Regression Discontinuity Designs in Social Sciences[1]

David S. Lee
Princeton University and NBER

Thomas Lemieux
University of British Columbia and NBER

May 2013

Abstract

This chapter provides an overview of Regression Discontinuity (RD) designs for social science researchers. It presents the conceptual framework behind the research design, explains when RD is likely to be valid or invalid, draws a parallel between RD and randomized experiments, and summarizes different ways of estimating a treatment effect in the presence of a RD design. Implementation issues are discussed in the context of an example from U.S. House elections (Lee (2008)).

[1]This chapter is an abridged and modified version of Lee and Lemieux (2010). We thank David Autor, David Card, John DiNardo, Guido Imbens, and Justin McCrary for suggestions, as well as for numerous illuminating discussions on the various topics we cover in this review. We also thank Henning Best, Christof Wolf, and Thornsten Kneip for their constructive comments on a first draft of this paper. Diane Alexander, Emily Buchsbaum, Mingyu Chen, Elizabeth Debraggio, Enkeleda Gjeci, Ashley Hodgson, Andrew Langan, Yan Lau, Steve Mello, Pauline Leung, Xiaotong Niu, and Zhuan Pei provided excellent research assistance.

1 Introduction

Regression Discontinuity (RD) designs were initially 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 "assignment" variable (also referred to in the literature as the "forcing" variable or the "running" variable) exceeds a known cutoff point. Thistlethwaite and Campbell (1960) analyzed the impact of merit awards on future academic outcomes in their original study, 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 over fifty years, it only attracted limited attention in economics, and social sciences more generally, until relatively recently.

Since the late 1990s, a burgeoning literature in economics has relied on RD designs to estimate program effects in a wide variety of 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. Following these early papers in the area of education, there has been a rapid growth over the last ten years in the number of studies using RD designs to examine a range of other question. 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 realization 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 assignment 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 and social questions. These two facts likely explain why the RD approach is rapidly becoming a major element in the toolkit of empirical economists and empirical social science researchers more generally.

The goal of this chapter is two fold. First, it seeks to provide the conceptual framework underneath RD designs – what assumptions they require, and their strengths and weaknesses. Second, it discusses the "nuts and bolts" of implementing RD designs in practice. Most of the issues discussed in this chapter are also covered in related pieces by Van der Klaauw (2008), Imbens and Lemieux (2008) and especially Lee and Lemieux (2010). Readers interested in learning more about conceptual and methodological issues should consult these studies, as we only briefly discuss these issues in this chapter.

The rest of the chapter is organized as follows. In Section 2, we introduce RD designs and discuss their main advantages and disadvantages. We 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. Section 3 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. We also provide a summary "checklist" highlighting our key recommendations. These implementation issues are illustrated using an example from U.S. House elections in Section 4. After discussing caveats and frequent errors in Section 5 we conclude by suggesting some further readings in Section 6.

2 Background and Conceptual Framework

In this section, we set the stage for the rest of the paper by discussing the origins and the conceptual framework underneath 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 using the potential outcomes framework. We show how RD designs can be viewed as local randomized experiments and discuss their generalizability. A key feature of RD designs is that they provide a very transparent way of graphically showing how the treatment effect is identified. We thus end the section by discussing how to graph the data in an informative way.

[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.

2.1 Origins and the potential outcomes approach

RD designs were 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. The 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 $X < c$.

Importantly, 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

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

where U 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. This case is depicted in Figure 1, which shows both the true underlying function and numerous realizations of U.

While this simple regression approach is intuitively appealing, it is useful to analyze RD designs more formally to illustrate the key assumptions that need to be satisfied for the design to be valid. A key contribution in this regard is the work of Hahn et al. (2001), who used the approach developed in the treatment effects literature to analyze RD designs. 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 estimating τ that did not assume underlying linearity, as we have assumed 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 individual 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

$$B - A = \lim_{\varepsilon \downarrow 0} E[Y_i | X_i = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} E[Y_i | X_i = c + \varepsilon],$$

which would equal

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

This is the "average treatment effect" at the cutoff c. Note that this particular treatment effect is different from the conventional average treatment effect (ATE) one typically seeks to estimate using a randomized experiment. For example, in Figure 2 we see that the treatment effect (the difference between the two potential outcome curves) depends on the assignment variable X. Therefore, the treatment effect identified at X=c may not be generalizable over the entire population, i.e. over the whole distribution of X.

Generalizability aside, inference is possible here because of the continuity of the underlying functions $E[Y_i(1)|X]$ and $E[Y_i(0)|X]$.[2] 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).

A key question is under which circumstances do we expect this continuity assumption to hold? As it turns out, continuity is a direct consequence of the fact that, under the weak assumptions discussed below, in a RD design we have local randomization around the cutoff point. From that point of view, RD designs are more closely related to randomized experiments, the "gold standard" of program evaluation methods, than to other commonly used methods such as matching on observables or instrumental variables (IV) methods.[3] We

[2]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.

[3]In the survey of Angrist and Krueger (1999), RD is viewed as an IV estimator, thus having essentially the same potential drawbacks and pitfalls. Here we argue that the assumptions required for RD designs to be valid are much weaker than what has to be imposed in the case of instrumental variables.

next explore the connection between RD designs and randomized experiments, and argue that RD designs can be analyzed and treated like randomized experiments.

2.2 RD design and local randomization

We consider a highly simplified example to illustrate the close connection between RD designs and randomized experiments. As we will explain later, the key results on local randomization can also be obtained in a much more general setting. More specifically, we assume that the treatment effect, τ, is constant for all individuals, and that potential outcomes are a linear function of baseline covariates, W, and an error term U:

$$\begin{aligned} Y(0) &= W\delta_1 + U, \\ Y(1) &= \tau + W\delta_1 + U, \end{aligned} \tag{2}$$

where we have omitted the subscript i to simplify the notation. Under these simplifying assumptions, we have a simple linear regression model for the observed outcome Y :

$$Y = (1-D)\cdot Y(0) + D\cdot Y(1) = D\tau + W\delta_1 + U. \tag{3}$$

The assignment variable, X, is assumed to depend linearly on the baseline covariates and a random component V :

$$X = W\delta_2 + V, \tag{4}$$

and treatment assignment is given by

$$D = 1[X \geq c] = 1[W\delta_2 + V \geq c],$$

where $1(.)$ is the indicator function.

Interestingly, a randomized experiment can be viewed as a special case of this model where $\delta_2 = 0$ and V is a randomly generated number used to divide individuals into treatments ($V \geq c$) and controls ($V < c$). Since treatment is randomly assigned, there are no systematic differences between the covariates W and the error term U between the treatment and control groups. In other words, W and U are "balanced" between

treatments and controls in the sense that:

$$\begin{aligned} E[W|D=0] &= E[W|D=1] = E[W], \\ E[U|D=1] &= E[U|D=0] = E[U]. \end{aligned}$$

It follows that

$$\begin{aligned} E[Y|D=1] &= \tau + E[W]\delta_1 + E[U], \\ E[Y|D=0] &= E[W]\delta_1 + E[U], \end{aligned}$$

and

$$\tau = E[Y|D=1] - E[Y|D=0].$$

The treatment effect τ can, therefore, be estimated as a simple difference between the mean outcomes for treatments ($E[Y|D=1]$) and controls ($E[Y|D=0]$). As is well known, one does not need to control for baseline covariates since those are not systematically different for treatment and controls. In the context of the simple regression model in equation (3), this means that failing to include W in a regression of Y on D does not result in an omitted variable bias since W is uncorrelated with D.

