NBER WORKING PAPER SERIES

REGRESSION DISCONTINUITY DESIGNS IN ECONOMICS

David S. Lee
Thomas Lemieux

Working Paper 14723
http://www.nber.org/papers/w14723

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2009

We thank David Autor, David Card, John DiNardo, Guido Imbens, and Justin McCrary for suggestions for this article, as well as for numerous illuminating discussions on the various topics we cover in this review. We also thank two anonymous referees for their helpful suggestions and comments. Emily Buchsbaum, Elizabeth Debraggio, Enkeleda Gjeci, Ashley Hodgson, Xiaotong Niu, and Zhuan Pei provided excellent research assistance. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Regression Discontinuity Designs in Economics
David S. Lee and Thomas Lemieux
NBER Working Paper No. 14723
February 2009
JEL No. C1,H0,I0,J0

ABSTRACT

This paper provides an introduction and "user guide" to Regression Discontinuity (RD) designs for empirical researchers. It presents the basic theory behind the research design, details when RD is likely to be valid or invalid given economic incentives, explains why it is considered a "quasi-experimental" design, and summarizes different ways (with their advantages and disadvantages) of estimating RD designs and the limitations of interpreting these estimates. Concepts are discussed using using examples drawn from the growing body of empirical research using RD.

David S. Lee
Industrial Relations Section
Princeton University
Firestone Library A-16-J
Princeton, NJ 08544
and NBER
davidlee@princeton.edu

Thomas Lemieux
Department of Economics
University of British Columbia
#997-1873 East Mall
Vancouver, BC V6T 1Z1
Canada
and NBER
tlemieux@interchange.ubc.ca

1 Introduction

Regression Discontinuity (RD) designs were first introduced by Thistlethwaite and Campbell (1960) as a way of estimating treatment effects in a non-experimental setting where treatment is determined by whether an observed "forcing" variable exceeds a known cutoff point. In their initial application of RD designs, Thistlethwaite and Campbell (1960) analyzed the impact of merit awards on future academic outcomes, using the fact that the allocation of these awards was based on an observed test score. The main idea behind the research design was that individuals with scores just below the cutoff (who did not receive the award) were good comparisons to those just above the cutoff (who did receive the award). Although this evaluation strategy has been around for almost fifty years, it did not attract much attention in economics until relatively recently.

Since the late 1990s, a growing number of studies have relied on RD designs to estimate program effects in a wide variety of economic contexts. Like Thistlethwaite and Campbell (1960), early studies by Van der Klaauw (2002) and Angrist and Lavy (1999) exploited threshold rules often used by educational institutions to estimate the effect of financial aid and class size, respectively, on educational outcomes. Black (1999) exploited the presence of discontinuities at the geographical level (school district boundaries) to estimate the willingness to pay for good schools. Following these early papers in the area of education, the past five years have seen a rapidly growing literature using RD designs to examine a range of questions. Examples include: the labor supply effect of welfare, unemployment insurance, and disability programs; the effects of Medicaid on health outcomes; the effect of remedial education programs on educational achievement; the empirical relevance of median voter models; and the effects of unionization on wages and employment.

One important impetus behind this recent flurry of research is a recognition, formalized by Hahn et al. (2001), that RD designs require seemingly mild assumptions compared to those needed for other non-experimental approaches. Another reason for the recent wave of research is the belief that the RD design is not "just another" evaluation strategy, and that causal inferences from RD designs are potentially more credible than those from typical "natural experiment" strategies (e.g. difference-in-differences or instrumental variables), which have been heavily employed in applied research in recent decades. This notion has a theoretical justification: Lee (2008) formally shows that one need not *assume* the RD design isolates treatment variation that is "as good as randomized"; instead, such randomized variation is a *consequence* of agents' inability to precisely control the forcing variable near the known cutoff.

So while the RD approach was initially thought to be "just another" program evaluation method with relatively little general applicability outside of a few specific problems, recent work in economics has shown quite the opposite.[1] In addition to providing a highly credible and transparent way of estimating program effects, RD designs can be used in a wide variety of contexts covering a large number of important economic questions. These two facts likely explain why the RD approach is rapidly becoming a major element in the toolkit of empirical economists.

Despite the growing importance of RD designs in economics, there is no single comprehensive summary of what is understood about RD designs – when they succeed, when they fail, and their strengths and weaknesses.[2] Furthermore, the "nuts and bolts" of implementing RD designs in practice are not (yet) covered in standard econometrics texts, making it difficult for researchers interested in applying the approach to do so. Broadly speaking, the main goal of this paper is to fill these gaps by providing an up-to-date overview of RD designs in economics and creating a guide for researchers interested in applying the method.

A reading of the most recent research reveals a certain body of "folk wisdom" regarding the applicability, interpretation, and recommendations of practically implementing RD designs. This article represents our attempt at identifying what we believe are the most important of these pieces of wisdom, while also dispelling misconceptions that could potentially (and understandably) arise for those new to the RD approach.

We will now briefly summarize what we see as the main points in the "folk wisdom" about RD designs to set the stage for the rest of the paper where we systematically discuss identification, interpretation, and estimation issues. Here, and throughout the paper, we refer to the forcing variable as X. Treatment is, thus, assigned to individuals (or "units") with a value of X greater than or equal to a cutoff value c.

- **RD designs can be invalid if individuals can precisely manipulate the "forcing variable".**

  When there is a payoff or benefit to receiving a treatment, it is natural for an economist to consider how an individual may behave to obtain such benefits. For example, if students could effectively "choose" their test score X through effort, those who chose a score c (and hence received the merit award) could be somewhat different from those who chose scores just below c. The important lesson here is that the existence of a treatment being a discontinuous function of a forcing variable is *not* sufficient to justify the validity of an RD design. Indeed, if anything, discontinuous rules may generate incentives,

[1] See Cook (2008) for an interesting history of the RD design in education research, psychology, statistics, and economics. Cook argues the resurgence of the RD design in economics is unique as it is still rarely used in other disciplines.

[2] See, however, two recent overview papers by Van der Klaauw (2008b) and Imbens and Lemieux (2008) that have begun bridging this gap.

causing behavior that would *invalidate* the RD approach.

- **If individuals – even while having some influence – are unable to *precisely* manipulate the forcing variable, *a consequence* of this is that the variation in treatment near the threshold is randomized as though from a randomized experiment.**

  This is a crucial feature of the RD design, since it is the reason RD designs are often so compelling. Intuitively, when individuals have imprecise control over the forcing variable, even if some are especially likely to have values of X *near* the cutoff, *every* individual will have approximately the same probability of having an X that is just above (receiving the treatment) or just below (being denied the treatment) the cutoff – similar to a coin-flip experiment. This result clearly differentiates the RD and IV approaches. When using IV for causal inference, one must *assume* the instrument is exogenously generated as if by a coin-flip. Such an assumption is often difficult to justify (except when an actual lottery was run, as in Angrist (1990), or if there were some biological process, e.g. gender determination of a baby, mimicking a coin-flip). By contrast, the variation that RD designs isolate is randomized *as a consequence* of individuals having imprecise control over the forcing variable.

- **RD designs can be analyzed – and tested – like randomized experiments.**

  This is the key implication of the local randomization result. If variation in the treatment near the threshold is approximately randomized, then it follows that all "baseline characteristics" – all those variables determined prior to the realization of the forcing variable – should have the same distribution just above and just below the cutoff. If there is a discontinuity in these baseline covariates, then at a minimum, the underlying identifying assumption of individuals' inability to precisely manipulate the forcing variable is unwarranted. Thus, the baseline covariates are used to *test* the validity of the RD design. By contrast, when employing an IV or a matching/regression-control strategy, assumptions typically need to be made about the relationship of these other covariates to the treatment and outcome variables.[3]

- **Graphical presentation of an RD design is helpful and informative, but the visual presentation should not be tilted toward either finding an effect or finding no effect.**

  It has become standard to summarize RD analyses with a simple graph showing the relationship be-

[3] Typically, one assumes that *conditional on the covariates*, the treatment (or instrument) is essentially "as good as" randomly assigned.

tween the outcome and forcing variables. This has several advantages. The presentation of the "raw data" enhances the transparency of the research design. A graph can also give the reader a sense of whether the "jump" in the outcome variable at the cutoff is unusually large compared to the bumps in the regression curve away from the cutoff. Also, a graphical analysis can help identify why different functional forms give different answers, and can help identify outliers, which can be a problem in any empirical analysis. The problem with graphical presentations, however, is that there is some room for the researcher to construct graphs making it seem as though there are effects when there are none, or hiding effects that truly exist. We suggest later in the paper a number of methods to minimize such biases in presentation.

- **Non-parametric estimation does not represent a "solution" to functional form issues raised by RD designs. It is therefore helpful to view it as a complement to – rather than a substitute for – parametric estimation.**

  When the analyst chooses a parametric functional form (say, a low-order polynomial) that is incorrect, the resulting estimator will, in general, be biased. When the analyst uses a non-parametric procedure such as local linear regression – essentially running a regression using only data points "close" to the cutoff – there will also be bias.[4] With a finite sample, it is impossible to know which case has a smaller bias without knowing something about the true function. There will be some functions where a low-order polynomial regression produces a smaller bias and other functions where a local linear regression will. In practice, there is often little difference in the results obtained using one estimator or the other. For example, the procedure of regressing the outcome Y on X and a treatment dummy D can be viewed as a parametric regression (as discussed above), or as a local linear regression with a very large bandwidth. Similarly, if one wanted to exclude the influence of data points in the tails of the X distribution, one could call the exact same procedure "parametric" after trimming the tails, or "non-parametric" by viewing the restriction in the range of X as a result of using a smaller bandwidth.[5] Our main suggestion in estimation is to not rely on one particular method or specification. In any

[4] Unless the underlying function is exactly linear in the area being examined.

[5] The main difference, then, between a parametric and non-parametric approach is not in the actual estimation, but rather in the discussion of the asymptotic behavior of the estimator as sample sizes tend to infinity. For example, standard non-parametric asymptotics considers what would happen if the bandwidth h – the width of the "window" of observations used for the regression – were allowed to shrink as the number of observations N tended to infinity. It turns out that if $h \to 0$ and $Nh \to \infty$ as $N \to \infty$, the bias will tend to zero. By contrast, with a parametric approach, when one is not allowed to make the model more flexible with more data points, the bias would generally remain – even with infinite samples.

empirical analysis, results that are stable across alternative and equally plausible specifications are generally viewed as more reliable than those that are sensitive to minor changes in specification. RD is no exception in this regard.

- **Goodness-of-fit and other statistical tests can help rule out overly restrictive specifications.** Often the consequence of trying many different specifications is that it may result in a wide range of estimates. Although there is no simple formula that works in all situations and contexts for weeding out inappropriate specifications, it seems reasonable, at a minimum, not to rely on an estimate resulting from a specification that can be rejected by the data when tested against a strictly more flexible specification. For example, it seems wise to place less confidence in results from a low-order polynomial model, when it is rejected in favor of a less restrictive model (e.g., separate means for each discrete value of X). Similarly, there seems little reason to prefer a specification that uses all the data, if using the same specification but restricting to observations closer to the threshold gives a substantially (and statistically) different answer.

The rest of the paper is organized as follows. In Section 2, we discuss the origins of the RD design and show how it has recently been formalized in economics using the potential outcome framework. We also introduce an important theme that we stress throughout the paper, namely that RD designs are particularly compelling because they are close cousins of randomized experiments. This theme is more formally explored in Section 3 where we discuss the conditions under which RD designs are "as good as a randomized experiment", how RD estimates should be interpreted, and how they compare with other commonly used approaches in the program evaluation literature. Section 4 goes through the main "nuts and bolts" involved in implementing RD designs and provides a "guide to practice" for researchers interested in using the design. Implementation issues in several specific situations (discrete forcing variable, panel data, etc.) are covered in Section 5. Based on a survey of the recent literature, Section 6 shows that RD designs have turned out to be much more broadly applicable in economics than was originally thought. We conclude in Section 7 by discussing recent progress and future prospects in using and interpreting RD designs in economics.

2 Origins and Background

In this section, we set the stage for the rest of the paper by discussing the origins and the basic structure of the RD design, beginning with the classic work of Thistlethwaite and Campbell (1960) and moving to the recent interpretation of the design using modern tools of program evaluation in economics (potential outcomes framework). One of the main virtues of the RD approach is that it can be naturally presented using simple graphs, which greatly enhances its credibility and transparency. In light of this, the majority of concepts introduced in this section are represented in graphical terms to help capture the intuition behind the RD design.

2.1 Origins

The RD design was first introduced by Thistlethwaite and Campbell (1960) in their study of the impact of merit awards on the future academic outcomes (career aspirations, enrollment in post-graduate programs, etc.) of students. Their study exploited the fact that these awards were allocated on the basis of an observed test score. Students with test scores, X, greater than or equal to a cutoff value c received the award, and those with scores below the cutoff were denied the award. This generated a sharp discontinuity in the "treatment" (receiving the award) as a function of the test score. Let the receipt of treatment be denoted by the dummy variable $D \in \{0,1\}$, so that we have $D = 1$ if $X \geq c$, and $D = 0$ if $X < c$.

At the same time, there appears to be no reason, other than the merit award, for future academic outcomes, Y, to be a discontinuous function of the test score. This simple reasoning suggests attributing the discontinuous jump in Y at c to the causal effect of the merit award. Assuming that the relationship between Y and X is otherwise linear, a simple way of estimating the treatment effect τ is by fitting the linear regression

$$Y = \alpha + D\tau + X\beta + \varepsilon \tag{1}$$

where ε is the usual error term that can be viewed as a purely random error generating variation in the value of Y around the regression line $\alpha + D\tau + X\beta$. This case is depicted in Figure 1, which shows both the true underlying function and numerous realizations of ε.

Thistlethwaite and Campbell (1960) provided some graphical intuition for why the coefficient τ could be viewed as an estimate of the causal effect of the award. We illustrate their basic argument in Figure 1.

Consider an individual whose score X is exactly c. To get the causal effect for a person scoring c, we need guesses for what her Y would be with and without receiving the treatment.

If it is "reasonable" to assume that all factors (other than the award) are evolving "smoothly" with respect to X, then B' would be a reasonable guess for the value of Y of an individual scoring c (and hence receiving the treatment). Similarly, A'' would be a reasonable guess for that same individual in the counterfactual state of not having received the treatment. It follows that $B' - A''$ would be the causal estimate. This illustrates the intuition that the RD estimates should use observations "close" to the cutoff (e.g. in this case, at points c' and c'').

There is, however, a limitation to the intuition that "the closer to c you examine, the better". In practice, one *cannot* "only" use data close to the cutoff. The narrower the area that is examined, the less data there are. In this example, examining data any closer than c' and c'' will yield no observations at all! Thus, in order to produce a reasonable guess for the treated and untreated states at $X = c$ with finite data, one has no choice but to use data *away* from the discontinuity.[6] Indeed, if the underlying function is truly linear, we know that the best linear unbiased estimator of τ is the coefficient on D from OLS estimation (using all of the observations) of Equation (1).

This simple heuristic presentation illustrates two important features of the RD design. First, in order for this approach to work, "all other factors" determining Y must be evolving "smoothly" with respect to X. If the other variables also jump at c, then the gap τ will potentially be biased for the treatment effect of interest. Second, since an RD estimate requires data away from the cutoff, the estimate will be dependent on the chosen functional form. In this example, if the slope β were (erroneously) restricted to equal zero, it is clear the resulting OLS coefficient on D would be a biased estimate of the true discontinuity gap.

2.2 RD Designs and the Potential Outcomes Framework

While the RD design was being imported into applied economic research by studies such as Van der Klaauw (2002), Black (1999), and Angrist and Lavy (1999), the identification issues discussed above were formalized in the theoretical work of Hahn et al. (2001), who described the RD evaluation strategy using the language of the treatment effects literature. Hahn et al. (2001) noted the key assumption of a valid RD design was that "all other factors" were "continuous" with respect to X, and suggested a non-parametric procedure for

[6] Interestingly, the very first application of the RD design by Thistlethwaite and Campbell (1960) was based on discrete data (interval data for test scores). As a result, their paper clearly points out that the RD design is fundamentally based on an extrapolation approach.

estimating τ that did not assume underlying linearity, as we have in the simple example above.

The necessity of the continuity assumption is seen more formally using the "potential outcomes framework" of the treatment effects literature, with the aid of a graph. It is typically imagined that for each individual i, there exists a pair of "potential" outcomes: $Y_i(1)$ for what would occur if the unit were exposed to the treatment and $Y_i(0)$ if not exposed. The causal effect of the treatment is represented by the difference $Y_i(1) - Y_i(0)$. The fundamental problem of causal inference is that we cannot observe the pair $Y_i(0)$ and $Y_i(1)$ simultaneously. We therefore typically focus on average effects of the treatment, that is, averages of $Y_i(1) - Y_i(0)$ over (sub-)populations, rather than on unit-level effects.

In the RD setting, we can imagine there are two underlying relationships between average outcomes and X, represented by $E[Y_i(1)|X]$ and $E[Y_i(0)|X]$, as in Figure 2. But by definition of the RD design, all individuals to the right of the cutoff ($c = 2$ in this example) are exposed to treatment, and all those to the left are denied treatment. Therefore, we only observe $E[Y_i(1)|X]$ to the right of the cutoff and $E[Y_i(0)|X]$ to the left of the cutoff, as indicated in the figure.

It is easy to see that with what is observable, we could try to estimate the quantity

$$B - A = \lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x],$$

which would equal

$$E[Y_i(1) - Y_i(0)|X = c].$$

This is the "average treatment effect" at the cutoff c.

This inference is possible because of the continuity of the underlying functions $E[Y_i(1)|X]$ and $E[Y_i(0)|X]$.[7] In essence, this continuity condition enables us to use the average outcome of those right below the cutoff (who are denied the treatment) as a valid counterfactual for those right above the cutoff (who received the treatment).

Although the potential outcome framework is very useful for understanding how RD designs work in a framework applied economists are used to dealing with, it also introduces some difficulties in terms of interpretation. First, while the continuity assumption sounds generally plausible, it is not completely clear

[7]The continuity of both functions is not the minimum that is required, as pointed out in Hahn et al. (2001). For example, identification is still possible even if only $E[Y_i(0)|X]$ is continuous, and only continuous at c. Nevertheless, it may seem more natural to assume that the conditional expectations are continuous for all values of X, since cases where continuity holds at the cutoff point but not at other values of X seem peculiar.

what it means from an economic point of view. The problem is that since continuity is not required in the more traditional applications used in economics (e.g. matching on observables), it is not obvious what assumptions about the behavior of economic agents are required to get continuity.

Second, RD designs are a fairly peculiar application of a "selection on observables" model. Indeed, the view in Heckman et al. (1999) was that "[r]egression discontinuity estimators constitute a special case of selection on observables," and that the RD estimator is "a limit form of matching at one point." In general, we need two crucial conditions for a matching/selection on observables approach to work. First, treatment must be randomly assigned conditional on observables (the *ignorability* or *unconfoundedness* assumption). In practice, this is typically viewed as a strong, and not particularly credible, assumption. For instance, in a standard regression framework this amounts to assuming that all relevant factors are controlled for, and that no omitted variables are correlated with the treatment dummy. In an RD design, however, this crucial assumption is trivially satisfied. When $X \geq c$, the treatment dummy D is always equal to 1. When $X < c$, D is always equal to 0. Conditional on X, there is no variation left in D, so it cannot, therefore, be correlated with any other factor.[8]

At the same time, the other standard assumption of *overlap* is violated since, strictly speaking, it is not possible to observe units with either $D = 0$ or $D = 1$ for a given value of the forcing variable X. This is the reason the continuity assumption is required - to compensate for the failure of the overlap condition. So while we cannot observe treatment and non-treatment for the same value of X, we can observe the two outcomes for values of X around the cutoff point that are arbitrarily close to each other.

2.3 RD design as a Local Randomized Experiment

When looking at RD designs the way we just did, one gets the impression that they require some assumptions to be satisfied, while other methods such as matching on observables and IV methods simply require other assumptions.[9] So determining which approach is preferable depends on the validity of assumptions that are not clearly comparable a priori. As we show in the rest of the paper, however, we don't think that this way of looking at RD designs does justice to their important advantages over most other existing methods. This point becomes much clearer once one also casts the "gold standard" of program evaluation methods,

[8] In technical term, the treatment dummy D follows a degenerate (concentrated at $D = 0$ or $D = 1$), but nonetheless random distribution conditional on X. Ignorability is thus trivially satisfied.

[9] For instance, in the survey of Angrist and Krueger (1999), RD is viewed as an IV estimator, thus having essentially the same potential drawbacks and pitfalls.

randomized experiments, in the potential outcomes framework. Doing so illustrates that the RD design is a much closer cousin of randomized experiment than other competing methods.

In a randomized experiment, units are typically divided into treatment and controls on the basis of a randomly generated number, v. For example, if v follows a uniform distribution over the range $[0,4]$, units with $v \geq 2$ are given the treatment while units with $v < 2$ are denied treatment. So the randomized experiment can be thought of as an RD design where the forcing variable is $X = v$ and the cutoff is $c = 2$. Figure 3 shows this special case in the potential outcomes framework, just as in the more general RD design case of Figure 2. The difference is that because the forcing variable X is now completely random, it is independent of the potential outcomes $Y_i(0)$ and $Y_i(1)$, and the curves $E[Y_i(1)|X]$ and $E[Y_i(0)|X]$ are flat. Since the curves are flat, it trivially follows that they are also continuous at the cutoff point $X = c$. In other words, continuity is a direct consequence of randomization.

The fact that the curves $E[Y_i(1)|X]$ and $E[Y_i(0)|X]$ are flat in a randomized experiment implies that, as is well known, the average treatment effect can be computed as the difference in the mean value of Y on the right and left hand side of the cutoff. One could also use an RD approach by running regressions of Y on X, but this would be less efficient since we know that if randomization were successful, then X is an irrelevant variable in this regression.

But now imagine that, for ethical reasons, people are compensated for having received a "bad draw" by getting a monetary compensation inversely proportional to the random number X. For example, the treatment could be job search assistance for the unemployed, and the outcome whether one found a job within a month of receiving the treatment. If people with a larger monetary compensation can afford to take more time looking for a job, the potential outcome curves will no longer be flat and will slope upward. The reason is that having a higher random number, i.e. a lower monetary compensation, increases the probability of finding a job. So in this "smoothly contaminated" randomized experiment, the potential outcome curves will instead look like the classical RD design case depicted in Figure 2.

Unlike a classical randomized experiment, in this contaminated experiment a simple comparison of means no longer yields a consistent estimate of the treatment effect. By focusing right around the threshold, however, an RD approach would still yield a consistent estimate of the treatment effect associated with job search assistance. The reason is that since people just above or below the cutoff receive (essentially) the same monetary compensation, we still have locally a randomized experiment around the cutoff point. Furthermore, as in a randomized experiment, it is possible to test whether randomization "worked" by comparing the local

values of baseline covariates on the two sides of the cutoff value.

Of course, this particular example is highly artificial. Since we know the monetary compensation is a continuous function of X, we also know the continuity assumption required for the RD estimates of the treatment effect to be consistent is also satisfied. The important result, due to Lee (2008), that we will show in the next section is that the conditions under which we locally have a randomized experiment (and continuity) right around the cutoff point are remarkably weak. Furthermore, in addition to being weak, the conditions for local randomization are testable in the same way global randomization is testable in a randomized experiment by looking at whether baseline covariates are balanced. It is in this sense that the RD design is a closer cousin to randomized experiments relative to other popular program evaluation methods such as matching on observables, difference-in-differences, and IV.

3 Identification and Interpretation

This section discusses a number of issues of identification and interpretation that arise when considering an RD design. Specifically, the applied researcher may be interested in knowing the answers to the following questions:

1. How do I know whether an RD design is appropriate for my context? When are the identification assumptions plausible or implausible?

2. Is there any way I can test those assumptions?

3. To what extent are results from RD designs generalizable?

On the surface , the answers to these questions seem straightforward: 1) "An RD design will be appropriate if it is plausible that all other unobservable factors are "continuously" related to the forcing variable," 2) "No, the continuity assumption is necessary, so there are no tests for the validity of the design," and 3) "The RD estimate of the treatment effect is only applicable to the sub-population of individuals at the discontinuity threshold, and uninformative about the effect anywhere else." These answers suggest the RD design is no more compelling than, say an instrumental variables approach, for which the analogous answers would be 1) "The instrument must be uncorrelated with the error in the outcome equation," 2) "The identification assumption is ultimately untestable," and 3) "The estimated treatment effect is applicable to the sub-population whose treatment was affected by the instrument." After all, who's to say whether one untestable design is

more "compelling" or "credible" than another untestable design? And it would seem that having a treatment effect for a vanishingly small sub-population (those at the threshold, in the limit) is hardly more (and probably much less) useful than that for a population "affected by the instrument."

As we describe below, however, a closer examination of the RD design reveals quite different answers to the above three questions:

1. "When there is a continuously distributed stochastic error component to the forcing variable – which can occur when optimizing agents do not have *precise* control over the forcing variable – then the variation in the treatment will be as good as randomized in a neighborhood around the discontinuity threshold."

2. "Yes. As in a randomized experiment, the distribution of observed baseline covariates should not change discontinuously at the threshold."

3. "The RD estimand can be interpreted as a weighted average treatment effect, where the weights are the relative *ex ante* probability that the value of an individual's forcing variable will be in the neighborhood of the threshold."

Thus, in many contexts, the RD design may have more in common with randomized experiments (or circumstances when an instrument is truly randomized) – in terms of their "internal validity" and how to implement them in practice – than with regression control or matching methods, instrumental variables, or panel data approaches. We will return to this point after first discussing the above three issues in greater detail.

3.1 Valid or Invalid RD?

Are individuals able to influence the forcing variable, and if so, what is the nature of this control? This is probably the most important question to ask when assessing whether a particular application should be analyzed as an RD design. If individuals have a great deal of control over the forcing variable and if there is a perceived benefit to a treatment, one would certainly expect individuals on one side of the threshold to be systematically different from those on the other side.

Consider the test-taking RD example. Suppose there are two types of students: *A* and *B*. Suppose type *A* students are more able than *B* types, and that *A* types are also keenly aware that passing the relevant threshold (50 percent) will give them a scholarship benefit, while *B* types are completely ignorant of the scholarship

and the rule. Now suppose that 50 percent of the questions are trivial to answer correctly, but due to random chance, students will sometimes make careless errors when they initially answer the test questions, but would certainly correct the errors if they checked their work. In this scenario, only type *A* students will make sure to check their answers before turning in the exam, thereby assuring themselves of a passing score. Thus, while we would expect those who barely passed the exam to be a mixture of type *A* and type *B* students, those who barely failed would exclusively be type *B* students. In this example, it is clear that the marginal failing students do *not* represent a valid counterfactual for the marginal passing students. Analyzing this scenario within an RD framework would be inappropriate.

On the other hand, consider the same scenario, except assume that questions on the exam are *not* trivial; there are no guaranteed passes, no matter how many times the students check their answers before turning in the exam. In this case, it seems more plausible that among those scoring near the threshold, it is a matter of "luck" as to which side of the threshold they land. Type *A* students can exert more effort – because they know a scholarship is at stake – but they do not know the exact score they will obtain. In this scenario, it would be reasonable to argue that those who marginally failed and passed would be otherwise comparable, and that an RD analysis *would* be appropriate and would yield credible estimates of the impact of the scholarship.

These two examples make it clear that one must have some knowledge about the mechanism generating the forcing variable, beyond knowing that if it crosses the threshold, the treatment is "turned on". It is "folk wisdom" in the literature to judge whether the RD is appropriate based on whether individuals could manipulate the forcing variable and *precisely* "sort" around the discontinuity threshold. The key word here is "precise", rather than "manipulate". After all, in both examples above, individuals do exert some control over the test score. And indeed in virtually every known application of the RD design, it is easy to tell a plausible story that the forcing variable is to some degree influenced by *someone*. But individuals will not always be able to have *precise* control over the forcing variable. It should perhaps seem obvious that it is necessary to rule out precise sorting to justify the use of an RD design . After all, individual self-selection into treatment or control regimes is exactly why simple comparison of means is unlikely to yield valid causal inferences. Precise sorting around the threshold is self-selection.

What is not obvious, however, is that when one formalizes the notion of having imprecise control over the forcing variable, there is a striking consequence: the variation in the treatment in a neighborhood of the threshold is "as good as randomized". We explain this below.

3.1.1 Randomized Experiments from Non-Random Selection

To see how the inability to precisely control the forcing variable leads to a source of randomized variation in the treatment, consider a simplified formulation of the RD design:[10]

$$\begin{aligned} Y &= D\tau + W\delta_1 + U \\ D &= 1[X \geq 0] \\ X &= W\delta_2 + V \end{aligned} \tag{2}$$

where Y is the outcome of interest, D is the binary treatment indicator, W is the vector of all pre-determined and observable characteristics of the individual that might impact the outcome and/or the forcing variable X, and where the discontinuity threshold has been normalized to be 0.

