Limitless Regression Discontinuity

September 18, 2014

1 Introduction

Randomization of treatment assignment is the gold-standard of causal inference in statistics: it is the only sure way to control all confounding. Of the options available when experiments are impossible, the "regression discontinuity design" (RDD) [Thistlethwaite and Campbell, 1960, Cook, 2008, Imbens and Lemieux, 2008, Lee and Lemieux, 2010] is among the most credible. In an RDD, the treatment assignment mechanism is known: each subject i has a "running variable" R_i, and those subjects whose R_i is greater (or less) than a pre-determined constant c are assigned to treatment. Lee [2008] argued that the RDD features "local randomization" of treatment assignment, and is therefore "a highly credible and transparent way of estimating program effects" [Lee and Lemieux, 2010, p. 282].

Lee's notion of local randomization is that, in a well-behaved RDD, one can recover the important advantages of randomized experiments—specifically, no confounding, unbiased estimation of average treatment effects (ATE) and covariate balance—by replacing statements about samples or populations with statements about limits. The conventional approach to RDDs [eg. Berk and Rauma, 1983, Angrist and Lavy, 1999, Oreopoulos, 2006], consistent with this "limit" understanding, uses regression to estimate the average functional relationship between the outcomes Y, the running variable R and treatment assignment Z.

However, formulating local randomization in terms of limits fails to realize some other benefits of experiments. In particular, randomized experiments allow researchers to estimate treatment effects for specific samples or sub-populations. In contradistinction, conventional RDD analysis estimates an ambiguously weighted population ATE [as in Lee, 2008], or the limit of sample ATEs as sub-populations shrink around a cutoff [as in Hahn et al., 2001]. Similarly, randomized experiments allow analysts access to distribution-free inferential methods, equally valid for large or small samples [Fisher, 1935, Pitman, 1937, Rosenbaum, 2002a].

This paper will suggest a method of modeling RDDs that realizes these benefits while also retaining some of the flavor of the conventional approach, using regression techniques to disentangle Y and R. The new approach has several advantages over the conventional approach. Its assumptions align better with the heuristic motivation that RDDs leverage natural randomization in the vicinity of a cut point. Since it relies on exact inference or permutation tests, it is suitable for small-sample inference. Similarly, the conceptual approach does not demand continuity in the running variable R, since it does not rely on taking limits of continuous functions of R. In fact, the data example we discuss below features a discrete R, which poses conceptual difficulties for the conventional approach.

The new approach allows the use of covariate adjustment not only to make estimates more precise but also to expand the notion of "local." That is, it estimates the treatment effect in a broader sample. Similarly, it suggests testable consequences of the model that can be wholly separate from estimation of treatment effects. In some scenarios, it allows researchers to use those consequences to make improvements to the model that would be incompatible with the conventional, limit-focused approach.

Our RDD analysis strategy unites the conventional, regression-based approach, with a randomization-based approach, as advanced primarily in Cattaneo et al. [2014]. Briefly, the idea is to use regression modeling to disentangle the outcome from the running variable,

producing "transformed" outcomes that are hopefully independent of treatment assignment, as (un-transformed) outcomes would be in an experiment. Then, use randomization inference techniques to test for, and estimate, an effect.

Like the conventional approach, it models outcomes and treatment effects as functions of the running variable, and typically fits these models with regression. Like randomization inference, its identification emerges from an ignorabililty assumption in line with what one might see from a randomized experiment, and its primary inferential tools are the same as in randomization inference. These links are technical, as well as conceptual. The approach we present here contains the method in Cattaneo et al. [2014] as a special case. Further, with certain modeling and inferential choices, it will reproduce a version of the conventional estimator. In that sense, this paper can be seen as giving a randomization-inference interpretation to conventional RDD analysis.