Now consider the RD design. To make the above equations more concrete, we work with a case similar to Thistlethwaite and Campbell (1960) where the assignment variable X is a test score that both depends on intrinsic ability, W, and on luck, V. Since future outcomes Y such as earnings, choice of major, etc. also likely depend on ability, we don't expect students above and below the cutoff c to be comparable. This means that, unlike in a randomized experiment, we have $E[W|D=1] \neq E[W|D=0]$ and $E[Y|D=1] - E[Y|D=0] \neq \tau$. But provided that the luck component, V, follows a continuous distribution $f(.)$, randomization will hold locally around the cutoff, and the potential outcomes will be continuous functions of the assignment variable X.

To see this formally, consider a further simplification where W is a dummy variable indicating whether the student is high ($W=1$) or low ($W=0$) ability. Since $X = W\delta_2 + V$, for any given value x of the test score (assignment variable) X , high ability students have a luck term $V = x - \delta_2$, while $V = x$ for low ability

students. Using a few manipulations it follows that:

$$\begin{aligned} E[W|X=x] &= Prob[W=1|X=x] \\ &= \frac{P_w \cdot Prob[X=x|W=1]}{P_w \cdot Prob[X=x|W=1] + (1-P_w) \cdot Prob[X=x|W=0]} \\ &= \frac{P_w \cdot f(x-\delta_2)}{P_w \cdot f(x-\delta_2) + (1-P_w) \cdot f(x)}, \end{aligned}$$

where $P_w = Prob[W=1]$ is the fraction of students who are high ability, $f(.)$ is the probability density function of V, and we have used the fact that $Prob[X=x|W] = Prob[V=x-W\delta_2] = f(x-W\delta_2)$. While it is clear that $E[W|X=x]$ is now a function of the assignment variable X, the function is also *continuous* in X since the probability density function of V, $f(.)$, is itself continuous. To simplify the notation we introduce the function $g(x)$ defined as:

$$g(x) \equiv E[W|X=x] = \frac{P_w \cdot f(x-\delta_2)}{P_w \cdot f(x-\delta_2) + (1-P_w) \cdot f(x)}.$$

When luck on the test, V, is unrelated to the error term U, it follows that

$$\begin{aligned} E[Y(0)|X] &= g(X)\delta_1 + E[U], \\ E[Y(1)|X] &= \tau + g(X)\delta_1 + E[U]. \end{aligned} \tag{5}$$

Since, $g(X)$ is a continuous function, the expected value of the potential outcomes are also continuous in X, thereby satisfying the condition in Hahn et al. (2001). This simple example shows that continuity of the potential outcome functions illustrated in Figure 2 is a consequence of the assumption that there is a random and continuously distributed component V in the assignment variable X.

Local randomization is also a direct consequence of this assumption. In a randomized experiment where 50 percent of individuals are assigned to the treatment and control groups, respectively, each individual is equally likely to be a treatment or a control. In the simple RD design discussed above, we also get that individuals are randomly split in a 50-50 way right around the cutoff point. To see this, consider the probabilities that $X = c+\varepsilon$ and $X = c-\varepsilon$ where ε is a small number. Since the density $f(V)$ is continuous in V it follows that:

$$\lim_{\varepsilon \to 0} \frac{Prob(X=c+\varepsilon)}{Prob(X=c-\varepsilon)} = \lim_{\varepsilon \to 0} \frac{f(c+\varepsilon-W\delta_2)}{f(c-\varepsilon-W\delta_2)} = 1.$$

Since this holds regardless of W and U, it follows that W and U are balanced on each side of the cutoff, i.e.:

$$\begin{aligned}\lim_{\varepsilon\to 0} E[W|X=c+\varepsilon] &= \lim_{\varepsilon\to 0} E[W|X=c-\varepsilon],\\ \lim_{\varepsilon\to 0} E[U|X=c+\varepsilon] &= \lim_{\varepsilon\to 0} E[U|X=c-\varepsilon],\end{aligned}$$

and, therefore:[4]

$$\lim_{\varepsilon\to 0} E[Y|X=c+\varepsilon] - \lim_{\varepsilon\to 0} E[Y|X=c-\varepsilon] = \tau.$$

The difference between randomized experiments and the RD design is that while randomization holds globally (for any value of X) in a randomized experiment, it only holds locally in a RD design. Therefore, while the treatment effect can be computed as simple difference of mean outcomes in a randomized experiment, regression methods have to be used to estimate local means right around the cutoff point in a RD design. In the simple model above, equation (5) yields the following model for observed outcomes:[5]

$$Y = D\tau + g(X)\delta_1 + U, \tag{6}$$

which can be estimated by running a regression of Y on D where X is controlled in a flexible way to account for the function $g(X)$. In the next section, we explain in detail how such flexible regressions can be estimated in practice.

But besides the need to use regression methods instead of comparisons of means, RD designs can be analyzed using the same set of standard procedures that are commonly used in the case of randomized experiments. This includes, for example, checking whether baseline covariates W are balanced on the two sides of the cutoff point. As in a randomized experiment, one also does not need to include the covariates W in a regression model since the mean value of W is locally the same on each side of the cutoff.[6]

[4]This last results follows from the fact that $\lim_{\varepsilon\to 0} E[Y|X=c+\varepsilon] = \tau + \lim_{\varepsilon\to 0} E[W|X=c+\varepsilon]\delta_1 + \lim_{\varepsilon\to 0} E[U|X=c+\varepsilon]$, $\lim_{\varepsilon\to 0} E[Y|X=c-\varepsilon] = \lim_{\varepsilon\to 0} E[W|X=c-\varepsilon]\delta_1 + \lim_{\varepsilon\to 0} E[U|X=c-\varepsilon]$, and, thus, $\lim_{\varepsilon\to 0} E[Y|X=c+\varepsilon] = \lim_{\varepsilon\to 0} E[Y|X=c-\varepsilon] = \tau$.

[5]From equation (5), it follows that $Y(0) = E[Y(0)|X] + U = g(X)\delta_1 + U$ and $Y(0) = E[Y(1)|X] + U = \tau + g(X)\delta_1 + U$. Thus, $Y = (1-D)\cdot Y(0) + D\cdot Y(1) = D\tau + g(X)\delta_1 + U$.

[6]Note that in the simple example we use here, since $g(X)$ is the fraction of high ability types (W=1), it would be fully captured by simply controlling for W in the regression model. But in a more realistic setting, the assignment variable X would also depend on other unobserved factors (e.g. unobserved ability) that are not captured by the covariates W, and are also likely correlated with the error term U. But since the above argument about the continuity of the potential outcomes in W and in X holds regardless of whether W is observed or not, the RD design remains valid and the treatment effect can still be estimated using a flexible regression model.

This simplified example can be easily generalized to a much richer setting where individuals have some control over the assignment variable, as shown in Lee (2008). To see this, consider again the test-taking example. When students know that scoring above a certain threshold (say 80 percent) will give them a scholarship benefit, we expect them to study harder and double check their answers more thoroughly than in a lower stake exam. Effort may well depend both on observed covariates and on the error term U in the outcome. For instance, high-ability students (high value of W) may have much better chances of scoring above 80 percent, which gives them a stronger incentive to try to score above 80 percent. Likewise, a student with a high value of U in the outcome equation may particularly benefit from the scholarship in terms of the program he/she will then be able to afford, etc.[7] Lee (2008) shows that the RD design remains valid in this setting as long as there is still a continuously distributed random component V in the assignment variable that remains beyond the control of the student. This is highly plausible in the test-taking example since students cannot perfectly control the grade they will get on an exam. More generally, one must have some knowledge about the mechanism generating the assignment 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 assignment variable and *precisely* "sort" around the discontinuity threshold. The key word here is "precise", rather than "manipulate". After all, in the above example, 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 assignment variable is to some degree influenced by *someone*.