This model looks like a standard endogenous dummy variable set-up, except that we observe the treatment determining variable, X. This allows us to relax most of the other assumptions usually made in this type of model. First, we allow W to be endogenously determined, as long as it is determined prior to V. Second, we take no stance as to whether some elements of δ_1 or δ_2 are zero (exclusion restrictions). Third, we make no assumptions about the correlations between W, U, and V.[11]

In this model, individual heterogeneity in the outcome is completely described by the pair of random variables (W,U); anyone with the same values of (W,U) will have one of two values for the outcome, depending on whether they receive treatment. Note that since RD designs are implemented by running regressions of Y on X, equation (2) looks peculiar since X is not included with W and U on the right hand side of the equation. We could add a function of X to the outcome equation, but this would not change anything to the model since we have not made any assumptions about the joint distribution of W, U, and V. For example, our setup allows for the case where $U = X\delta_3 + U'$, which yields the outcome equation $Y = D\tau + W\delta_1 + X\delta_3 + U'$. For the sake of simplicity, we work with the simple case where X is not included on the right hand side of the equation.[12]

[10] We use a simple linear endogenous dummy variable setup to describe the results in this section, but all of the results could be stated within the standard potential outcomes framework, as in Lee (2008).

[11] This is much less restrictive than textbook descriptions of endogenous dummy variable systems. It is typically assumed that (U,V) is independent of (W,D) .

[12] When RD designs are implemented in practice, the estimated effect of X on Y can either reflect a true causal effect of X on Y, or a spurious correlation between X and the unobservable term U. Since it is not possible to distinguish between these two effects in practice, we simplify the setup by implicitly assuming that X only comes into equation (2) indirectly through its (spurious) correlation with U.

Now consider the distribution of X, conditional on a particular pair of values $W = w$, $U = u$. It is equivalent (up to a translational shift) to the distribution of V conditional on $W = w, U = u$. If an individual has complete and exact control over X, we would model it as having a degenerate distribution, conditional on $W = w, U = u$. That is, in repeated trials, this individual would choose the same score. This is depicted in Figure 4 as the thick line.

If there is some room for error, but individuals can nevertheless have precise control about whether they will fail to receive the treatment, then we would expect the density of X to be zero just below the threshold, but positive just above the threshold, as depicted in Figure 4 as the truncated distribution. This density would be one way to model the first example described above for the type A students. Since type A students know about the scholarship, they will double-check their answers and make sure they answer the easy questions, which comprise 50 percent of the test. How high they score above the passing threshold will be determined by some randomness.

Finally, if there is stochastic error in the forcing variable and individuals do *not* have precise control over the forcing variable, we would expect the density of X (and hence V), conditional on $W = w, U = u$ to be continuous at the discontinuity threshold, as shown in Figure 4 as the untruncated distribution.[13] It is important to emphasize that in this final scenario, the individual still has control over X : through her efforts, she can choose to shift the distribution to the right. This is the density for someone with $W = w, U = u$, but may well be different – with a different mean, variance, or shape of the density – for other individuals, with different levels of ability, who make different choices. We are assuming, however, that all individuals are unable to precisely control the score just around the threshold.

> **Definition:** We say individuals have imprecise control over X when conditional on $W = w$ and $U = u$, the density of V (and hence X) is continuous.

When individuals have imprecise control over X this leads to the striking implication that variation in treatment status will be randomized in a neighborhood of the threshold. To see this, note that by Bayes' Rule, we have

$$\Pr[W = w, U = u | X = x] = f(x | W = w, U = u) \frac{\Pr[W = w, U = u]}{f(x)} \tag{3}$$

where $f(\cdot)$ and $f(\cdot|\cdot)$ are marginal and conditional densities for X. So when $f(x|W = w, U = u)$ is contin-

[13] For example, this would be plausible when X is a test score modeled as a sum of Bernoulli random variables, which is approximately normal by the central limit theorem.

uous in x, the right hand side will be continuous in x, which therefore means that the distribution of W,U conditional on X will be continuous in x. That is, *all observed and unobserved pre-determined characteristics will have identical distributions on either side of $x = 0$ – in the limit, as we examine smaller and smaller neighborhoods of the threshold.*

In sum,

> **Local Randomization:** If individuals have imprecise control over X as defined above, then $\Pr[W = w, U = u | X = x]$ is continuous in x: the treatment is "as good as" randomly assigned around the cutoff.

In other words, the behavioral assumption that individuals do not precisely manipulate X around the threshold has the *prediction* that treatment is locally randomized.

This is perhaps why RD designs can be so compelling. A deeper investigation into the real-world details of how X (and hence D) is determined can help assess whether it is plausible that individuals have precise or imprecise control over X. By contrast, with most non-experimental evaluation contexts, learning about how the treatment variable is determined will rarely lead one to conclude that it is "as good as" randomly assigned.

3.2 Consequences of Local Random Assignment

There are three practical implications of the above local random assignment result.

3.2.1 Identification of the Treatment Effect

First and foremost, it means that the discontinuity gap at the cutoff identifies the treatment effect of interest. Specifically, we have

$$\begin{aligned} E[Y|X=0] - \lim_{\varepsilon \uparrow 0} E[Y|X=\varepsilon] &= \tau + \sum_{w,u} (w\delta_1 + u) \Pr[W=w, U=u|X=0] \\ &\quad - \lim_{\varepsilon \uparrow 0} \sum_{w,u} (w\delta_1 + u) \Pr[W=w, U=u|X=\varepsilon] \\ &= \tau \end{aligned}$$

where the last line follows from the continuity of $\Pr[W = w, U = u | X = x]$.

As we mentioned earlier, nothing changes if we augment the model by adding a direct impact of X itself in the outcome equation, as long as the effect of X on Y does not jump at the cutoff. For example, in the example of Thistlethwaite and Campbell (1960), we can allow higher test scores to improve future academic outcomes (perhaps by raising the probability of admission to higher quality schools), as long as that probability does not jump at precisely the same cutoff used to award scholarships.

3.2.2 Testing the Validity of the RD design

An almost equally important implication of the above local random assignment result is that it makes it possible to empirically assess the relevance of the prediction that $\Pr[W = w, U = u|X = x]$ is continuous in x. Although it is impossible to test this directly – since U is unobserved – it is nevertheless possible to assess whether $\Pr[W = w|X = x]$ is continuous in x at the threshold. A discontinuity would indicate a failure of the identifying assumption.

This is akin to the tests performed to empirically assess whether the randomization was carried out properly in randomized experiments. It is standard in these analyses to demonstrate that treatment and control groups are similar in their observed baseline covariates. It is similarly impossible to test whether unobserved characteristics are balanced in the experimental context, so the most favorable statement that can be made about the experiment is that the data "failed to reject" the assumption of randomization.

Subjecting the analysis to this kind of test is arguably more important in the RD design than in the experimental context. After all, the true nature of individuals' control over the forcing variable – and whether it is precise or imprecise – may well be somewhat debatable, even after a great deal of investigation into the exact treatment-assignment mechanism (which itself is always advisable to do). Imprecision of control will often be nothing more than a conjecture, but thankfully, it has observable predictions.

There is a complementary, and arguably more direct and intuitive test of the imprecision of control over the forcing variable: examination of the density of X itself, as suggested in McCrary (2008). If the density of X for each individual is continuous, then the marginal density of X over the population should be continuous as well. A jump in the density at the threshold is probably the most direct evidence of some degree of sorting around the threshold, and should provoke serious skepticism about the appropriateness of the RD design.

This test is also a partial one. Whether each individual's *ex ante* density of X is continuous is fundamentally untestable, since for each individual we only observe one realization of X. Thus, in principle, at the threshold some individuals' densities may jump up while others may sharply fall, so that in the aggregate

positives and negatives offset each other making the density appear continuous. In recent applications of RD, such occurrences do seem far-fetched, and even if this were the case, one would typically expect to observe heterogeneity in the discontinuous densities, stratified by observable characteristics; which would be detected by performing the local randomization test described above.

3.2.3 Irrelevance of Including Baseline Covariates

A consequence of a randomized experiment is that the assignment to treatment is, by construction, independent of the baseline covariates. As such, it is not necessary to include them to obtain consistent estimates of the treatment effect. In practice, however, researchers will include them in regressions, because doing so can reduce the sampling variability in the estimator. Arguably the greatest potential for this occurs when one of the baseline covariates is a pre-random-assignment observation on the dependent variable, which may likely be highly correlated with the post-assignment outcome variable of interest.

The local random assignment result allows us to apply these ideas to the RD context. For example, since the lagged dependent variable (determined prior to the realization of X) has a continuous relationship with X, then performing an RD analysis on y minus its lagged value should also yield the treatment effect of interest. The hope, however, is that the differenced outcome measure will have a sufficiently lower variance than the level of the outcome, so as to lower the variance in the RD estimator.

More formally, we have

$$\begin{aligned} E\left[y - W\pi | X = 0\right] - \lim_{\varepsilon \uparrow 0} E\left[y - W\pi | X = \varepsilon\right] &= \tau + \sum_{w,u} \left(w\left(\delta_1 - \pi\right) + u\right) \Pr\left[W = w, U = u | X = 0\right] \\ &\quad - \lim_{\varepsilon \uparrow 0} \sum_{w,u} \left(w\left(\delta_1 - \pi\right) + u\right) \Pr\left[W = w, U = u | X = \varepsilon\right] \\ &= \tau \end{aligned} \tag{4}$$

where $W\pi$ is *any* linear function, and W can include a lagged dependent variable, for example. We return to how to implement this in practice in Section 4.4.

3.3 Generalizability: the RD Gap as a Weighted Average Treatment Effect

In the presence of heterogeneous treatment effects, the discontinuity gap in an RD design can be interpreted as a kind of average treatment effect across *all* individuals. This is somewhat contrary to the temptation

to conclude that the RD design only delivers a credible treatment effect applying exclusively to the sub-population of individuals at the threshold, and saying nothing about the treatment effect "away from the threshold". This is perhaps an overly simplistic and pessimistic assessment.

Consider the scholarship test example again, and define the "treatment" as "receiving a scholarship by scoring 50 percent or greater on the scholarship exam." Recall that the pair W,U characterizes individual heterogeneity. We now let $\tau(w,u)$ denote the treatment effect for an individual with $W=w$ and $U=u$, so that the outcome equation in (2) is instead given by

$$y = D\tau(W,U) + W\delta_1 + U.$$

This is essentially a model of completely unrestricted heterogeneity in the treatment effect. Following the same line of argument as above, we obtain

$$\begin{aligned} E[y|X=0] - \lim_{\varepsilon \uparrow 0} E[y|X=\varepsilon] &= \sum_{w,u} \tau(w,u) \Pr[W=w,U=u|X=0] \\ &= \sum_{w,u} \tau(w,u) \frac{f(0|W=w,U=u)}{f(0)} \Pr[W=w,U=u] \end{aligned} \tag{5}$$

where the second line follows from Equation (3).

The discontinuity gap then, is a particular kind of average treatment effect *across all individuals*. If not for the term $\frac{f(0|W=w,U=u)}{f(0)}$, it would be the overall average treatment effect for the entire population. The presence of the ratio $\frac{f(0|W=w,U=u)}{f(0)}$ implies the discontinuity is instead a *weighted* average treatment effect where the weights are directly proportional to the *ex ante* likelihood that an individual's realization of X will be close to the threshold. All individuals could get some weight, and the similarity of the weights across individuals is ultimately untestable, since again we only observe one realization of X per person and do not know anything about the *ex ante* probability distribution of X for any one individual. The weights may be relatively similar across individuals, in which case the RD gap would be closer to the overall average treatment effect; but, if the weights are highly varied and also related to the magnitude of the treatment effect, then the RD gap would be very different from the overall average treatment effect.

It is important to emphasize the RD gap is not informative about the treatment if it were defined as "receipt of a scholarship by scoring *90 percent* or higher on the scholarship exam." This is not so much a "drawback" of the RD design as a limitation shared with even a carefully controlled randomized experiment.

For example, if we randomly assigned financial aid awards to low-achieving students, whatever treatment effect we estimate may not be informative about the effect of financial aid for high-achieving students.

In some contexts, the treatment effect "away from the discontinuity threshold" may not make much practical sense. Consider the RD analysis of incumbency in congressional elections of Lee (2008). When the treatment is "being the incumbent party," it is implicitly understood that incumbency entails winning the previous election by obtaining at least 50 percent of the vote.[14] In the election context, the treatment "being the incumbent party by virtue of winning an election, whereby 90 percent of the vote is required to win" simply does not apply to any real-life situation. Thus, in this context, it is awkward to interpret the RD gap as "the effect of incumbency that exists at 50 percent vote-share threshold" (as if there is an effect at a 90 percent threshold). Instead it is more natural to interpret the RD gap as estimating a weighted average treatment effect of incumbency across all districts, where more weight is given to those districts in which a close election race was expected.

3.4 Variations on the Regression Discontinuity Design

To this point, we have focused exclusively on the "classic" RD design introduced by Thistlethwaite and Campbell (1960), whereby there is a single binary treatment and the forcing variable perfectly predicts treatment receipt. We now discuss two variants of this base case: 1) when there is so-called "imperfect compliance" of the rule, and 2) when the treatment of interest is a continuous variable.

3.4.1 Imperfect Compliance: the "Fuzzy" RD

In many settings of economic interest, treatment is determined partly by whether the forcing variable crosses a cutoff point. This situation is very important in practice for a variety of reasons, including cases of imperfect take-up by program participants or when factors other than the threshold rule affect the probability of program participation. Starting with Trochim (1984), this setting has been referred to as a "fuzzy" RD design. In the case we have discussed so far – the "sharp" RD design – the probability of treatment jumps from 0 to 1 when X crosses the threshold c. The fuzzy RD design allows for a smaller jump in the probability of assignment to the treatment at the threshold and only requires

$$\lim_{x \downarrow c} \Pr(D = 1|X = x) \neq \lim_{x \uparrow c} \Pr(D = 1|X = x).$$

[14] For this example, consider the simplified case of a two-party system.

Since the probability of treatment jumps by less than one at the threshold, the jump in the relationship between Y and X can no longer be interpreted as an average treatment effect. As in an instrumental variable setting however, the treatment effect can be recovered by dividing the jump in the relationship between Y and X at c by the fraction induced to take-up the treatment at the threshold – in other words, the discontinuity jump in the relation between D and X. In this setting, the treatment effect can be written as

$$\tau_{\text{F}} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y|X=x] - \lim_{x \uparrow c} \mathbb{E}[Y|X=x]}{\lim_{x \downarrow c} \mathbb{E}[D|X=x] - \lim_{x \uparrow c} \mathbb{E}[D|X=x]},$$

where the subscript "F" refers to the fuzzy RD design.

There is a close analogy between how the treatment effect is defined in the fuzzy RD design and in the well-known "Wald" formulation of the treatment effect in an instrumental variables setting. Hahn et al. (2001) were the first to show this important connection and to suggest estimating the treatment effect using two-stage least-squares (TSLS) in this setting. We discuss estimation of fuzzy RD designs in greater detail in Section 4.3.3.

Hahn et al. (2001) furthermore pointed out that the interpretation of this ratio as a causal effect requires the same assumptions as in Imbens and Angrist (1994). That is, one must assume "monotonicity" (i.e. X crossing the cutoff cannot simultaneously *cause* some units to take up and others to reject the treatment) and "excludability" (i.e. X crossing the cutoff cannot impact Y except through impacting receipt of treatment). When these assumptions are made, it follows that[15]

$$\tau_{\text{F}} = \mathbb{E}[Y(1) - Y(0)|\text{unit is complier}, X = c],$$

where "compliers" are units that receive the treatment when they satisfy the cutoff rule ($X_i \geq c$), but would not otherwise receive it.

In summary, if there is local random assignment (e.g. due to the plausibility of individuals' imprecise control over X), then we can simply apply all of what is known about the assumptions and interpretability of instrumental variables. The difference between the "sharp" and "fuzzy" RD design is exactly parallel to the difference between the randomized experiment with perfect compliance and the case of imperfect compliance, when only the "intent to treat" is randomized.

For example, in the case of imperfect compliance, even if a proposed binary instrument Z is randomized,

[15] See Imbens and Lemieux (2008) for a more formal exposition.

it is necessary to rule out the possibility that Z affects the outcome, outside of its influence through treatment receipt, D. Only then will the instrumental variables estimand – the ratio of the reduced form effects of Z on Y and of Z on D – be properly interpreted as a causal effect of D on Y. Similarly, supposing that individuals do not have precise control over X, it is necessary to assume that whether X crosses the threshold c (the instrument) has no impact on y except by influencing D. Only then will the ratio of the two RD gaps in Y and D be properly interpreted as a causal effect of D on Y.

In the same way that it is important to verify a strong first-stage relationship in an IV design, it is equally important to verify that a discontinuity exists in the relationship between D and X in a fuzzy RD design.

Furthermore, in this binary-treatment-binary-instrument context with unrestricted heterogeneity in treatment effects, the IV estimand is interpreted as the average treatment effect "for the sub-population affected by the instrument," (or LATE). Analogously, the ratio of the RD gaps in Y and D (the "fuzzy design" estimand) can be interpreted as a *weighted* LATE, where the weights reflect the ex-ante likelihood the individual's X is near the threshold. In both cases, an exclusion restriction and monotonicity condition must hold.

3.4.2 Continuous Endogenous Regressor

In a context where the "treatment" is a continuous variable – call it T – and there is a randomized binary instrument (that can additionally be excluded from the outcome equation), an IV approach is an obvious way of obtaining an estimate of the impact of T on Y. The IV estimand is the reduced-form impact of Z on Y divided by the first-stage impact of Z on T.

The same is true for an RD design when the regressor of interest is continuous. Again, the causal impact of interest will still be the ratio of the two RD gaps (i.e. the discontinuities in Y and T).

To see this more formally, consider the model

$$\begin{aligned} Y &= T\gamma + W\delta_1 + U_1 \\ T &= D\phi + W\gamma + U_2 \\ D &= 1[X \geq 0] \\ X &= W\delta_2 + V \end{aligned} \tag{6}$$

which is the same set-up as before, except with the added second equation, allowing for imperfect compliance or other factors (observables W or unobservables U_2) to impact the continuous regressor of interest T. If $\gamma = 0$

and $U_2 = 0$, then the model collapses to a "sharp" RD design (with a continuous regressor).

Note that we make no additional assumptions about U_2 (in terms of its correlation with W or V). We do continue to assume imprecise control over X (conditional on W and U_1, the density of X is continuous).[16]

Given the discussion so far, it is easy to show that

$$E[Y|X=0] - \lim_{\varepsilon \uparrow 0} E[Y|X=\varepsilon] = \left\{ E[T|X=0] - \lim_{\varepsilon \uparrow 0} E[T|X=\varepsilon] \right\} \gamma \tag{7}$$

The left hand side is simply the "reduced form" discontinuity in the relation between y and X. The term preceding γ on the right hand side is the "first-stage" discontinuity in the relation between T and X, which is also estimable from the data. Thus, analogous to the exactly-identified instrumental variable case, the ratio of the two discontinuities yields the parameter γ: the effect of T on Y. Again, because of the added notion of imperfect compliance, it is important to assume that D (X crossing the threshold) does not directly enter the outcome equation.

In some situations, more might be known about the rule determining T. For example, in Angrist and Lavy (1999) and Urquiola and Verhoogen (2007), class size is an increasing function of total school enrollment, except for discontinuities at various enrollment thresholds. But additional information about characteristics such as the slope and intercept of the underlying function (apart from the magnitude of the discontinuity) generally adds nothing to the identification strategy.

To see this, change the second equation in 6 to $T = D\phi + g(X)$ where $g(\cdot)$ is any continuous function in the forcing variable. Equation (7) will remain the same, and thus knowledge of the function $g(\cdot)$is irrelevant for identification.[17]

There is also no need for additional theoretical results in the case when there is individual-level heterogeneity in the causal effect of the continuous regressor T. The local random assignment result allows us to borrow from the existing IV literature and interpret the ratio of the RD gaps as in Angrist and Krueger (1999), except that we need to add the note that all averages are weighted by the *ex ante* relative likelihood that the individual's X will land near the threshold.

[16] Although it would be unnecessary to do so for the identification of γ, it would probably be more accurate to describe the situation of imprecise control with the continuity of the density of X conditional on the three variables (W, U_1, U_2). This is because U_2 is now another variable characterizing heterogeneity in individuals.

[17] And similar to that noted in 3.2.1, the inclusion of a direct effect of X in the outcome equation will not change identification of τ.

3.5 Summary: A Comparison of RD and Other Evaluation Strategies

We conclude this section by comparing the RD design with other evaluation approaches. We believe it is helpful to view the RD design as a distinct approach, rather than as a special case of either IV or matching/regression-control. Indeed, in important ways the RD design is more similar to a randomized experiment, which we illustrate below.

Consider a randomized experiment, where subjects are assigned a random number X, and are given the treatment if $X \geq 0$. By construction, X is independent and not systematically related to any observable or unobservable characteristic determined prior to the randomization. This situation is illustrated in Panel A of Figure 5. The first column shows the relationship between the treatment variable D and X, a step function, going from 0 to 1 at the $X = 0$ threshold. The second column shows the relationship between the observables W and X. This is flat because X is completely randomized. The same is true for the unobservable variable U, depicted in the third column. These three graphs capture the appeal of the randomized experiment: treatment varies while all other factors are kept constant (on average). And even though we cannot directly test whether there are no treatment-control differences in U, we can test whether there are such differences in the observable W.

Now consider an RD (Panel B of Figure 5) where individuals have imprecise control over X. Both W and U may be systematically related to X, perhaps due to the actions taken by units to increase their probability of receiving treatment. Whatever the shape of the relation, as long as individuals have imprecise control over X, the relationship will be continuous. And therefore, as we examine Y near the $X = 0$ cutoff, we can be assured that like an experiment, treatment varies (the first column) while other factors are kept constant (the second and third columns). And, like an experiment, we can test this prediction by assessing whether observables truly are continuous with respect to X (the second column).[18]

We now consider two other commonly-used non-experimental approaches, referring to the model (2):

$$Y = D\tau + W\delta_1 + U$$

$$D = 1[X \geq 0]$$

$$X = W\delta_2 + V$$

[18] We thank an anonymous referee for suggesting these illustrative graphs.

3.5.1 Selection on Observables: Regression Control

The basic idea of the "selection on observables" approach is to adjust for differences in the Ws between treated and control individuals. It is usually motivated by the fact that it seems "implausible" that the unconditional mean Y for the control group represents a valid counterfactual for the treatment group. So it is argued that, *conditional on* W, treatment-control contrasts may identify the (W-specific) treatment effect.

The underlying assumption is that conditional on W, U and V are independent. From this it is clear that

$$\begin{aligned} E[Y|D=1,W=w]-E[Y|D=0,W=w] &= \tau+E[U|W=w,V\geq -w\delta_2]-E[U|W=w,V<-w\delta_2] \\ &= \tau \end{aligned}$$

Two issues arise when implementing this approach. The first is one of functional form: how exactly to control for the Ws? When the Ws take on discrete values, one possibility is to compute treatment effects for each distinct value of W, and then average these effects across the constructed "cells". This will not work when W has continuous elements though, in which case it is necessary to implement multivariate matching, propensity score, or re-weighting procedures.

Irrespective of the functional form issue, there is arguably a more fundamental question of which Ws use in the analysis. While it is tempting to answer "all of them" and hope that more Ws will lead to less biased estimates, this is obviously not necessarily the case. For example, consider estimating the economic returns to graduating high school (versus dropping out). It seems natural to include variables like parents' socioeconomic status, family income, year, and place of birth in the regression. Including more and more family-level Ws will ultimately lead to a "within-family" sibling analysis; extending it even further by including date of birth leads to a "within-twin-pair" analysis. And researchers have been critical – justifiably so – of this source of variation in education. The same reasons causing discomfort about the twin analyses should also cause skepticism about "kitchen sink" multivariate matching/propensity score/regression control analyses.[19]

It is also tempting to believe that if the Ws do a "good job" in predicting D, the selection on observables approach will "work better." But the opposite is true: in the extreme case when the Ws perfectly predict X (and hence D), it is *impossible* to construct a treatment-control contrast for virtually all observations. For each value of W, the individuals will either all be treated or all control. In other words, there will be literally

[19] We thank David Card for pointing out this connection.

no overlap in the support of the propensity score for the treated and control observations. The propensity score would take the values of either 1 or 0.

The "selection on observables" approach is illustrated in Panel C of Figure 5. Observables W can help predict the probability of treatment (first column), but ultimately one must assume that unobservable factors U must be the same for treated and control units for every value of W. That is, the crucial assumption is that the two lines in the third column be on top of each other. Importantly, there is no comparable graph in the second column because there is no way to test the design since all the Ws are used for estimation.

3.5.2 Selection on Unobservables: Instrumental Variables and "Heckit"

A less restrictive modeling assumption is to allow U and V to be correlated, conditional on W. But because of the arguably "more realistic"/flexible data generating process, another assumption is needed to identify τ. One such assumption is that some elements of W (call them Z) enter the selection equation, but not the outcome equation and are also uncorrelated with U. An instrumental variables approach utilizes the fact that

$$\begin{aligned} E[Y|W^*=w^*,Z=z] &= E[D|W^*=w^*,Z=z]\,\tau + w^*\gamma + E[U|W^*=w^*,Z=z] \\ &= E[D|W^*=w^*,Z=z]\,\tau + w^*\gamma + E[U|W^*=w^*] \end{aligned}$$

where W has been split up into W^* and Z. Conditional on $W^*=w^*$, Y only varies with Z because of how D varies with Z. Thus, one identifies τ by "dividing" the reduced form quantity $E[D|W^*=w^*,Z=z]\,\tau$ (which can be obtained by examining the expectation of Y conditional on Z for a particular value w^* of W^*) by $E[D|W^*=w^*,Z=z]$, which is also provided by the observed data. It is common to model the latter quantity as a linear function in Z, in which case the IV estimator is (conditional on W^*) the ratio of coefficients from regressions of Y on Z and D on Z. When Z is binary, this appears to be the only way to identify τ without imposing further assumptions.

When Z is continuous, there is an additional approach to identifying τ. The "Heckit" approach uses the fact that

$$\begin{aligned} E[Y|W^*=w^*,Z=z,D=1] &= \tau + E[U|W=w,V\geq -w\delta_2] \\ E[Y|W^*=w^*,Z=z,D=0] &= E[U|W=w,V<-w\delta_2] \end{aligned}$$

If we further assume a functional form for the joint distribution of U,V, conditional on W^* and Z, then the "control function" terms $E[U|W=w,V\geq -w\delta_2]$ and $E[U|W=w,V<-w\delta_2]$ are functions of observed variables, with the parameters then estimable from the data. It is then possible, for any value of $W=w$, to identify τ as

$$\begin{aligned}(E[Y|W^*=w^*,Z=z,D=1]-E[Y|W^*=w^*,Z=z,D=0])-\\ (E[U|W=w,V\geq -w\delta_2]-E[U|W=w,V<-w\delta_2])\end{aligned} \tag{8}$$

Even if the joint distribution of U,V is unknown, in principle it is still possible to identify τ, if it were possible to choose two different values of Z such that $-w\delta_2$ approaches $-\infty$ and ∞. If so, the last two terms in (8) approach $E[U|W=w]$, and hence cancel one another. This is known as "identification at infinity".

Perhaps the most important assumption that any of these approaches require is the existence of a variable Z that is (conditional on W^*) independent of U.[20] There does not seem to be any way of testing the validity of this assumption. Different, but equally "plausible" Zs may lead to different answers in the same way that including different sets of Ws may lead to different answers in the selection on observables approach.

Even when there is a mechanism that justifies an instrument Z as "plausible," it is often unclear which covariates W^* to include in the analysis. Again, when different sets of W^* lead to different answers, the question becomes which is more plausible: Z is independent of U conditional on W^*, or Z is independent of U conditional on a *subset* of the variables in W^*? While there may be some situations where knowledge of the mechanism dictates which variables to include, in other contexts, it may not be obvious.

The situation is illustrated in Panel D of Figure 5. It is necessary that the instrument Z is related to the treatment (as in the first column). The crucial assumption is regarding the relation between Z and the unobservables U (the third column). In order for an IV or a "Heckit" approach to work, the function in the third column needs to be flat. Of course, we cannot observe whether this is true. Furthermore, in most cases, it is unclear how to interpret the relation between W and Z (second column). Some might argue the observed relation between W and Z should be flat if Z is truly exogenous, and that if Z is highly correlated with W, then it casts doubt on Z being uncorrelated with U. Others will argue that using the second graph as a test is

[20] For IV, violation of the assumption essentially means that Z varies with Y for reasons other than its influence on D. For the textbook "Heckit" approach, it is typically assumed that U,V have the same distribution for any value of Z. It is also clear that the "identification at infinity" approach will only work if Z is uncorrelated with U, otherwise the last two terms in equation (8) would not cancel. See also the framework of Heckman and Vytlacil (2005), which maintains the assumption of the independence of the error terms and Z, conditional on W^*.

only appropriate when Z is truly randomized, and that the assumption invoked is that Z is uncorrelated with U, *conditional on* W . In this latter case, the design seems fundamentally untestable, since all the remaining observable variables (the Ws) are being "used up" for identifying the treatment effect.

3.5.3 RD as "Design" not "Method"

RD designs can be valid under the more general "selection on unobservables" environment, allowing an arbitrary correlation among U, V, and W, but at the same time not requiring an instrument. As discussed above, all that is needed is that conditional on W, U, the density of V is continuous, and the local randomization result follows.