Our case study is the regression discontinuity design found in Lindo, Sanders, and Oreopoulos [2010] (hereafter LSO). In many colleges and universities, struggling students are put on "academic probation" (AP); the school administration monitors these students and devotes additional resources to them. In addition, if their grade-point-averages (GPAs) fail to improve, they are subject to suspension. But does AP actually help these students? What is the effect of AP on students' subsequent GPAs? LSO realized that at a certain large Canadian university, AP status was function of students' GPAs: students with GPAs below 1.5 or 1.6, depending on the campus, were put on AP. The GPA-AP system, then, is a classical RDD. One complication, though, is that GPA is measured in increments of 1/100, and is therefore discrete–in fact, ties abound. This fact is somewhat at odds with the conventional, limit-focused interpretation of RDD analysis, but is entirely compatible with our approach.

The following section will discuss our novel approach to RDD analysis, discussing testing a strict null hypothesis and estimating effects. Section three will discuss two approaches to protect causal inference from model misspecification: limiting the focus of the analysis

to a window around the cutoff, and covariate balance tests. Section four will compare and contrast our approach with the standard method, the randomization-inference approach in Cattaneo et al. [2014], and the suggestion in Angrist and Rokkanen [2012] for estimating causal effects away from the cutoff. Section five will demonstrate the new approach on the LSO dataset, and section six will conclude.

1.1 Notation: the Rubin Causal Model

In a randomized experiment with a binary treatment, let $Z_i \in \{0, 1\}$ be a random variable coding subject i's treatment status ($Z = 1$ signifies treatment). Let Y represent the outcome of interest. Then each subject has two (possibly random) values: Y_{Ci} is subject i's outcome *if subject i is* not *treated* and Y_{Ti} is subject i's outcome *if subject i is treated* (Rubin [1974], Splawa-Neyman et al. [1990]). For each i, only one of these two values is observed, dependent on Z_i; subject i's observed outcome is $Y_i = Z_i Y_{Ti} + (1 - Z_i)Y_{Ci}$. The values Y_{Ti} and Y_{Ci} are called "counterfactuals" or "potential outcomes." This notation implicitly assumes *non-interference* [Cox, 1958] (part of the "stable unit treatment value assumption" in Rubin 1978): for alternative vectors of treatment assignments $\boldsymbol{Z}$ and $\boldsymbol{Z'}$, if $Z_i = Z'_i$ then $Y_{\boldsymbol{Z}} = Y_{\boldsymbol{Z'}}$. The only element of $\boldsymbol{Z}$ that affects Y_i is the ith.

In addition, each subject i has a vector of pre-treatment covariates X_i.

2 Transformed Ignorability: A Basis for Inference in RDDs

In a RDD, each subject i is characterized by a running variable value R_i. In a "sharp" RDD, subject i's treatment is a deterministic function of R_i: $Z_i = R_i > c$ for a known constant c. Of course, the inequality need not be strict, (e.g. $R \geq c$) or may go in the opposite direct

(e.g. $R \leq c$) but the theory remains the same; we will focus on the strict case $R > c$.

RDDs differ from other observational studies in two important ways [See Angrist and Rokkanen, 2012, p. 10]. First, in RDDs there is no threat of omitted variable bias: conditional on R, treatment assignment is trivially ignorable. In this sense, causal inference from RDDs is easier than in other observational studies. The second difference cuts the other way. In sharp RDDs, by definition, there is no covariate overlap—R values are necessarily different for treated and untreated subjects. As a result, most matching techniques are invalid for RDDs.[1] For this reason, conventional approaches to analyzing RDDs have relied on modeling the relationship between R and $\{Y_C, Y_T\}$, with the goal of estimating a treatment effect for subjects at the margin [Imbens and Lemieux, 2008, Angrist and Pischke, 2009] or a weighted average treatment effect for a wider group [Lee, 2005]. The method we are presenting here also relies on modeling potential outcomes as a function of R, but under an alternative conceptualization that, in some circumstances, may be more appropriate than the conventional story.

