Chapter 6

Getting a Little Jumpy: Regression Discontinuity Designs

> But when you start exercising those rules, all sorts of processes start to happen and you start to find out all sorts of stuff about people . . . Its just a way of thinking about a problem, which lets the shape of the problem begin to emerge. The more rules, the tinier the rules, the more arbitrary they are, the better.
>
> Douglas Adams, *Mostly Harmless* (1995)

Regression discontinuity (RD) research designs exploit precise knowledge of the rules determining treatment. RD identification is based on the idea that in a highly rule-based world, some rules are arbitrary and therefore provide good experiments. RD comes in two styles, fuzzy and sharp. The sharp design can be seen as a selection-on-observables story. The fuzzy design leads to an instrumental-variables-type setup.

6.1 Sharp RD

Sharp RD is used when treatment status is a deterministic and discontinuous function of a covariate, x_i. Suppose, for example, that

$$\text{D}_i = \begin{cases} 1 & \text{if } x_i \geq x_0 \\ 0 & \text{if } x_i < x_0 \end{cases}. \tag{6.1.1}$$

where x_0 is a known threshold or cutoff. This assignment mechanism is a deterministic function of x_i because once we know x_i we know D_i. It's a discontinuous function because no matter how close x_i gets to x_0, treatment is unchanged until $x_i = x_0$.

This may seem a little abstract, so here is an example. American high school students are awarded National Merit Scholarship Awards on the basis of PSAT scores, a test taken by most college-bound high

school juniors, especially those who will later take the SAT. The question that motivated the first discussions of RD is whether students who win these awards are more likely to finish college (Thistlewaithe and Campbell, 1960; Campbell, 1969). Sharp RD compares the college completion rates of students with PSAT scores just above and just below the National Merit Award thresholds. In general, we might expect students with higher PSAT scores to be more likely to finish college, but this effect can be controlled by fitting a regression to the relationship between college completion and PSAT scores, at least in the neighborhood of the award cutoff. In this example, jumps in the relationship between PSAT scores and college attendance in the neighborhood of the award threshold are taken as evidence of a treatment effect. It is this jump in regression lines that gives RD its name.[1]

An interesting and important feature of RD, highlighted in a recent survey of RD by Imbens and Lemieux (2008), is that there is *no* value of x_i at which we get to observe both treatment and control observations. Unlike full-covariate matching strategies, which are based on treatment-control comparisons conditional on covariate values where there is some overlap, the validity of RD turns on our willingness to extrapolate across covariate values, at least in a neighborhood of the discontinuity. This is one reason why sharp RD is usually seen as distinct from other control strategies. For this same reason, we typically cannot afford to be as agnostic about regression functional form in the RD world as in the world of Chapter 3.

Figure 6.1.1 illustrates a hypothetical RD scenario where those with $x_i \geq 0.5$ are treated. In Panel A, the trend relationship between Y_i and x_i is linear, while in Panel B, it's nonlinear. In both cases, there is a discontinuity in the relation between $E[\text{Y}_{0i}|x_i]$ and x_i around the point x_0.

A simple model formalizes the RD idea. Suppose that in addition to the assignment mechanism, (6.1.1), potential outcomes can be described by a linear, constant-effects model

$$\begin{aligned} E[\text{Y}_{0i}|x_i] &= \alpha + \beta x_i \\ \text{Y}_{1i} &= \text{Y}_{0i} + \rho \end{aligned}$$

This leads to the regression,

$$\text{Y}_i = \alpha + \beta x_i + \rho \text{D}_i + \eta_i, \tag{6.1.2}$$

where ρ is the causal effect of interest. The key difference between this regression and others we've used to estimate treatment effects (e.g., in Chapter 3) is that D_i, the regressor of interest, is not only correlated with x_i, it is a deterministic function of x_i. RD captures causal effects by distinguishing the nonlinear and discontinuous function, $1(x_i \geq x_0)$, from the smooth and (in this case) linear function, x_i.

[1] The basic structure of RD designs appears to have emerged simultaneously in a number of disciplines but has only recently become important in applied econometrics. Cook (2008) gives an intellectual history. In a recent paper using Lalonde (1986) style within-study comparisons, Cook and Wong (2008) find that RD generally does a good job of reproducing the results from randomized trials.