How is an RD design able to achieve this, given these weaker assumptions? The answer lies in what is absolutely necessary in an RD design: observability of the latent index X. Intuitively, given that both the "selection on observables" and "selection on unobservables" approaches rely heavily on modeling X and its components (e.g. which Ws to include, and the properties of the unobservable error V and its relation to other variables, such as an instrument Z), actually *knowing* the value of X ought to help.

In contrast to the "selection on observables" and "selection on unobservables" modeling approaches, with the RD design the researcher can avoid taking any strong stance about what Ws to include in the analysis, since the design *predicts* that the Ws are irrelevant and unnecessary for identification. Having data on Ws is, of course, of some use, as they allow testing of the underlying assumption (described in Section 4.4).

For this reason, it may be more helpful to consider RD designs as a description of a particular *data generating process*, rather than a "method" or even an "approach". In virtually any context with an outcome variable Y, treatment status D, and other observable variables W, in principle a researcher can construct a regression-control or instrumental variables (after designating one of the W variables a valid instrument) estimator, and state that the identification assumptions needed are satisfied.

This is not so with an RD design. Either the situation is such that X is observed, or it is not. If not, then the RD design simply does not apply. If X is observed, then the RD design *forces* the researcher to analyze it in a particular way, and there is little room for researcher discretion – at least from an identification standpoint. The design forces one to examine the discontinuity gap, and predicts that the inclusion of Ws in the analysis should be irrelevant. It also forces the researcher to examine the density of X or the distribution of Ws, conditional on X for discontinuities as a test for validity.

The analogy of the truly randomized experiment is again helpful. Once the researcher is faced with what

she thinks is a properly carried out randomized controlled trial, the analysis is quite straightforward. Even before running the experiment, most researchers agree it would be helpful to display the treatment-control contrasts in the Ws to test whether the randomization was carried out properly, then to show the simple mean comparisons, and finally to verify the inclusion of the Ws make little difference in the analysis, even if they might reduce sampling variability in the estimates.

4 Presentation, Estimation, and Inference

In this section, we systematically discuss the nuts and bolts of implementing RD designs in practice. An important virtue of RD designs is that they provide a very transparent way of graphically showing how the treatment effect is identified. We thus begin the section by discussing how to graph the data in an informative way. We then move to arguably the most important issue in implementing an RD design: the choice of the regression model. We address this by presenting the various possible specifications, discussing how to choose among them, and showing how to compute the standard errors.

We then move to a number of other practical issues that often arise in RD designs. Examples of questions discussed include whether we should control for other covariates and what to do when the running variable is discrete. Finally, we discuss a number of tests to assess the validity of the RD designs, which examine whether covariates are "balanced" on the two sides of the threshold, and whether the density of the running variable is continuous at the threshold.

Throughout this section, we illustrate the various concepts using an empirical example from Lee (2008) who uses an RD design to estimate the causal effect of incumbency in U.S. House elections. We use a sample of 6,558 elections over the 1946-98 period (see Lee (2008) for more detail). The forcing variable in this setting is the fraction of votes awarded to Democrats in the previous election. When the fraction exceeds 50 percent, a Democrat is elected and the party becomes the incumbent party in the next election. Both the share of votes and the probability of winning the next election are considered as outcome variables.

4.1 Graphical Presentation

A major advantage of the RD design over competing methods is its transparency, which can be illustrated using graphical methods. A standard way of graphing the data is to divide the forcing variable into a number of bins, making sure there are two separate bins on each side of the cutoff point (to avoid having treated and

untreated observations mixed together in the same bin). Then, the average value of the outcome variable can be computed for each bin and graphed against the mid-points of the bins.

More formally, for some bandwidth h, and for some number of bins K_0 and K_1 to the left and right of the cutoff value, respectively, the idea is to construct bins $(b_k, b_{k+1}]$, for $k = 1, \ldots, K = K_0 + K_1$, where

$$b_k = c - (K_0 - k + 1) \cdot h.$$

The average value of the outcome variable in the bin is

$$\overline{Y}_k = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Y_i \cdot 1\{b_k < X_i \leq b_{k+1}\}.$$

It is also useful to calculate the number of observations in each bin

$$N_k = \sum_{i=1}^{N} 1\{b_k < X_i \leq b_{k+1}\},$$

to detect a possible discontinuity in the forcing variable at the threshold, which would suggest manipulation.

There are several important advantages in graphing the data this way before starting to run regressions to estimate the treatment effect. First, the graph provides a simple way of visualizing what the functional form of the regression function looks like on either side of the cutoff point. Since the mean of Y in a bin is, for non-parametric kernel regression estimators, evaluated at the bin mid-point using a rectangular kernel, the set of bin means literally represent non-parametric estimates of the regression function. Seeing what the non-parametric regression looks like can then provide useful guidance in choosing the functional form of the regression models.

A second advantage is that comparing the mean outcomes just to the left and right of the cutoff point provides an indication of the magnitude of the jump in the regression function at this point, i.e. of the treatment effect. Since an RD design is "as good as a randomized experiment" right around the cutoff point, the treatment effect could be computed by comparing the average outcomes in "small" bins just to the left and right of the cutoff point. If there is no visual evidence of a discontinuity in a simple graph, it is unlikely the formal regression methods discussed below will yield a significant treatment effect.

A third advantage is that the graph also shows whether there are unexpected comparable jumps at other points . If such evidence is clearly visible in the graph and cannot be explained on substantive grounds, this

calls into question the interpretation of the jump at the cutoff point as the causal effect of the treatment. We discuss below several ways of testing explicitly for the existence of jumps at points other than the cutoff .

Note that the visual impact of the graph is typically enhanced by also plotting a relatively flexible regression model, such as a polynomial model, which is a simple way of smoothing the graph. The advantage of showing both the flexible regression line and the unrestricted bin means is that the regression line better illustrates the shape of the regression function and the size of the jump at the cutoff point, and laying this over the unrestricted means gives a sense of the underlying noise in the data.

Of course, if bins are too narrow the estimates will be highly imprecise. If they are too wide, the estimates may be biased as they fail to account for the slope in the regression line (negligible for very narrow bins). More importantly, wide bins make the comparisons on both sides of the cutoff less credible, as we are no longer comparing observations just to the left and right of the cutoff point.

This raises the question of how to choose the bandwidth (the width of the bin) . In practice, this is typically done informally by trying to pick a bandwidth that makes the graphs look informative in the sense that bins are wide enough to reduce the amount of noise, but narrow enough to compare observations "close enough" on both sides of the cutoff point. While it is certainly advisable to experiment with different bandwidths and see how the corresponding graphs look, it is also useful to have more formal guidance in the selection process.

One approach to bandwidth choice is based on the fact that, as discussed above, the mean outcomes by bin correspond to kernel regression estimates with a rectangular kernel. There is a well-developed literature on the choice of bandwidth in non-parametric regression models that we discuss below in the context of local linear regressions, arguing that a method of choice is a cross-validation (leave one out) procedure. Since the standard kernel regression is a special case of a local linear regression where the slope term is equal to zero, the cross-validation procedure described in more detail in section 4.3.1 can also be used here by constraining the slope term to equal zero.[21]

In practice, however, a range a bandwidths often yield similar values of the cross-validation function

[21] In Section 4.3.1, we consider the cross-validation function $\text{CV}_Y(h) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}(X_i)\right)^2$ where $\widehat{Y}(X_i)$ is the predicted value of Y_i based on a regression using observations with a bin of width h on either the left (for observations on left of the cutoff) or the right (for observations on the right of the cutoff) of observation i, but not including observation i itself. In the context of the graph discussed here, the only modification to the cross-validation function is that the predicted value $\widehat{Y}(X_i)$ is based only on a regression with a constant term, which means $\widehat{Y}(X_i)$ is the average value of Y among all observations in the bin (excluding observation i). Note that this is slightly different from the standard cross-validation procedure in kernel regressions where the left-out observation is in the middle instead of the edge of the bin (see, for example, Blundell and Duncan (1998)). Our suggested procedure is arguably better suited to the RD context since estimation of the treatment effect takes place at boundary points.

(see below). A researcher may, therefore, want to use some discretion in choosing a bandwidth that provides a particularly compelling illustration of the RD design. An alternative approach is to choose a bandwidth based on a more heuristic visual inspection of the data, and then perform some formal tests to make sure this informal choice is not clearly rejected .

We suggests two such tests. Consider the case where one has decided to use K' bins based on a visual inspection of the data. The first test is a standard F-test comparing the fit of a regression model with K' bin dummies to one where we further divide each bin into two equal sized smaller bins, i.e. increase the number of bins to $2K'$ (reduce the bandwidth from h' to $h'/2$). Since the model with K' bins is nested in the one with $2K'$ bins, a standard F-test with K' degrees of freedom can used. If the null hypothesis is not rejected, this means we are not oversmoothing the data by using only K' bins.

Another test is based on the idea that if the bins are "narrow enough", then there should not be a systematic relationship between Y and X within each bin. Otherwise, this suggests the bin is too wide and that the the mean value of Y over the whole bin is not representative of the value of Y at the boundaries of the bin. In particular, when this happens in the two bins next to the cutoff point, a simple comparison of the two bin means yields a biased estimate of the treatment effect. A simple test for this consists of adding a set of interactions between the bin dummies and X to a base regression of Y on the set of bin dummies, and testing whether the interactions are jointly significant. The test statistic once again follows a F distribution with K' degrees of freedom.

Figures 6 and 7 show the graphs for the share of Democrat vote in the next election and the probability of Democrats winning the next election, respectively. Three sets of graphs with different bandwidths are reported using a bandwidth of 0.02 in Figures 6a and 7a, 0.01 in Figures 6b and 7b, and 0.005 in Figures 6c and 7c. In all cases, we also show the fitted values from a quartic regression model estimated separately on each side of the cutoff point. Note that the forcing variable is normalized as the difference between the share of vote to Democrats and Republicans in the previous election. This means that a Democrat is the incumbent when the forcing variable exceeds zero. We also limit the range of the graphs to winning margins of 50 percent or less (in absolute terms) as data become relatively sparse for larger winning (or losing) margins.

All graphs show clear evidence of a discontinuity at the cutoff point. While the graphs are all quite informative, the ones with the smallest bandwidth (0.005, Figure 6c and 7c) are more noisy and likely provide too many data points (200) for optimal visual impact.

The results of the bandwidth selection procedures are presented in Table 1. Panel A shows the cross-

validation procedure always suggests using a bandwidth of 0.02 or more, which corresponds to similar or wider bins than those used in Figures 6a and 7a (those with the largest bins) . This is true irrespective of whether we pick a separate bandwidth on each side of the cutoff (first two rows of the panel), or pick the bandwidth that minimizes the cross-validation function for the entire date range on both the left and right sides of the cutoff. In the case where the outcome variable is winning the next election, the cross-validation procedure for the data to the right of the cutoff point and for the entire range suggests using a very wide bin (0.049) that would only yield about 10 bins on each side of the cutoff.

As it turns out, the cross-validation function for the entire data range has two local minima at 0.021 and 0.049 that correspond to the optimal bandwidths on the left and right hand side of the cutoff. This is illustrated in Appendix Figure A2, which plots the cross-validation function as a function of the bandwidth. By contrast, the cross-validation function is better behaved and shows a global minimum around 0.020 when the outcome variable is the vote share (Figure A1). For both outcome variables, the value of the cross-validation function grows quickly for bandwidths smaller than 0.02, suggesting that the graphs with narrower bins (Figures 6b, 6c, 7b, and 7c) are too noisy.

Panel B of Table 1 shows the results of our two suggested specification tests. The tests based on doubling the number of bins and running regressions within each bin yield remarkably similar results. Generally speaking, the results indicate that only fairly wide bins are rejected. Looking at both outcome variables, the tests systematically reject models with bandwidths of 0.05 or more (20 bins over the -0.5 to 0.5 range). The models are never rejected for either outcome variable once we hit bandwidths of 0.02 (50 bins) or less. In practice, the testing procedure rules out bins that are larger than those reported in Figures 6 and 7.

At first glance, the results in the two panels of Table 1 appear to be contradictory. The cross-validation procedure suggests bandwidths ranging from 0.02 to 0.05, while the bin and regression tests suggests than almost all bandwidth of less than 0.05 is acceptable. The reason for this discrepancy is that while the cross-validation procedure tries to balance precision and bias, the bin and regression tests only deal with the "bias" part of the equation by checking whether the value of Y is more or less constant within a given bin. Models with small bins easily pass this kind of test, although they may yield a very noisy graph. One alternative approach is to choose the largest possible bandwidth that passes the bin and the regression test, which turns out to be 0.033 in Table 1, a bandwidth that is within the range of those suggested by the cross-validation procedure.

From a practical point of view, it seems to be the case that formal procedures, and in particular cross-

validation, suggest bandwidths that are wider than those one would likely choose based on a simple visual examination of the data. In particular, both Figure 6b and 7b (bandwidth of 0.01) look visually acceptable but are clearly not recommended on the basis of the cross-validation procedure. This likely reflects the fact that one important goal of the graph is to show how the raw data look, and too much smoothing would defy the purpose of such a data illustration exercise. Furthermore, the regression estimates of the treatment effect accompanying the graphical results are a formal way of smoothing the data to get precise estimates. This suggests that there is probably little harm in undersmoothing (relative to what formal bandwidth selection procedures would suggest) to better illustrate the variation in the raw data when graphically illustrating an RD design.

4.2 Regression Methods

4.2.1 Parametric or Non-parametric Regressions?

When we introduced the RD design in Section 2, we followed Thistlethwaite and Campbell (1960) in assuming that the underlying regression model was linear in the forcing variable X:

$$Y = \alpha + D\tau + X\beta + \varepsilon.$$

In general, as in any other setting, there is no particular reason to believe that the true model is linear. The consequences of using an incorrect functional form are more serious in the case of RD designs however, since misspecification of the functional form typically generates a bias in the treatment effect, τ.[22] This explains why, starting with Hahn et al. (2001), the estimation of RD designs have generally been viewed as a nonparametric estimation problem.

This being said, applied papers using the RD design often just report estimates from parametric models. Does this mean that these estimates are incorrect? Should all studies use non-parametric methods instead? As we pointed out in the introduction, we think that the distinction between parametric and non-parametric methods has sometimes been a source of confusion to practitioners. Before covering in detail the practical issues involved in the estimation of RD designs, we thus provide some background to help clarify the in-

[22] By contrast, when one runs a linear regression in a model where the true functional form is nonlinear, the estimated model can still be interpreted as a linear predictor that minimizes specification errors. But since specification errors are only minimized globally, we can still have large specification errors at specific points including the cutoff point and, therefore, a large bias in RD estimates of the treatment effect.

sights provided by non-parametric analysis, while also explaining why, in practice, RD designs can still be implemented using "parametric" methods.

Going beyond simple parametric linear regressions when the true functional form is unknown is a well-studied problem in econometrics and statistics. A number of non-parametric methods have been suggested to provide flexible estimates of the regression function. As it turns out, however, the RD setting poses a particular problem because we need to estimate regressions at the cutoff point. This results in a "boundary problem" that causes some complications for non-parametric methods.

From an applied perspective, a simple way of relaxing the linearity assumption is to include polynomials functions of X in the regression model. This corresponds to the series estimation approach often used in non-parametric analysis. A possible disadvantage of the approach, however, is that it provides global estimates of the regression function over all values of X, while the RD design depends instead on local estimates of the regression function at the cutoff point. The fact that polynomial regression models use data far away from the cutoff point to predict the value of Y at the cutoff point is not intuitively appealing. This being said, trying more flexible specification by adding polynomials in X as regressors is an important and useful way of assessing the robustness of the RD estimates of the treatment effect.

The other leading non-parametric approach is kernel regressions. Unlike series (polynomial) estimators, the kernel regression is fundamentally a local method well suited for estimating the regression function at a particular point. Unfortunately, this property does not help very much in the RD setting because the cutoff represents a boundary point where kernel regressions perform poorly.

These issues are illustrated in Figure 2, which shows a situation where the relationship between Y and X (under treatment or control) is non-linear. First, consider the point D located away from the cutoff point. The kernel estimate of the regression of Y on X at $X = X_D$ is simply a local mean of Y for values of X close to X_D. The kernel function provides a way of computing this local average by putting more weight on observations with values of X close to X_D than on observations with values of X far away from X_D . Following Imbens and Lemieux (2008), we focus on the convenient case of the rectangular kernel. In this setting, computing kernel regressions simply amounts to computing the average value of Y in the bin illustrated in Figure 2. The resulting local average is depicted as the horizontal line EF, which is very close to true value of Y evaluated at $X = X_D$ on the regression line.

Applying this local averaging approach is problematic, however, for the RD design. Consider estimating the value of the regression function just on the right of the cutoff point. Clearly, only observations on the

right of the cutoff point that receive the treatment should be used to compute mean outcomes on the right hand side. Similarly, only observations on the left of the cutoff point that do not receive the treatment should be used to compute mean outcomes on the left hand side. Otherwise, regression estimates would mix observations with and without the treatment, which would invalidate the RD approach.

In this setting, the best thing is to compute the average value of Y in the bin just to the right and just to the left of the cutoff point. These two bins are shown in Figure 2. The RD estimate based on kernel regressions is then equal to $B' - A'$. In this example where the regression lines are upward sloping, it is clear, however, that the estimate $B' - A'$ overstates the true treatment effect represented as the difference $B - A$ at the cutoff point. In other words, there is a systematic bias in kernel regression estimates of the treatment effect. Hahn et al. (2001) provide a more formal derivation of the bias (see also Imbens and Lemieux (2008) for a simpler exposition when the kernel is rectangular). In practical terms, the problem is that in finite samples the bandwidth has to be large enough to encompass enough observations to get a reasonable amount of precision in the estimated average values of Y. Otherwise, attempts to reduce the bias by shrinking the bandwidth will result in extremely noisy estimates of the treatment effect.[23]

As a solution to this problem, Hahn et al. (2001) suggests running local linear regressions to reduce the importance of the bias. In our setup with a rectangular kernel, this suggestion simply amounts to running standard linear regressions within the bins on both sides of the cutoff point to better predict the value of the regression function right at the cutoff point. In this example, the regression lines within the bins around the cutoff point are close to linear. It follows that the predicted values of the local linear regressions at the cutoff point are very close to the true values of A and B. Intuitively, this means that running local linear regressions instead of just computing averages within the bins reduces the bias by an order of magnitude. Indeed, Hahn et al. (2001) show that the remaining bias is of an order of magnitude lower, and is comparable to the usual bias in kernel estimation at interior points like D (the small difference between the horizontal line EF and the true value of the regression line evaluated at D).

In the literature on non-parametric estimation at boundary points, local linear regressions have been introduced as a means of reducing the bias in standard kernel regression methods.[24] One of the several

[23]The trade-off between bias and precision is a fundamental feature of kernel regressions. A larger bandwidth yields more precise, but potentially biased, estimates of the regression. In an interior point like D, however, we see that the bias is of an order of magnitude lower that at the cutoff (boundary) point. In more technical terms, it can be shown (see Hahn et al. (2001) or Imbens and Lemieux (2008)) that the usual bias is of order h^2 at interior points, but of order h at boundary point, where h is the bandwidth. In other words, the bias dies off much more quickly when h goes to zero when we are at interior, as opposed to boundary, points.

[24]See Fan and Gijbels (1996).

contributions of Hahn et al. (2001) is to show how the same bias-reducing procedure should also be applied to the RD design. We have shown here that, in practice, this simply amounts to applying the original insight of Thistlethwaite and Campbell (1960) to a narrower window of observations around the cutoff point. When one is concerned the regression function is not linear over the whole range of X, a highly sensible procedure is, thus, to restrict the estimation range to values closer to the cutoff point where the linear approximation of the regression line is less likely to result in large biases in the RD estimates. In practice, many applied papers present RD estimates with varying window widths to illustrate the robustness (or lack thereof) of the RD estimates to specification issues. It is comforting to know that this common empirical practice can be justified on more formal econometric grounds like those presented by Hahn et al. (2001). The main conclusion we draw from this discussion of non-parametric methods is that it is essential to explore how RD estimates are robust to the inclusion of higher order polynomial terms (the series or polynomial estimation approach) and to changes in the window width around the cutoff point (the local linear regression approach).

4.3 Estimating the Regression

A simple way of implementing RD designs in practice is to estimate two separate regressions on each side of the cutoff point. In terms of computations, it is convenient to subtract the cutoff value from the covariate, i.e. transform X to $X - c$, so the intercepts of the two regressions yield the value of the regression functions at the cutoff point.

The regression model on the left hand side of the cutoff point ($X < c$) is

$$Y = \alpha_l + f_l(X - c) + \varepsilon,$$

while the regression model on the right hand side of the cutoff point ($X \geq c$) is

$$Y = \alpha_r + f_r(X - c) + \varepsilon,$$

where $f_l(\cdot)$ and $f_r(\cdot)$ are functional forms that we discuss later. The treatment effect can then be computed as the difference between the two regressions intercepts, α_r and α_l, on the two sides of the cutoff point. A more direct way of estimating the treatment effect is to run the pooled regression on both sides of the cutoff

point:

$$Y = \alpha_l + \tau \cdot D + f(X-c) + \varepsilon,$$

where $\tau = \alpha_r - \alpha_l$ and $f(X-c) = f_l(X-c) + D \cdot [f_r(X-c) - f_l(X-c)]$. One advantage of the pooled approach is that it directly yields estimates and standard errors of the treatment effect τ. Note, however, that it is recommended to let the regression function differ on both sides of the cutoff point by including interaction terms between D and X. For example, in the linear case where $f_l(X-c) = \beta_l \cdot (X-c)$ and $f_r(X-c) = \beta_r \cdot (X-c)$, the pooled regression would be

$$Y = \alpha_l + \tau \cdot D + \beta_l \cdot (X-c) + (\beta_r - \beta_l) \cdot D \cdot (X-c) + \varepsilon.$$

The problem with constraining the slope of the regression lines to be the same on both sides of the cutoff ($\beta_r = \beta_l$) is best illustrated by going back to the separate regressions above. If we were to constrain the slope to be identical on both sides of the cutoff, this would amount to using data on the right hand side of the cutoff to estimate α_l, and vice versa. Remember from Section 2 that in an RD design, the treatment effect is obtained by comparing conditional expectations of Y when approaching from the left ($\alpha_l = \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]$) and from the right ($\alpha_r = \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x]$) of the cutoff. Constraining the slope to be the same would thus be inconsistent with the spirit of the RD design as data from the right of the cutoff would be used to estimate α_l, which is defined as a limit when approaching from the left of the cutoff, and vice versa.

In practice, however, estimates where the regression slope or, more generally, the regression function $f(X-c)$ are constrained to be the same on both sides of the cutoff point are often reported. One possible justification for doing so is that if the functional form is indeed the same on both sides of the cutoff, then more efficient estimates of the treatment effect τ are obtained by imposing that constraint. Such a constrained specification should only be viewed, however, as an additional estimate to be reported for the sake of completeness. It should not form the core basis of the empirical approach.

4.3.1 Local Linear Regressions and Bandwidth Choice

As discussed above, local linear regressions provide a non-parametric way of consistently estimating the treatment effect in an RD design (Hahn et al. (2001), Porter (2003)). Following Imbens and Lemieux (2008), we focus on the case of a rectangular kernel which amounts to estimating a standard regression over a window of width h on both sides of the cutoff point. While other kernels (triangular, Epanechnikov, etc.) could also

be used, the choice of kernel typically has little impact in practice. As a result, the convenience of working with a rectangular kernel compensates for efficiency gains that could be achieved using more sophisticated kernels.[25]

The regression model on the left hand side of the cutoff point is

$$Y = \alpha_l + \beta_l \cdot (X - c) + \varepsilon, \text{where } c - h \leq X < c,$$

while the regression model on the right hand side of the cutoff point is

$$Y = \alpha_r + \beta_r \cdot (X - c) + \varepsilon, \text{where } c \leq X \leq c + h.$$

As before, it is also convenient to estimate the pooled regression

$$Y = \alpha_l + \tau \cdot D + \beta_l \cdot (X - c) + (\beta_r - \beta_l) \cdot D \cdot (X - c) + \varepsilon, \text{where } c - h \leq X \leq c + h,$$

since the standard error of the estimated treatment effect can be directly obtained from the regression.

While it is straightforward to estimate the linear regressions within a given window of width h around the cutoff point, a more difficult question is how to choose this bandwidth. In general, choosing a bandwidth in non-parametric estimation involves finding an optimal balance between precision and bias. One the one hand, using a larger bandwidth yields more precise estimates as more observations are available to estimate the regression. On the other hand, the linear specification is less likely to be accurate when a larger bandwidth is used, which can bias the estimate of the treatment effect. If the underlying conditional expectation is not linear, the linear specification will provide a close approximation over a limited range of values of X (small bandwidth), but an increasingly bad approximation over a larger range of values of X (larger bandwidth).

As the number of observations available increases, it becomes possible to use an increasingly small bandwidth since linear regressions can be estimated relatively precisely over even a small range of values

[25] It has been shown in the statistics literature (Fan and Gijbels (1996)) that a triangular kernel is optimal for estimating local linear regressions at the boundary. As it turns out, the only difference between regressions using a rectangular or a triangular kernel is that the latter puts more weight (in a linear way) on observations closer to the cutoff point. It thus involves estimating a weighted, as opposed to an unweighted, regression within a bin of width h. An arguably more transparent way of putting more weight on observations close to the cutoff is simply to re-estimate a model with a rectangular kernel using a smaller bandwidth. In practice, it is therefore simpler and more transparent to just estimate standard linear regressions (rectangular kernel) with a variety of bandwidths, instead of trying out different kernels corresponding to particular weighted regressions that are more difficult to interpret.

of X. As it turns out, Hahn et al. (2001) show the optimal bandwidth is proportional to $N^{-1/5}$, which corresponds to a fairly slow rate of convergence to zero. For example, this suggests that the bandwidth should only be cut in half when the sample size increases by a factor of 32 (2^5). For technical reasons, however, it is preferable to undersmooth by shrinking the bandwidth at a faster rate requiring that $h \propto N^{-\delta}$ with $1/5 < \delta < 2/5$, in order to eliminate an asymptotic bias that would remain when $\delta = 1/5$. In the presence of this bias, the usual formula for the variance of a standard least square estimator would be invalid.[26]

In practice however, knowing at what rate the bandwidth should shrink in the limit does not really help since only one actual sample with a given number of observations is available. The importance of undersmoothing only has to do with a thought experiment that is important for computing asymptotically correct standard errors, but not for choosing a particular bandwidth in a particular sample.[27]

In the econometrics and statistics literature, two procedures are generally considered for choosing bandwidths. The first procedure consists of characterizing the optimal bandwidth in terms of the unknown joint distribution of all variables. The relevant components of this distribution can then be estimated and plugged into the optimal bandwidth function.[28] In the context of local linear regressions, Fan and Gijbels (1996) show this involves estimating a number of parameters including the curvature of the regression function. In practice, this can be done in two steps. In step one, a rule-of-thumb (ROT) bandwidth is estimated over the whole relevant data range. In step two, the ROT bandwidth is used to estimate the optimal bandwidth right at the cutoff point. For the rectangular kernel, the ROT bandwidth is given by:

$$h_{\text{ROT}} = 2.702 \cdot \left(\frac{\widetilde{\sigma}^2 R}{\sum_{i=1}^{N} \{\widetilde{m}''(x_i)\}^2} \right)^{1/5},$$

where $\widetilde{m}''(\cdot)$ is the second derivative (curvature) of an estimated regression of Y on X, $\widetilde{\sigma}$ is the estimated

[26] See Hahn et al. (2001) and Imbens and Lemieux (2008) for more details.

[27] The main purpose of asymptotic theory is to use the large sample properties of estimators to approximate the distribution of an estimator in the real sample being considered. The issue is a little more delicate in a non-parametric setting where one also has to think about how fast the bandwidth should shrink when the sample size approaches infinity. The point about undersmoothing is simply that one unpleasant property of the optimal bandwidth is that it does not yield the convenient least squares variance formula. But this can be fixed by shrinking the bandwidth a little faster as the sample size goes to infinity. Strictly speaking, this is only a technical issue with how to perform the thought experiment (what happens when the sample size goes to infinity?) required for using asymptotics to approximate the variance of the RD estimator in the actual sample. This does not say anything about what bandwidth should be chosen in the actual sample available for implementing the RD design.

[28] A well known example of this procedure is the "rule-of-thumb" bandwidth selection formula in kernel density estimation where an estimate of the dispersion in the variable (standard deviation or the interquartile range), $\widehat{\sigma}$, is plugged into the formula $0.9 \cdot \widehat{\sigma} \cdot N^{-1/5}$. Silverman (1986) shows that this formula is the closed form solution for the optimal bandwidth choice problem when both the actual density and the kernel are Gaussian. See also Imbens and Kalyanaraman (2009), who derive an optimal bandwidth for this RD setting, and propose a data-dependent method for choosing the bandwidth.

standard error of the regression, R is the range of the forcing variable over which the regression is estimated, and the constant 2.702 is a number specific to the rectangular kernel. A similar formula can be used for the optimal bandwidth, except both the regression standard error and the average curvature of the regression function are estimated locally around the cutoff point. For the sake of simplicity, we only compute the ROT bandwidth in our empirical example. Following the common practice in studies using these bandwidth selection methods, we also use a quartic specification for the regression function.[29]

The second approach is based on a cross-validation procedure. In the case considered here, Ludwig and Miller (2005) and Imbens and Lemieux (2008) have proposed a "leave one out" procedure aimed specifically at estimating the regression function at the boundary. The basic idea behind this procedure is the following. Consider an observation i. To see how well a linear regression with a bandwidth h fits the data, we run a regression with observation i left out and use the estimates to predict the value of Y at $X = X_i$. In order to mimic the fact that RD estimates are based on regression estimates at the boundary, the regression is estimated using only observations with values of X on the left of X_i ($X_i - h \leq X < X_i$) for observations on the left of the cutoff point ($X_i < c$). For observations on the right of the cutoff point ($X_i \geq c$), the regression is estimated using only observations with values of X on the right of X_i ($X_i < X \leq X_i + h$).