The main take away points from our discussion of local randomization are the following:

- **RD designs can be invalid if individuals can precisely manipulate the "assignment variable".**
  When there is a payoff or benefit to receiving a treatment, it is natural 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 that is a discontinuous function of an assignment variable is *not* sufficient to justify the validity of an RD design. Indeed, if anything, discontinuous rules may generate incentives, causing

[7] In a more realistic setting we would expect the error term to take on different value $U(0)$ and $U(1)$ in the two potential outcome equations. Since individuals with higher values of $U(1)-U(0)$ gain more from the treatment (higher treatment effect), they would likely put more effort into trying to score high enough to indeed receive the treatment. Lee (2008) shows that local randomization still holds in that setting provided, once again, that individuals have imperfect control over the assignment variable (some randomness in the test score in the example considered here).

behavior that would *invalidate* the RD approach.

- **If individuals – even while having some influence – are unable to *precisely* manipulate the assignment 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 assignment 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 isolates is randomized *as a consequence* of individuals having imprecise control over the assignment 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 assignment 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 assignment 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.[8]

2.3 Fuzzy RD designs

The above discussion is based on what is called a "sharp" RD design, where all individuals above the cutoff receive the treatment, while none of those below the cutoff get treated. However, in many interesting settings,

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

treatment is only determined partly by whether the assignment 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 "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

$$\lim_{\varepsilon \downarrow 0} Prob\left[D = 1 | X = c + \varepsilon\right] \neq \lim_{\varepsilon \uparrow 0} Prob\left[D = 1 | X = c + \varepsilon\right].$$

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 discontinuous jump in the relation between D and X. In this setting, the treatment effect can be written as

$$\tau_{\mathrm{F}} = \frac{\lim_{\varepsilon \downarrow 0} E[Y|X = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} E[Y|X = c + \varepsilon]}{\lim_{\varepsilon \downarrow 0} E[D|X = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} E[D|X = c + \varepsilon]},$$

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 3.3.3.

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[9]

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

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

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, 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 local average treatment effect (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.

2.4 Generalizability

As we pointed out while discussing Figure 2, in a RD design we can only identify the treatment effect right at the cutoff point c. In the fuzzy RD design, this means we can only estimate a local average treatment effect for individuals who are both marginally affected by the instrument (the usual LATE issue) and are right at the cutoff.

Depending on the context, this may be an overly simplistic and pessimistic assessment of how informative the treatment effect estimated using a RD design is for at least two reasons. First, the treatment effect

"right at the cutoff" is often the parameter of policy interest. Going back to the test score example, let us say that students with a GPA of at least 85 are offered a generous scholarship, and that a RD design is used to analyze its impact on future outcomes such as college attendance and earnings. A relevant policy question may be whether it is worth investing more into the program by allowing students with a GPA of 83 and 84 to also get the scholarship. In such a case, the average treatment effect for these students would likely be very close to the RD estimates obtained using the cutoff at a GPA of 85. In such a situation, the average treatment effect estimated using the RD design would be more policy relevant than the average treatment effect (ATE) for the whole population.

A second point, discussed in more detail in Lee and Lemieux (2010), is that the treatment effect estimated using a RD design is a weighted average of the individual treatment effect over the whole population. To see this, remember the treatment assignment rule introduced above: $D = 1[W\delta_2 + V \geq c]$. Since V is random, individuals right around the cutoff point c will have different values of the covariates W depending on the value of V they draw. In particular, individuals drawing a high value of V will tend to have a low value of $W\delta_2$, and vice versa. The treatment effect estimated using the RD design is, therefore, a weighted average of individual treatment effects where the weights are proportional to the conditional probability density function of X given W and U.[10] While it is not possible to know how close the resulting RD gap is from the overall average treatment effect, it remains the case that the treatment effect estimated using a RD design is averaged over a larger population than one would have anticipated from a purely "cut-off" interpretation.

2.5 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 assignment 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

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

[10] See Lee and Lemieux (2010) for more details.

The average value of the outcome variable in the bin is

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

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

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

to detect a possible discontinuity in the assignment variable at the threshold, which would suggest manipulation (see Section 3.4.1).

There are several important advantages in graphing the data this way before performing 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, in Lee and Lemieux (2010) we also discuss formal procedures for selecting the bandwidth.

3 Estimation and Inference

In this section, we systematically discuss the nuts and bolts of implementing RD designs in practice. We first discuss what is, 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 one should control for other covariates and how to assess the validity of the RD design. We then summarize our recommendations for implementing the RD design.

3.1 Regression Methods: Parametric or Non-parametric Regressions?

When we introduced the RD design in Section 2, we used a simple example where the resulting regression model is a non-linear function in the assignment variable X:

$$Y = \alpha + D\tau + g(X)\delta_1 + U,$$

where we have also added an intercept α to the model. Finding a good approximation for the functional form is fairly critical in RD designs since misspecification of the functional form typically generates a bias in the estimated treatment effect, τ.[11] Accordingly, the estimation of RD designs have generally been viewed as a nonparametric estimation problem. In particular, Hahn et al. (2001) suggest running local linear regressions to reduce the importance of the bias. As in many nonparametric estimation problems, one has to choose a particular kernel function. Following Imbens and Lemieux (2008) and Lee and Lemieux (2010), we only look at the case of a rectangular kernel. In practice, this means we can simply run standard linear regressions within a given bin on both sides of the cutoff point to better predict the value of the regression function right at the cutoff point.

The other important implementation issue in nonparametric estimation is the choice of the bandwidth (bin size). With a small bandwidth, the linear approximation will be highly accurate and the bias in the estimated treatment effect will be small. However, the drawback of a small bandwidth is that it yields more imprecise estimates. Therefore, we face a tradeoff between precision and bias, and optimal bandwidth selection procedures seek to find a balance between these two factors.

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 discuss in more detail in Lee and Lemieux (2010), we think that the distinction between parametric and non-parametric methods has sometimes been a source of confusion to practitioners. In practice, it is typically more important 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), than seeking to formally determine what is the "best" specification to use for implementing the RD design. With this in mind, we next explain how to estimate these various regression models.

3.2 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 assignment

[11] 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.

variable, 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

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

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

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

where $g_l(\cdot)$ and $g_r(\cdot)$ 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:

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

where $\tau = \alpha_r - \alpha_l$ and $g(X-c) = g_l(X-c) + D \cdot [g_r(X-c) - g_l(X-c)]$. 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 $g_l(X-c) = \beta_l \cdot (X-c)$ and $g_r(X-c) = \beta_r \cdot (X-c)$, the pooled regression would be

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

If we were to constrain the slope to be identical on both sides of the cutoff ($\beta_r = \beta_l$), 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 ($\alpha_l = \lim_{x \uparrow c} E[Y_i | X_i = x]$) and from the right ($\alpha_r = \lim_{x \downarrow c} E[Y_i | X_i = x]$) of the cutoff. Constraining the slope to be the same would thus be inconsistent with the spirit of the RD design.

In practice, however, estimates where the regression slope or, more generally, the regression function $g(X-c)$ 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.