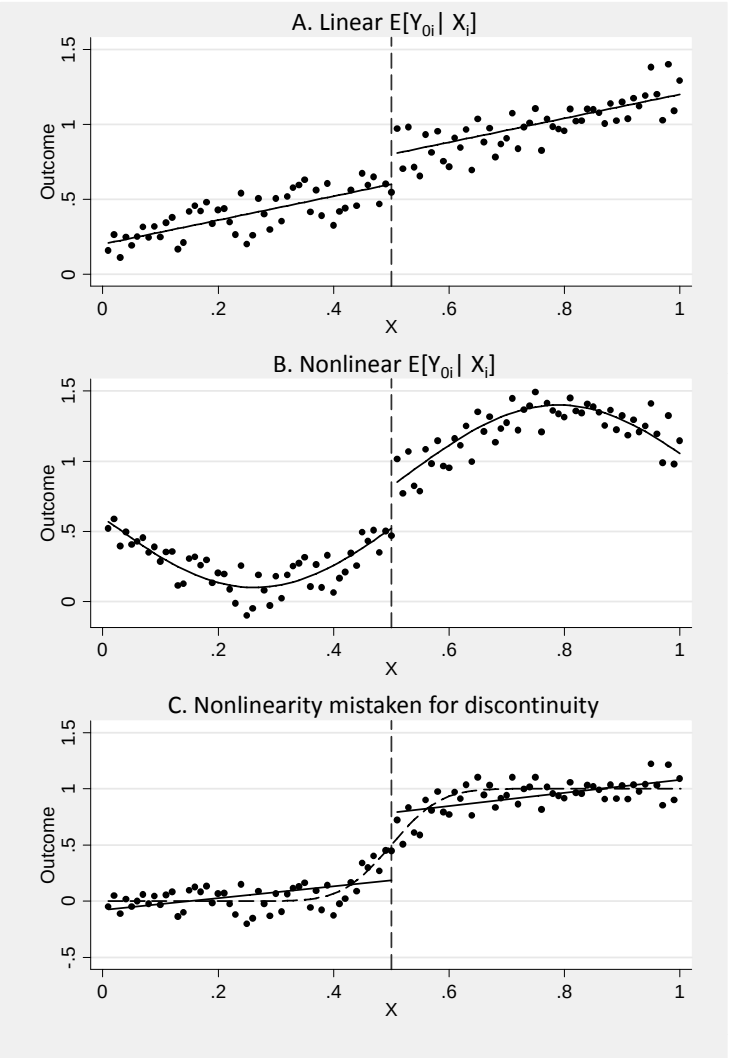

Figure 6.1.1: The sharp regression discontinuity design

But what if the trend relation, $E[\text{Y}_{0i}|x_i]$, is nonlinear? To be precise, suppose that $E[\text{Y}_{0i}|x_i] = f(x_i)$ for some reasonably smooth function, $f(x_i)$. Panel B in Figure 6.1.1 suggests there is still hope even in this more general case. Now we can construct RD estimates by fitting

$$\text{Y}_i = f(x_i) + \rho \text{D}_i + \eta_i, \tag{6.1.3}$$

where again, $\text{D}_i = 1(x_i \geq x_0)$ is discontinuous in x_i at x_0. As long as $f(x_i)$ is continuous in a neighborhood of x_0, it should be possible to estimate a model like (6.1.3), even with a flexible functional form for $f(x_i)$. For example, modeling $f(x_i)$ with a p^{th}-order polynomial, RD estimates can be constructed from the regression

$$\text{Y}_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \rho \text{D}_i + \eta_i. \tag{6.1.4}$$

A generalization of RD based on (6.1.4) allows different trend functions for $E[\text{Y}_{0i}|x_i]$ and $E[\text{Y}_{1i}|x_i]$. Modeling both of these CEFs with p^{th}-order polynomials, we have

$$\begin{aligned} E[\text{Y}_{0i}|x_i] &= f_0(x_i) = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + ... + \beta_{0p}\tilde{x}_i^p \\ E[\text{Y}_{1i}|x_i] &= f_1(x_i) = \alpha + \rho + \beta_{11}\tilde{x}_i + \beta_{12}\tilde{x}_i{}^2 + ... + \beta_{1p}\tilde{x}_i{}^p, \end{aligned}$$

where $\tilde{x}_i \equiv x_i - x_0$. Centering x_i at x_0 is just a normalization; it ensures that the treatment effect at $x_i = x_0$ is still the coefficient on D_i in the regression model with interactions.

To derive a regression model that can be used to estimate the effects interest in this case, we use the fact that D_i is a deterministic function of x_i to write

$$E[\text{Y}_i|x_i] = E[\text{Y}_{0i}|x_i] + E[\text{Y}_{1i} - \text{Y}_{0i}|x_i]\text{D}_i.$$

Substituting polynomials for conditional expectations, we then have

$$\begin{aligned} \text{Y}_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + ... + \beta_{0p}\tilde{x}_i^p \\ &\quad + \rho \text{D}_i + \beta_1^* \text{D}_i\tilde{x}_i + \beta_2^* \text{D}_i\tilde{x}_i{}^2 + ... + \beta_p^* \text{D}_i\tilde{x}_i{}^p + \eta_i, \end{aligned} \tag{6.1.6}$$

where $\beta_1^* = \beta_{11} - \beta_{01}$, $\beta_2^* = \beta_{12} - \beta_{02}$, and $\beta_p^* = \beta_{1p} - \beta_{0p}$ and the error term, η_i, is the CEF residual.

Equation (6.1.4) is a special case of (6.1.6) where $\beta_1^* = \beta_2^* = \beta_p^* = 0$. In the more general model, the treatment effect at $x_i - x_0 = c > 0$ is $\rho + \beta_1^* c + \beta_2^* c^2 + ... + \beta_p^* c^p$, while the treatment effect at x_0 is ρ. The model with interactions has the attraction that it imposes no restrictions on the underlying conditional mean functions But in our experience, RD estimates of ρ based on the simpler model, (6.1.4), usually turn out to be similar to those based on (6.1.6).