Choosing an appropriate statistical model is always difficult, and RDDs are no exception. We put off a detailed discussion of this important point until the next section, where we will provide some guidance on model choice, as well as robustness results. One tool, common in the RDD literature [Imbens and Lemieux, 2008, DesJardins and McCall, 2008, Imbens and Kalyanaraman, 2012, Cattaneo et al., 2014], is to fit the model to a limited window of analysis surrounding the cutoff. That is, for a constant "bandwidth" b,[2] define the set of subjects $\mathcal{W}$ as

$$i \in \mathcal{W} \text{ if } c - b \leq R_i \leq c + b. \tag{1}$$

In our approach, $\mathcal{W}$ will define both the data that we will use for estimation, and the target

[1] See, however, Hansen [2008] which suggests matching on prognostic scores. Additionally, the method suggested in Cattaneo et al. [2014] can be conceptualized as an approximate matching scheme with a caliper: subjects close to the cutoff, on either side, are "matched" in one stratum, and the rest are discarded.

[2] The window of analysis need not be symmetric about c, but here we consider symmetric windows for simplicity.

population for inference.

2.1 Testing Fisher's Null Hypothesis

Testing for causal inference typically requires some form of ignorability assumption, the strongest form of which is

$$Y_C \perp\!\!\!\perp Z. \tag{2}$$

Typically, Z is not ignorable in RDDs, since it is a function of R, which often correlates with Y_C. Indeed, (2) may not be plausible even in relatively small windows around the cutoff $\mathcal{W}$, if Y_C is strongly related to R. To weaken the ignorability assumption (2), let $Y_{Ci} = f(R_i; \boldsymbol{\theta}) + \epsilon_i$, $i \in \mathcal{W}$, be a model relating Y_C to R, where $\boldsymbol{\theta}$ is a set of parameters. Here, ϵ_i may be modeled as either drawn from a random population or fixed; in the latter case, the model fit f can be thought of as algorithmic, as opposed to probabilistic, in the sense of Rosenbaum [2002b]. For example, one may model $Y_{Ci} = \alpha + \beta R_i + \epsilon_i$, a linear model. A simpler example is $Y_{Ci} = \epsilon_i$, in which there is no relationship between R and Y_C for subjects in $\mathcal{W}$; in this case, our method reduces to the method described in Cattaneo et al. [2014].

An anonymous reviewer pointed out that there is a trade-off between the flexibility of f and the power to detect effects. In the extreme case, if f is allowed to jump arbitrarily at c, any treatment effect would be modeled as part of f, and not as a result of Z. For this reason, we will favor continuous models f.

Given $\boldsymbol{Y_C}$, fit the model $Y_{Ci} = f(R_i; \boldsymbol{\theta}) + \epsilon_i$ to subjects $i \in \mathcal{W}$, estimating $\boldsymbol{\theta}$. Then define

$$\boldsymbol{E_C} = \boldsymbol{Y_C} - f(\boldsymbol{R}; \hat{\boldsymbol{\theta}}) \tag{3}$$

Then make the following assumption:

Assumption (**Transformed Ignorability**). For subjects $i \in \mathcal{W}$

$$E_{Ci} \perp\!\!\!\perp R_i \tag{4}$$

This assumption states that, though Y_C is not independent of Z, it can be transformed, using model f, into E that is independent of R, and is therefore also independent of Z.

In many cases, transformed ignorability, using a model f that explicitly accounts for R, will be more appropriate than (2), even when (2) is only assumed in a small window around c, as in Cattaneo et al. [2014]. In order for (2) to approximately hold, either R and Y_C would have to be unrelated in general, or the researchers would have to pick such a small window of analysis $\mathcal{W}$ that, within $\mathcal{W}$, the relationship between R and Y_C is negligible. For an example in which this is unlikely to be the case, consider Wong et al. [2007], which use RDDs to study the effects of pre-school on children's pre-literacy skills. In those studies, the running variable is a youngster's age: children born before a certain date are eligible for government pre-school assistance, and those born after are not. The pre-school-age years of a child's life compose a period of rapid cognitive development. It is hard to imagine that, on average, even slightly older students would not perform substantially better on pre-literacy exams than their slightly-younger peers. A window $\mathcal{W}$ would have to be quite small indeed, in this case, for children's' ages to be ignorable. On the other hand, if researchers choose a sensible model f to transform Y_C into E_C, children's' rapid growth can be suitably accounted for, and Transformed Ignorability may, indeed, approximately hold.