Repeating the exercise for each and every observation, we get a whole set of predicted values of Y that can be compared to the actual values of Y. The optimal bandwidth can be picked by choosing the value of h that minimizes the mean square of the difference between the predicted and actual value of Y.

More formally, let $\widehat{Y}(X_i)$ represent the predicted value of Y obtained using the regressions described above. The cross-validation criterion is defined as

$$\text{CV}_Y(h) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}(X_i)\right)^2, \tag{9}$$

with the corresponding cross-validation choice for the bandwidth

$$h_{\text{CV}}^{\text{opt}} = \arg\min_h \text{CV}_Y(h).$$

Imbens and Lemieux (2008) discuss this procedure in more detail and point out that since we are primarily

[29] See McCrary and Royer (2003) for an example where the bandwidth is selected using the ROT procedure (with a triangular kernel), and McCall and Desjardins (2008) for an example where the second step optimal bandwidth is computed (for the Epanechnikov kernel). Both papers use a quartic regression function $m(x) = \beta_0 + \beta_1 x + ... + \beta_4 x^4$, which means that $m''(x) = 2\beta_2 + 6\beta_3 x + 12\beta_4 x^2$. Note that the quartic regressions are estimated separately on both sides of the cutoff.

interested in what happens around the cutoff, it may be advisable to only compute $CV_Y(h)$ for a subset of observations with values of X close enough to the cutoff point. For instance, only observations with values of X between the median value of X to the left and right of the cutoff could be used to perform the cross-validation.

The second row of Tables 2a and 2b shows the local linear regression estimates of the treatment effect for the two outcome variables (share of vote and winning the next election). We show the estimates for a wide range of bandwidths going from the entire data range (bandwidth of 1 on each side of the cutoff) to a very small bandwidth of 0.01 (winning margins of one percent or less). As expected, the precision of the estimates declines quickly as we approach smaller and smaller bandwidths. Notice also that estimates based on very wide bandwidths (0.5 or 1) are systematically larger than those for the smaller bandwidths (in the 0.05 to 0.25 range) that are still large enough for the estimates to be reasonably precise. A closer examination of Figures 6 and 7 also suggests that the estimates for very wide bandwidths are larger than what the graphical evidence would suggest.[30] This is consistent with a substantial bias for these estimates linked to the fact that the linear approximation does not hold over a wide data range. This is particularly clear in the case of winning the next election where Figure 7 shows some clear curvature in the regression function.

Table 3 shows the optimal bandwidth obtained using the ROT and cross-validation procedure. Consistent with the above discussion, the suggested bandwidth range from 0.14 to 0.28, which is large enough to get precise estimates, but narrow enough to minimize the bias. Two interesting patterns can be observed in Table 3. First, the bandwidth chosen by cross-validation tends to be a bit larger than the one based on the rule-of-thumb. Second, the bandwidth is generally smaller for winning the next election (second column) than for the vote share (first column). This is particularly clear when the optimal bandwidth is constrained to be the same on both sides of the cutoff point. This is consistent with the graphical evidence showing more curvature for winning the next election than the vote share, which calls for a smaller bandwidth to reduce the estimation bias linked to the linear approximation.

Figures A3 and A4 plot the value of the cross-validation function over a wide range of bandwidths. In the case of the vote share where the linearity assumption appears more accurate (Figure 6), the cross-validation function is fairly flat over a sizable range of values for the bandwidth (from about 0.16 to 0.29).

[30] In the case of the vote share, the quartic regression shown in Figure 6 implies a treatment effect of 0.066, which is substantially smaller than the local linear regression estimates with a bandwidth of 0.5 (0.090) or 1 (0.118). Similarly, the quartic regression shown in Figure 7 for winning the next election implies a treatment effect of 0.375, which is again smaller than the local linear regression estimates with a bandwidth of 0.5 (0.566) or 1 (0.689).

This range includes the optimal bandwidth suggested by cross-validation (0.282) at the upper end, and the ROT bandwidth (0.180) at the lower end. In the case of winning the next election (Figure A4), the cross-validation procedure yields a sharper suggestion of optimal bandwidth around 0.15, which is quite close to both the optimal cross-validation bandwidth (0.172) and the ROT bandwidth (0.141).

The main difference between the two outcome variables is that larger bandwidths start getting penalized more quickly in the case of winning the election (Figure A4) than in the case of the vote share (Figure A3). This is consistent with the graphical evidence in Figures 6 and 7. Since the regression function looks fairly linear for the vote share, using larger bandwidths does not get penalized as much since they improve efficiency without generating much of a bias. But in the case of winning the election where the regression function exhibits quite a bit of curvature, larger bandwidths are quickly penalized for introducing an estimation bias. Since there is a real tradeoff between precision and bias, the cross-validation procedure is quite informative. By contrast, there is not much of a tradeoff when the regression function is more or less linear, which explains why the optimal bandwidth is larger in the case of the vote share.

This example also illustrates the importance of first graphing the data before running regressions and trying to choose the optimal bandwidth. When the graph shows a more or less linear relationship, it is natural to expect different bandwidths to yield similar results and the bandwidth selection procedure not to be terribly informative. But when the graph shows substantial curvature, it is natural to expect the results to be more sensitive to the choice of bandwidth and that bandwidth selection procedures will play a more important role in selecting an appropriate empirical specification.

4.3.2 Order of Polynomial in Local Polynomial Modeling

In the case of polynomial regressions, the equivalent to bandwidth choice is the choice of the order of the polynomial regressions. As in the case of local linear regressions, it is advisable to try and report a number of specifications to see to what extent the results are sensitive to the order of the polynomial. For the same reason mentioned earlier, it is also preferable to estimate separate regressions on the two sides of the cutoff point.

The simplest way of implementing polynomial regressions and computing standard errors is to run a

pooled regression. For example, in the case of a third order polynomial regression, we would have

$$Y = \alpha_l + \tau \cdot D + \beta_{l1} \cdot (X-c) + \beta_{l2} \cdot (X-c)^2 + \beta_{l3} \cdot (X-c)^3$$
$$+ (\beta_{r1} - \beta_{l1}) \cdot D \cdot (X-c) + (\beta_{r2} - \beta_{l2}) \cdot D \cdot (X-c)^2 + (\beta_{r3} - \beta_{l3}) \cdot D \cdot (X-c)^3 + \varepsilon.$$

While it is important to report a number of specifications to illustrate the robustness of the results, it is often useful to have some more formal guidance on the choice of the order of the polynomial. Starting with Van der Klaauw (2002), one approach has been to use a generalized cross-validation procedure suggested in the literature on non-parametric series estimators .[31] One special case of generalized cross-validation used by Black et al. (2007) that we also use in our empirical example is the well known Akaike information criterion (AIC) of model selection. In a regression context, the AIC is given by

$$AIC = N \ln(\widehat{\sigma}^2) + 2p,$$

where $\widehat{\sigma}$ is the standard error of the regression, and p is the number of parameters in the regression model (order of the polynomial plus one for the intercept).

One drawback of this approach is that it does not provide a very good sense of how a particular parametric model (say a cubic model) compares relative to a more general non-parametric alternative. In the context of the RD design, a natural non-parametric alternative is the set of unrestricted means of the outcome variable by bin used to graphically depict the data in Section 4.1. Since one virtue of polynomial regressions is that they provide a smoothed version of the graph, it is natural to ask how well the polynomial model fits the unrestricted graph. A simple way of implementing the test is to add the set of bin dummies to the polynomial regression and jointly test the significance of the bin dummies. For example, in a first order polynomial model (linear regression), the test can be computed by including $K-2$ bin dummies B_k, for $k = 2$ to $K-1$, in the model

$$Y = \alpha_l + \tau \cdot D + \beta_{l1} \cdot (X-c) + (\beta_{r1} - \beta_{l1}) \cdot D \cdot (X-c) + \sum_{k=2}^{K-1} \phi_k B_k + \varepsilon,$$

and testing the null hypothesis that $\phi_2 = \phi_3 = ... = \phi_{K-1} = 0$. Note that two of the dummies are excluded because of collinearity with the constant and the treatment dummy, D.[32] In terms of specification choice

[31] See Blundell and Duncan (1998) for a more general discussion of series estimators.

[32] While excluding dummies for the two bins next to the cutoff point yields more interpretable results (τ remains the treatment effect), the test is invariant to the excluded bin dummies, provided that one excluded dummy is on the left of the cutoff point and

procedure, the idea is to add a higher order term to the polynomial until the bin dummies are no longer jointly significant.

Another major advantage of this procedure is that testing whether the bin dummies are significant turns out to be a test for the presence of discontinuities in the regression function at points other than the cutoff point. In that sense, it provides a falsification test of the RD design by examining whether there are other unexpected discontinuities in the regression function at randomly chosen points (the bin thresholds) . To see this, rewrite $\sum_{k=1}^{K} \phi_k B_k$ as

$$\sum_{k=1}^{K} \phi_k B_k = \phi_1 + \sum_{k=2}^{K} (\phi_k - \phi_{k-1}) B_k^+,$$

where $B_k^+ = \sum_{j=k}^{K} B_j$ is a dummy variable indicating that the observation is in bin k or above, i.e. that the forcing variable X is above the bin cutoff b_k. Testing whether all the $\phi_k - \phi_{k-1}$ are equal to zero is equivalent to testing that all the ϕ_k are the same (the above test), which amounts to testing that the regression line does not jump at the bin thresholds b_k.

Table 2a and 2b show the estimates of the treatment effect for the voting example. For the sake of completeness, a wide range of bandwidths and specifications are presented, along with the corresponding p-values for the goodness-of fit test discussed above (a bandwidth of 0.01 is used for the bins used to construct the test). We also indicate at the bottom of the tables the order of the polynomial selected for each bandwidth using the AIC. Note that the estimates of the treatment effect for the "order zero" polynomials are just comparisons of means on the two sides of the cutoff point, while the estimates for the "order one" polynomials are based on (local) linear regressions.

Broadly speaking, the goodness-of-fit tests do a very good job ruling out clearly misspecified models, like the zero order polynomials with large bandwidths that yield upward biased estimates of the treatment effect. Estimates from models that pass the goodness-of-fit test mostly fall in the 0.05-0.10 range for the vote share (Table 2a) and 0.37-0.57 for the probability of winning (Table 2b). One set of models the goodness-of-fit test does not rule out, however, is higher order polynomial models with small bandwidths that tend to be imprecisely estimated as they "overfit" the data.

Looking informally at both the fit of the model (goodness-of-fit test) and the precision of the estimates (standard errors) suggests the following strategy: use higher order polynomials for large bandwidths of 0.50 and more, lower order polynomials for bandwidths between 0.05 and 0.50, and zero order polynomials

the other one on the right (something standard regression packages will automatically do if all K dummies are included in the regression).

(comparisons of means) for bandwidths of less than 0.05, since the latter specification passes the goodness-of-fit test for these very small bandwidths. Interestingly, this informal approach more or less corresponds to what is suggested by the AIC. In this specific example, it seems that given a specific bandwidth, the AIC provides reasonable suggestions on which order of the polynomial to use.

4.3.3 Estimation in the Fuzzy RD Design

As discussed earlier, in both the "sharp" and the "fuzzy" RD designs, the probability of treatment jumps discontinuously at the cutoff point. Unlike the case of the sharp RD where the probability of treatment jumps from 0 to 1 at the cutoff though, the probability jumps by less than one in the fuzzy RD case. In other words, treatment is not solely determined by the strict cutoff rule in the fuzzy RD design. For example, even if eligibility for a treatment solely depends on a cutoff rule, not all the eligibles may get the treatment because of imperfect compliance. Similarly, program eligibility may be extended in some cases even when the cutoff rule is not satisfied. For example, while Medicare eligibility is mostly determined by a cutoff rule (age 65 or older), some disabled individuals under the age of 65 are also eligible.

Since we have already discussed the interpretation of estimates of the treatment effect in a fuzzy RD design in Section 3.4.1, here we just focus on estimation and implementation issues . The key message to remember from the earlier discussion is that, as in a standard IV framework, the estimated treatment effect can be interpreted as a local average treatment effect provided monotonicity holds.

In the fuzzy RD design, we can write the probability of treatment as

$$\Pr(D = 1|X = x) = \gamma + \delta T + g(x - c),$$

where $T = 1[X \geq c]$ indicates whether the forcing variable exceeds the eligibility threshold c.[33] Note that the sharp RD is a special case where $\gamma = 0$, $g(\cdot) = 0$, and $\delta = 1$. It is advisable to draw a graph for the treatment dummy D as a function of the forcing variable X using the same procedure discussed in Section 4.1. This provides an informal way of seeing how large the jump in the treatment probability δ is at the cutoff point, and what the functional form $g(\cdot)$ looks like.

Since $D = \Pr(D = 1|X = x) + v$, where v is an error term independent of X, the fuzzy RD design can be

[33] Note that this formulation does not impose any restrictions on the probability model since $g(x - c)$ is unrestricted on both sides of the cutoff c, while T is a dummy variable. So there is no need to write the model using a probit or logit formulation.

described by the two equation system:

$$Y = \alpha + \tau D + f(X - c) + \varepsilon, \tag{10}$$

$$D = \gamma + \delta T + g(X - c) + \nu. \tag{11}$$

Looking at these equations suggests estimating the treatment effect τ by instrumenting the treatment dummy D with T. Note also that substituting the treatment determining equation into the outcome equation yields the reduced form

$$Y = \alpha_r + \tau_r T + f_r(X - c) + \varepsilon_r, \tag{12}$$

where $\tau_r = \tau \cdot \delta$. In that setting, τ_r can be interpreted as an "intent-to-treat" effect.

Estimation in the fuzzy RD design can be performed using either the local linear regression approach or polynomial regressions. Since the model is exactly identified, 2SLS estimates are numerically identical to the ratio of reduced form coefficients τ_r/δ, provided that the same bandwidth is used for equations (11) and (12) in the local linear regression case, and that the same order of polynomial is used for $g(\cdot)$ and $f(\cdot)$ in the polynomial regression case.

In the case of the local linear regression, Imbens and Lemieux (2008) recommend using the same bandwidth in the treatment and outcome regression. When we are close to a sharp RD design, the function $g(\cdot)$ is expected to be very flat and the optimal bandwidth to be very wide. In contrast, there is no particular reason to expect the function $f(\cdot)$ in the outcome equation to be flat or linear, which suggests the optimal bandwidth would likely be less than the one for the treatment equation. As a result, Imbens and Lemieux (2008) suggest focusing on the outcome equation for selecting bandwidth, and then using the same bandwidth for the treatment equation.

While using a wider bandwidth for the treatment equation may be advisable on efficiency grounds, there are two practical reasons that suggest not doing so. First, using different bandwidths complicates the computation of standard errors since the outcome and treatment samples used for the estimation are no longer the same, meaning the usual 2SLS standard errors are no longer valid. Second, since it is advisable to explore the sensitivity of results to changes in the bandwidth, "trying out" separate bandwidths for each of the two equations would lead to a large and difficult-to-interpret number of specifications.

The same broad arguments can be used in the case of local polynomial regressions. In principle, a lower

order of polynomial could be used for the treatment equation (11) than for the outcome equation (12). In practice, however, it is simpler to use the same order of polynomial and just run 2SLS (and use 2SLS standard errors).

4.3.4 How to compute standard errors?

As discussed above, for inference in the sharp RD case we can use standard least squares methods. As usual, it is recommended to use heteroskedasticity-robust standard errors instead of standard least squares standard errors. One additional reason for doing so in the RD case is to ensure the standard error of the treatment effect is the same when either a pooled regression or two separate regressions on each side of the cutoff are used to compute the standard errors. As we just discussed, it is also straightforward to compute standard errors in the fuzzy RD case using 2SLS methods, although robust standard errors should also be used in this case. Imbens and Lemieux (2008) propose an alternative way of computing standard errors in the fuzzy RD case, but nonetheless suggest using 2SLS standard errors readily available in econometric software packages.

One small complication that arises in the non-parametric case of local linear regressions is that the usual (robust) standard errors from least squares are only valid provided that $h \propto N^{-\delta}$ with $1/5 < \delta < 2/5$. As we mentioned earlier, this is not a very important point in practice, and the usual standard errors can be used with local linear regressions.

4.4 Implementing Empirical Tests of RD Validity and Using Covariates

In this part of the section, we describe how to implement tests of the validity of the RD design and how to incorporate covariates in the analysis.

4.4.1 Inspection of the Histogram of the Forcing Variable

Recall that the underlying assumption that generates the local random assignment result is that each individual has imprecise control over the forcing variable, as defined in Section 3.1.1. We cannot test this directly (since we will only observe one observation on the forcing variable per individual at a given point in time), but an intuitive test of this assumption is whether the *aggregate* distribution of the forcing variable is discontinuous, since a mixture of individual-level continuous densities is itself a continuous density.

McCrary (2008) proposes a simple two-step procedure for testing whether there is a discontinuity in the density of the forcing variable. In the first step, the forcing variable is partitioned into equally spaced bins

and frequencies are computed within those bins. The second step treats the frequency counts as a dependent variable in a local linear regression. See McCrary (2008), who adopts the non-parametric framework for asymptotics, for details on this procedure for inference.

As McCrary (2008) points out, this test can fail to detect a violation of the RD identification condition if for some individuals there is a "jump" up in the density, offset by jumps "down" for others, making the aggregate density continuous at the threshold. McCrary (2008) also notes it is possible the RD estimate could remain unbiased, even when there is important manipulation of the forcing variable causing a jump in the density. It should be noted, however, that in order to rely upon the RD estimate as unbiased, one needs to invoke other identifying assumptions and cannot rely upon the mild conditions we focus on in this article.[34]

One of the examples McCrary uses for his test is the voting model of Lee (2008) that we used in the earlier empirical examples. Figure 8 shows a graph of the raw densities computed over bins with a bandwidth of 0.005 (200 bins in the graph), along with a smooth second order polynomial model. Consistent with McCrary (2008), the graph shows no evidence of discontinuity at the cutoff. McCrary also shows that a formal test fails to reject the null hypothesis of no discontinuity in the density at the cutoff.

4.4.2 Inspecting Baseline Covariates

An alternative approach for testing the validity of the RD design is to examine whether the observed baseline covariates are "locally" balanced on either side of the threshold, which should be the case if the treatment indicator is locally randomized.

A natural thing to do is conduct both a graphical RD analysis and a formal estimation, replacing the dependent variable with each of the observed baseline covariates in W. A discontinuity would indicate a violation in the underlying assumption that predicts local random assignment. Intuitively, if the RD design is valid, we *know* that the treatment variable cannot influence variables determined prior to the realization of the forcing variable and treatment assignment; if we observe it does, something is wrong in the design.

If there are many covariates in W, even abstracting from the possibility of misspecification of the functional form, some discontinuities will be statistically significant by random chance. It is thus useful to combine the multiple tests into a single test statistic to see if the data are consistent with no discontinuities for any of the observed covariates. A simple way to do this is with a Seemingly Unrelated Regression (SUR)

[34] McCrary (2008) discusses an example where students who barely fail a test are given extra points so that they barely pass. The RD estimator can remain unbiased if one assumes that those who are given extra points were chosen randomly from those who barely failed.

where each equation represents a different baseline covariate, and then perform an χ^2 test for the discontinuity gaps in all questions being zero. For example, supposing the underlying functional form is linear, one would estimate the system

$$\begin{aligned} w_1 &= \alpha_1 + D\beta_1 + X\gamma_1 + \varepsilon_1 \\ &\vdots \\ w_K &= \alpha_K + D\beta_K + X\gamma_K + \varepsilon_K \end{aligned}$$

and test the hypothesis that $\beta_1, \ldots, \beta_K$ are jointly equal to zero, where we allow the εs to be correlated across the K equations. Alternatively, one can simply use the OLS estimates of $\beta_1, \ldots, \beta_K$ obtained from a "stacked" regression where all the equations for each covariate are pooled together, while D and X are fully interacted with a set of K dummy variables (one for each covariate w_k). Correlation in the error terms can then be captured by clustering the standard errors on individual observations (which appear in the stacked dataset K times). Under the null hypothesis of no discontinuities, the Wald test statistic $N\hat{\beta}'\hat{V}^{-1}\hat{\beta}$ (where $\hat{\beta}$ is the vector of estimates of $\beta_1, \ldots, \beta_K$, and $\hat{V}$ is the cluster-and-heteroskedasticity consistent estimate of the asymptotic variance of $\hat{\beta}$) converges in distribution to a χ^2 with K degrees of freedom.

Of course, the importance of functional form for RD analysis means a rejection of the null hypothesis tells us either that the underlying assumptions for the RD design are invalid, or that at least some of the equations are sufficiently mis-specified and too restrictive, so that nonzero discontinuities are being estimated, even though they do not exist in the population. One could use the parametric specification tests discussed earlier for each of the individual equations to see if mis-specification of the functional form is an important problem. Alternatively, the test could be performed only for observations within a narrower window around the cutoff point, such as the one suggested by the bandwidth selection procedures discussed in Section 4.3.1.

Figure 9 shows the RD graph for a baseline covariate, the Democratic vote share in the election prior to the one used for the forcing variable (four years prior to the current election). Consistent with Lee (2008), there is no indication of a discontinuity at the cutoff. The actual RD estimate using a quartic model is -0.004 with a standard error of 0.014. Very similar results are obtained using winning the election as outcome variable instead (RD estimate of -0.003 with a standard error of 0.017).

4.5 Incorporating Covariates in Estimation

If the RD design is valid, the other use for the baseline covariates is to reduce the sampling variability in the RD estimates. A simple way to do this is to first "residualize" the dependent variable – subtracting from y a prediction based on the baseline covariates W – and then to conduct an RD analysis on the residuals. Intuitively, this procedure nets out the portion of the variation in y we could have predicted using the pre-determined characteristics, making the question whether the treatment variable can explain the remaining residual variation in y. The important thing to keep in mind is that if the RD design is valid, this procedure provides a consistent estimate of the same RD parameter of interest. Indeed, any combination of covariates can be used, and abstracting from funcational form issues, the estimator will be consistent for the same parameter, as discussed above in equation (4).

Specifically, one can regress y on the baseline variables W, and then treat the residual as the dependent variable in an RD analysis. This two-step approach also allows one to perform a graphical analysis of the residual. In order to compute the standard errors accounting for the fact that the first-step was estimated, note that this procedure is equivalent to estimating the system of equations:

$$y - W\pi = D\tau + f(X;\gamma) + \varepsilon$$
$$y = W\pi + u$$

where $f(X;\gamma)$ is a polynomial, and ε and u are assumed to be orthogonal to the regressors in each equation, but possibly correlated with one another. Therefore, as in Section 4.4.2, this amounts to estimating a stacked regression and computing standard errors consistent with arbitrary correlation between ε and u.

This procedure can be justified if one believed that the conditional expectation function of y with respect to X can be characterized by a polynomial (plus treatment dummy), and that the conditional expectation function of W with respect to X can be characterized by a polynomial of the same order. It is clear that if this is the case, then $y - W\pi$ (for any value π) can be characterized by a polynomial of the same order, plus a treatment dummy.

Since this procedure works for any value of π, another natural way to proceed is to simply include W directly on the right hand side of the regression as in the specification:

$$y = D\tau + f(X;\gamma) + W\delta + \varepsilon$$

Note that this procedure is slightly different from the approach suggested above where we only residualize the outcome variable y. While this specification seems particularly restrictive, the procedure is justified under the same assumption as the "residualizing" procedure described above. This is because the coefficient τ can be equivalently obtained, by running a regression of $y - W\delta$ on D and the polynomial $f(X;\gamma)$, where δ is the coefficient from running the full regression. Here, δ replaces π above. Again, this procedure is justified as long as the conditional expectation functions of y and (each element of) W with respect to X can be characterized as the same order polynomial.

To better understand the difference between the two approaches, consider a simplified case where $f(X;\gamma)$ is just a linear function $X\gamma$. In the "residualized" approach one runs a regression of $y - W\pi$ on D and X. In the OLS approach, we (implicitly) run a regression of $y - W\pi$ on $D - W\pi_D$ and $X - W\pi_X$, where π_D and π_W are the coefficients from the regression of D and X, respectively, on W.

Recall, however, that in the RD design treatment is randomly assigned in a close neighborhood around the cutoff point. This means both D and X are locally independent of W, and that $\pi_D = \pi_X = 0$. As a result, the residualized and OLS approaches become equivalent as we near the cutoff point. Furthermore, if the true model is $y = D\tau + X\gamma + W\delta + \varepsilon$, then $\pi = \delta$ since W is uncorrelated with both D and X. In these circumstances, when we residualize y we subtract $W\pi = W\delta$ from y, which is precisely the component generating extra noise on the right hand side of the regression equation.

But as we just discussed in the estimation section, we cannot get too close to the cutoff point because of efficiency issues. As we move further away from the cutoff point, nothing prevents the baseline covariates W from being correlated with both X and D. For instance, consider the case where W represents family background in the test-taking example from Section 3. For a wide enough range of test scores X, we expect family background to be correlated with the test result and, thus, with whether the student passes the test ($D = 1$).

In these circumstances, it is generally more efficient to include the covariates directly in the regression model even though both estimation approaches yield consistent results for the reason mentioned in Section 3.2.3. The problem with the residualized approach is that the regression coefficient of y on W is now given by $\pi = \tau\pi_D + \gamma\pi_X + \delta \neq \delta$. So subtracting $W\pi$ from y no longer just removes the component $W\delta$ responsible for the extra noise on the right hand side of the regression. This explains why OLS is generally expected to yield more efficient estimates than the residualized approach.[35]

[35]If the effect of W is truly linear in the model $y = D\tau + f(X;\gamma) + W\delta + \varepsilon$, then OLS is efficient under the usual assumptions

For the voting example used throughout this paper, Lee (2008) shows that adding a set of covariates essentially has no impact on the RD estimates in the model where the outcome variable is winning the next election. Doing so does not have a large impact on the standard errors either, at least up to the third decimal. Using the procedure based on residuals instead actually slightly increases the second step standard errors - a possibility mentioned above. Therefore in this particular example, the main advantage of using baseline covariates is to help establish the validity of the RD design, as opposed to improving the efficiency of the estimators.

5 Special Cases

In this section, we discuss how the RD design can be implemented in a number of specific cases beyond the one considered up to this point (that of a single cross-section with a continuous forcing variable).

5.1 Discrete Forcing Variable and Specification Errors

Up until now, we have assumed the forcing variable was continuous. In practice, however, X is often discrete. For example, age or date of birth are often only available at a monthly, quarterly, or annual frequency level. Studies relying on an age-based cutoff thus typically rely on discrete values of the age variable when implementing an RD design.

Lee and Card (2008) study this case in detail and make a number of important points. First, with a discrete forcing variable, it is not possible to compare outcomes in very narrow bins just to the right and left of the cutoff point. Consequently, one must use regressions to estimate the conditional expectation of the outcome variable at the cutoff point by extrapolation. As discussed in Section 4, however, in practice we always extrapolate to some extent, even in the case of a continuous forcing variable. So the fact we must do so in the case of a discrete running variable does not introduce particular complications from an econometric point of view, provided the discrete variable is not too coarsely distributed.

Additionally, the various estimation and graphing techniques discussed in Section 4 can readily be used in the case of a discrete running variable. For instance, as with a continuous forcing variable, either local linear regressions or polynomial regressions can be used to estimate the jump in the regression function at the

(ε is $i.i.d.$). However, as in the case of X, there is no particular reason to expect the effect of W is linear. In this case, OLS still globally minimizes specification error in the model, but there is no reason to expect the treatment effect will be estimated efficiently. Therefore, it is no longer clear that OLS necessarily performs better at removing the noise due to the covariates than the residualized approach.

cutoff point. Furthermore, the discreteness of the forcing variable simplifies the problem of bandwidth choice when graphing the data, since in most cases one can simply compute and graph the mean of the outcome variable for each value of the discrete running variable. The fact the variable is discrete also provides a natural way of testing whether the regression model is well specified by comparing the fitted model to the raw dispersion in mean outcomes at each value of the forcing variable. Lee and Card (2008) show that, when errors are homoskedastic, the model specification can be tested using the standard goodness-of-fit statistic

$$G \equiv \frac{(ESS_R - ESS_{UR})/(J-K)}{ESS_{UR}/(N-J)}$$

where ESS_R is the estimated sum of squares of the restricted model (e.g. low order polynomial), while ESS_R is the estimated sum of squares of the unrestricted model where a full set of dummies (for each value of the forcing variable) are included. In this unrestricted model, the fitted regression corresponds to the mean outcome in each cell. G follows a $F(J-K, N-J)$ distribution where J is the number of values taken by the forcing variables and K is the number of parameters of the restricted model.