The validity of RD estimates based on (6.1.4) or (6.1.6) turns on whether polynomial models provide an adequate description of $E[\mathrm{Y}_{0i}|\mathrm{X}_i]$. If not, then what looks like a jump due to treatment might simply be an unaccounted-for nonlinearity in the counterfactual conditional mean function. This possibility is illustrated in Panel C of Figure 6.1.1, which shows how a sharp turn in $E[\mathrm{Y}_{0i}|x_i]$ might be mistaken for a jump from one regression line to another. To reduce the likelihood of such mistakes, we can look only at data in a neighborhood around the discontinuity, say the interval $[x_0 - \delta, x_0 + \delta]$ for some small number δ. Then we have

$$\begin{aligned} E\left[\mathrm{Y}_i|x_0 - \delta < x_i < x_0\right] &\simeq E[\mathrm{Y}_{0i}|x_i = x_0] \\ E\left[\mathrm{Y}_i|x_0 < x_i < x_0 + \delta\right] &\simeq E[\mathrm{Y}_{1i}|x_i = x_0], \end{aligned}$$

so that

$$\lim_{\delta \to 0} E\left[\mathrm{Y}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\mathrm{Y}_i|x_0 - \delta < x_i < x_0\right] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|x_i = x_0]. \tag{6.1.7}$$

In other words, comparisons of average outcomes in a small enough neighborhood to the left and right of x_0 should provide an estimate of the treatment effect that does not depend on the correct specification of a model for $E[\mathrm{Y}_{0i}|x_i]$. Moreover, the validity of this nonparametric estimation strategy does not turn on the constant effects assumption, $\mathrm{Y}_{1i} - \mathrm{Y}_{0i} = \rho$; the estimand in (6.1.7) is the average causal effect, $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|x_i = x_0]$.

The nonparametric approach to RD requires good estimates of the mean of Y_i in small neighborhoods to the right and left of x_0. Obtaining such estimates is tricky. The first problem is that working in a small neighborhood of the cutoff means that you don't have much data. Also, the sample average is biased for the population average in the neighborhood of a boundary (in this case, x_0). Solutions to these problems include the use of a non-parametric version of regression called local linear regression (Hahn, Todd, and van der Klaauw, 2001) and the partial-linear and local-polynomial estimators developed by Porter (2003). Local linear regression amounts to weighted least squares estimation of an equation like (6.1.6), with linear terms only and more weight given to points close to the cutoff.

Sophisticated nonparametric RD methods have not yet found wide application in empirical practice; most applied RD work is still parametric. But the idea of focusing on observations near the cutoff value - what Angrist and Lavy (1999) call a "discontinuity sample" - suggests a valuable robustness check: Although RD estimates get less precise as the window used to select a discontinuity sample gets smaller, the number of polynomial terms needed to model $f(x_i)$ should go down. Hopefully, as you zero in on x_0 with fewer and fewer controls, the estimated effect of D_i remains stable.[2] A second important check looks at the behavior of

[2] Hoxby (2000) also uses this idea to check RD estimates of class size effects. A fully nonparametric approach requires data-driven rules for selection of the width of the discontinuity-sample window, also known as "bandwidth". The bandwidth must shrink with the sample size at a rate sufficiently slow so as to ensure consistent estimation of the underlying conditional mean functions. See Imbens and Lemieux (2007) for details. We prefer to think of estimation using (6.1.4) or (6.1.6) as essentially parametric: in any given sample, the estimates are only as good as the model for $E[\mathrm{Y}_{0i}|x_i]$ that you happen to be

pre-treatment variables near the discontinuity. Since pre-treatment variables are unaffected by treatment, there should be no jump in the CEF of these variables at x_0.

Lee's (2008) study of the effect of party incumbency on re-election probabilities illustrates the sharp RD design. Lee is interested in whether the Democratic candidate for a seat in the U.S. House of Representatives has an advantage if his party won the seat last time. The widely-noted success of House incumbents raises the question of whether representatives use the privileges and resources of their office to gain advantage for themselves or their parties. This conjecture sounds plausible, but the success of incumbents need not reflect a real electoral advantage. Incumbents - by definition, candidates and parties who have shown they can win - may simply be better at satisfying voters or getting the vote out.

To capture the causal effect of incumbency, Lee looks at the likelihood a Democratic candidate wins as a function of relative vote shares in the previous election. Specifically, he exploits the fact that an election winner is determined by $\text{D}_i = 1(x_i \geq .0)$, where x_i is the *vote share margin of victory* (e.g., the difference between the Democratic and Republican vote shares when these are the two largest parties). Note that, because D_i is a deterministic function of x_i, there are no confounding variables other than x_i. This is a signal feature of the RD setup.

Figure 6.1.2a, from Lee (2008), shows the sharp RD design in action. This figure plots the probability a Democrat wins against the difference between Democratic and Republican vote shares in the previous election. The dots in the figure are local averages (the average win rate in non-overlapping windows of share margins that are .005 wide); the lines in the figure are fitted values from a parametric model with a discontinuity at zero.[3] The probability of a democratic win is an increasing function of past vote share. The most important feature of the plot is the dramatic jump in win rates at the 0 percent mark, the point where a Democratic candidate gets more votes. Based on the size of the jump, incumbency appears to raise party re-election probabilities by about 40 percentage points.

