and then average them using the conditional distribution of the covariates given $X_i = 0$. Such strategy may be hard to implement, however, unless the covariate dimension L is very low, since it involves running kernel regression with $L + 1$ right-hand side variables. A simpler approach is to add the covariates linearly, regressing Y_i onto $m(X_i)$ and W_i for observations within an estimation window. This leads to the estimator

$$\tilde{\tau}_{Y,h} = \tilde{\beta}_{Y,h,1}, \quad \tilde{\beta}_{Y,h} = \left(\sum_{i=1}^{n} K(X_i/h) \tilde{m}(X_i, W_i) \tilde{m}(X_i, W_i)' \right)^{-1} \sum_{i=1}^{n} K(X_i/h) \tilde{m}(X_i, W_i) Y_i, \quad (7)$$

where $\tilde{m}(x, w) = (I\{x \geq 0\}, I\{x \geq 0\}x, 1, x, w)'$. Denote the coefficient on W_i in this regression by $\tilde{\gamma}_{Y,h}$; this corresponds to the last L elements of $\tilde{\beta}_{Y,h}$. As in the case without covariates, we first take the bandwidth h as given, and defer bandwidth selection choice to the end of this subsection.

Remark 5 (Interactions). It is tempting to interact the covariates with the treatment, in analogy to how covariates are included when estimating the average treatment effect (ATE) under unconfoundedness. That is, we regress Y_i onto an intercept, $m(X_i)$, $(1 - D_i)W_i$, and $D_i W_i$ for observations within an estimation window. As the window shrinks to zero, this is equivalent to the difference in intercepts when projecting Y_i onto a constant and W_i for units just above vs just below the cutoff. By standard regression results, the intercepts in these regressions are given by $E[Y_i(1) \mid X_i = 0] - \mu_W(0)\gamma_{I,+}$ and $E[Y_i(0) \mid X_i = 0] - \mu_W(0)\gamma_{I,-}$, where $\gamma_{I,+} = E[Y_i(1)(W_i - \mu_W(0)) \mid X_i = 0] / \operatorname{var}(W_i \mid X_i = 0)$ and $\gamma_{I,-} = E[Y_i(1)(W_i - \mu_W(0)) \mid X_i = 0] / \operatorname{var}(W_i \mid X_i = 0)$ are projections of the outcome on demeaned covariates. If these projections are different for $Y_i(1)$ and $Y_i(0)$, then this strategy won't yield a consistent estimator of τ_Y, as pointed out in Calonico et al. (2019). One solution to this issue is to demean the covariates, but Calonico et al. (2019) argue that because the demeaning has to be local this has a negative effect on the large-sample variance and convergence rates of the estimator.

To motivate the estimator in eq. (7), we need to formalize the assumption that the covariates are predetermined (without any assumptions on the covariates, it is optimal to ignore the covariates and use the unadjusted estimator $\hat{\tau}_{Y,h}$). Let $\mu_W(x) = E[W_i \mid X_i = x]$ denote the regression function from regressing the covariates on the running variable, and let

$$\Sigma_{WW}(x) = \operatorname{var}(W_i \mid X_i = x), \qquad \Sigma_{WY}(x) = \operatorname{cov}(W_i, Y_i \mid X_i = x).$$

We assume that the variance and covariance functions are continuous, except possibly at zero. Let $\gamma_Y = (\Sigma_{WW}(0_+) + \Sigma_{WW}(0_-))^{-1}(\Sigma_{WY}(0_+) + \Sigma_{WY}(0_-))$ denote the coefficient on W_i when we regress Y_i onto W_i for observations at the cutoff. Let $\tilde{Y}_i := Y_i - W_i'\gamma_Y$ denote the covariate-adjusted outcome. To formalize the assumption that the covari-

ates are pre-determined, we assume that $\tau_W = \lim_{x \downarrow 0} f_W(0) - \lim_{x \uparrow 0} f_W(0) = 0$, which implies that τ_Y can be identified as the jump in the covariate-adjusted outcome $\tilde{Y}_i$ at 0. Following Appendix B.1 in Armstrong and Kolesár (2018), we also assume that the covariate-adjusted outcome varies smoothly with the running variable (except for a possible jump at the cutoff), in that the second derivative of

$$\tilde{\mu}(x) := \mu_Y(x) - \mu_W(x)'\gamma_Y$$

is bounded by a known constant $\tilde{M}$. In addition, we assume μ_W has bounded second derivatives.