This test is similar to the test in Section 4 where we suggested including a full set of bin dummies in the regression model and testing whether the bin dummies were jointly significant. The procedure is even simpler here as bin dummies are replaced by dummies for each value of the discrete running variable. In the presence of heteroskedasticity, the goodness-of-fit test can be computed by estimating the model and testing whether a set of dummies for each value of the discrete forcing variable are jointly significant. In that setting, the test statistic follows a chi-square distribution with $J-K$ degrees of freedom.

In Lee and Card (2008), the difference between the true conditional expectation $E[Y|X=x]$ and the estimated regression function forming the basis of the goodness-of-fit test is interpreted as a random specification error that introduces a group structure in the standard errors. One way of correcting the standard errors for group structure is to run the model on cell means.[36] Another way is to "cluster" the standard errors. Note that in this setting, the goodness-of-fit test can also be interpreted as a test of whether standard errors should be adjusted for the group structure. In practice, it is nonetheless advisable to either group the data or cluster the standard errors in micro-data models irrespective of the results of the goodness-of-fit test. The main purpose of the test should be to help choose a reasonably accurate regression model.

[36] When the discrete forcing variable –and the "treatment" dummy solely dependent on this variable– is the only variable used in the regression model, standard OLS estimates will be numerically equivalent to those obtained by running a weighted regression on the cell means, where the weights are the number of observations (or the sum of individual weights) in each cell.

Lee and Card (2008) also discuss a number of issues including what to do when specification errors under treatment and control are correlated, and how to possibly adjust the RD estimates in the presence of specification errors. Since these issues are beyond the scope of this paper, interested readers should consult Lee and Card (2008) for more detail.

5.2 Panel Data and Fixed Effects

In some situations, the RD design will be embedded in a panel context, whereby period by period, the treatment variable is determined according to the realization of the forcing variable X. Again, it seems natural to propose the model

$$y_{it} = D_{it}\tau + f(X_{it};\gamma) + a_i + \varepsilon_{it}$$

(where i and t denote the individuals and time, respectively), and simply estimate a fixed effects regression by including individual dummy variables to capture the unit-specific error component, a_i. It is important to note, however, that including fixed effects is unnecessary for identification in an RD design. This sharply contrasts with a more traditional panel data setting where the error component a_i is allowed to be correlated with the observed covariates, including the treatment variable D_{it}, in which case including fixed effects is essential for consistently estimating the treatment effect τ.

An alternative is to simply conduct the RD analysis for the entire pooled-cross-section dataset, taking care to account for within-individual correlation of the errors over time using clustered standard errors. The source of identification is a comparison between those just below and above the threshold, and can be carried out with a single cross-section. Therefore, imposing a specific dynamic structure introduces more restrictions without any gain in identification.

Time dummies can also be treated like any other baseline covariate. This is apparent by applying the main RD identification result: conditional on what period it is, we are assuming the density of X is continuous at the threshold, and hence, conditional on X, the probability of an individual observation coming from a particular period is also continuous.

We note that it becomes a little bit more awkward to use the justification proposed in Sub-section 4.5 for directly including dummies for individuals and time periods on the right hand side of the regression. This is because the assumption would have to be that the probability that an observation belonged to each individual (or the probability that an observation belonged to each time period) was a polynomial function in X, and

strictly speaking, nontrivial polynomials are not bounded between 0 and 1.

A more practical concern is that inclusion of individual dummy variables may lead to an *increase* in the variance of the RD estimator for another reason. If there is little "within-unit" variability in treatment status, then the variation in the main variable of interest (treatment after partialling out the individual heterogeneity) may be quite small.

Overall, since the RD design is still valid ignoring individual or time effects, then the only rationale for including them is to reduce sampling variance. But there are other ways to reduce sampling variance by exploiting the structure of panel data. For instance, we can treat the lagged dependent variable y_{it-1} as simply another baseline covariate in period t. In cases where y_{it} is highly persistent over time, y_{it-1} may well be a very good predictor and has a very good chance of reducing the sampling error. As we have also discussed earlier, looking at possible discontinuities in baseline covariates is an important test of the validity of the RD design. In this particular case, since y_{it} can be highly correlated with y_{it-1}, finding a discontinuity in y_{it} but not in y_{it-1} would be a strong piece of evidence supporting the validity of the RD design.

In summary, one can utilize the panel nature of the data, by conducting an RD analysis on the entire dataset, using lagged variables as baseline covariates for inclusion as described in Sub-section 4.5. The primary caution in doing this is to ensure that for each period, the included covariates are the variables determined *prior to* the present period's realization of X_{it}.

6 Applications of RD Designs in Economics

In what areas has the RD design been applied in economic research? Where do discontinuous rules come from and where might we expect to find them? In this section, we provide some answers to these questions by providing a survey of the areas of applied economic research that have employed the RD design. Furthermore, we highlight some examples from the literature that illustrate what we believe to be the most important elements of a compelling, "state-of-the-art" implementation of RD.

6.1 Areas of Research Using RD

As we suggested in the introduction, the notion that the RD design has limited applicability to a few specific topics is inconsistent with our reading of existing applied research in economics. Table 4 summarizes our survey of empirical studies on economic topics that have utilized the RD design. In compiling this list, we

searched economics journals as well as listings of working papers from economists, and chose any study that recognized the potential use of an RD design in their given setting. We also included some papers from non-economists when the research was closely related to economic work.

Even with our undoubtedly incomplete compilation of over 60 studies, Table 4 illustrates that RD designs have been applied in many different contexts. Table 4 summarizes the context of the study, the outcome variable, the treatment of interest, and the forcing variable employed.

While the categorization of the various studies into broad areas is rough and somewhat arbitrary, it does appear that a large share come from the area of education, where the outcome of interest is often an achievement test score and the forcing variable is also a test score, either at the individual or group (school) level. The second clearly identifiable group are studies that deal with labor market issues and outcomes. This probably reflects that, within economics, the RD design has so far primarily been used by labor economists, and that the use of quasi-experiments and program evaluation methods in documenting causal relationships is more prevalent in labor economics research.

There is, of course, nothing in the structure of the RD design tying it specifically to labor economics applications. Indeed, as the rest of the table shows, the remaining half of the studies are in the areas of political economy, health, crime, environment, and other areas.

Table 4: Regression Discontinuity Applications in Economics

Study	Context	Outcome(s)	Treatment(s)	Forcing variable(s)
Education				
Angrist and Lavy (1999)	Public Schools (Grades 3-5), Israel	Test scores	Class size	Student Enrollment
Asadullah (2005)	Secondary Schools,Bangladesh	Examination Pass Rate	Class size	Student Enrollment
Bayer et al. (2007)	Valuation of schools and neighborhoods, Northern California	Housing prices, school test scores, demographic characteristics	Inclusion in school attendance region	Geographic location
Black (1999)	Valuation of school quality, Massachusetts	Housing prices	Inclusion in school attendance region	Geographic location
Canton and Blom (2004)	Higher Education, Mexico	University enrollment, GPA, Part-time Employment, Career choice	Student Loan Receipt	Economic need index
Cascio and Lewis (2005)	Teenagers, United States	AFQT test scores	Age at school entry	Birthdate
Chay et al. (2005)	Elementary Schools, Chile	Test scores	Improved infrastructure, more resources	School averages of test scores
Guryan (2001)	State-level equalization: Elementary, Middle Schools, Massachusetts	Spending on schools, test scores	State education aid	Relative average property values
Hoxby (2000)	Elementary Schools, Connecticut	Test scores	Class size	Student Enrollment
Kane (2003)	Higher Education, California	College attendance	Financial aid receipt	Income, Assets, GPA
Lavy (2004)	Secondary Schools, Israel	Test scores	Pay-for-performance incentives	School matriculation rates
Lavy (2006)	Secondary Schools, Tel Aviv	Dropout rates, test scores	School choice	Geographic location

Jacob and Lefgren (2004a)	Elementary Schools, Chicago	Test scores	Teacher training	School averages on test scores
Jacob and Lefgren (2004)	Elementary Schools, Chicago	Test scores	Summer school attendance, grade retention	Standardized test scores
Leuven et al. (Forthcoming)	Primary Schools, Netherlands	Test scores	Extra funding	Percent disadvantaged minority pupils
Matsudaira (2008)	Elementary Schools, Northeastern United States	Test scores	Summer school, grade promotion	Test scores
Urquiola (2006)	Elementary Schools, Bolivia	Test scores	Class size	Student Enrollment
Urquiola and Verhoogen (2007)	Primary Schools, Chile	Household income	Class size	Student Enrollment
Van der Klaauw (2002, 1997)	College enrollment, East Coast College	Enrollment	Financial Aid Offer	SAT scores, GPA
Van der Klaauw (2008a)	Elementary/Middle Schools, New York City	Test scores, student attendance	Title I federal funding	Poverty rates
Labor Market				
Battistin and Rettore (2002)	Job Training,Italy	Employment Rates	Training program (computer skills)	Attitudinal test score
Black et al. (2003, 2007)	UI Claimants, Kentucky	Earnings, Benefit receipt/duration	Mandatory reemployment services (job search assistance)	Profiling score (expected benefit duration)
Card et al. (2007)	Unemployment Benefits, Austria	Unemployment duration	Lump-sum severence pay, extended UI benefits	Months employed, job tenure
Chen and Wilbert van der Klaauw (2008)	Disability Insurance Beneficiaries, United States	Labor force participation	Disability insurance benefits	Age at disability decision
Giorgi (2005)	Welfare-to-work program, United Kingdom	Re-employment probability	Job search assistance, training, education	Age at end of unemployment spell

DiNardo and Lee (2004)	Unionization, United States	Wages, Employment, Output	Union victory in NLRB election	Vote share
Dobkin and Ferreira (2007)	Individuals, California and Texas	Educational Attainment, Wages	Age at school entry	Birthdate
Edmonds (2004)	Child labor supply and school attendance, South Africa	Child labor supply, school attendance	Pension receipt of oldest family member	Age
Hahn et al. (1999)	Discrimination, United States	Minority employment	Coverage of federal antidiscrimination law	Number of employees at firm
Lalive (2008)	Unemployment Benefits, Austria	Unemployment duration	Maximum benefit duration	Age at start of unemployment spell, geographic location
Leuven and Oosterbeek (2004)	Employers, Netherlands	Training, Wages	Business tax deduction, training	Age of employee
Lemieux and Milligan (2008)	Welfare, Canada	Employment, marital status, living arrangements	Cash benefit	Age
Oreopoulos (2006)	Returns to Education, UK	Earnings	Coverage of compulsory schooling law	Birth year
Political Economy				
Albouy (2007b)	Congress, United States	Federal Expenditures	Party control of seat	Vote share in election
Albouy (2007a)	Senate, United States	Roll call votes	Incumbency	Initial vote share
Ferreira and Gyourko (2009)	Mayoral Elections, United States	Local Expenditures	Incumbency	Initial vote share
Lee (2008, 2001)	Congressional elections, United States	Vote share in next election	Incumbency	Initial vote share
Lee et al. (2004)	House of Representatives, United States	Roll call votes	Incumbency	Initial vote share

McCrary (2008)	House of Representatives, United States	N/A	Passing of resolution	Share of roll call vote "Yeay"
Pettersson-Lidbom (2006b)	Local Governments, Sweden and Finland	Expenditures, Tax Revenues	Number of council seats	Population
Pettersson-Lidbom (2006a)	Local Governments, Sweden	Expenditures, Tax Revenues	Left-, right-wing bloc	Left-wing parties' share
Health				
Card and Shore-Sheppard (2004)	Medicaid, United States	Overall insurance coverage	Medicaid Eligibility	Birthdate
Card et al. (2009)	Medicare, United States	Health care utilization	Coverage under Medicare	Age
Carpenter and Dobkin (2009)	Alcohol and Mortality, United States	Mortality	Attaining Minimum Legal Drinking Age	Age
Ludwig and Miller (2007)	Head Start, United States	Child mortality, educational attainment	Head Start funding	County poverty rates
McCrary and Royer (2003)	Maternal Education, United States, California and Texas	Infant health, fertility timing	Age of school entry	Birthdate
Crime				
Berk and DeLeeuw (1999)	Prisoner behavior in California	Inmate misconduct	Prison security levels	Classification score
Berk and Rauma (1983)	Ex-prisoners recidivism, California	Arrest, parole violation	Unemployment insurance benefit	Reported hours of work
Chen and Shapiro (2004)	Ex-prisoners recidivism, United States	Arrest rates	Prison security levels	Classification score
Lee and McCrary (2005)	Criminal Offenders, Florida	Arrest rates	Severity of Sanctions	Age at arrest
Pintoff (2005)	Juvenile Offenders, Washington State	Recidivism	Sentence length	Criminal history score
Environment				

Chay and Greenstone (2003)	Health Effects of Pollution, United States	Infant Mortality	Regulatory status	Pollution levels
Chay and Greenstone (2005)	Valuation of Air Quality, United States	Housing prices	Regulatory status	Pollution levels
Greenstone and Gallagher (2005)	Hazardous Waste, United States	Housing prices	Superfund clean-up status	Ranking of level of hazard
Other				
Battistin and Rettore (2008)	Mexican anti-poverty program (PROGRESA)	School Attendance	Cash grants	Pre-assigned probability of being poor
Buddelmeyer and Skoufias (2004)	Mexican anti-poverty program (PROGRESA)	Child Labor and School Attendance	Cash grants	Pre-assigned probability of being poor
Buettner (2006)	Fiscal Equalization across municipalities, Germany	Business tax rate	Implicit marginal tax rate on grants to localities	Tax base
Card et al. (2008)	Racial segregation, United States	Changes in census tract racial composition	Minority share exceeding “tipping” point	Initial minority share
Edmonds et al. (2004)	Household structure, South Africa	Household composition	Pension receipt of oldest family member	Age
Ferreira (2007)	Residential Mobility, California	Household mobility	Coverage of tax benefit	Age
Pence (2006)	Mortgage credit, United States	Size of Loan	State mortgage credit laws	Geographical location
Pitt and Khandker (1998)	Poor Households, Bangladesh	Labor supply, children school enrollment	Group-based credit program	Acreage of land
Pitt et al. (1999)	Poor Households, Bangladesh	Contraceptive Use, Childbirth	Group-based credit program	Acreage of land

6.2 Sources of Discontinuous Rules

Where do discontinuous rules come from and in what situations would we expect to encounter them? As Table 4 shows, there is a wide variety of contexts where discontinuous rules determine treatments of interest. There are, nevertheless, some patterns that emerge. We organize the various discontinuous rules below.

Before doing so, we emphasize that a good RD analysis – as with any other approach to program evaluation – is careful in clearly spelling out exactly what the treatment is, and whether it is of any real salience, independent of whatever effect it might have on the outcome. For example, when a pre-test score is the forcing variable, we could always define a "treatment" as being "having passed the exam" (with a test score of 50 percent or higher), but this is not a very interesting "treatment" to examine, since it seems nothing more than an arbitrary label. On the other hand, if failing the exam meant not being able to advance to the next grade in school, the actual experience of treated and control individuals is observably different, no matter how large or small the impact on the outcome.

As another example, in the U.S. Congress, a Democrat obtaining the most votes in an election means something real – the Democratic candidate becomes a representative in Congress; otherwise, the Democrat has no official role in the government. But in a three-way electoral race, the treatment of the Democrat receiving the *second-most* number of votes (versus receiving the lowest number) is not likely a treatment of interest: only the first-place candidate is given any legislative authority. In principle, stories could be concocted about the psychological effect of placing second, rather than third in an election, but this would be an example where the salience of the treatment is more speculative than when treatment is a concrete and observable event (e.g. a candidate becoming the sole representative of a constituency).

6.2.1 Necessary Discretization

Many discontinuous rules come about because resources cannot, for all practical purposes, be provided in a continuous manner. For example, a school can only have a whole number of classes per grade. For a fixed level of enrollment, the moment a school adds a single class, the average class size drops. As long as the number of classes is an increasing function of enrollment, there will be discontinuities at enrollments where a teacher is added. If there is a mandated maximum for the student to teacher ratio, this means that these discontinuities will be expected at enrollments that are exact multiples of the maximum. This is the essence of the discontinuous rules used in the analyses of Angrist and Lavy (1999), Asadullah (2005), Hoxby (2000),

Urquiola (2006), and Urquiola and Verhoogen (2007).

Another example of necessary discretization arises when children begin their schooling years. Although there are certainly exceptions, school districts typically follow a guideline that aims to group children together by age, leading to a grouping of children born in year-long intervals, determined by a single calendar date (e.g. Sept. 1). This means children who are essentially of the same age (e.g. those born on Aug. 31 and Sept. 1), start school one year apart. This allocation of students to grade cohorts is used in Cascio and Lewis (2005), Dobkin and Ferreira (2007), and McCrary and Royer (2003).

Choosing a single representative by way of an election is yet another example. When the law or constitution calls for a single representative of some constituency and there are many competing candidates, the choice can be made via a "first-past-the-post" or "winner-take-all" election. This is the typical system for electing government officials at the local, state, and federal level in the United States. The resulting discontinuous relationship between win/loss status and the vote share is used in the context of the U.S. Congress in Lee (2001, 2008), Lee et al. (2004), Albouy (2007b), Albouy (2007a), and in the context of mayoral elections in Ferreira and Gyourko (2009). The same idea is used in examining the impacts of union recognition, which is also decided by a secret ballot election (DiNardo and Lee, 2004).

6.2.2 Intentional Discretization

Sometimes resources could potentially be allocated on a continuous scale, but in practice are instead done in discrete levels. Among the studies we surveyed, we identified three broad motivations behind the use of these discontinuous rules.

First, a number of rules seem driven by a compensatory or equalizing motive. For example, in Chay et al. (2005), Leuven et al. (Forthcoming), and Van der Klaauw (2008a), extra resources for schools were allocated to the neediest communities, either on the basis of school-average test scores, disadvantaged minority proportions, or poverty rates. Similarly, Ludwig and Miller (2007), Battistin and Rettore (2008), and Buddelmeyer and Skoufias (2004) study programs designed to help poor communities, where the eligibility of a community is based on poverty rates. In each of these cases, one could imagine providing the most resources to the neediest and gradually phasing them out as the need index declines, but in practice this is not done, perhaps because it was impractical to provide very small levels of the treatment, given the fixed costs in administering the program.

A second motivation for having a discontinuous rule is to allocate treatments on the basis of some mea-

sure of merit. This was the motivation behind the merit award from the analysis of Thistlethwaite and Campbell (1960), as well as recent studies of the effect of financial aid awards on college enrollment, where the forcing variable is some measure of student achievement or test score, as in Kane (2003) and Van der Klaauw (2002).

Finally, we have observed that a number of discontinuous rules are motivated by the need to most effectively target the treatment. For example, environmental regulations or clean-up efforts naturally will focus on the most polluted areas, as in Chay and Greenstone (2003), Chay and Greenstone (2005), and Greenstone and Gallagher (2005). In the context of criminal behavior, prison security levels are often assigned based on an underlying score that quantifies potential security risks, such rules were used in Berk and DeLeeuw (1999) and Chen and Shapiro (2004).

6.3 Non-randomized Discontinuity Designs

Throughout this article, we have focused on regression discontinuity designs that follow a certain structure and timing in the assignment of treatment. First, individuals or communities – potentially in anticipation of the assignment of treatment – make decisions and act, potentially altering their probability of receiving treatment. Second, there is a stochastic shock due to "nature," reflecting that the units have incomplete control over the forcing variable. And finally, the treatment (or the intention to treat) is assigned on the basis of the forcing variable.

We have focused on this structure because in practice most RD analyses can be viewed along these lines, and also because of the similarity to the structure of a randomized experiment. That is, subjects of a randomized experiment may or may not make decisions in anticipation to participating in a randomized controlled trial (although their actions will ultimately have no influence on the probability of receiving treatment). Then the stochastic shock is realized (the randomization). Finally, the treatment is administered to one of the groups.

A number of the studies we surveyed though, did not seem to fit the spirit or essence of a randomized experiment. Since it is difficult to think of the treatment as being locally randomized in these cases, we will refer to the two research designs we identified in this category as "non-randomized" discontinuity designs.

6.3.1 Discontinuities in Age with Inevitable Treatment

Sometimes program status is turned on when an individual reaches a certain age. Receipt of pension benefits is typically tied to reaching a particular age (see Edmonds (2004); Edmonds et al. (2004)), and in the United States eligibility for the Medicare program begins at age 65 (see Card et al. (2009)) and young adults reach the legal drinking age at 21 (see Carpenter and Dobkin (2009)). Similarly, one is subject to the less punitive juvenile justice system until the age of majority (typically, eighteen) (see Lee and McCrary (2005)).

These cases stand apart from the typical RD designs discussed above because here assignment to treatment is essentially inevitable, as all subjects will eventually age into the program (or, conversely, age out of the program). One cannot, therefore, draw any parallels with a randomized experiment, which necessarily involves some *ex ante* uncertainty about whether a unit ultimately receives treatment (or the intent to treat).

Another important difference is that the tests of smoothness in baseline characteristics will generally be uninformative. Indeed, if one follows a single cohort over time, all characteristics determined prior to reaching the relevant age threshold are *by construction* identical just before and after the cutoff.[37] Note that in this case, *time* is the forcing variable, and therefore cannot be manipulated.

This design and the standard RD share the necessity of interpreting the discontinuity as the combined effect of *all* factors that switch on at the threshold . In the example of Thistlethwaite and Campbell (1960), if passing a scholarship exam provides the symbolic honor of passing the exam *as well as* a monetary award, the true treatment is a package of the two components, and one cannot attribute any effect to only one of the two . Similarly, when considering an age-activated treatment, one must consider the possibility that the age of interest is causing eligibility for potentially many other programs, which could affect the outcome.

There are at least two new issues that are irrelevant for the standard RD, but are important for the analysis of age discontinuities. First, even if there is truly an effect on the outcome, if the effect is not immediate, it generally will not generate a discontinuity in the outcome. For example, suppose the receipt of Social Security benefits has no immediate impact, but does have a long-run impact on labor force participation. Examining the labor force behavior as a function of age will not yield a discontinuity at age 67 (the full retirement age for those born after 1960), even though there may be a long-run effect. It is infeasible to estimate long-run effects because by the time we examine outcomes five years after receiving the treatment, for

[37] There are exceptions to this. There could be attrition over time, so that in principle, the number of observations could discontinuously drop at the threshold, changing the composition of the remaining observations. Alternatively, when examining a cross-section of different birth cohorts at a given point in time, it is possible to have sharp changes in the characteristics of individuals with respect to birthdate.

example, those individuals who were initially just below and just above age 67 will be exposed to essentially the same length of time of treatment (e.g. five years).[38]

The second important issue is that because treatment is inevitable with the passage of time, individuals may fully anticipate the change in the regime, and therefore they may behave in certain ways prior to the time when treatment is turned on. Optimizing behavior in anticipation of a sharp regime change may either accentuate or mute observed effects. For example, simple life-cycle theories, assuming no liquidity constraints, suggest that the path of consumption will exhibit no discontinuity at age 67, when Social Security benefits commence payment. On the other hand, some medical procedures are too expensive for an under-65-year-old, but would be covered under Medicare upon turning 65. In this case, individuals' greater awareness of such a predicament will tend to *increase* the size of the discontinuity in utilization of medical procedures with respect to age (e.g. see Card et al. (2009)).

At this time we are unable to provide any more specific guidelines for analyzing these age/time discontinuities, since it seems that how one models expectations, information, and behavior in anticipation of sharp changes in regimes will be highly context-dependent. But it does seem important to recognize these designs as being distinct from the standard RD design.

We conclude by emphasizing that when distinguishing between age-triggered treatments and a standard RD design, the involvement of age as a forcing variable is not as important as whether the receipt of treatment – or analogously, entering the control state – is inevitable. For example, on the surface, the analysis of the Medicaid expansions in Card and Shore-Sheppard (2004) appears to be an age-based discontinuity, since effective July 1991, U.S. law requires states to cover children born after September 30, 1983, implying a discontinuous relationship between coverage and age, where the discontinuity in July 1991 was around 8 years of age. This design actually fits quite easily into the standard RD framework we have discussed throughout this paper though.

First, note that treatment receipt is *not* inevitable for those individuals born near the September 30, 1983 threshold. Those born strictly after that date were covered from July 1991 until their 18th birthday, while those born on or before the date received no such coverage. Second, the data generating process does follow the structure discussed above. Parents do have some influence regarding when their children are born, but with only imprecise control over the exact date (and at any rate, it seems implausible that parents would have

[38] By contrast, there is no such limitation with the standard RD design. One can examine outcomes defined at an arbitrarily long time period after the assignment to treatment.

anticipated that such a Medicaid expansion would have occurred 8 years in the future, with the particular birthdate cutoff chosen). Thus the treatment is assigned based on the forcing variable, which is the birthdate in this context.

Examples of other age-based discontinuities where neither the treatment nor control state is guaranteed with the passage of time that can also be viewed within the standard RD framework include studies by Cascio and Lewis (2005), McCrary and Royer (2003), Dobkin and Ferreira (2007), and Oreopoulos (2006).

6.3.2 Discontinuities in Geography

Another "non-randomized" RD design is one involving the location of residences, where the discontinuity threshold is a boundary that demarcates regions. Black (1999) and Bayer et al. (2007) examine housing prices on either side of school attendance boundaries to estimate the implicit valuation of different schools. Lavy (2006) examines adjacent neighborhoods in different cities, and therefore subject to different rules regarding student busing . Lalive (2008) compares unemployment duration in regions in Austria receiving extended benefits to adjacent control regions. Pence (2006) examines census tracts along state borders to examine the impact of more borrower-friendly laws on mortgage loan sizes.

In each of these cases, it is awkward to view either houses or families as locally randomly assigned. Indeed this is a case where economic agents have quite precise control over where to place a house or where to live. The location of houses will be planned in response to geographic features (rivers, lakes, hills) and in conjunction with the planning of streets, parks, commercial development, etc. In order for this to resemble a more standard RD design, one would have to imagine the relevant boundaries being set in a "random" way, so that it would be simply luck determining whether a house ended up on either side of the boundary. The concern over the endogeneity of boundaries is clearly recognized by Black (1999), who "...[b]ecause of concerns about neighborhood differences on opposite sides of an attendance district boundary, ... was careful to omit boundaries from [her] sample if the two attendance districts were divided in ways that seemed to clearly divide neighborhoods; attendance districts divided by large rivers, parks, golf courses, or any large stretch of land were excluded." As one could imagine, the selection of which boundaries to include could quickly turn into more of an art than a science.

We have no uniform advice on how to analyze geographic discontinuities, because it seems that the best approach would be particularly context-specific. It does, however, seem prudent for the analyst, in assessing the internal validity of the research design, to carefully consider three sets of questions. First, what is the

process that led to the location of the boundaries? Which came first: the houses or the boundaries? Were the boundaries a response to some pre-existing geographical or political constraint? Second, how might sorting of families or the endogenous location of houses affect the analysis? And third, what are all the things differing between the two regions *other than the treatment of interest?* An exemplary analysis and discussion of these latter two issues in the context of school attendance zones is found in Bayer et al. (2007).

7 Concluding Remarks on RD Designs in Economics: Progress and Prospects

Our reading of the existing and active literature is that – after being largely ignored by economists for almost 40 years – there have been significant inroads made in understanding the properties, limitations, interpretability , and perhaps most importantly, useful application of RD designs to a wide variety of empirical questions in economics. These developments have for the most part occurred within a short period of time, beginning in the late 1990s.

Here we highlight what we believe are the most significant recent contributions of the economics literature to the understanding and application of RD designs. We believe these are helpful developments in guiding applied researchers implement RD designs, and we also illustrate them with a few examples from the literature.

- **Sorting and Manipulation of the Forcing Variable**: Economists consider how self-interested individuals or optimizing organizations may behave in response to rules that allocate resources. It is therefore unsurprising that the discussion of how endogenous sorting around the discontinuity threshold *can invalidate* the RD design has been found (to our knowledge, exclusively) in the economics literature. By contrast, textbook treatments outside economics on RD do not discuss this sorting or manipulation, and give the impression that the knowledge of the assignment rule is sufficient for the validity of the RD.[39]

  We believe a "state-of-the-art" RD analysis today will consider carefully the possibility of endogenous sorting. A recent analysis that illustrates this standard is that of Urquiola and Verhoogen (2007), who

[39] For example, Trochim (1984) characterizes the three central assumptions of the RD design as: 1) perfect adherence to the cutoff rule, 2) having the correct functional form, and 3) no other factors (other than the program of interest) cause the discontinuity. More recently, Shadish et al. (2002) claim on page 243 that the proof of the unbiasedness of RD primarily follows from the fact that treatment is known perfectly once the forcing variable is known. They go on to argue that this deterministic rule implies omitted variables will not pose a problem. But Hahn et al. (2001) make it clear that the existence of a deterministic rule for the assignment of treatment is *not* sufficient for unbiasedness, and it is necessary to *assume* the influence of all other factors (omitted variables) are the same on either side of the discontinuity threshold (i.e. their continuity assumption).

examine the class size cap RD design pioneered by Angrist and Lavy (1999) in the context of Chile's highly liberalized market for primary schools. In a certain segment of the private market, schools receive a fixed payment per student from the government . However, each school faces a very high marginal cost (hiring one extra teacher) for crossing a multiple of the class size cap. Perhaps unsurprisingly, they find striking discontinuities in the *histogram* of the forcing variable (total enrollment in the grade), with an undeniable "stacking" of schools at the relevant class size cap cutoffs. They also provide evidence that those families in schools just to the left and right of the thresholds are systematically different in family income, suggesting some degree of sorting. For this reason, they conclude that an RD analysis in this particular context is most likely inappropriate.[40]

This study, as well as the analysis of Bayer et al. (2007) reflects a heightened awareness of a sorting issue recognized since the beginning of the recent wave of RD applications in economics.[41] From a practitioner's perspective, an important recent development is the notion that we can empirically examine the degree of sorting, and one way of doing so is suggested in McCrary (2008).