3.2.1 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 (Imbens and Lemieux (2008)). As a result, the convenience of working with a rectangular kernel compensates for efficiency gains that could be achieved using more sophisticated kernels.

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

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

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. 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 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.[12] 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 of how much the bandwidth should shrink if the sample size were larger so that one obtains asymptotically correct standard errors, and does not help one choose a particular bandwidth in a particular sample.

In the econometrics and statistics literature, there are two "benchmark" approaches commonly considered for choosing optimal 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 (such as the curvature of the regression function) can then be estimated and plugged into the optimal bandwidth function. Imbens and Kalyanaraman (2012) derive such an optimal bandwidth in the case of the RD design.

The second approach is based on a cross-validation procedure. In the case considered here, Ludwig and Miller (2007) 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 is to see how well a regression estimated over a window of width h fits data points (X_i, Y_i) just to the right or to the left of the window.[13] 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.

Let $\widehat{Y}(X_i)$ represent the predicted value of Y obtained using these regressions. The cross-validation criterion is defined as:

$$\text{CV}_Y(h) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}(X_i)\right)^2, \tag{7}$$

with the corresponding cross-validation choice for the bandwidth

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

[12] For technical reasons, however, it would be 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 squares estimator would be invalid. See Hahn et al. (2001) and Imbens and Lemieux (2008) for more details.

[13] 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 ($X_i < X \leq X_i + h$).

Imbens and Lemieux (2008) discuss this procedure in more detail and point out that since we are primarily interested in what happens around the cutoff, it may be advisable to only compute $\text{CV}_Y(h)$ for a subset of observations with values of X close enough to the cutoff point.

3.2.2 Order of Polynomial in Local Polynomial Modeling

In the case of polynomial regressions, for a given bandwidth (typically a large one) one needs to choose 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 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

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

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.[14] 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). Note that while this procedure is useful for choosing one polynomial specification over another, it does not provide a direct indication of how well a particular polynomial model fits the data. For instance, even if a cubic model is preferred to a quadratic model, this

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

does not necessarily mean that the cubic model fits the data well right around the cutoff point. Lee and Lemieux (2010) address this issue by proposing a goodness-of-fit type test that compares how well a given regression model fits the data compared to a fully non-parametric model based on local averages of the outcome variable within a rich set of narrow bins (in X).

3.2.3 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 2.3, here we 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

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

where $T = 1[X \geq c]$ indicates whether the assignment variable exceeds the eligibility threshold c.[15] Note that the sharp RD is a special case where $\gamma = 0$, $g_D(\cdot) = 0$, and $\delta = 1$. It is advisable to draw a graph for the treatment dummy D as a function of the assignment variable X using the same procedure discussed in Section 3.1. 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 $g_D(\cdot)$ looks like.

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

[15] Although the probability of treatment is modeled as a linear probability model here, this does not impose any restrictions on the probability model since $g_D(x - c)$ 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.

be described by the two equation system:

$$Y = \alpha + \tau D + g(X - c) + \varepsilon, \tag{8}$$

$$D = \gamma + \delta T + g_D(X - c) + \nu. \tag{9}$$

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

$$Y = \alpha_r + \tau_r T + g_r(X - c) + \varepsilon_r, \tag{10}$$

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 the reduced form coefficients, τ_r/δ, provided that the same bandwidth is used for equations (9) and (10) in the local linear regression case, and that the same order of polynomial is used for $g_D(\cdot)$ and $g(\cdot)$ in the polynomial regression case.

3.2.4 How to compute standard errors?

As discussed above, for inference in the sharp RD case, we can use standard least squares methods. It is recommended to use heteroskedasticity-robust standard errors (White, 1980) instead of standard least squares standard errors, as usual.[16] 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.

[16] 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 $1/5 < \delta < 2/5$. This is not a very important point in practice, and the usual standard errors can be used with local linear regressions.

3.3 Implementing Empirical Tests of RD Validity and Using Covariates

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

3.3.1 Inspection of the Histogram of the Assignment Variable

Recall that the underlying assumption that generates the local random assignment result is that each individual has imprecise control over the assignment variable, as defined in Section 2.2. We cannot test this directly (since we will only observe one observation on the assignment variable per individual at a given point in time), but an intuitive test of this assumption is whether the *aggregate* distribution of the assignment 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 assignment variable. In the first step, the assignment 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 assignment 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.[17]

3.3.2 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

[17] 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.

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 assignment variable and treatment assignment; if we observe it does, something is wrong in the design.

Lee and Lemieux (2010) also discuss how to jointly test if the data are consistent with no discontinuities for any of the observed covariates. With many covariates, some discontinuities will be statistically significant by random chance. Lee and Lemieux (2010) suggest estimating a Seemingly Unrelated Regression (SUR) system where each equation represents a different baseline covariate, and then performing an χ^2 test for the discontinuity gaps in all equations being zero.

3.4 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, just as in the case of randomized experiments. The simplest way to do so is to add the covariates W to the regression. Unlike the case of X, where we have to be careful when choosing the right functional form for the regression equation, here we can simply include W linearly in the regression. Intuitively, including or not including W in the regression does not affect the consistency of the estimates of τ since W is continuous at the cutoff. So, for the purpose of variance reduction, one can simply use a linear specification in W.

Lee and Lemieux (2010) also suggest a second procedure for incorporating covariates in the estimation based on "residualizing" the dependent variable – subtracting from Y a prediction of Y based on the baseline covariates W – and then conducting 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 one of 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 functional form issues, the estimator will be consistent for the same parameter, just as the estimator for the treatment effect in a randomized experiment will be consistent for the same parameter, no matter what combination of covariates is included. Importantly, this two-step approach also allows one to perform a graphical analysis of the residual. See Lee and Lemieux (2010) for more detail.

3.5 A Recommended "Checklist" for Implementation

Below is a brief summary of our recommendations for the analysis, presentation, and estimation of RD designs. The make the "checklist" more concrete, we refer to specific tables and figures that we discuss in the empirical example of Section 4.

1. **To assess the possibility of manipulation of the assignment variable, show its distribution.** The most straightforward thing to do is to present a histogram of the assignment variable, using a fixed number of bins (see Figure 4). The bin widths should be as small as possible, without compromising the ability to visually see the overall shape of the distribution. The bin-to-bin jumps in the frequencies can provide a sense of whether any jump at the threshold is "unusual". For this reason, we recommend *against* plotting a smooth function comprised of kernel density estimates. A more formal test of a discontinuity in the density can be found in McCrary (2008).

2. **Present the main RD graph using binned local averages.** As with the histogram, we recommend using a fixed number of non-overlapping bins, as described in Subsection 2.5 and illustrated in Figure 3. The non-overlapping nature of the bins for the local averages is important; we recommend against simply presenting a continuum of nonparametric estimates (with a single break at the threshold), as this will naturally tend to give the impression of a discontinuity even if there does not exist one in the population. We recommend generally "undersmoothing", while at the same time avoiding "overly narrow" bins that produce a scatter of data points, from which it is difficult to see the shape of the underlying function. Indeed, we recommend against simply plotting the raw data without a minimal amount of local averaging.

3. **Graph a benchmark polynomial specification.** Super-impose onto the graph the predicted values from a low-order polynomial specification (see Figure 3). One can often informally assess, by comparing the two functions, whether a simple polynomial specification is an adequate summary of the data. If the local averages represent the most flexible "non-parametric" representation of the function, the polynomial represents a "best-case" scenario in terms of the variance of the RD estimate, since if the polynomial specification is correct, under certain conditions, the least squares estimator is efficient.