Under these assumptions, if γ_Y was known and hence $\tilde{Y}_i$ was directly observable, we could estimate τ as in the case without covariates, replacing M with $\tilde{M}$ and Y_i with $\tilde{Y}_i$. Furthermore, such approach would be optimal under homoskedasticity assumptions. Although γ_Y is unknown, it turns out that the estimator $\tilde{\tau}_{Y,h}$ has the same large sample behavior as the infeasible estimator $\hat{\tau}_{\tilde{Y},h}$.

Proof. To show this, note that by standard regression algebra, $\tilde{\tau}_{Y,h}$ can equivalently be written as

$$\tilde{\tau}_{Y,h} = \hat{\tau}_{Y - W'\tilde{\gamma}_{Y,h},h} = \hat{\tau}_{\tilde{Y},h} - \sum_{k=1}^{K} \hat{\tau}_{W_k,h}(\tilde{\gamma}_{Y,h,k} - \gamma_{Y,k}).$$

The first equality says that covariate-adjusted estimate is the same as an unadjusted estimate that replaces the original outcome Y_i with the covariate-adjusted outcome $Y_i - W_i'\tilde{\gamma}_{Y,h}$. The second equality uses the decomposition $Y_i - W_i'\tilde{\gamma}_{Y,h} = \tilde{Y}_i - W_i'(\tilde{\gamma}_{Y,h} - \gamma_Y)$ to write the estimator as a sum of the infeasible estimator and a linear combination of placebo RD estimators $\hat{\tau}_{W_k,h}$ that replace Y_i in the outcome equation with the kth element of W_i.

Since μ_W has bounded second derivatives, these placebo estimators converge to zero, with rate that is at least as fast as the rate of convergence of the infeasible estimator $\hat{\tau}_{\tilde{Y},h}$: $\hat{\tau}_{W_k,h} = O_p(B_{\tilde{M},h} + \mathrm{sd}(\hat{\tau}_{\tilde{Y},h}))$. Furthermore, under regularity conditions, $\tilde{\gamma}_{Y,h}$ converges to γ_Y, so that the second term in the previous display is asymptotically negligible relative to the first. □

Consequently, we can form bias-aware CIs based on $\tilde{\tau}_{Y,h}$ as in the case without covariates, treating the covariate-adjusted outcome $Y_i - W_i'\tilde{\gamma}_Y$ as the outcome,

$$\tilde{\tau}_{Y,h} \pm \mathrm{cv}_{1-\alpha}(B_{\tilde{M},h}/\operatorname{var}(\hat{\tau}_{\tilde{Y},h})^{1/2})\operatorname{var}(\hat{\tau}_{\tilde{Y},h})^{1/2}, \operatorname{var}(\hat{\tau}_{\tilde{Y},h}) = \sum_{i=1}^{n} w(X_i,h)^2\sigma^2_{\tilde{Y}}(x_i),$$

where $\sigma^2_{\tilde{Y}}(x_i) = \sigma^2_Y(x_i) + \gamma_Y'\Sigma_{WW}(x_i)\gamma_Y - 2\gamma_Y'\Sigma_{WY}(x_i)$. If the covariates are effective at explaining variation in the outcomes, then $\sum_i w(X_i,h)^2 \cdot \left(\gamma_Y'\Sigma_{WW}(x_i)\gamma_Y - 2\gamma_Y'\Sigma_{WY}(x_i)\right)$ will be negative, and $\mathrm{sd}(\hat{\tau}_{\tilde{Y},h}) \leq \mathrm{sd}(\hat{\tau}_{Y,h})$. If the smoothness of the covariate-adjusted conditional mean function $\mu_Y - \mu_W'\gamma_Y$ is greater than the smoothness of the unadjusted conditional mean function μ_Y, so that $\tilde{M} \leq M$, then using the covariates will help tighten the confidence intervals.