Figure 6.1.2b checks the sharp RD identification assumptions by looking at Democratic victories *before* the last election. Democratic win rates in older elections should be unrelated to the cutoff in the last election, a specification check that works out well and increases our confidence in the RD design in this case. Lee's investigation of pre-treatment victories is a version of the idea that covariates should be balanced by treatment status in a (quasi-) randomized trial. A related check examines the density of x_i around the discontinuity, looking for bunching in the distribution of x_i near x_0. The concern here is that individuals with a stake in D_i might try to manipulate x_i near the cutoff, in which case observations on either side may not be comparable (McCrary 2008 proposes a formal test for this). Until recently, we would have said this is unlikely in election studies like Lee's. But the recount in Florida after the 2000 presidential election suggests we probably should worry about manipulable vote shares when U.S. elections are close.

---

using. Promises about how you might change the model if you had more data should be irrelevant.

[3] The fitted values in this figure are from a Logit model for the probability of winning as a function of the cutoff indicator $\text{D}_i = 1(x_i \geq 0)$, a 4^{th}-order polynomial in x_i, and interactions between the polynomial terms and D_i.

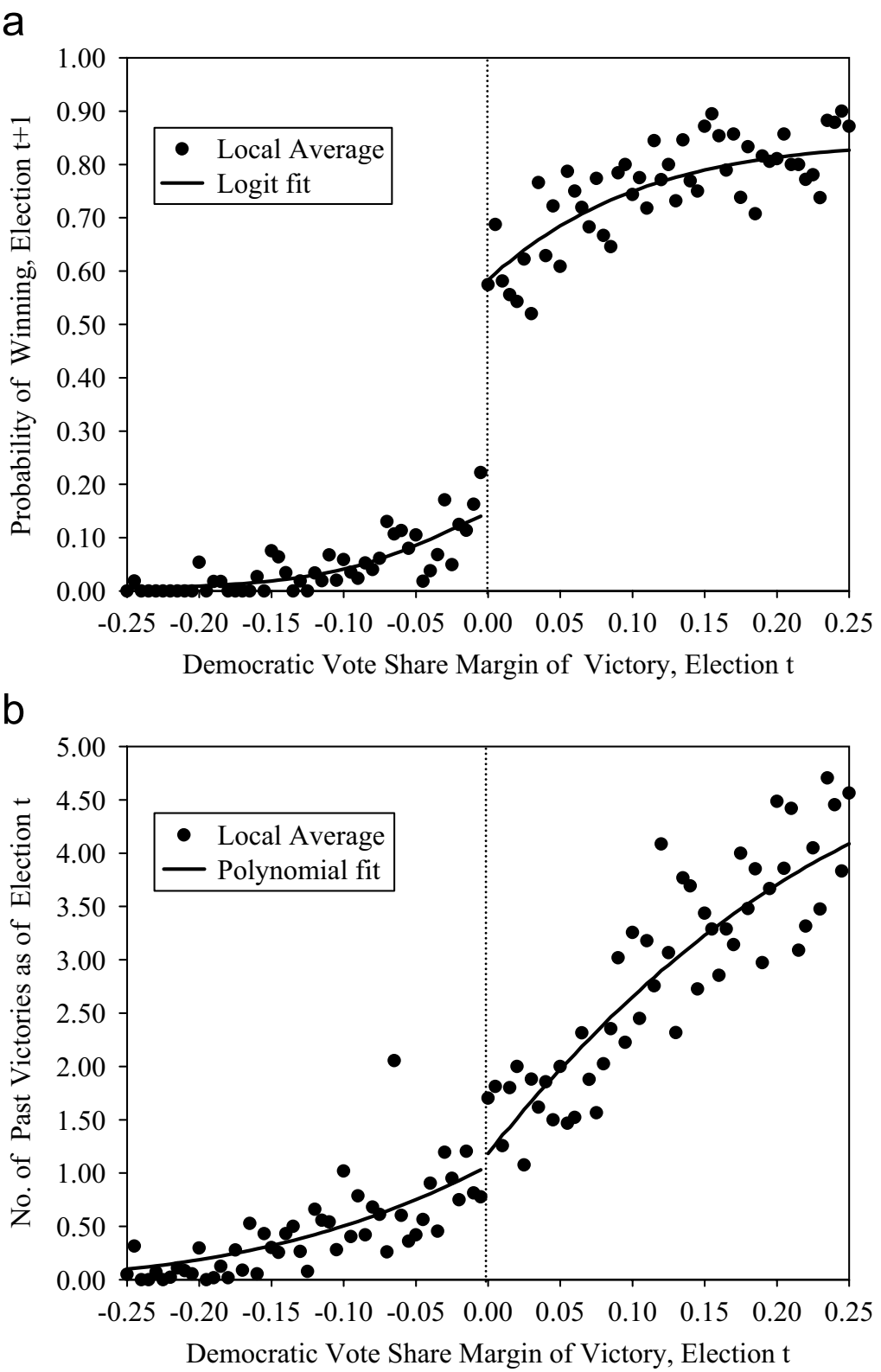

Figure 6.1.2: Probability of winning an election by past and future vote share (from Lee, 2008). (a) Candidate's probability of winning election $t+1$, by margin of victory in election t: local averages and parametric fit. (b) Candidate's accumulated number of past election victories, by margin of victory in election t: local averages and parametric fit.