4. **Explore the sensitivity of the results to a range of bandwidths, and a range of orders to the polynomial.** For an example, see Table 1. It is useful to supplement the table with information on

optimal bandwidths selected using a plug-in or cross-validation procedure for local linear regression, as well as the AIC-implied optimal order of the polynomial. A useful graphical device for illustrating the sensitivity of the results to bandwidths is to plot the local linear discontinuity estimate against a continuum of bandwidths. For an example of such a presentation, see Figure 18 in Lee and Lemieux (2010).

5. **Conduct a parallel RD analysis on the baseline covariates.** If the assumption that there is no precise manipulation or sorting of the assignment variable is valid, then there should be no discontinuities in variables that are determined prior to the assignment (see Figure 5).

6. **Explore the sensitivity of the results to the inclusion of baseline covariates.** The inclusion of baseline covariates – no matter how highly correlated they are with the outcome – should not affect the estimated discontinuity, if the no-manipulation assumption holds. If the estimates do change in an important way, it may indicate a potential sorting of the assignment variable that may be reflected in a discontinuity in one or more of the baseline covariates. In terms of implementation, in Subsection 3.4, we suggest, after choosing a suitable order of polynomial, simply including the covariates directly. An alternative "residualizing" procedure could also be used.

We recognize that due to space limitations, researchers may be unable to present every permutation of presentation within an published article. Nevertheless, we do believe that documenting the sensitivity of the results to these array of tests and alternative specifications – even if they only appear in unpublished, online appendices – is an important component of a thorough RD analysis.

4 Empirical example

In this section, we illustrate the various concepts discussed above 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 assignment 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. The outcome variable is the share of votes in the next election. In the presence of "incumbency effects", we should observe a jump in the outcome variable for Democrats who barely won the previous election – and

are just above the 50 percent cutoff – relative to those who barely lost it. The data and Stata programs used for this empirical example are provided along with this chapter.

Starting with the graphical representation of the data, Figure 3 shows the bin means using a bandwidth of 0.01, along with the fitted values from a quartic regression model estimated separately on each side of the cutoff point. Note that the assignment 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 assignment 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. The graph shows clear evidence of a discontinuity at the cutoff point.

Turning to the local linear regressions, an important question is how to choose the bandwidth. Using the same voting data, Lee and Lemieux (2010) show that the optimal bandwidth chosen using a cross-validation procedure is equal to 0.282. Imbens and Kalyanaraman (2012)'s plug-in procedure yield an optimal bandwidth in the 0.26-0.29 range (for the same data), which is similar to the bandwidth selected using the cross-validation procedure.

Table 1 shows the estimates of the treatment effect for a rich set of specifications (up to a quartic) and bandwidths. Local linear regression estimates of the treatment effect are reported in the second row of the table. 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 slightly 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. The fact that the estimates from bandwidths around .25 (near the calculations of the "optimal bandwidth") are quite similar is consistent with the graphical evidence in Figure 3, which suggest relatively little curvature in the regression function. Since the linear approximation is accurate over a relatively wide range of values of X , bias is not much of an issue and it is sensible to use a relatively large bandwidth to increase precision.

As in Lee (2008), the estimates reported in Table 1 suggest large incumbency effects in the 0.05 to 0.10 range. When Democrats barely win a congressional election they get, on average, an additional 5 to 10 percent share of the vote in the next election relative to Democrats who barely lose an election. This large causal effect of incumbency may come from a variety of channels such as more exposure in the media and community, more ability to raise money, etc.

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 – as is the case here – 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.

In the case of polynomial regressions, the order of the polynomial selected using the AIC for each bandwidth is presented at the bottom of Table 1. The p-values of Lee and Lemieux (2010)'s goodness-of-fit tests are reported in square brackets. Broadly speaking, the goodness-of-fit tests do a very good job ruling out clearly misspecified models, like the zero order polynomials (simple comparison of means on each side of the cutoff) with large bandwidths that yield upward biased estimates of the treatment effect. Estimates of τ from models that pass the goodness-of-fit test mostly fall in the 0.05-0.10 range.

Looking informally at 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 (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.

Turning to possible evidence of manipulation, Figure 4 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) who also uses Lee (2008)'s data in his paper, 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.

There is also no evidence of discontinuity in baseline covariates in the voting data. For instance, Figure 5 considers the case where the Democratic vote share in the election prior to the one used for the assignment variable (four years prior to the current election) is used as baseline covariate. 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.

5 Caveats and Frequent Errors

We now comment on two implementation issues that seem both innocuous at a superficial level, but pose problems in correctly interpreting evidence from a regression discontinuity design. First is the issue of relying on a single, or "the" optimal bandwidth in computing RD estimates. While it is tempting to think there is a single, best bandwidth formula that "should" be used in all RD analyses, the fact is that "optimality" can be defined in so many different ways (see discussion in Imbens and Kalyanaraman (2012)), and different notions of optimality will lead to different recommended bandwidth formulas. In our view, it is a mistake in empirical analyses to only report the estimate from a single bandwidth, no matter its claim to be optimal. At some level, there is no escaping that one limitation of RD designs is that the underlying functional form is unknown, and for every problem there will be a number of different and equally sensible specifications (or bandwidths), and – especially with relatively smaller sample sizes – those specifications may lead to somewhat different answers. Our recommendation is, in all cases, to report estimates using a range of bandwidths (or specifications) in addition to computing "benchmark" bandwidths commonly computed in the literature for reference. In a sense, exploring the sensitivity of the empirical analysis to equally plausible specifications is common practice in quality empirical work. The case of RD analysis is no exception. The only difference is that the question of the "right" specification does not have anything to do with what variables to include; instead it has everything to do with how best to model the shape of the relation between the outcome and the assignment variable.

A second, rather pernicious problem has to do with sample selection and missing values. For example, suppose there is missing data in the voting data that we use in this chapter, and that whether the data is missing is related to whether the district was won in period t by a Democrat or a Republican. Then, selecting only data for which there are non-missing values for outcomes in $t+1$ (the outcome variable) will necessarily induce a classic sample selection problem. So when one examines, for example, the baseline characteristics, one might be tempted to interpret a discontinuity as evidence of precise sorting or manipulation of the assignment variable. Unfortunately, this is potentially an erroneous inference, since a discontinuity in the baseline covariates could be completely consistent with a valid RD design (i.e. local randomization), with the discontinuity being driven by sample selection. To see why this is the case, it is helpful to consider the analogy of the randomized experiment. Clearly, if treatment status affects attrition propensities, then for the *selected* sample, the treatment and control no longer have a similar composition, even if the treatment was

perfectly randomized at the outset. Another way such non-random sample selection can occur is through the merging of extra variables from other datasets. As an example, suppose one wanted to add campaign expenditure data to the election data used in this chapter, but the availability of such data was partially influenced by whether a district was won by a Democrat or Republican. After merging the data, the researcher is tempted to simply select the data for which there are no missing values for all of the variables. This seemingly innocuous merging of data can also generate sample selection bias for the same reasons discussed above.

Unfortunately, just as in a randomized experiment with non-random sample selection, there are no assumption-free, bullet-proof methods to adjust for this. In the case where missing values are only a problem for the outcome variable, one could impute the missing values (e.g. with the mean of the outcome variable) to ensure that the full sample (not a selected one) is being utilized in the analysis. This may not eliminate any bias in the estimated causal effect of the treatment on the outcome, but at least a full sample would ensure that the test of continuity of the baseline covariates would still be informative about the imprecise control/manipulation assumption, rather than a test that confounds an invalid RD design and sample selection.