Implementation of covariate-adjustment requires a choice of $\tilde{M}$. We can first estimate the model without covariates (using a rule of thumb to calibrate M, the bound on the second derivative of μ_Y), and compute the bandwidth $\check{h}$ that's MSE optimal without

covariates. Based on this bandwidth, we compute a preliminary estimate $\tilde{\gamma}_{Y,\check{h}}$ of γ_Y, and use this preliminary estimate to compute a preliminary covariate-adjusted outcome $Y_i - W_i'\tilde{\gamma}_{Y,\check{h}}$. We then calibrate $\tilde{M}$ using the rule of thumb, using this preliminary covariate-adjusted outcome as the outcome.

5.2. *Other extensions*

There are a number of extensions that we don't have time to talk about. Here are some of these extensions with relevant references:

- Sharp and fuzzy regression kink designs: Card et al. (2015).
- Bunching designs: Blomquist et al. (2021) and Bertanha, McCallum, and Seegert (2023)
- Multiple cutoffs. See Bertanha (2020), Cattaneo et al. (2016).
- Cutoffs based on multi-dimensional running variable.
- Extrapolating treatment effects away from the cutoff. See Angrist and Rokkanen (2015), Dong and Lewbel (2015), Bertanha and Imbens (2020), Cattaneo et al. (2021).

REFERENCES

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74, no. 1 (January): 235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x.

Abadie, Alberto, Guido W. Imbens, and Fanyin Zheng. 2014. "Inference for Misspecified Models With Fixed Regressors." *Journal of the American Statistical Association* 109 (508): 1601–1614. https://doi.org/10.1080/01621459.2014.928218.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114, no. 2 (May): 533–575. https://doi.org/10.1162/003355399556061.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides' Rule Redux." *American Economic Review: Insights* 1, no. 3 (December): 309–324. https://doi.org/10.1257/aeri.20180120.

Angrist, Joshua D., and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff." *Journal of the American Statistical Association* 110, no. 512 (October): 1331–1344. https://doi.org/10.1080/01621459.2015.1012259.

Armstrong, Timothy B., and Michal Kolesár. 2018. "Optimal Inference in a Class of Regression Models." *Econometrica* 86, no. 2 (March): 655–683. https://doi.org/10.3982/ECTA14434.

———. 2020. "Simple and Honest Confidence Intervals in Nonparametric Regression." *Quantitative Economics* 11, no. 1 (January): 1–39. https://doi.org/10.3982/QE1199.

Battistin, Erich, Agar Brugiavini, Enrico Rettore, and Guglielmo Weber. 2009. "The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach." *American Economic Review* 99, no. 5 (December): 2209–2226. https://doi.org/10.1257/aer.99.5.2209.

Bertanha, Marinho. 2020. "Regression Discontinuity Design with Many Thresholds." *Journal of Econometrics* 218, no. 1 (September): 216–241. https://doi.org/10.1016/j.jeconom.2019.09.010.

Bertanha, Marinho, and Guido W. Imbens. 2020. "External Validity in Fuzzy Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 38, no. 3 (July): 593–612. https://doi.org/10.1080/07350015.2018.1546590.

Bertanha, Marinho, Andrew H. McCallum, and Nathan Seegert. 2023. "Better Bunching, Nicer Notching." *Journal of Econometrics* 237, no. 2 (December): 105512. https://doi.org/10.1016/j.jeconom.2023.105512.

Bleemer, Zachary, and Aashish Mehta. 2022. "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major." *American Economic Journal: Applied Economics* 14, no. 2 (April): 1–22. https://doi.org/10.1257/app.20200447.

Blomquist, Sören, Whitney K. Newey, Anil Kumar, and Che-Yuan Liang. 2021. "On Bunching and Identification of the Taxable Income Elasticity." *Journal of Political Economy* 129, no. 8 (August): 2320–2343. https://doi.org/10.1086/714446.

Bugni, Federico A., and Ivan Alexis Canay. 2021. "Testing Continuity of a Density via G-Order Statistics in the Regression Discontinuity Design." *Journal of Econometrics* 221, no. 1 (March): 138–159. https://doi.org/10.1016/j.jeconom.2020.02.004.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2019. "Regression Discontinuity Designs Using Covariates." *The Review of Economics and Statistics* 101, no. 3 (July): 442–451. https://doi.org/10.1162/rest_a_00760.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82, no. 6 (November): 2295–2326. https://doi.org/10.3982/ECTA11757.