6.2 Fuzzy RD is IV

Fuzzy RD exploits discontinuities in the *probability or expected value* of treatment conditional on a covariate. The result is a research design where the discontinuity becomes an instrumental variable for treatment status instead of deterministically switching treatment on or off. To see how this works, let D_i denote the treatment as before, though here D_i is no longer deterministically related to the threshold-crossing rule, $x_i \geq x_0$. Rather, there is a jump in the probability of treatment at x_0, so that

$$P[\text{D}_i = 1|x_i] = \begin{cases} g_0(x_i) & \text{if } x_i \geq x_0 \\ g_{1(}x_i) & \text{if } x_i < x_0 \end{cases}, \text{ where } g_1(x_0) \neq g_0(x_0).$$

The functions $g_0(x_i)$ and $g_1(x_i)$ can be anything as long as they differ (and the more the better) at x_0. We'll assume $g_1(x_0) > g_0(x_0)$, so $x_i \geq x_0$ makes treatment more likely. We can write the relation between the probability of treatment and x_i as

$$E\left[\text{D}_i|x_i\right] = P\left[\text{D}_i = 1|x_i\right] = g_0(x_i) + [g_1(x_i) - g_0(x_i)]\text{T}_i,$$

where

$$\text{T}_i = 1(x_i \geq x_0).$$

The dummy variable T_i indicates the point of discontinuity in $E\left[\text{D}_i|x_i\right]$.

Fuzzy RD leads naturally to a simple 2SLS estimation strategy. Assuming that $g_0(x_i)$ and $g_1(x_i)$ can be described by *pth*-order polynomials as in (6.1.4), we have

$$\begin{aligned} E\left[\text{D}_i|x_i\right] &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + ... + \gamma_{0p}x_i^p \\ &\quad + [\gamma_0^* + \gamma_1^* x_i + \gamma_2^* x_i^2 + ... + \gamma_p^* x_i^p]\text{T}_i \\ &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + ... + \gamma_{0p}x_i^p \\ &\quad + \gamma_0^*\text{T}_i + \gamma_1^* x_i\text{T}_i + \gamma_2^* x_i^2\text{T}_i + ... + \gamma_p^* x_i^p\text{T}_i. \end{aligned} \tag{6.2.1}$$

From this we see that T_i, as well as the interaction terms $\{x_i\text{T}_i, x_i^2\text{T}_i, \ldots x_i^p\text{T}_i\}$ can be used as instruments for D_i in (6.1.4).[4]

The simplest fuzzy RD estimator uses only T_i as an instrument, without the interaction terms (with the

[4] The idea of using jumps in the probability of assignment as a source of identifying information appears to originate with Trochim (1984), although the IV interpretation came later. Not everyone agrees that fuzzy RD is IV, but this view is catching on. In a recent history of the RD idea, Cook (2008) writes about the fuzzy design: "In many contexts, the cutoff value can function as an IV and engender unbiased causal conclusions . . . fuzzy assignment does not seem as serious a problem today as earlier."

interaction terms in the instrument list, we might also like to allow for interactions in the second stage as in 6.1.6). The resulting just-identified IV estimator has the virtues of transparency and good finite-sample properties. The first stage in this case is

$$\mathrm{D}_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + ... + \gamma_p x_i^p + \pi \mathrm{T}_i + \xi_{1i}, \tag{6.2.2}$$

where T_i is the excluded instrument that provides identifying power with a first-stage effect given by π.

The fuzzy RD reduced form is obtained by substituting (6.2.2) into (6.1.4):

$$\mathrm{Y}_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + ... + \kappa_p x_i^p + \rho\pi \mathrm{T}_i + \xi_{2i}, \tag{6.2.3}$$

where $\mu = \alpha + \rho\gamma_0$ and $\kappa_j = \beta_1 + \rho\gamma_j$ for $j = 1, ..., p$. As with sharp RD, identification in the fuzzy case turns on the ability to distinguish the relation between Y_i and the discontinuous function, $\mathrm{T}_i = 1(x_i \geq x_0)$, from the effect of polynomial controls included in the first and second stage. In one of the first RD studies in applied econometrics, van der Klaauw (2002) used a fuzzy design to evaluate the effects of university financial aid awards on college enrollment. In van der Klaauw's study, D_i is the size of the financial aid award offer, and T_i is a dummy variable indicating applicants with an ability index above pre-determined award-threshold cutoffs.[5]