One diagnostic to gauge whether sample selection is a problem is to perform a RD analysis of the dummy variable that is equal to 1 if the outcome data is non-missing, and 0 otherwise. Inspecting whether there is a discontinuity in the probability of sample selection, would at least give some evidence on whether the treatment did impact selection into the sample. Failing to reject continuity would not prove an absence of sample selection, but a rejection would strongly suggest a problem. As an example, Caughey and Sekhon (2011) use the U.S. House election data from 1942-2008, and for years that overlap, use a sub-sample of the data used in Lee (2008). They find, in contrast to Lee (2008), discontinuities in key baseline covariates (as well as discontinuities in new covariates that were merged on to the original Lee (2008) data). They interpret this as evidence of sorting or manipulation of the assignment variable. As mentioned above, a simpler explanation is that the data selection process and/or merging of new variables with missing values induced sample selection. Figure 6 plots the estimated probability that an observation in the Lee (2008) data has a non-missing democratic vote share (period $t-1$) in the Caughey and Sekhon (2011) data. As the figure shows, over the same time frame, Caughey and Sekhon (2011) use a strict subset of the data. More importantly, the selection (into the Caughey and Sekhon (2011) data) probability drops discontinuously from about .9 to about .8 (the point estimate is -0.114 with a standard error of 0.035): the data do seem to indicate

a sample selection problem of the sort described above.

6 Further readings

Most of the issues discussed in this chapter are addressed in more detail in Lee and Lemieux (2010) who also provide an extensive survey of recent applications of the RD design in economics. Cook (2008) provides some complementary coverage on the history of the RD design in other disciplines. Other recent surveys include Van der Klaauw (2008) and Imbens and Lemieux (2008). Finally, a more thorough coverage of the methodological issues discussed in Section 2 is provided by Hahn et al. (2001) and Lee (2008).

References

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.

Black, Dan A., Jose Galdo, and Jeffrey A. Smith, "Evaluating the Worker Profiling and Reemployment Services System Using a Regression Discontinuity Approach," *American Economic Review*, 2007, *97(2)*, 104–107.

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

Caughey, Devin and Jasjeet S. Sekhon, "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008," *Political Analysis*, 2011, *19*, 385–408.

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.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, January 2001, *69* (1), 201–209.

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," *Review of Economic Studies*, 2012, *79* (3), 933–960.

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

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

_ **and Thomas Lemieux**, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, June 2010, *48* (2), 281–355.

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

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

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

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.

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

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

White, Halbert, "A Heteroskedasticty-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 1980, *48* (4), 817–838.

Table 1: 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)