- **RD Designs as Locally Randomized Experiments:** Economists are hesitant to apply methods that have not been rigorously formalized within an econometric framework, and where crucial identifying assumptions have not been clearly specified. This is perhaps one of the reasons why RD designs were under-utilized by economists for so long, since it is only relatively recently that the underlying assumptions needed for the RD were formalized.[42] In the recent literature, RD designs were initially viewed as a special case of matching (Heckman et al., 1999), or alternatively as a special case of IV (Angrist and Krueger, 1999), and these perspectives may have provided empirical researchers a famil-

[40] Urquiola and Verhoogen (2007) emphasize the sorting issues may well be specific to the liberalized nature of the Chilean primary school market, and that they may or may not be present in other countries.

[41] See, for example, footnote 23 in Van der Klaauw (1997) and page 549 in Angrist and Lavy (1999)

[42] An example of how economists'/econometricians' notion of a proof differs from that in other disciplines is found in Cook (2008), who views the discussion in Goldberger (1972a) and Goldberger (1972b) as the first "proof of the basic design", quoting the following passage in Goldberger (1972a) (brackets from Cook (2008)): "The explanation for this serendipitous result [no bias when selection is on an observed pretest score] is not hard to locate. Recall that z [a binary variable representing the treatment contrast at the cutoff] is completely determined by pretest score x [an obtained ability score]. It cannot contain any information about x^* [true ability] that is not contained within x . Consequently, when we control on x as in the multiple regression, z has no explanatory power with respect to y [the outcome measured with error]. More formally, the partial correlation of y and z controlling on x vanishes although the simple correlation of y and z is nonzero".

After reading the article, an econometrician will recognize the discussion above not as a proof of the validity of the RD, but rather as a re-statement of the consequence of z being an indicator variable determined by an observed variable x, in a specific parametrized example. Today we know the existence of such a rule is *not sufficient* for a valid RD design, and a crucial necessary assumption is the continuity of the influence of all other factors, as shown in Hahn et al. (2001). In Goldberger (1972a), the role of the continuity of omitted factors was not mentioned (although it is implicitly assumed in the stylized model of test scores involving normally distributed and independent errors). Indeed, apparently Goldberger himself later clarified that he did not set out to propose the RD design, and was instead interested in the issues related to selection on observables and unobservables (Cook, 2008).

iar econometric framework within which identifying assumptions could be more carefully discussed. Today, RD is increasingly recognized in applied research as a distinct design that is a close relative to a randomized experiment. Formally shown in (Lee, 2008), even when individuals have some control over the forcing variable, as long as this control is imprecise – that is, the *ex ante* density of the forcing variable is continuous – the consequence will be local randomization of the treatment. So in a number of non-experimental contexts where resources are allocated based on a sharp cutoff rule, there may indeed be a hidden randomized experiment to utilize. And furthermore, as in a randomized experiment, this implies that all observable baseline covariates will locally have the same distribution on either side of the discontinuity threshold – an empirically testable proposition.

We view the testing of the continuity of the baseline covariates as an important part of assessing the validity of any RD design – particularly in light of the incentives that can potentially generate sorting – and as something that truly sets RD apart from other evaluation strategies. Examples of this kind of testing of the RD design include, Matsudaira (2008), Card et al. (2007), DiNardo and Lee (2004), Lee et al. (2004), McCrary and Royer (2003), Greenstone and Gallagher (2005), and Urquiola and Verhoogen (2007).

- **Graphical Analysis and Presentation:** The graphical presentation of an RD analysis is not a contribution of economists,[43] but it is safe to say that the body of work produced by economists has led to a kind of "industry standard" that the transparent identification strategy of the RD be accompanied by an equally transparent graph showing the empirical relation between the outcome and the forcing variable. Graphical presentations of RD are so prevalent in applied research, it is tempting to guess that studies not including the graphical evidence are ones where the graphs are not compelling or well-behaved.

  In an RD analysis, the graph is indispensable because it can summarize a great deal of information in one picture. It can give a rough sense of the range of the both the forcing variable and the outcome variable, as well as the overall shape of the relationship between the two, thus indicating what functional forms are likely to make sense. It can also alert the researcher to potential outliers in both the forcing and outcome variables. A graph of the raw means – in non-overlapping intervals, as discussed in Section 4.1 – also gives a rough sense of the likely sampling variability of the RD gap estimate

[43] Indeed the original article of Thistlethwaite and Campbell (1960) included a graphical analysis of the data.

itself, since one can compare the size of the jump at the discontinuity to natural "bumpiness" in the graph away from the discontinuity.

Our reading of the literature is that the most informative graphs are ones that simultaneously allow the raw data "to speak for themselves" in revealing a discontinuity if there is one, yet at the same time treat data near the threshold the same as data away from the threshold.[44] There are many examples that follow this general principle; recent ones include Matsudaira (2008), Card et al. (2007), Card et al. (2009), McCrary and Royer (2003), Lee (2008), and Ferreira and Gyourko (2009).

- **Applicability:** Soon after the introduction of RD, in a chapter in a book on research methods, Campbell and Stanley (1963) wrote that the RD design was "very limited in range of possible applications". The emerging body of research produced by economists in recent years has proven quite the opposite. Our survey of the literature suggests that there are many kinds of discontinuous rules that can help answer important questions in economics and related areas. Indeed, one may go so far as to guess that whenever a scarce resource is rationed for individual entities, if the political climate demands a transparent way of distributing that resource, it is a good bet there is an RD design lurking in the background. In addition, it seems that the approach of using changes in laws that disqualify older birth cohorts based on their date of birth (as in Card and Shore-Sheppard (2004) or Oreopoulos (2006)) may well have much wider applicability.

  One way to understand both the applicability and limitations of the RD is to recognize its relation to a standard econometric policy evaluation framework, where the main variable of interest is a potentially endogenous binary treatment variable (as considered in Heckman (1978), or more recently discussed in Heckman and Vytlacil (2005)). This selection model applies to a great deal of economic problems. As we pointed out in Section 3, the RD design describes a situation where you are able to *observe* the latent variable that determines treatment. As long as the density of that variable is continuous for each individual, the benefit of observing the latent index is that one neither needs to make exclusion restrictions nor assume any variable (i.e. an instrument) is independent of errors in the outcome equation.

  From this perspective, for the class of problems that fit into the standard treatment evaluation problem, RD designs can be seen as a subset since there is an institutional, index-based rule playing a role in determining treatment. Among this subset, the binding constraint of RD lies in obtaining the necessary

[44] For example, graphing a smooth conditional expectation function everywhere *except* at the discontinuity threshold violates this principle.

data: readily available public-use household survey data, for example, will often only contain variables that are correlated with the true forcing variable (e.g. reported income in a survey, as opposed to the income used for allocation of benefits), or are measured too coarsely (e.g. years rather than months or weeks) to detect a discontinuity in the presence of a regression function with significant curvature. This is where there can be a significant payoff to investing in securing high quality data, which is evident in most of the studies listed in Table 4.

7.1 Extensions

We conclude by discussing two natural directions in which the RD approach can be extended. First, we have discussed the "fuzzy" RD design as an important departure from the "classic" RD design where treatment is a deterministic function of the forcing variable, but there are other departures that could be practically relevant but not as well understood. For example, even if there is perfect compliance of the discontinuous rule, it may be that the researcher does not directly observe the forcing variable, but instead possesses and a slightly noisy measure of the variable. Understanding the effects of this kind of measurement error further expand the applicability of RD. In addition, there may be situations where the researcher both suspects and statistically detects some degree of precise sorting around the threshold, but that the sorting may appear to be relatively minor, even if statistically significant (based on observing discontinuities in baseline characteristics). The challenge, then, is to specify under what conditions one can correct for small amounts of this kind of contamination.

Second, so far we have discussed the sorting or manipulation issue as a potential problem or nuisance to the general program evaluation problem. But there is another way of viewing this sorting issue. Economists are interested in studying behavioral relationships, and at a minimum would like to empirically distinguish true behavioral responses to incentives from the multitude of confounding factors that may simply be correlated with behaviors. Thus, when we observe discontinuities in the *frequency distribution* of grade enrollment (as in Urquiola and Verhoogen (2007)) or similarly in the distribution of roll call votes (as in McCrary (2008)), the observed discontinuities seem to be due to behavior and not to correlation of unobserved characteristics.

These cases, as well as the age/time and boundary discontinuities discussed above, do not fit into the "standard" RD framework, but nevertheless can tell us something important about behavior, and are certainly in the spirit of the more standard RD design. These designs further expand the kinds of questions that can be

addressed by exploiting discontinuous rules to identify meaningful economic parameters of interest.

References

Albouy, David, "Do Voters Affect or Elect Policies? A New Perspective with Evidence from the U.S. Senate," July 2007.

_, "Partisan Representation in Congress and the Geographic Distribution of Federal Funds," May 2007.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, June 1990, *80* (3), 313–336.

_ and Alan B. Krueger, "Empirical Strategies in Labor Economics," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A, Elsevier Science, 1999.

_ and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, May 1999, *114* (2), 533–575.

Asadullah, M. Niaz, "The Effect of Class Size on Student Achievement: Evidence from Bangladesh," *Applied Economic Letters*, March 2005, *12* (4), 217–221.

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

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

Bayer, Patrick, Fernando Ferreira, and Robert McMillan, "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 2007, *115* (4), 588–638.

Berk, Richard A. and David Rauma, "Capitalizing on Nonrandom Assignment to Treatments: A Regression-Discontinuity Evaluation of a Crime-Control Program," *Journal of the American Statistical Association*, March 1983, *78* (381), 21–27.

_ and Jan DeLeeuw, "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design," *Journal of the American Statistical Association*, December 1999, *94* (448), 1045–1052.

Black, Dan A., Jeffrey A. Smith, Mark C. Berger, and Brett J. Noel, "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System," *American Economic Review*, November 2003, *93* (4), 1313–1327.

_ , Jose Galdo, and Jeffrey A. Smith, "Evaluating the Regression Discontinuity Design Using Experimental Data," Working Paper, University of Michigan 2007.

Black, Sandra, "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, May 1999, *114* (2), 577–599.

Blundell, Richard and Alan Duncan, "Kernel Regression in Empirical Microeconomics," *Journal of Human Resources*, Winter 1998, *33* (1), 62–87.

Buddelmeyer, Hielke and Emmanuel Skoufias, "An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA," Working Paper 3386, The World Bank 2004.

Buettner, Theiss, "The incentive effect of fiscal equalization transfers on tax policy," *Journal of Public Economics*, February 2006, *90* (3), 477–497.

Campbell, D. T. and J. C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," in N. L. Gage, ed., *Handbook of Research on Teaching*, Chicago: Rand McNally, 1963.

Canton, Erik and Andreas Blom, "Can Student Loans Improve Accessibility to Higher Education and Student Performance? An Impact Study of the Case of SOFES, Mexico," Working Paper 3425, The World Bank 2004.

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

_ , C. Dobkin, and Nicole Maestas, "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare," *American Economic Review*, forthcoming 2009.

Card, David and Lara D. Shore-Sheppard, "Using Discontinuous Eligibility Rules to Identify the Effects of the Federal Medicaid Expansions on Low-Income Children," *Review of Economics and Statistics*, August 2004, *86* (3), 752–766.

_ , Raj Chetty, and Andrea Weber, "Cash-on-Hand and Competing Models of Intertemporal Behavior: New Evidence from the Labor Market," *Quarterly Journal of Economics*, November 2007, *122* (4), 1511–1560.

Carpenter, Christopher and Carlos Dobkin, "The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age," *American Economic Journal: Applied Economics*, in press 2009, *1*.

Cascio, Elizabeth and Ethan Lewis, "Schooling and the AFQT: Evidence from School Entry Laws," Working Paper 11113, National Bureau of Economic Research February 2005.

Chay, Kenneth Y. and Michael Greenstone, "Air Quality, Infant Mortality, and the Clean Air Act of 1970," 2003.

_ and _ , "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, April 2005, *113* (2), 376–424.

_ , Patrick J. McEwan, and Miguel Urquiola, "The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools," *American Economic Review*, September 2005, *95* (4), 1237–1258.

Chen, M. Keith and Jesse M. Shapiro, "Does Prison Harden Inmates? A Discontinuity-based Approach," Working Paper 1450, Cowles Foundation 2004.

Chen, Susan and Wilbert van der Klaauw, "The Work Disincentive Effects of the Disability Insurance Program in the 1990s," *Journal of Econometrics*, February 2008, *142* (2), 757–784.

Cook, T.D., ""Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics," *Journal of Econometrics*, February 2008, *142* (2), 636–654.

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

Dobkin, Carlos and Fernando Ferreira, "Do School Entry Laws Affect Educational Attainment and Labor Market Outcomes?," July 2007.

Edmonds, Eric V., "Does Illiquidity Alter Child Labor and Schooling Decisions? Evidence from Household Responses to Anticipated Cash Transfers in South Africa," Working Paper 10265, National Bureau of Economic Research 2004.

_ , Kristen Mammen, and Douglas L. Miller, "Rearranging the Family? Income Support and Elderly Living Arrangements in a Low-Income Country," Working Paper 10306, National Bureau of Economic Research 2004.

Fan, J. and I. Gijbels, *Local Polynomial Modelling and Its Applications*, (Chapman and Hall, London), 1996.

Ferreira, Fernando, "You Can Take It with You: Proposition 13 Tax Benefits, Residential Mobility, and Willingness to Pay for Housing Amenities," July 2007.

_ and Joseph Gyourko, "Do Political Parties Matter? Evidence from Cities," *Quarterly Journal of Economics*, 2009, *124* (1).

Giorgi, Giacomo De, "Long-term effects of a mandatory multistage program: the New Deal for young people in the UK," Working Paper W05/08 2005, Institute for Fiscal Studies Working Paper 2005.

Goldberger, A. S., *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*, Unpublished manuscript, Madison, WI: Institute for Research on Poverty, 1972a.

_ , *Selection Bias in Evaluating Treatment Effects: The case of interaction*, Madison, WI: Institute for Research on Poverty, 1972b.

Greenstone, Michael and Justin Gallagher, "Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program," Working Paper 11790, National Bureau of Economic Research 2005.

Guryan, Jonathan, "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts," Technical Report, National Bureau of Economic Research 2001.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, "Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design," 1999, (7131).

_ , _ , and _ , "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, January 2001, *69* (1), 201–209.

Heckman, James J., "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 1978, *46* (4), 931–59.

_ and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

_ , Robert J. Lalonde, and Jeffrey A. Smith, "The Economics and Econometrics of Active Labor Market Programs," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A, Elsevier Science, 1999.

Hoxby, Caroline M., "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, November 2000, *115* (4), 1239–1285.

Imbens, G. W. and J. D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrics*, 1994, *61(2)*, 467–476.

Imbens, Guido and Karthik Kalyanaraman, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," January 2009.

_ and Thomas Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, February 2008, *142* (2), 615–635.

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

_ and _ , "The Impact of Teacher Training on Student Achievement: Quasi-experimental Evidence from School Reform Efforts in Chicago," *Journal of Human Resources*, Winter 2004a, *39* (1), 50–79.

Kane, Thomas J., "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going," Working Paper 9703, National Bureau of Economic Research 2003.

Lalive, R., "How do Extended Benefits affect Unemployment Duration? A Regression Discontinuity Approach," *Journal of Econometrics*, February 2008, *142* (2), 785–806.

Lavy, Victor, "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," Working Paper 10622, National Bureau of Economic Research 2004.

_ , "From Forced Busing to Free Choice in Public Schools: Quasi-Experimental Evidence of Individual and General Effects," Working Paper 11969, National Bureau of Economic Research 2006.

Lee, David S., "The Electoral Advantage to Incumbency and Voters' Valuation of Politicians' Experience: A Regression Discontinuity Analysis of Close Elections," Working Paper 31, Center for Labor Economics April 2001.

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

_ and David Card, "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, February 2008, *142* (2), 655–674.

_ and Justin McCrary, "Crime, Punishment, and Myopia," Working Paper 11491, National Bureau of Economic Research June 2005.

_ , Enrico Moretti, and Matthew Butler, "Do Voters Affect or Elect Policies? Evidence from the U.S. House," *Quarterly Journal of Economics*, August 2004, *119* (3), 807–859.

Lemieux, Thomas and Kevin Milligan, "Incentive Effects of Social Assistance: A Regression Discontinuity Approach," *Journal of Econometrics*, February 2008, *142* (2), 807–828.

Leuven, E., M. Lindahl, H. Oosterbeek, and D. Webbink, "The Effect of Extra Funding for Disadvantaged Pupils on Achievement," *The Review of Economics and Statistics*, Forthcoming.

Leuven, Edwin and Hessel Oosterbeek, "Evaluating the Effect of Tax Deductions on Training," *Journal of Labor Economics*, April 2004, *22* (2), 461–488.

Ludwig, J. and D. Miller, "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," Working Paper 11702, National Bureau of Economic Research October 2005.

_ and _ , "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 2007, *122(1)*, 159–208.

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

McCall, Brian and Stephen L. Desjardins, "The Impact of the Gates Millennium Scholars Program on the Retention, College Finance- and Work-Related Choices, and Future Educational Aspirations of Low-Income Minority Students," Working Paper, University of Michigan 2008.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, February 2008, *142* (2), 698–714.

_ and Heather Royer, "Does Maternal Education Affect Infant Health? A Regression Discontinuity Approach Based on School Age Entry Laws," Unpublished Paper November 2003.

Oreopoulos, Phillip, "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*, March 2006, *96* (1), 152–175.

Pence, Karen M., "Foreclosing on opportunity: State laws and mortgage credit," *Review of Economics and Statistics*, February 2006, *88* (1), 177–182.

Pettersson-Lidbom, Per, "Do Parties Matter for Economic Outcomes? A Regression-Discontinuity Approach," Working Paper, Stockholm University February 2006.

_, "Does the Size of the Legislature Affect the Size of Government? Evidence from Two Nautral Experiments," January 2006.

Pintoff, Randy, "Juvenile Jails: A Path to the Straight and Narrow or Hardened Criminality," November 2005.

Pitt, Mark M. and Shahidur R. Khandker, "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?," *The Journal of Political Economy*, October 1998, *106* (5), 958–996.

_, _, Signe-Mary McKernan, and M. Abdul Latif, "Credit Programs for the Poor and Reproductive Behavior in Low-Income Countries: Are the Reported Causal Relationships the Result of Heterogeneity Bias?," *Demography*, February 1999, *36* (1), 1–21.

Porter, J., "Estimation in the Regression Discontinuity Model," *Department of Economics, University ofWisconsin*, 2003.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*, Boston: Houghton Mifflin, 2002.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, London: Chapman Hall., 1986.

Thistlethwaite, Donald L. and Donald T. Campbell, "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment," *Journal of Educational Psychology*, December 1960, *51*, 309–317.

Trochim, William M. K., *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, Sage Publications, Beverly Hills, CA, 1984.

Urquiola, Miguel, "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia," *Review of Economics and Statistics*, 2006, *88*, 171–177.

_ **and Eric A. Verhoogen**, "Class Size and Sorting in Market Equilibrium: Theory and Evidence," July 2007.

Van der Klaauw, Wilbert, "A Regression-discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment," Working Paper 10, New York University C. V. Start Center 1997.

_ , "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, November 2002, *43* (4), 1249–1287.

_ , "Breaking the Link Between Poverty and Low Student Achievement: An Evaluation of Title I," *Journal of Econometrics*, February 2008, *142* (2), 731–756.

_ , "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, June 2008, *22* (2), 219–245.

Table 1: Choice of Bandwidth in Graph for Voting Example

A. Optimal bandwidth selected by cross-validation

Side of cutoff	Share of vote	Win next election
Left	0.021	0.049
Right	0.026	0.021
Both	0.021	0.049

B. P-values of tests for the numbers of bins in RD graph

		Share of vote		Win next election	
No. of bins	Bandwidth	Bin test	Regr. test	Bin test	Regr. test
10	0.100	0.000	0.000	0.001	0.000
20	0.050	0.000	0.000	0.026	0.049
30	0.033	0.163	0.390	0.670	0.129
40	0.025	0.157	0.296	0.024	0.020
50	0.020	0.957	0.721	0.477	0.552
60	0.017	0.159	0.367	0.247	0.131
70	0.014	0.596	0.130	0.630	0.743
80	0.013	0.526	0.740	0.516	0.222
90	0.011	0.815	0.503	0.806	0.803
100	0.010	0.787	0.976	0.752	0.883

Notes: Estimated over the range of the forcing variable (Democrat to Republican difference in the share of vote in the previous election) ranging between -0.5 and 0.5. The "bin test" is computed by comparing the fit of a model with the number of bins indicated in the table to an alternative where each bin is split in 2. The "regression test" is a joint test of significance of bin-specific regression estimates of the outcome variable on the share of vote in the previous election.

Table 2a: RD estimates of the effect of winning the previous election on the share of votes in the next election

Bandwidth:	1.00	0.50	0.25	0.15	0.10	0.05	0.04	0.03	0.02	0.01
Polynomial of order:										
Zero	0.347	0.257	0.179	0.143	0.125	0.096	0.080	0.073	0.077	0.088
	(0.003)	(0.004)	(0.004)	(0.005)	(0.006)	(0.009)	(0.011)	(0.012)	(0.014)	(0.015)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.003]	[0.047]	[0.778]	[0.821]	[0.687]	
One	0.118	0.090	0.082	0.077	0.061	0.049	0.067	0.079	0.098	0.096
	(0.006)	(0.007)	(0.008)	(0.011)	(0.013)	(0.019)	(0.022)	(0.026)	(0.029)	(0.028)
	[0.000]	[0.332]	[0.423]	[0.216]	[0.543]	[0.168]	[0.436]	[0.254]	[0.935]	
Two	0.052	0.082	0.069	0.050	0.057	0.100	0.101	0.119	0.088	0.098
	(0.008)	(0.010)	(0.013)	(0.016)	(0.020)	(0.029)	(0.033)	(0.038)	(0.044)	(0.045)
	[0.000]	[0.335]	[0.371]	[0.385]	[0.458]	[0.650]	[0.682]	[0.272]	[0.943]	
Three	0.111	0.068	0.057	0.061	0.072	0.112	0.119	0.092	0.108	0.082
	(0.011)	(0.013)	(0.017)	(0.022)	(0.028)	(0.037)	(0.043)	(0.052)	(0.062)	(0.063)
	[0.001]	[0.335]	[0.524]	[0.421]	[0.354]	[0.603]	[0.453]	[0.324]	[0.915]	
Four	0.077	0.066	0.048	0.074	0.103	0.106	0.088	0.049	0.055	0.077
	(0.013)	(0.017)	(0.022)	(0.027)	(0.033)	(0.048)	(0.056)	(0.067)	(0.079)	(0.063)
	[0.014]	[0.325]	[0.385]	[0.425]	[0.327]	[0.560]	[0.497]	[0.044]	[0.947]	
Optimal order of the polynomial	6	3	1	2	1	2	0	0	0	0
Observations	6558	4900	2763	1765	1209	610	483	355	231	106

Notes: Standard errors in parentheses. P-values from the goodness-of-fit test in square brackets. The goodness-of-fit test is obtained by jointly testing the significance of a set of bin dummies included as additional regressors in the model. The bin width used to construct the bin dummies is .01. The optimal order of the polynomial is chosen using Akaike's criterion (penalized cross-validation)

Table 2b: RD estimates of the effect of winning the previous election on probability of winning the next election

Bandwidth:	1.00	0.50	0.25	0.15	0.10	0.05	0.04	0.03	0.02	0.01
Polynomial of order:										
Zero	0.814	0.777	0.687	0.604	0.550	0.479	0.428	0.423	0.459	0.533
	(0.007)	(0.009)	(0.013)	(0.018)	(0.023)	(0.035)	(0.040)	(0.047)	(0.058)	(0.082)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.011]	[0.201]	[0.852]	[0.640]	[0.479]	
One	0.689	0.566	0.457	0.409	0.378	0.378	0.472	0.524	0.567	0.453
	(0.011)	(0.016)	(0.026)	(0.036)	(0.047)	(0.073)	(0.083)	(0.099)	(0.116)	(0.157)
	[0.000]	[0.000]	[0.126]	[0.269]	[0.336]	[0.155]	[0.400]	[0.243]	[0.125]	
Two	0.526	0.440	0.375	0.391	0.450	0.607	0.586	0.589	0.440	0.225
	(0.016)	(0.023)	(0.039)	(0.055)	(0.072)	(0.110)	(0.124)	(0.144)	(0.177)	(0.246)
	[0.075]	[0.145]	[0.253]	[0.192]	[0.245]	[0.485]	[0.367]	[0.191]	[0.134]	
Three	0.452	0.370	0.408	0.435	0.472	0.566	0.547	0.412	0.266	0.172
	(0.021)	(0.031)	(0.052)	(0.075)	(0.096)	(0.143)	(0.166)	(0.198)	(0.247)	(0.349)
	[0.818]	[0.277]	[0.295]	[0.115]	[0.138]	[0.536]	[0.401]	[0.234]	[0.304]	
Four	0.385	0.375	0.424	0.529	0.604	0.453	0.331	0.134	0.050	0.168
	(0.026)	(0.039)	(0.066)	(0.093)	(0.119)	(0.183)	(0.214)	(0.254)	(0.316)	(0.351)
	[0.965]	[0.200]	[0.200]	[0.173]	[0.292]	[0.593]	[0.507]	[0.150]	[0.244]	
Optimal order of the polynomial	4	3	2	1	1	2	0	0	0	1
Observations	6558	4900	2763	1765	1209	610	483	355	231	106

Notes: Standard errors in parentheses. P-values from the goodness-of-fit test in square brackets. The goodness-of-fit test is obtained by jointly testing the significance of a set of bin dummies included as additional regressors in the model. The bin width used to construct the bin dummies is .01. The optimal order of the polynomial is chosen using Akaike's criterion (penalized cross-validation)

Table 3: Optimal Bandwidth for Local Linear Regressions, Voting Example

	Share of vote	Win next election
A. Rule-of-thumb bandwidth		
Left	0.162	0.164
Right	0.208	0.130
Both	0.180	0.141
B. Optimal bandwidth selected by cross-validation		
Left	0.192	0.247
Right	0.282	0.141
Both	0.282	0.172

Notes: Estimated over the range of the forcing variable (Democrat to Republican difference in the share of vote in the previous election) ranging between -0.5 and 0.5. See the text for a description of the rule-of-thumb and cross-validation procedures for choosing the optimal bandwidth.