Fuzzy RD estimates with treatment effects that change as a function of x_i can be constructed by 2SLS estimation of an equation with treatment-covariate interactions. Here, the second stage model with interaction terms is the same as (6.1.6), while the first stage is similar to (6.2.1), except that to match the second-stage parametrization, we center polynomial terms at x_0. In this case, the excluded instruments are $\{\mathrm{T}_i, \tilde{x}_i\mathrm{T}_i, \tilde{x}_i^2\mathrm{T}_i, \ldots \tilde{x}_i^p\mathrm{T}_i\}$ while the variables $\{\mathrm{D}_i, \tilde{x}_i\mathrm{D}_i, \mathrm{D}_i\tilde{x}_i{}^2, \ldots \mathrm{D}_i\tilde{x}_i{}^p\}$ are treated as endogenous. The first stage for D_i becomes

$$\begin{aligned} \mathrm{D}_i \quad = \quad & \gamma_{00} + \gamma_{01}\tilde{x}_i + \gamma_{02}\tilde{x}_i^2 + ... + \gamma_{0p}\tilde{x}_i^p \\ & + \gamma_0^*\mathrm{T}_i + \gamma_1^*\tilde{x}_i\mathrm{T}_i + \gamma_2^*\tilde{x}_i^2\mathrm{T}_i + ... + \gamma_p^*\tilde{x}_i^p\mathrm{T}_i + \xi_{1i}. \end{aligned} \tag{6.2.4}$$

An analogous first stage is constructed for each of the polynomial interaction terms in the set $\{\tilde{x}_i\mathrm{D}_i, \mathrm{D}_i\tilde{x}_i{}^2, \ldots \mathrm{D}_i\tilde{x}_i{}^p\}$.[6]

The nonparametric version of fuzzy RD consists of IV estimation in a small neighborhood around the discontinuity. The reduced-form conditional expectation of Y_i near x_0 is

[5] van der Klaauw's original working paper circulated in 1997. Note that the fact that (6.2.2) is only an approximation of $E[\mathrm{D}_i|x_i]$ is not very important; second-stage estimates are still consistent.

[6] Alternately, center neither the first or second stage. In this case, however, ρ no longer captures the treatment effect at the cutoff.

$$E\left[\text{Y}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{Y}_i|x_0 - \delta < x_i < x_0\right] \simeq \rho\gamma_0^*.$$

Similarly, for the first stage for D_i, we have

$$E\left[\text{D}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{D}_i|x_0 - \delta < x_i < x_0\right] \simeq \gamma_0^*.$$

Therefore

$$\lim_{\delta \to 0} \frac{E\left[\text{Y}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{Y}_i|x_0 - \delta < x_i < x_0\right]}{E\left[\text{D}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{D}_i|x_0 - \delta < x_i < x_0\right]} = \rho. \tag{6.2.5}$$

The sample analog of (6.2.5) is a Wald estimator of the sort discussed in Section **??**, in this case using T_i as an instrument for D_i in a $\delta-$neighborhood of x_0. As with other dummy-variable instruments, the result is a local average treatment effect. In particular, the Wald estimand for fuzzy RD captures the causal effect on compliers defined as individuals whose treatment status changes as we move the value of x_i from just to the left of x_0 to just to the right of x_0. This interpretation of fuzzy RD was introduced by Hahn, Todd, and van der Klaauw (2001). Note, however, that there is another sense in which this version of LATE is local: the estimates are for compliers with $x_i = x_0$, a feature of sharp nonparametric estimates as well.

Finally, note that as with the nonparametric version of sharp RD, the finite-sample behavior of the sample analog of (6.2.5) is not likely to be very good. Hahn, Todd, and van der Klaauw (2001) develop a nonparametric IV procedure using local linear regression to estimate the top and bottom of the Wald estimator with less bias. This takes us back to a 2SLS model with linear or polynomial controls, but the model is fit in a discontinuity sample using a data-driven bandwidth. The idea of using discontinuity samples informally also applies in this context: start with a parametric 2SLS setup in the full sample, say, based on (6.1.4). Then restrict the sample to points near the discontinuity and get rid of most or all of the polynomial controls. Ideally, 2SLS estimates in the discontinuity samples with few controls will be broadly consistent with the more precise estimates constructed using the larger sample.

Angrist and Lavy (1999)use a fuzzy RD research design to estimate the effects of class size on children's test scores, the same question addressed by the STAR experiment discussed in Chapter 2. Fuzzy RD is an especially powerful and flexible research design, a fact highlighted by the Angrist and Lavy study, which generalizes fuzzy RD in two ways relative to the discussion above. First, the causal variable of interest, class size, takes on many values. So the first stage exploits jumps in *average* class size instead of probabilities. Second, the Angrist and Lavy (1999) research design uses multiple discontinuities.

The Angrist and Lavy study begins with the observation that class size in Israeli schools is capped at 40. Students in a grade with up to 40 students can expect to be in classes as large as 40, but grades with 41 students are split into two classes, grades with 81 students are split into three classes, and so on. Angrist

and Lavy call this "Maimonides Rule" since a maximum class size of 40 was first proposed by the medieval Talmudic scholar Maimonides. To formalize Maimonides Rule, let m_{sc} denote the predicted class size (in a given grade) assigned to class c in school s, where enrollment in the grade is denoted e_s. Assuming grade cohorts are split up into classes of equal size, the predicted class size that results from a strict application of Maimonides' Rule is

$$m_{sc} = \frac{e_s}{int[\frac{(e_s-1)}{40}] + 1}$$

where $int(x)$ is the integer part of a real number, x. This function, plotted with dotted lines in Figure 6.2.1 for fourth and fifth graders, has a sawtooth pattern with discontinuities (in this case, sharp drops in predicted class size) at integer multiples of 40. At the same time, m_{sc} is clearly an increasing function of enrollment, e_s, making the enrollment variable an important control.