Figure 1: Simple Linear RD Setup

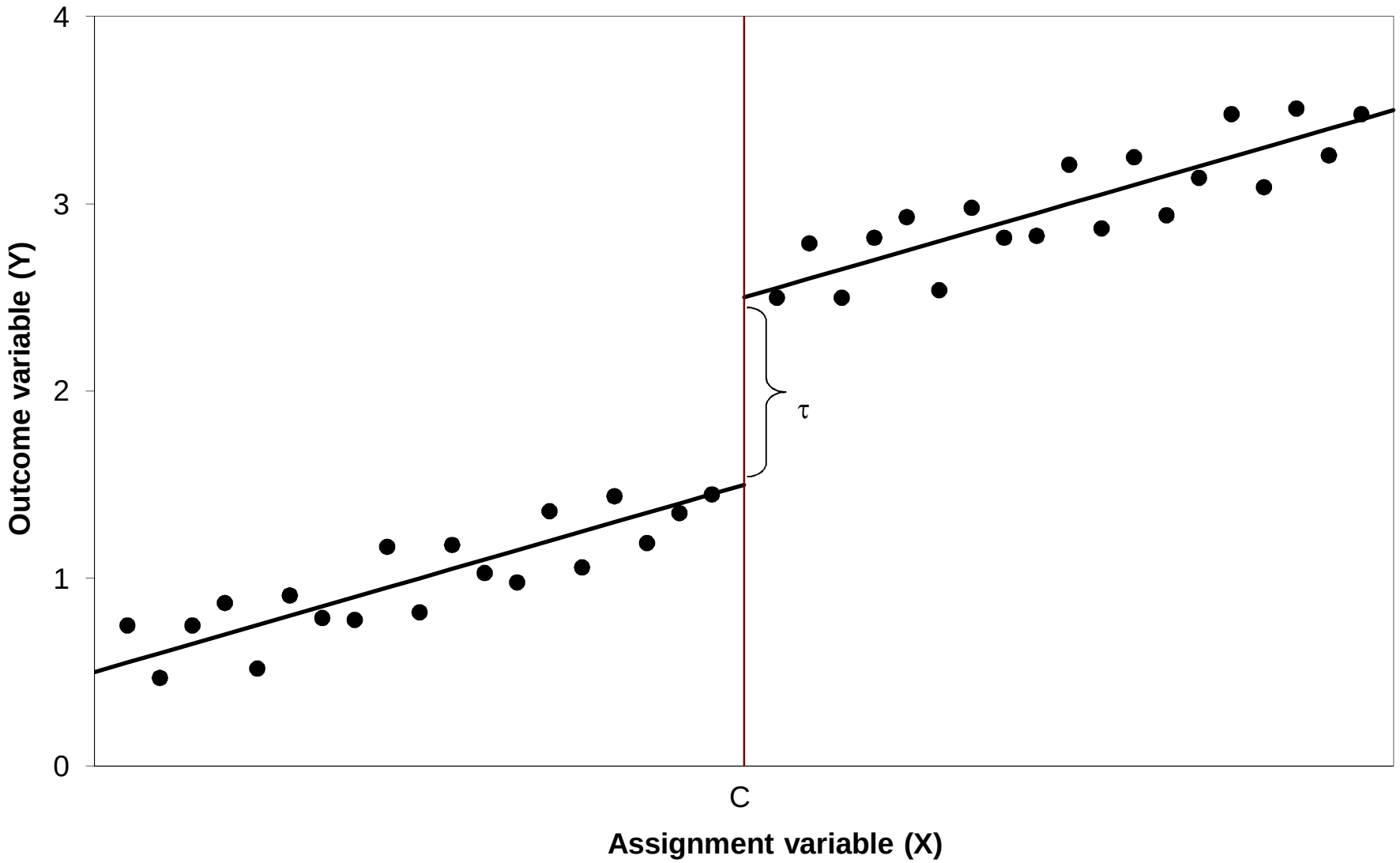

Figure 2: Nonlinear RD

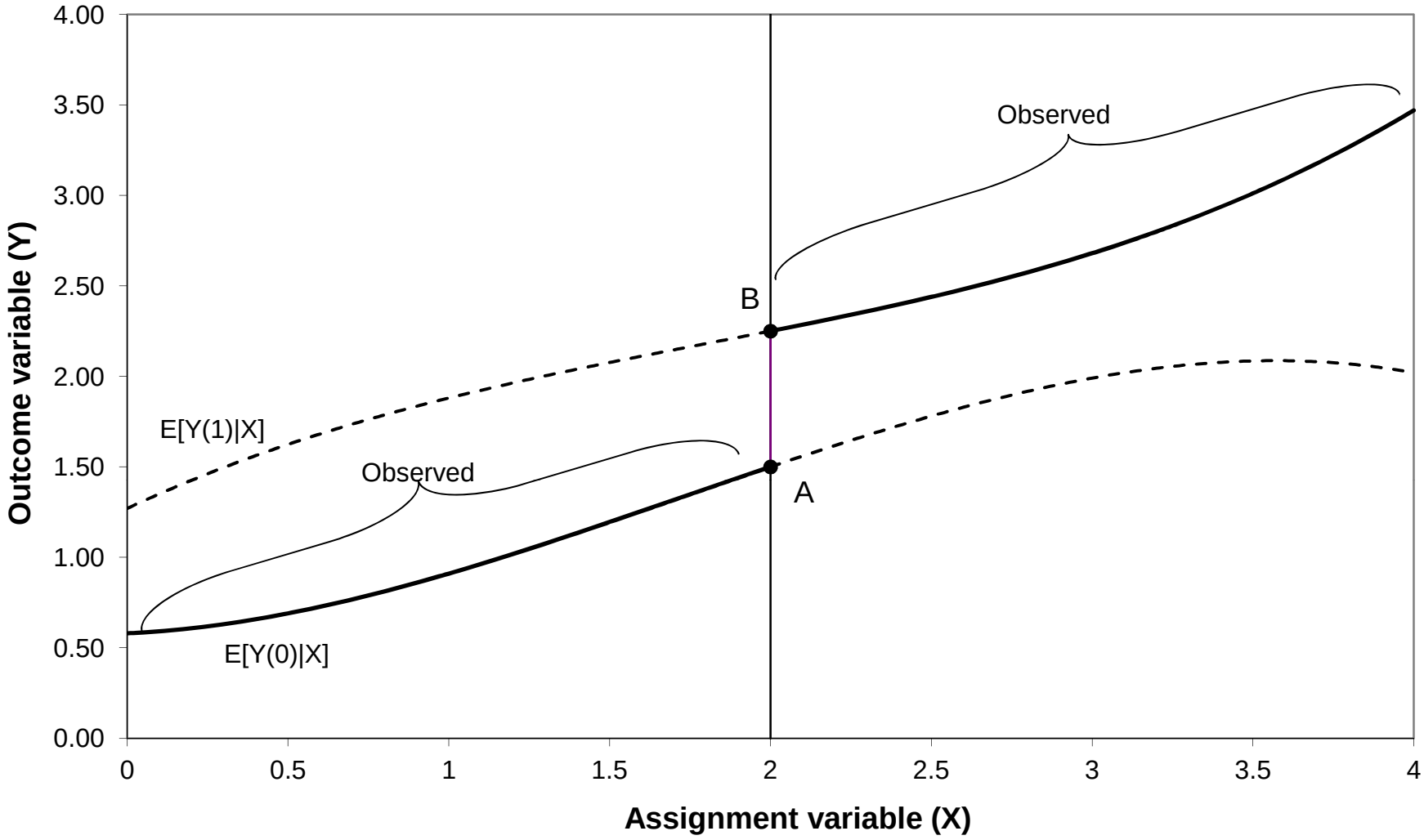

Figure 3: Share of vote in next election, bandwidth of 0.01 (100 bins)

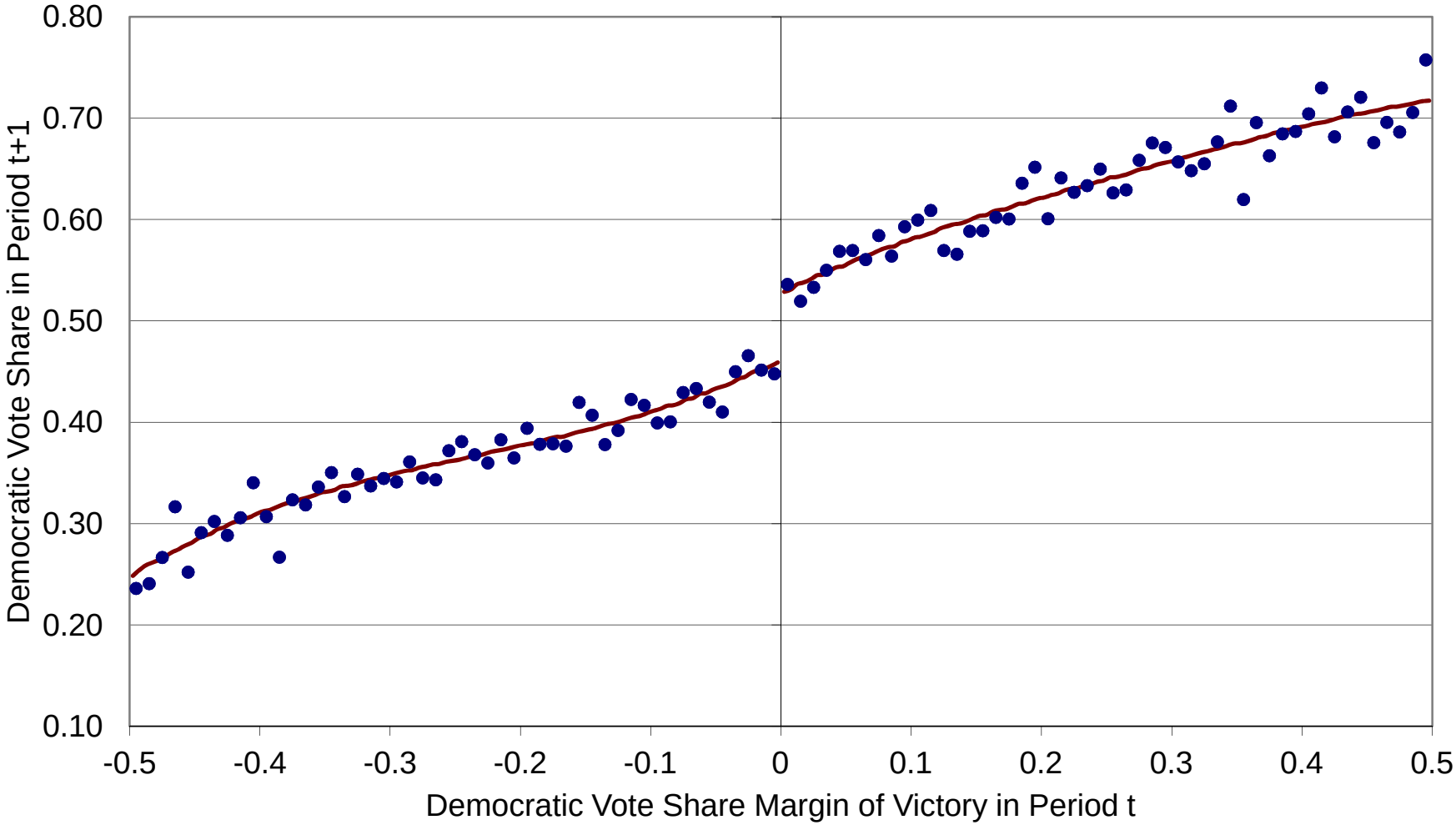

Figure 4: Density of the assignment variable (vote share in previous election)