Canay, Ivan Alexis, and Vishal Kamat. 2018. "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design." *The Review of Economic Studies* 85, no. 3 (July): 1577–1608. https://doi.org/10.1093/restud/rdx062.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica* 83, no. 6 (November): 2453–2483. https://doi.org/10.3982/ECTA11224.

Cattaneo, Matias D., Brigham R. Frandsen, and Rocío Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate." *Journal of Causal Inference* 3, no. 1 (January): 1–24. https://doi.org/10.1515/jci-2013-0010.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2020. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association* 115, no. 531 (September): 1449–1455. https://doi.org/10.1080/01621459.2019.1635480.

Cattaneo, Matias D., Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. 2016. "Interpreting Regression Discontinuity Designs with Multiple Cutoffs." *The Journal of Politics* 78, no. 4 (October): 1229–1248. https://doi.org/10.1086/686802.

———. 2021. "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs." *Journal of the American Statistical Association* 116, no. 536 (October): 1941–1952. https://doi.org/10.1080/01621459.2020.1751646.

Cheng, Ming-Yen, Jianqing Fan, and J. S. Marron. 1997. "On Automatic Boundary Corrections." *The Annals of Statistics* 25, no. 4 (August): 1691–1708. https://doi.org/10.1214/aos/1031594737.

Dong, Yingying, and Arthur Lewbel. 2015. "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models." *Review of Economics and Statistics* 97, no. 5 (December): 1081–1092. https://doi.org/10.1162/REST_a_00510.

Fan, Jianqing, and Irène Gijbels. 1996. *Local Polynomial Modelling and Its Applications.* Monographs on Statistics and Applied Probability 66. New York, NY: Chapman & Hall/CRC. https://doi.org/10.1201/9780203748725.

Frölich, Markus, and Martin Huber. 2019. "Including Covariates in the Regression Discontinuity Design." *Journal of Business & Economic Statistics* 37, no. 4 (October): 736–748. https://doi.org/10.1080/07350015.2017.1421544.

Ganong, Peter, and Simon Jäger. 2018. "A Permutation Test for the Regression Kink Design." *Journal of the American Statistical Association* 113, no. 522 (April): 494–504. https://doi.org/10.1080/01621459.2017.1328356.

Gelman, Andrew, and Guido W. Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 37, no. 3 (July): 447–456. https://doi.org/10.1080/07350015.2017.1366909.

Hahn, Jinyong, Petra Elisabeth Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69, no. 1 (January): 201–209. https://doi.org/doi.org/10.1111/1468-0262.00183.

Hall, Peter. 1992. "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density." *The Annals of Statistics* 20, no. 2 (June): 675–694. https://doi.org/10.1214/aos/1176348651.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62, no. 2 (March): 467–475. https://doi.org/10.2307/2951620.

Imbens, Guido W., and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies* 79, no. 3 (July): 933–959. https://doi.org/10.1093/restud/rdr043.

Imbens, Guido W., and Stefan Wager. 2019. "Optimized Regression Discontinuity Designs." *The Review of Economics and Statistics* 101, no. 2 (May): 264–278. https://doi.org/10.1162/rest_a_00793.

Kolesár, Michal, and Christoph Rothe. 2018. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review* 108, no. 8 (August): 2277–2304. https://doi.org/10.1257/aer.20160945.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142, no. 2 (February): 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004.

Lee, David S., and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142, no. 2 (February): 655–674. https://doi.org/10.1016/j.jeconom.2007.05.003.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142, no. 2 (February): 698–714. https://doi.org/10.1016/j.jeconom.2007.05.005.

Otsu, Taisuke, Ke-Li Xu, and Yukitoshi Matsushita. 2013. "Estimation and Inference of Discontinuity in Density." *Journal of Business & Economic Statistics* 31, no. 4 (October): 507–524. https://doi.org/10.1080/07350015.2013.818007.

van der Klaauw, Wilbert. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43, no. 4 (November): 1249–1287. https://doi.org/10.1111/1468-2354.t01-1-00055.