Angrist and Lavy exploit the discontinuities in Maimonides Rule by constructing 2SLS estimates of an equation like

$$\mathrm{Y}_{isc} = \alpha_0 + \alpha_1 pd_s + \beta_1 e_s + \beta_2 e_s^2 + ... + \beta_p e_s^p + \rho n_{sc} + \eta_{isc}, \tag{6.2.6}$$

where Y_{isc} is $i's$ test score in school s and class c, n_{sc} is the size of this class, and e_s is enrollment. In this version of fuzzy RD, m_{sc} plays the role of T_i, e_s plays the role of x_i, and class size, n_{sc} plays the role of D_i. Angrist and Lavy also include a non-enrollment covariate, pd_s, to control for the proportion of students in the school from a disadvantaged background. This is not necessary for RD, since the only source of omitted variables bias in the RD model is e_s, but it makes the specification comparable to the model used to construct a corresponding set of OLS estimates.[7]

Figure 6.2.1 from Angrist and Lavy (1999) plots the average of actual and predicted class sizes against enrollment in fourth and fifth grade. Maimonides' Rule does not predict class size perfectly because some schools split grades at enrollments lower than 40. This is what makes the RD design fuzzy. Still, there are clear drops in class size at enrollment levels of enrollment levels of 40, 80, and 120. Note also that the m_{sc} instrument neatly combines both discontinuities and slope-discontinuity interactions such as $\tilde{x}_i \mathrm{T}_i$ in (6.2.4) in a single variable. This compact parametrization comes from a specific understanding of the institutions and rules that determine Israeli class size.

Estimates of equation (6.2.6) for fifth-grade Math scores are reported in Table 6.2.1, beginning with OLS. With no controls, there is a strong positive relationship between class size and test scores. Most of this vanishes however, when the percent disadvantaged in the school is included as a control. The correlation between class size and test scores shrinks to insignificance when enrollment is added as an additional control, as can be seen in column 3. Still, there is no evidence that smaller classes are better, as we might believe based on the results from the Tennessee STAR randomized trial.

[7] The Angrist and Lavy (1999) study differs modestly from the description here in that the data used to estimate equation (6.2.6) are class averages. But since the covariates are all defined at the class or school level, the only difference between student-level and class-level estimation is the implicit weighting by number of students in the student-level estimates.

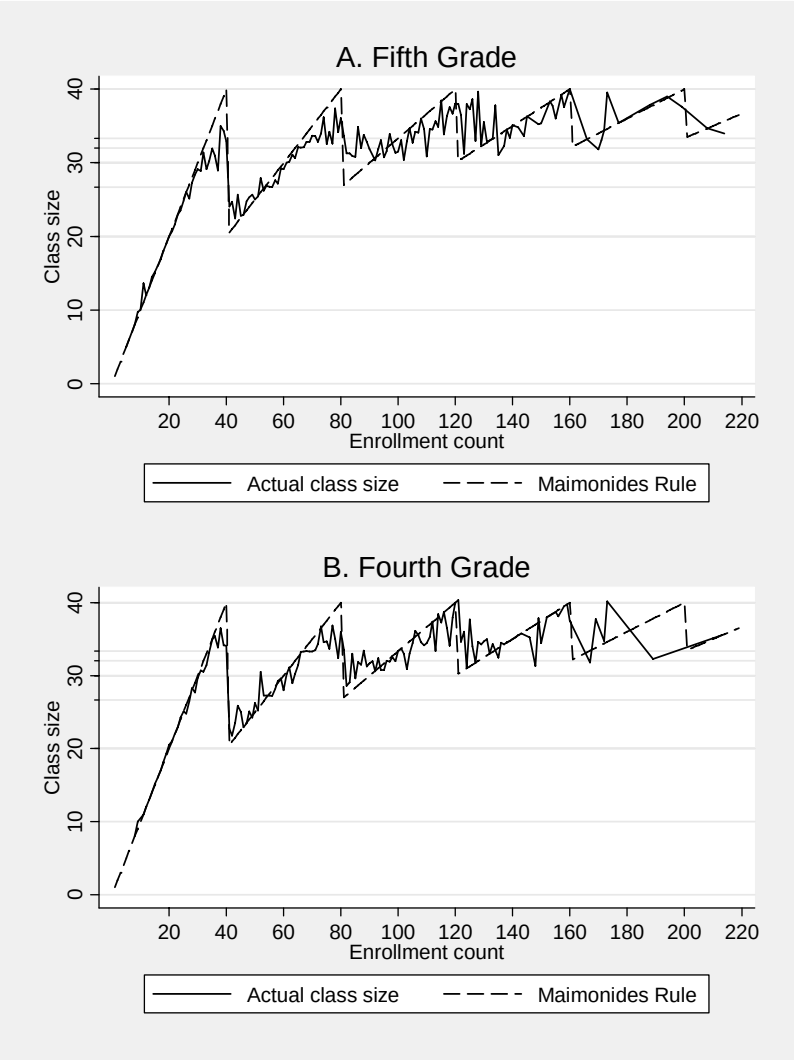

Figure 6.2.1: The fuzzy-RD first-stage for regression-discontinuity estimates of the effect of class size on pupils' test scores (from Angrist and Lavy, 1999)