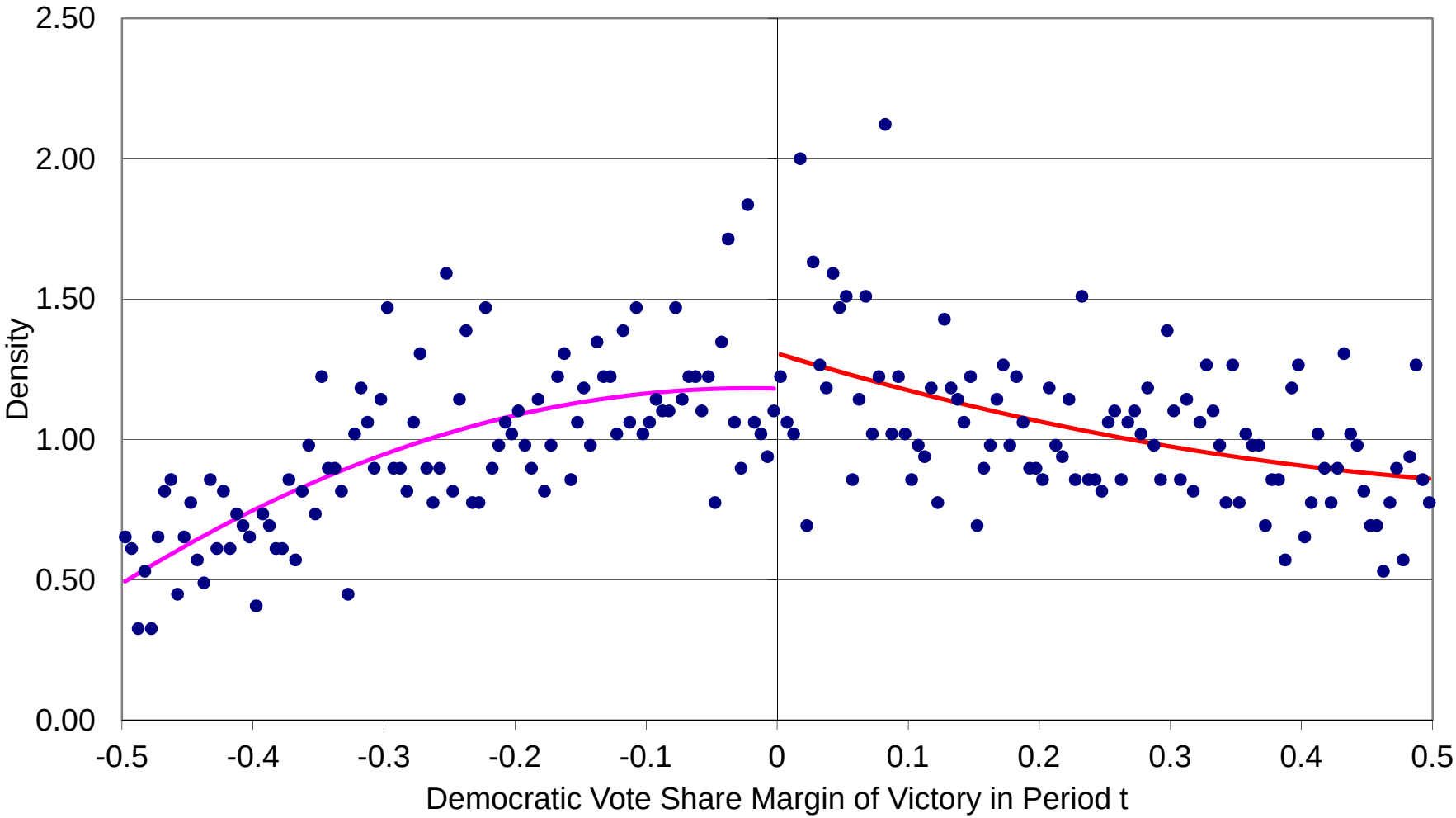

Figure 5: Discontinuity in baseline covariate (share of vote in prior election)

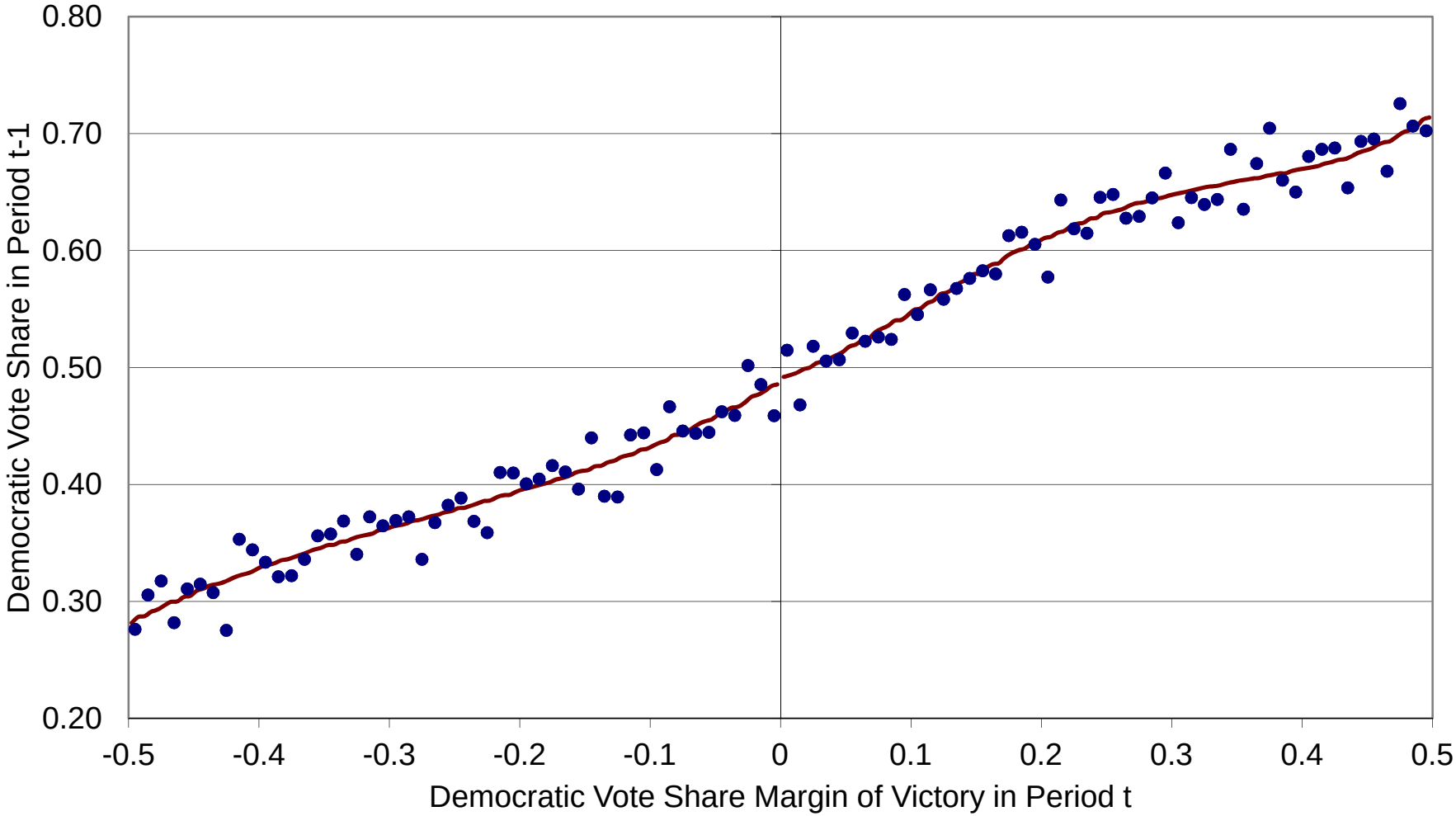

Figure 6: Discontinuity in the probability that an observation is non-missing in Caughey and Sekhon (2011) data