Figure 1: Simple Linear RD Setup

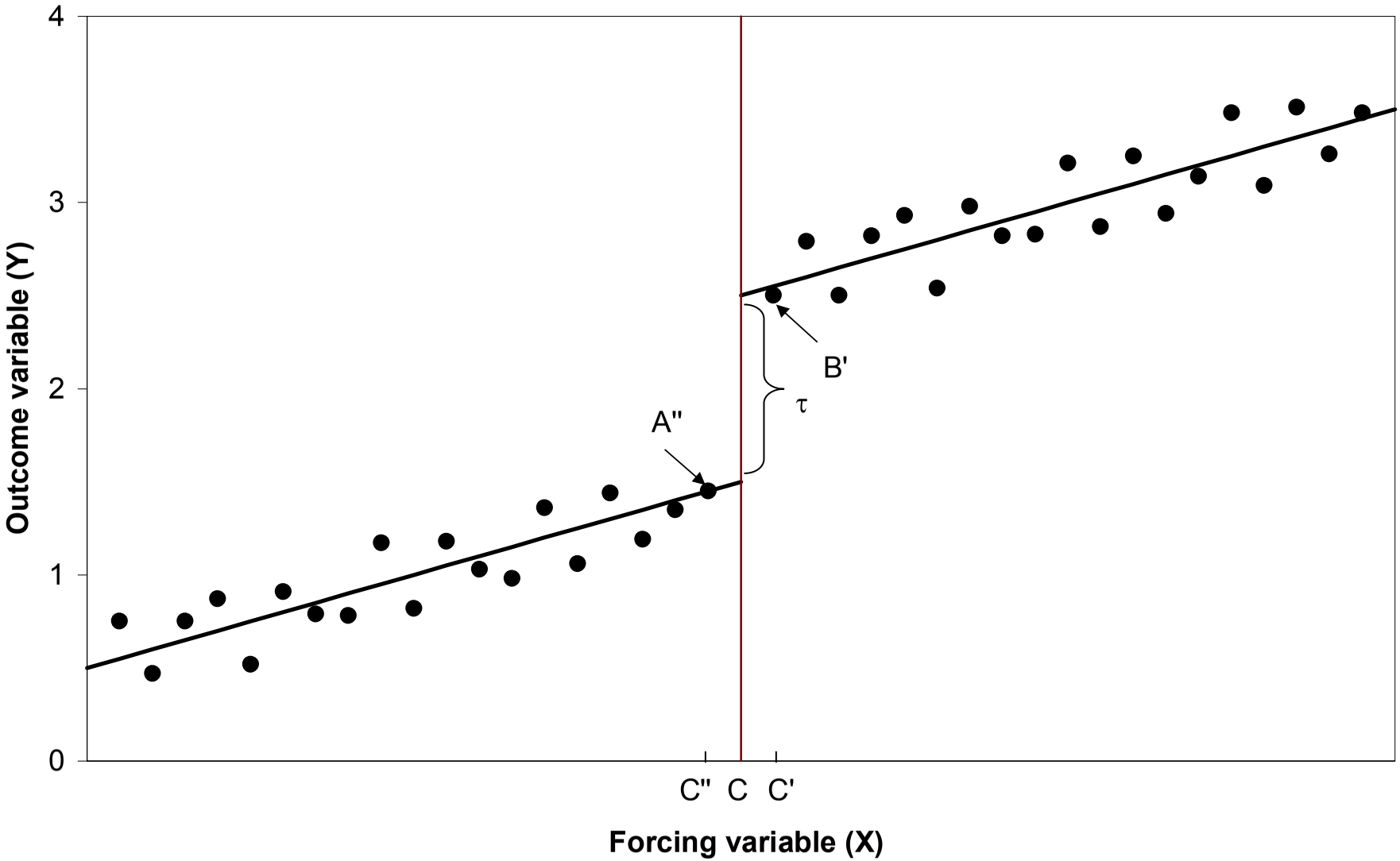


Figure 2: Nonlinear RD

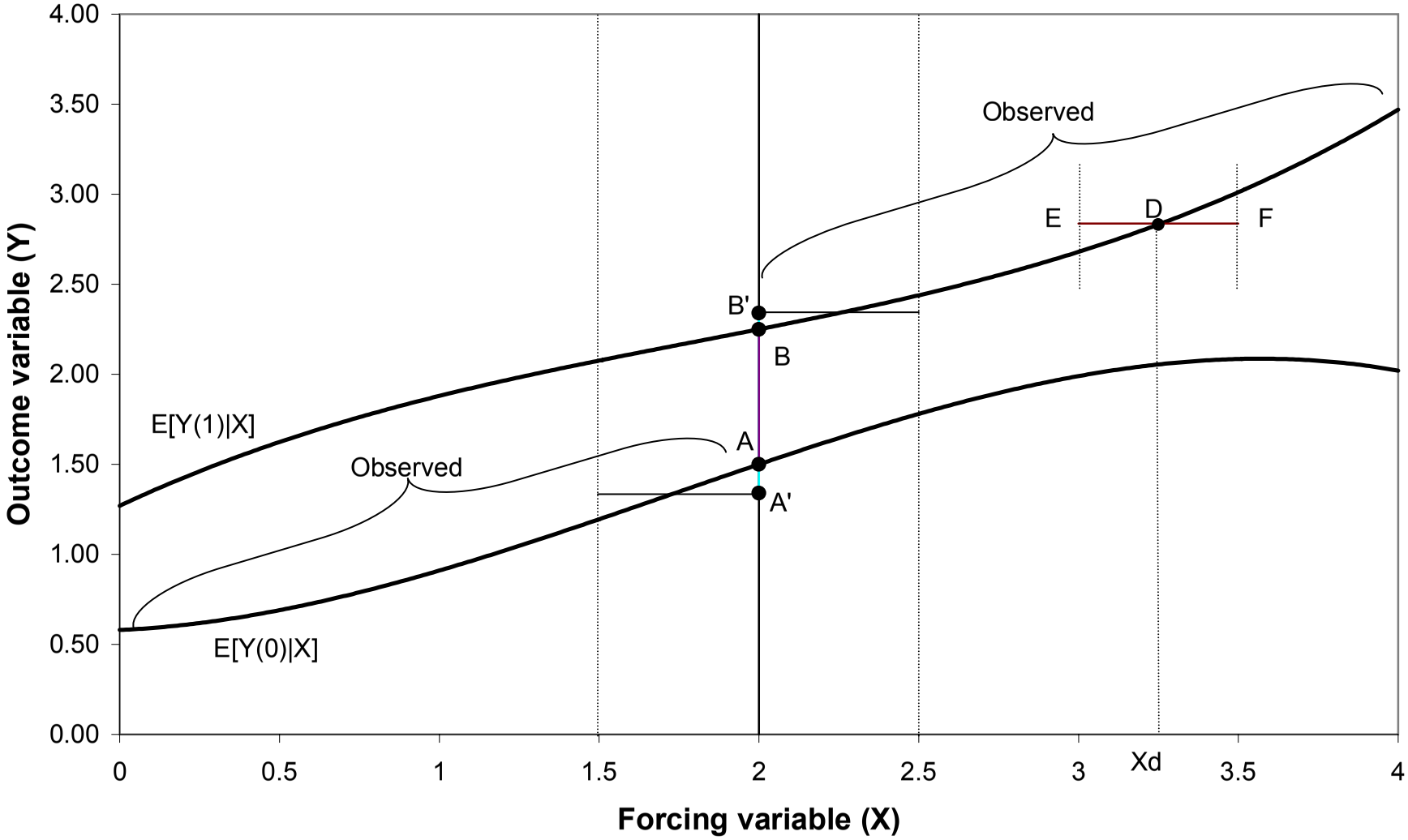


Figure 3: Randomized Experiment as a RD Design

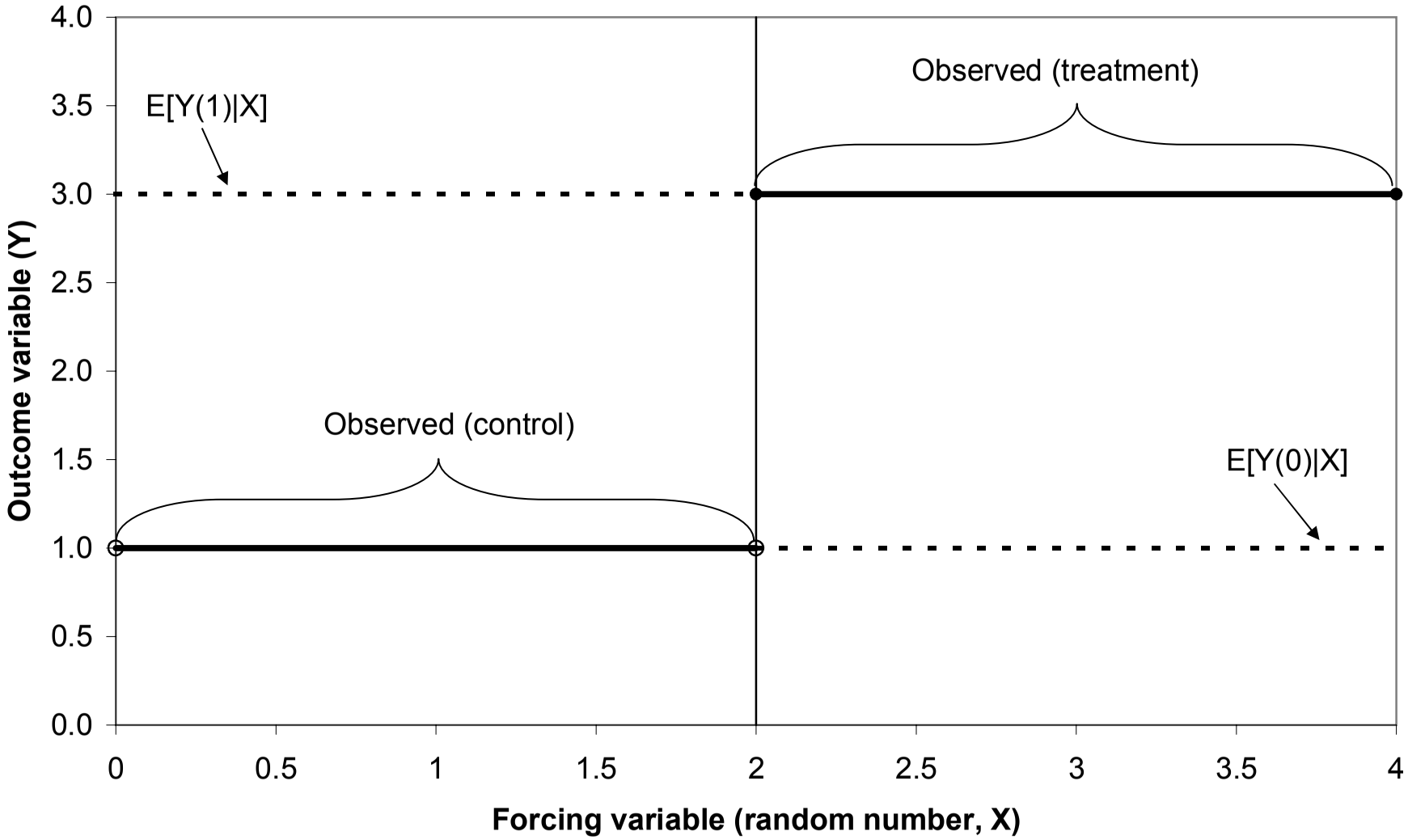


Figure 4: Density of Forcing Variable Conditional on W=w, U=u

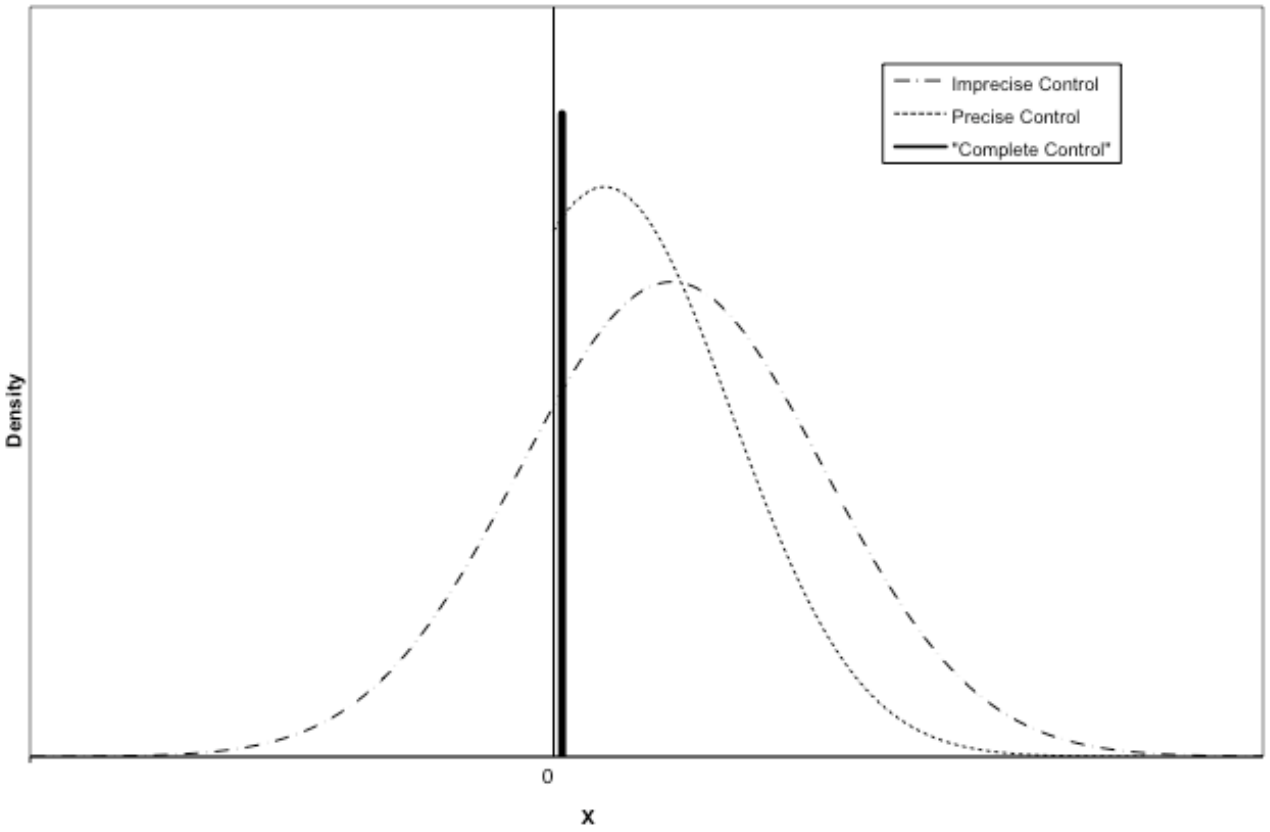
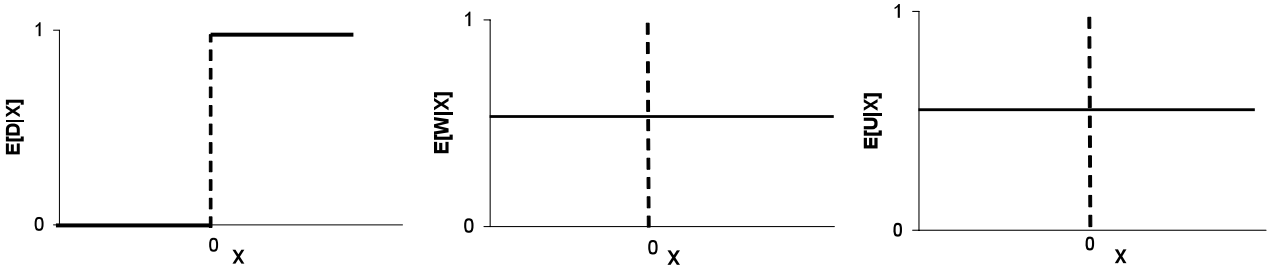
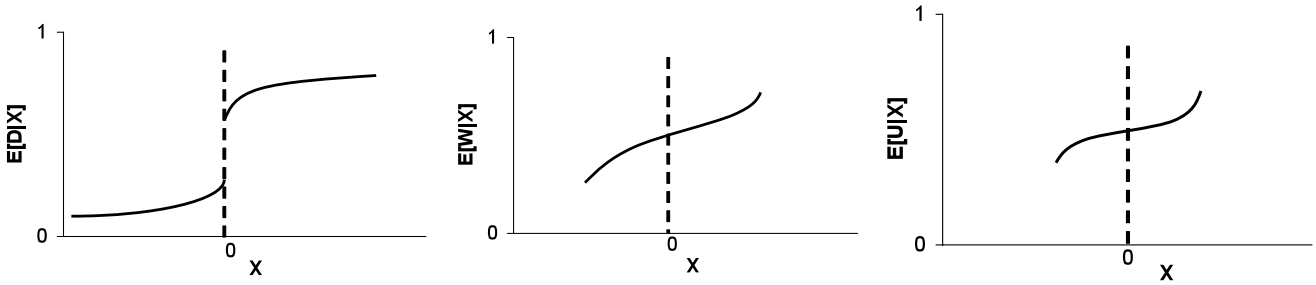


Figure 5. Treatment, Observables, and Unobservables in four research designs.

A. Randomized Experiment



B. Regression Discontinuity Design



C. Matching on Observables

D. Instrumental Variables

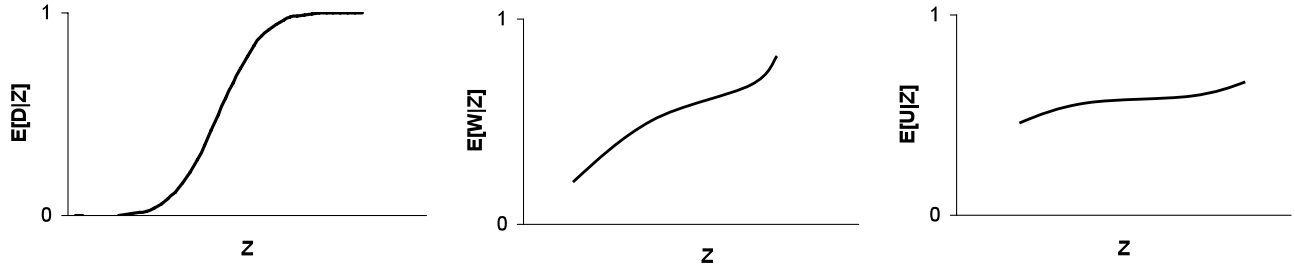


Figure 6a: Share of vote in next election, bandwidth of 0.02 (50 bins)

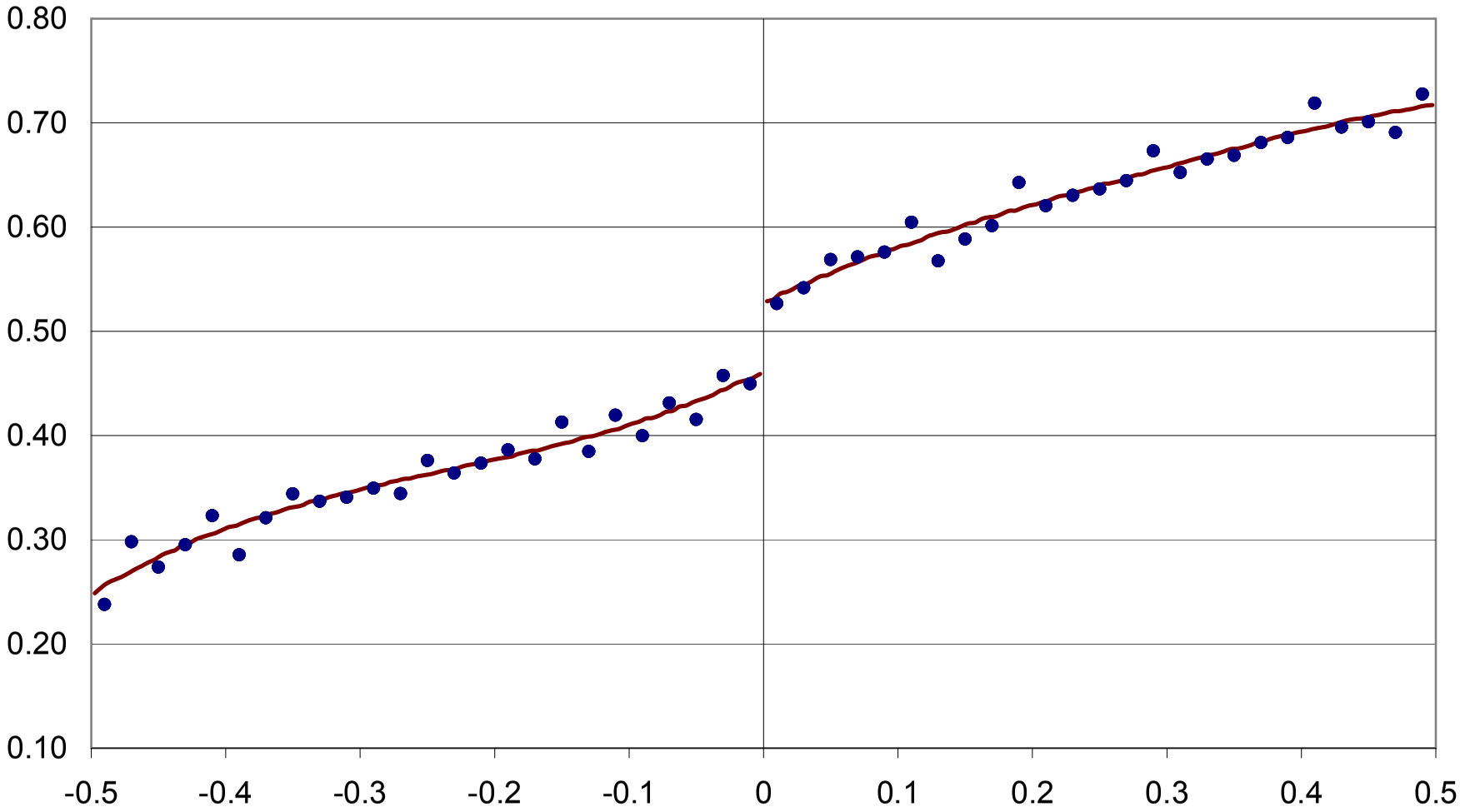


Figure 6b: Share of vote in next election, bandwidth of 0.01 (100 bins)

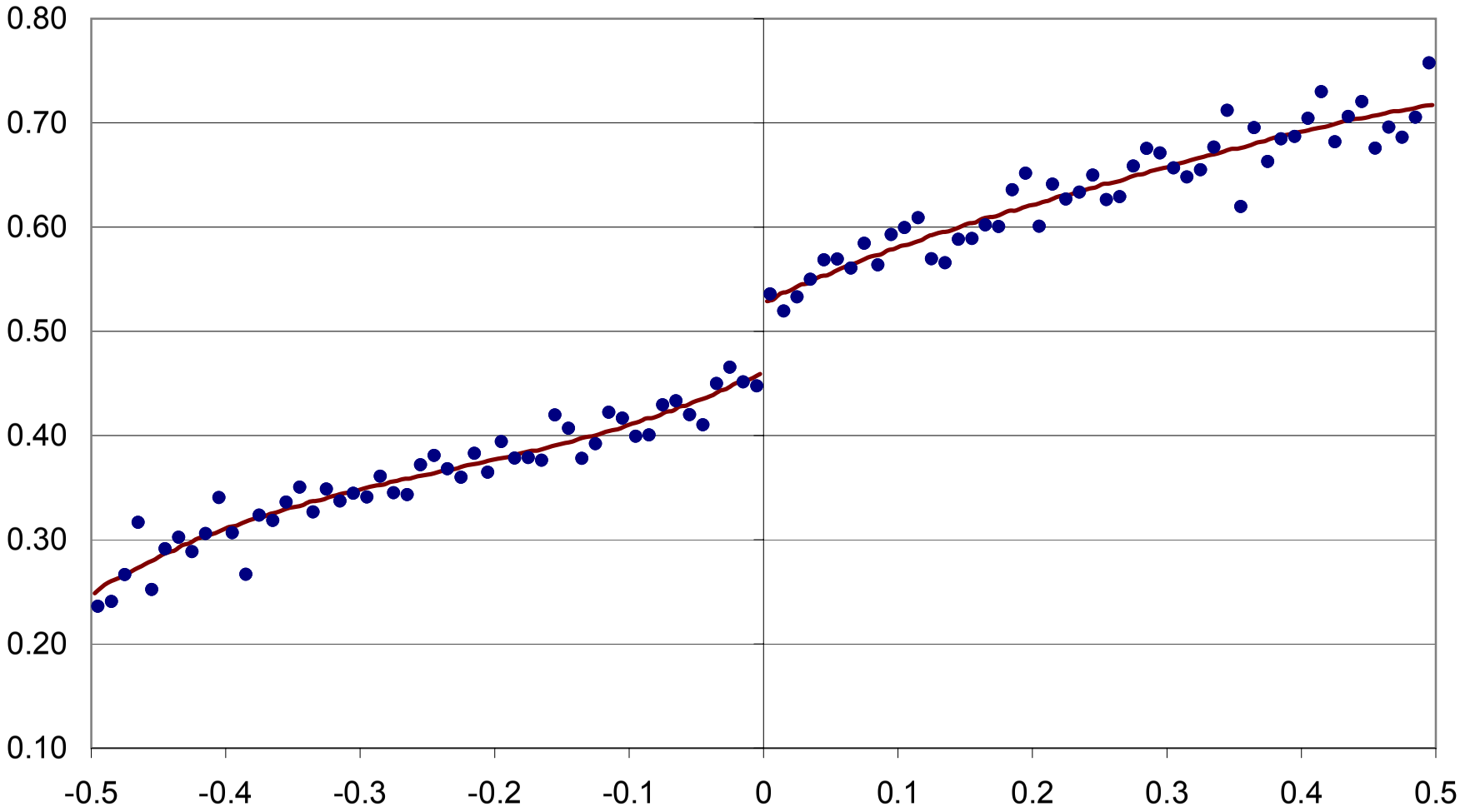


Figure 6c: Share of vote in next election, bandwidth of 0.005 (200 bins)

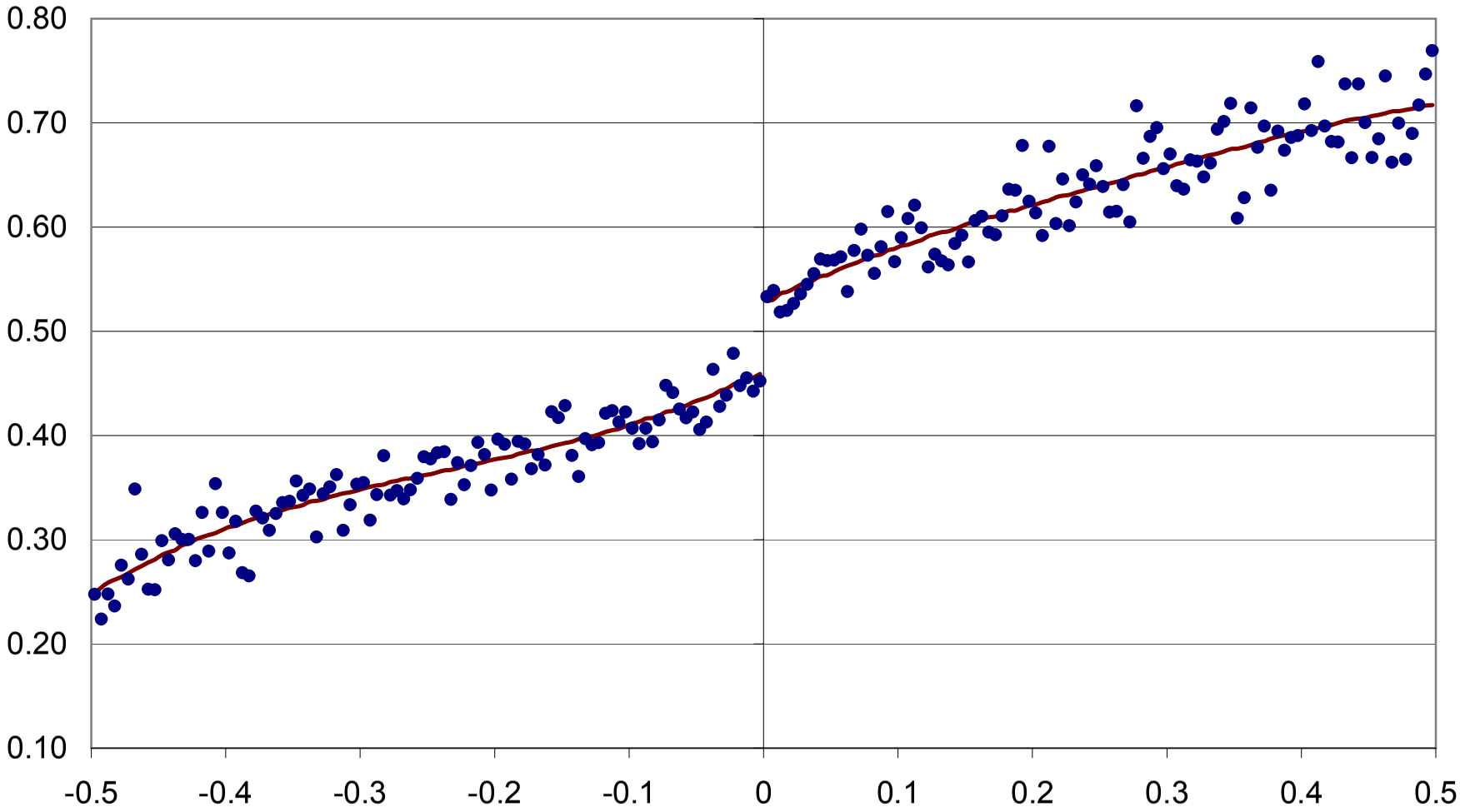


Figure 7a: Winning the next election, bandwidth of 0.02 (50 bins)