In contrast with the OLS estimates in column 3, 2SLS estimates of similar specification using m_{sc} as an instrument for n_{sc} strongly suggest that smaller classes increase test scores. These results, reported in column 4 for models that include a linear enrollment control and in column 5 for models that include a quadratic enrollment control range from -.23 to -.26 with standard error around .1. These results suggest a 7-student reduction in class size (as in Tennessee STAR) raises Math scores by about 1.75 points, for an effect size of $.18\sigma$, where σ is the standard deviation of class average scores. This is not too far from the Tennessee estimates.

Importantly, the functional form of the enrollment control does not seem to matter very much (though estimates with no controls - not reported in the table - come out much smaller and insignificant). Columns 6 and 7 check the robustness of the main findings using a +/-5 discontinuity sample. Not surprisingly, these results are much less precise than those reported in columns 5 and 6 since they were estimated with only about one-quarter of the data used to construct the full-sample estimates. Still, they bounce around the -.25 mark. Finally, the last column shows the results of estimation using an even narrower discontinuity sample limited to schools with plus or minus an enrollment of 3 students around the discontinuities at 40, 80, and 120 (with dummy controls for which of these discontinuities is relevant). These are Wald estimates in the spirit of Hahn, Todd, and van der Klaauw (2001) and formula (6.2.5); the instrument used to construct these estimates is a dummy for being in a school with enrollment just to the right of the relevant discontinuity. The result is an imprecise -.270 (s.e.=.281), but still strikingly similar to the other estimates in the table. This set of estimates illustrates the high price to be paid in terms of precision when we shrink the sample around the discontinuities. Happily, however, the picture that emerges from Table (6.2.1) is fairly clear.

Table 6.2.1: OLS and fuzzy RD estimates of the effects of class size on fifth grade math scores

	OLS			2SLS				
				Full sample		Discontinuity samples		
						+/- 5		+/- 3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean score		67.3		67.3		67.0		67.0
(s.d.)		(9.6)		(9.6)		(10.2)		(10.6)
Regressors								
Class size	.322	.076	.019	-.230	-.261	-.185	-.443	-.270
	(.039)	(.036)	(.044)	(.092)	(.113)	(.151)	(.236)	(.281)
Percent disadvantaged		-.340	-.332	-.350	-.350	-.459	-.435	
		(.018)	(.018)	(.019)	(.019)	(.049)	(.049)	
Enrollment			.017	.041	.062		.079	
			(.009)	(.012)	(.037		(.036)	
Enrollment squared/100					-.010			
					(.016)			
Segment 1								-12.6
(enrollment 36-45)								(3.80)
Segment 2								-2.89
(enrollment 76-85)								(2.41)
Root MSE	9.36	8.32	8.30	8.40	8.42	8.79	9.10	10.2
R-squared	.048	.249	.252					
N		2,018		2,018		471		302

Notes: Adapted from Angrist and Lavy (1999). The table reports estimates of equation (6.2.6) in the text using class averages. Standard errors, reported in parentheses, are corrected for within-school correlation.

Chapter 7

Quantile Regression

> Here's a prayer for you. Got a pencil? . . . 'Protect me from knowing what I don't need to know. Protect me from even knowing that there are things to know that I don't know. Protect me from knowing that I decided not to know about the things I decided not to know about. Amen.' There's another prayer that goes with it. 'Lord, lord, lord. Protect me from the consequences of the above prayer.'
>
> Douglas Adams, *Mostly Harmless* (1995)

Rightly or wrongly, 95 percent of applied econometrics is concerned with averages. If, for example, a training program raises average earnings enough to offset the costs, we are happy. The focus on averages is partly because obtaining a good estimate of the average causal effect is hard enough. And if the dependent variable is a dummy for something like employment, the mean describes the entire distribution. But many variables, like earnings and test scores, have continuous distributions. These distributions can change in ways not revealed by an examination of averages, for example, they can spread out or become more compressed. Applied economists increasingly want to know what's happening to an entire distribution, to the relative winners and losers, as well as to averages.

Policy-makers and labor economists have been especially concerned with changes in the wage distribution. We know, for example, that flat average real wages are only a small part of what's been going on in the labor market for the past 25 years. Upper earnings quantiles have been increasing, while lower quantiles have been falling. In other words, the rich are getting richer and the poor are getting poorer. But that's not all - recently, inequality has grown asymmetrically; for example, among college graduates, it's mostly the rich getting richer, with wages at the lower decile unchanging. The complete story of the changing wage distribution is fairly complicated and would seem to be hard to summarize.

Quantile regression is a powerful tool that makes the task of modeling distributions easy, even when the underlying story is complex and multi-dimensional. We can use this tool to see whether participation in a training program or membership in a labor union affects earnings inequality as well as average earnings. We