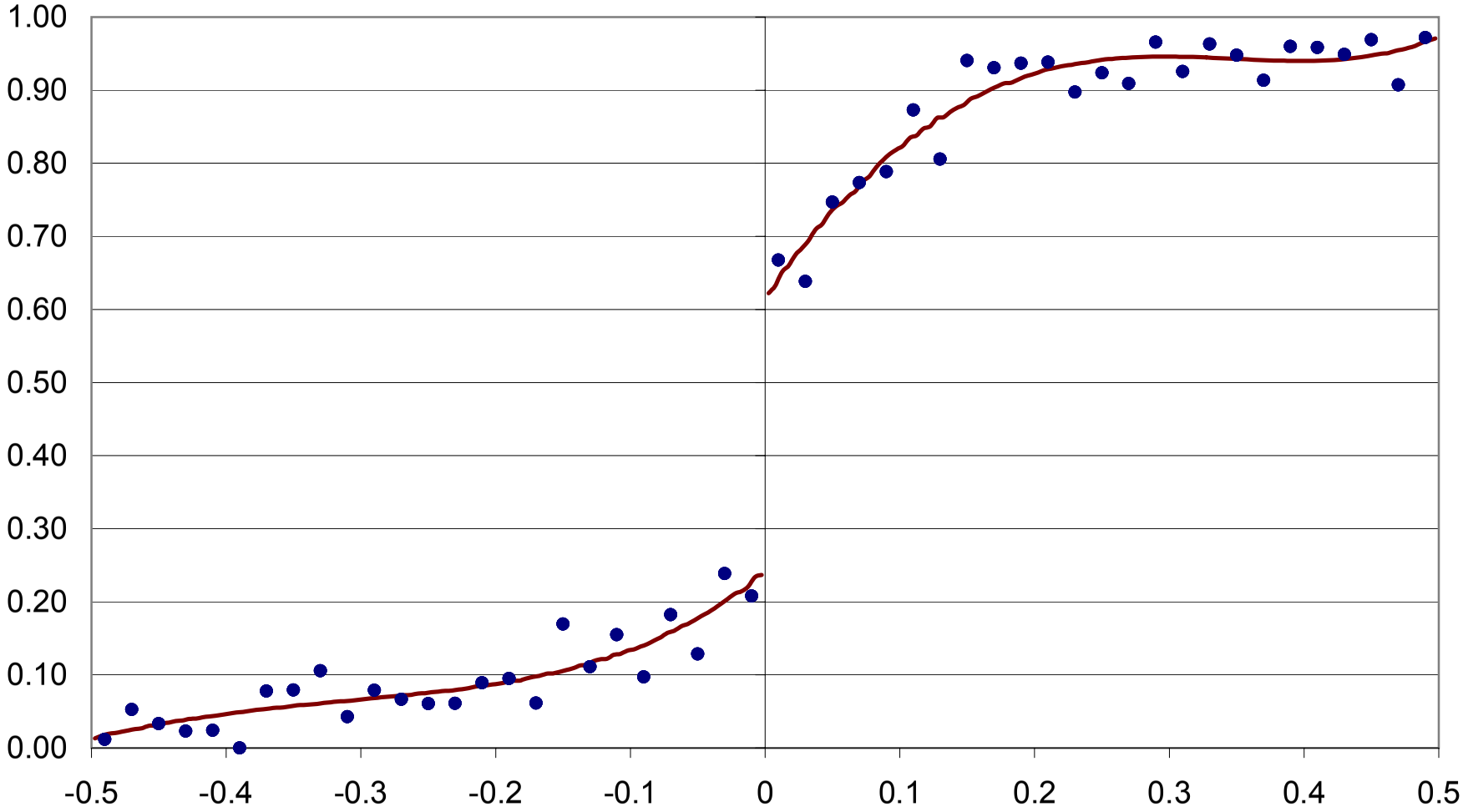


Figure 7b: Winning the next election, bandwidth of 0.01 (100 bins)

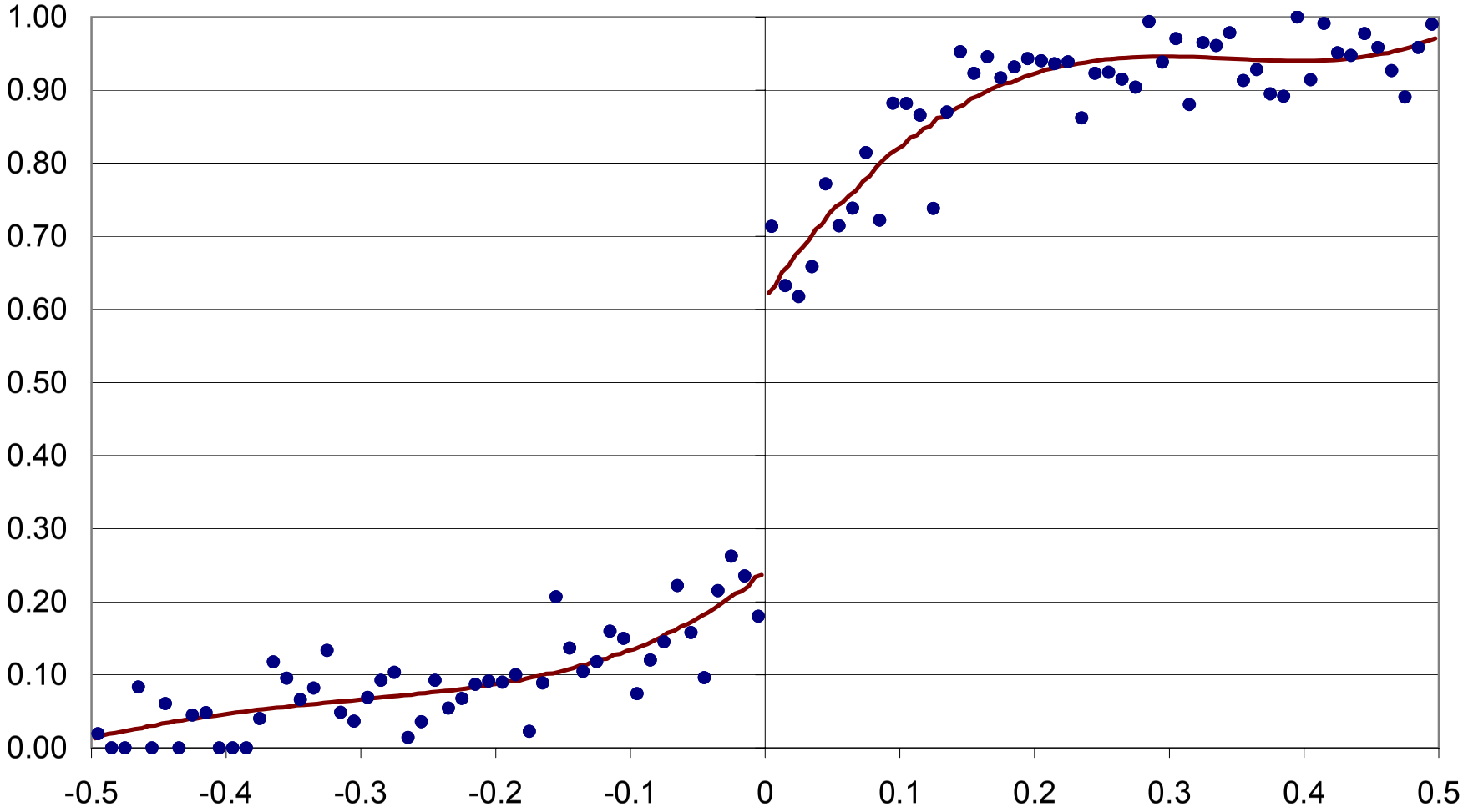


Figure 7c: Winning the next election, bandwidth of 0.005 (200 bins)

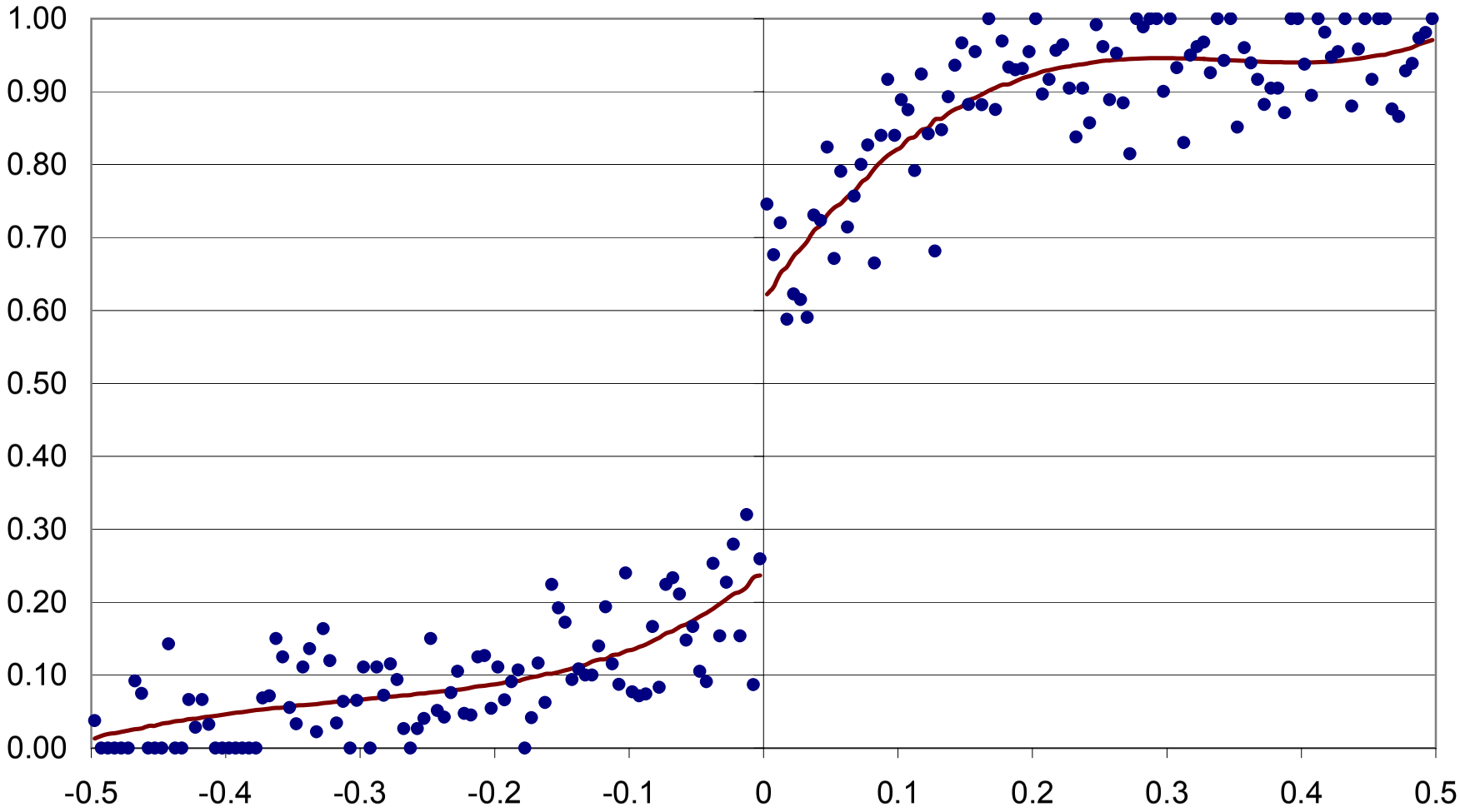


Figure 8: Density of the forcing variable (vote share in previous election)

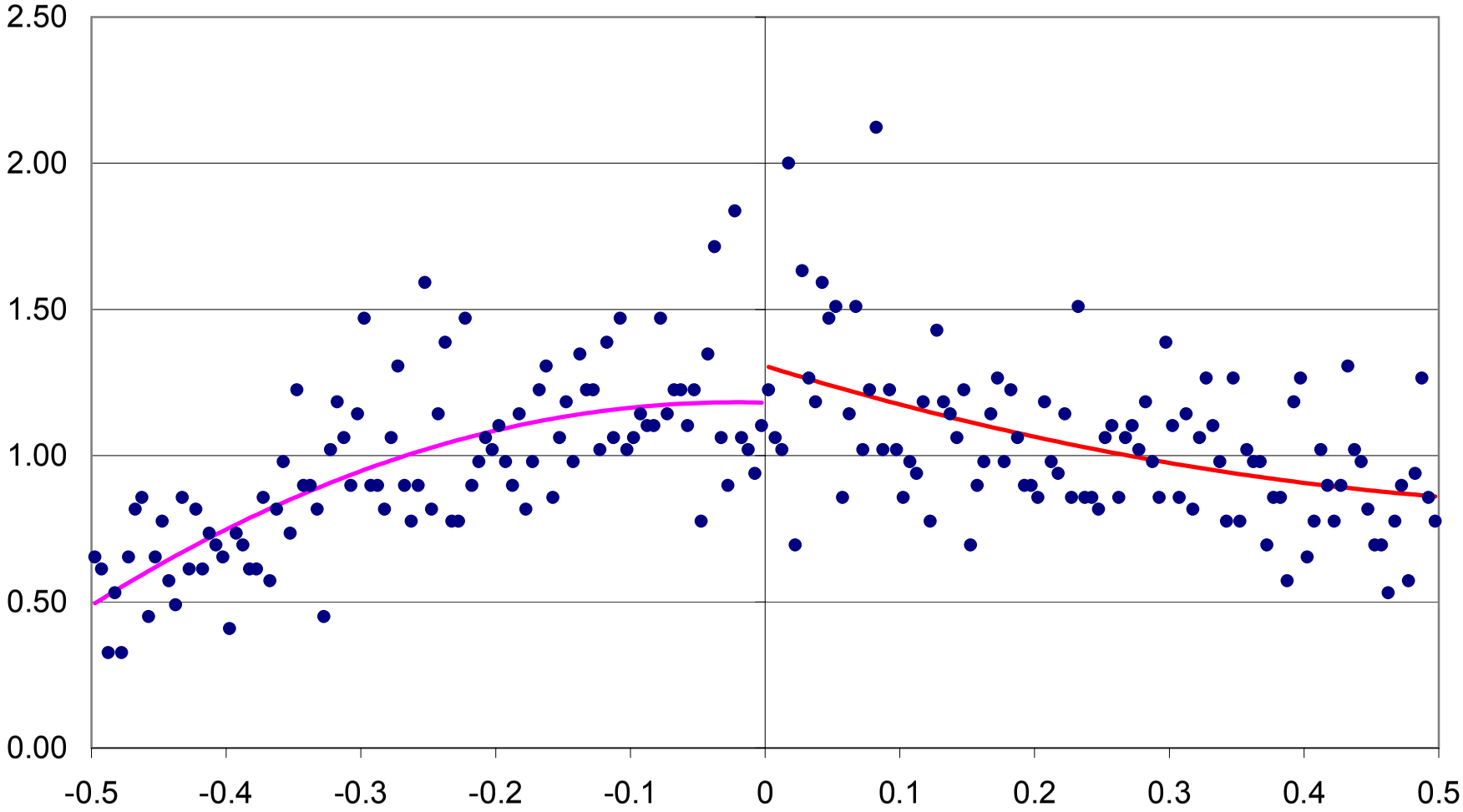


Figure 9: Discontinuity in baseline covariate (share of vote in prior election)

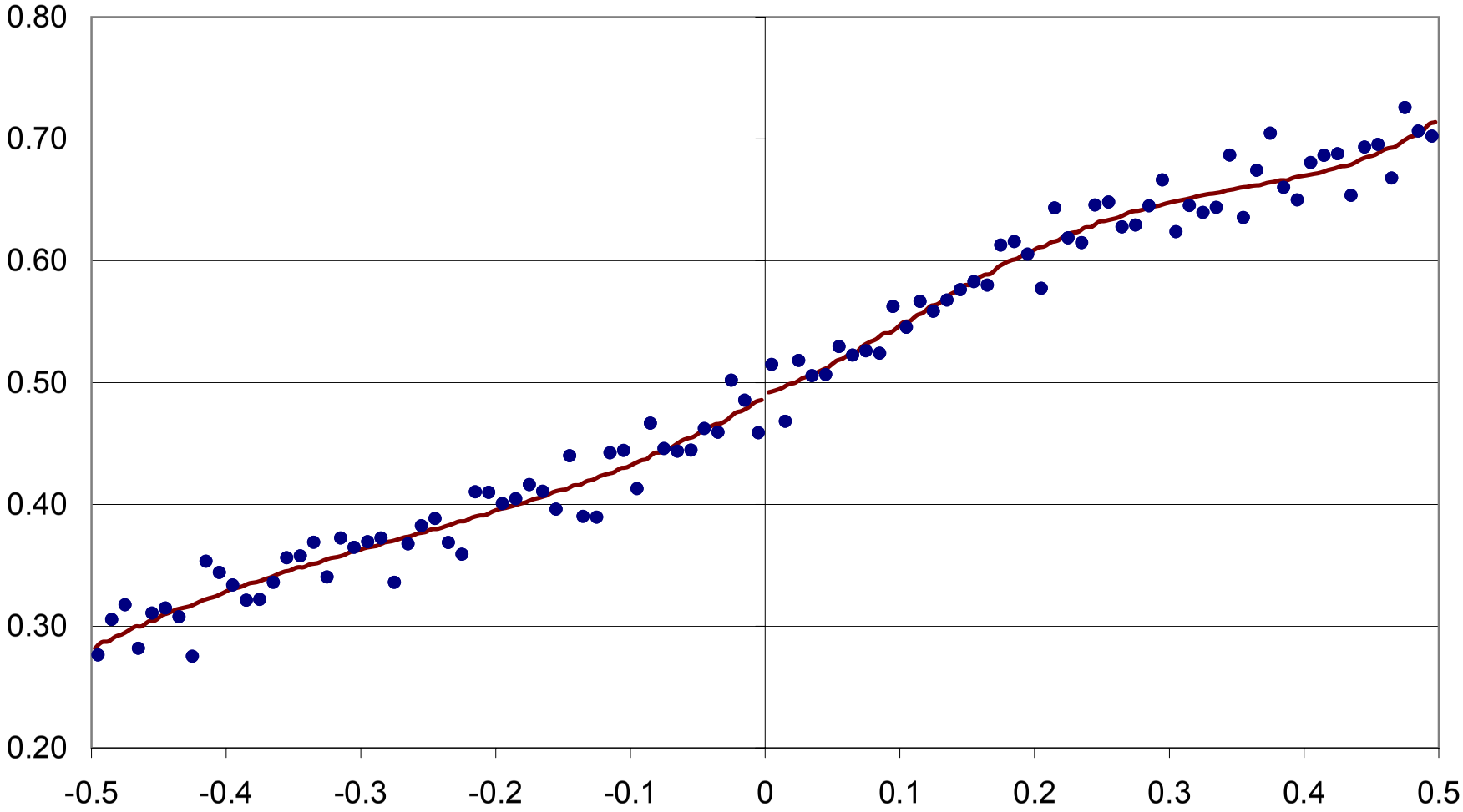


Figure A1: Cross-validation Function for Choosing Bandwidth in a RD Graph: Share of Vote at Next Election

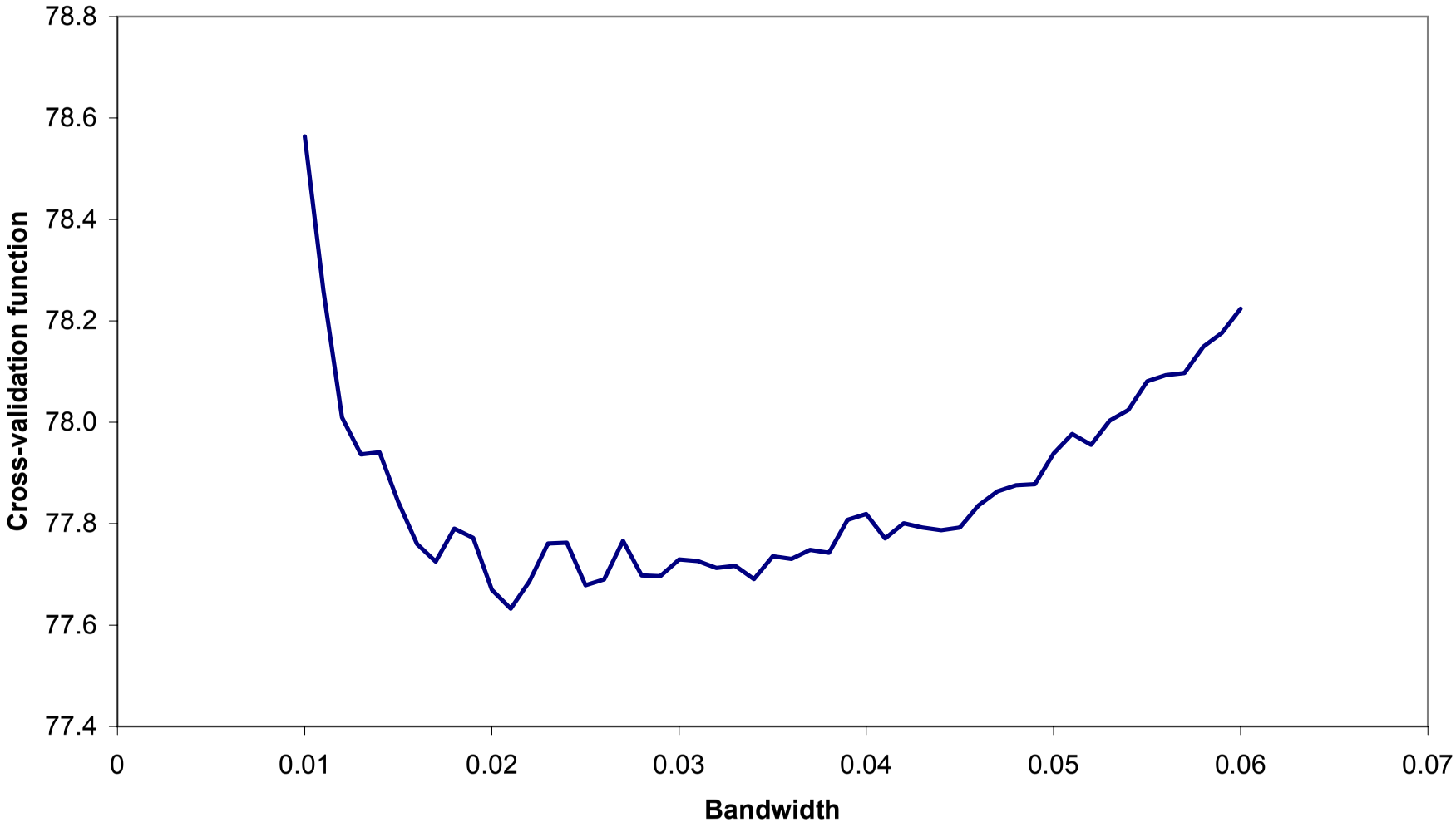


Figure A2: Cross-validation Function for Choosing the Bandwidth in a RD Graph: Winning the Next Election

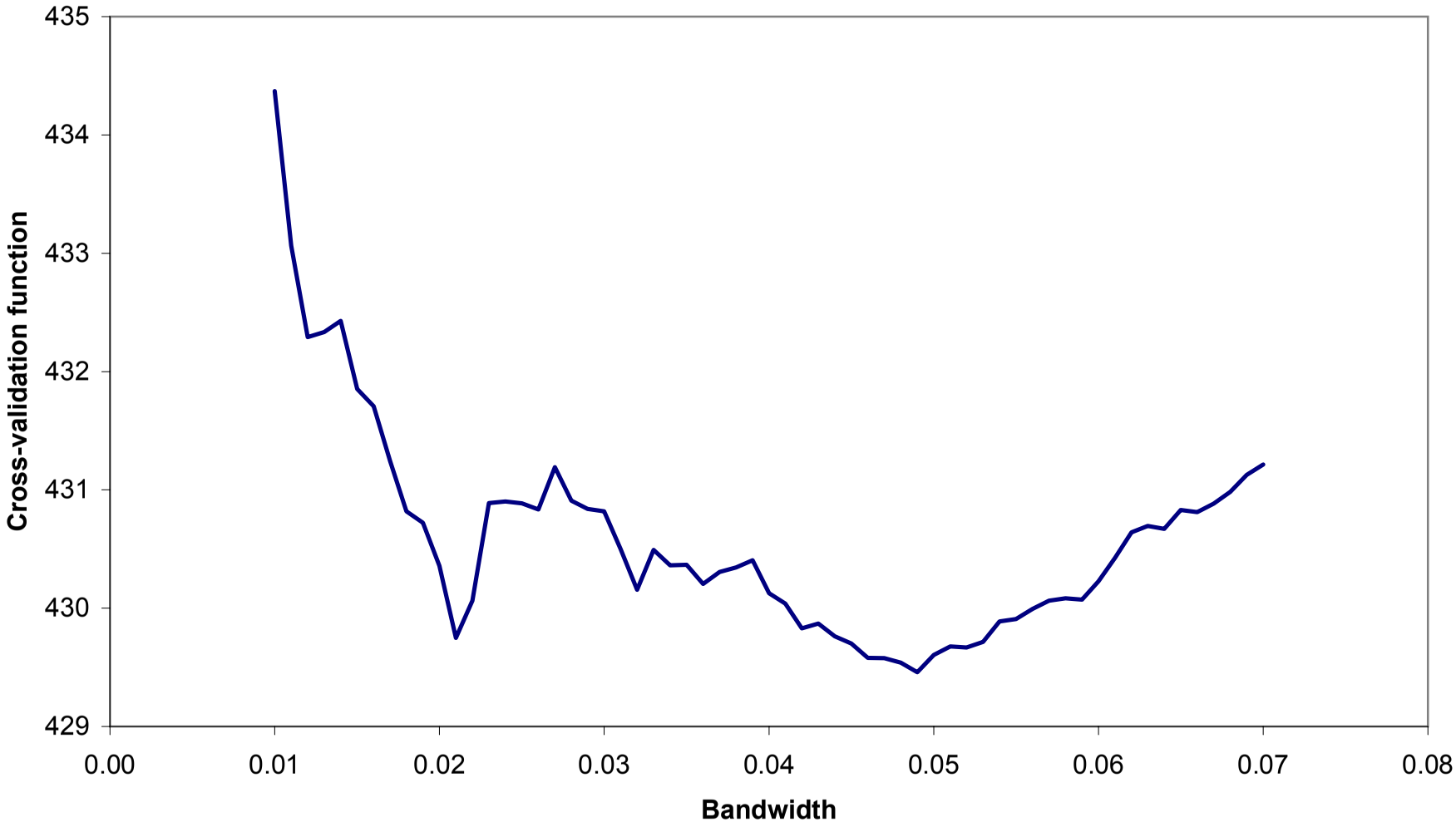


Figure A3: Cross-validation Function for Local Linear Regression: Share of Vote at Next Election

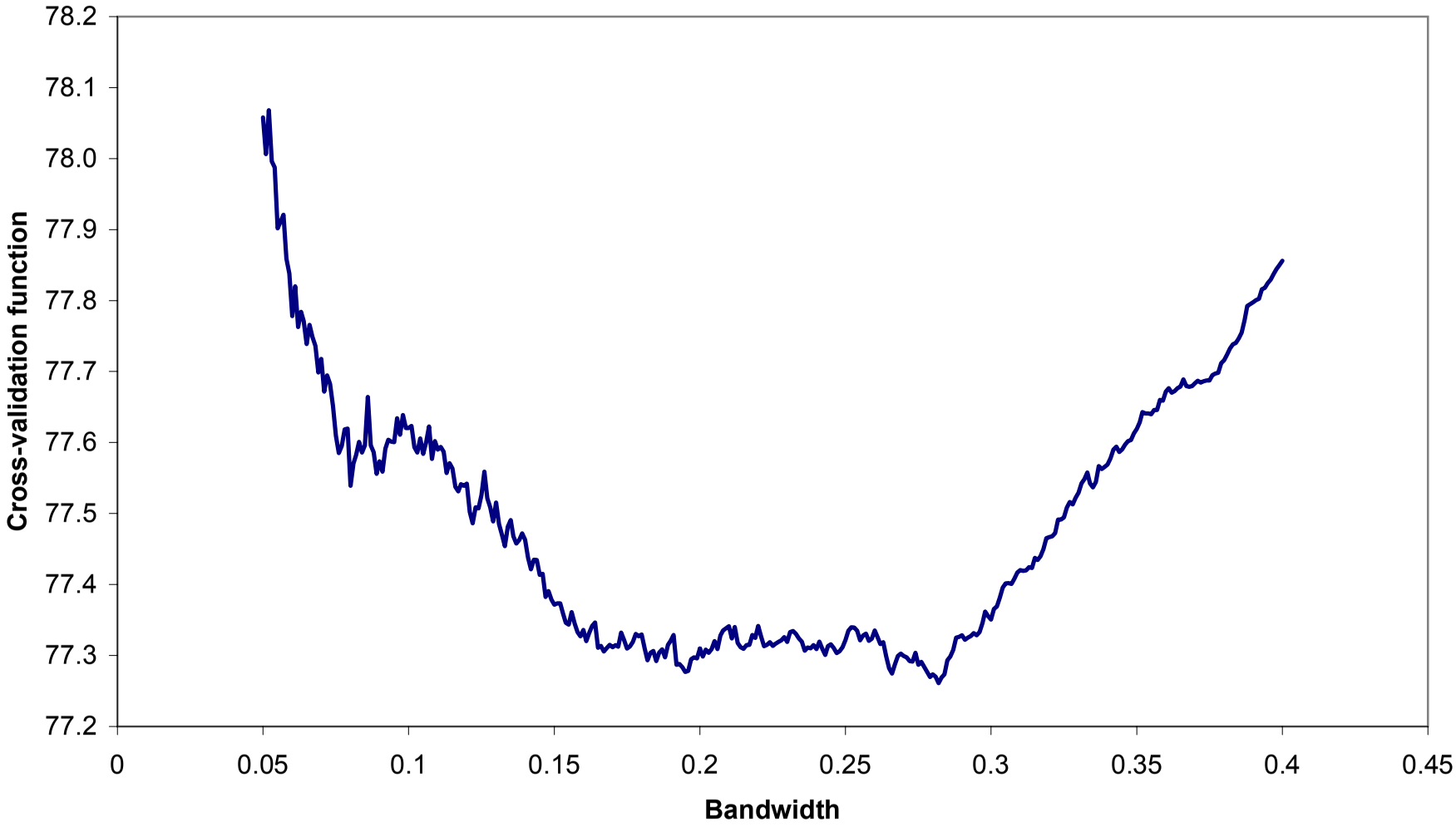


Figure A4: Cross-validation Function for Local Linear Regression: Winning the Next Election