More broadly, RDDs typically feature running variables that are correlated with outcomes, and researchers ignore this correlation at their peril. In our discussion, below, of the academic probation data, we will discuss the relative strengths of transformed ignorability and (2) in some detail, and argue that the former is more plausible. In fact, we will estimate the effect of probation under both assumptions, with appropriately varying bandwidths b,

and argue that the results are consistent with a bias resulting from (2).

Fisher's strict null hypothesis states that the treatment has no effect at all: $H_0 : Y_{Ci} = Y_{Ti}$ for all subjects i [Fisher, 1935, Rosenbaum, 2002a]. Under H_0, $Y_i = Y_{Ci} = Y_{Ti}$ for all i—the observed outcomes are identical to the outcomes that would have been observed in the absence of treatment. Therefore, under H_0, researchers can estimate E, using observed Y values in place of Y_C. Then, researchers can test H_0 using permutation or randomization-based techniques. That is, begin by specifying a test statistic T, and compute its null distribution under H_0 in one of the following ways: enumerating or randomly sampling from all permutations of Z or R and Y_C, calculating the distribution from first principals, or using large-sample approximations. A number of possible randomization schemes are compatible with transformed ignorability. In an analogous circumstance, Cattaneo et al. [2014] suggests choosing a scheme based on substantive concerns. Alternatively, researchers can conduct their inference after conditioning on $\sum_i Z_i = n_T$, thereby inducing a hypergeometric model for $\boldsymbol{Z}$ [Rosenbaum, 2002b].

A number of statistical tests are then available to test H_0, such as the Kolmagorov-Smirnov test [Massey Jr, 1951], the Wilcoxon-Rank-Sum test [Wilcoxon et al., 1970], or a permutation test based on the difference in means between the treated and untreated E values or equivalently $\boldsymbol{Z}^t\boldsymbol{E_0}$. Any of them might be used to test the plausibility of H_0.

If a researcher wants a test that is sensitive to relationships between hypothetical Y_C and R, as well as between Y_C and Z, other test statistics are possible. For instance, an analyst can regress the vector $\boldsymbol{Y_C}$ on vectors $\boldsymbol{Z}$ and $\boldsymbol{R}$, and extract goodness-of-fit statistics, such as the omnibus F-statistic. To compute the permutation of the statistic, she would then permute the $\boldsymbol{R}$ vector, each time calculating a new vector of treatment dummies based on the permuted R values.

2.2 Estimating Effects: Recovering Y_C by Modeling and Hypothesizing $\boldsymbol{\tau}$

Analogous to the model $Y_{Ci} = f(R_i; \boldsymbol{\theta}) + \epsilon_i$, researchers will choose a model for the treatment effects $\boldsymbol{\tau}$ for subjects $i \in \mathcal{W}$. That is, let $\tau_i = g(R_i; \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a set of parameters to estimate—the causal estimands of interest. There are other techniques, some of which are compatible with the framework we present here, that will allow researchers to estimate average treatment effects (ATE) or treatment effects on the treated (ETT) for subjects in $\mathcal{W}$ without specifying a model.[3] However, if the treatment effect τ varies with R, the true ATE or ETT will vary with the selection of $\mathcal{W}$. If this variation is not modeled as we are suggesting here, then the overall estimate may be difficult to interpret.

That being said, the simplest treatment effect model is $\tau_i = \tau_0$, a constant treatment effect. An alternative is $\tau_i = g(R_i; \nu) = \tau_0 + \nu R_i$, which allows the treatment effect to vary linearly with R. Both of these models are deterministic, in that given a model fit and data, a subject's treatment effect is known exactly. For the moment, we will stick to this interpretation, but in the following subsection we will show that our approach may be made robust to random variation in the treatment model.

Given g, and a window of analysis $\mathcal{W}$, researchers can specify a hypothesis $H_{\boldsymbol{\phi_0}} : \boldsymbol{\phi} = \boldsymbol{\phi}_0$, for $\boldsymbol{\phi}$, the parameters of g. The hypothesis $H_{\boldsymbol{\phi_0}}$ implies a vector of hypothetical treatment effects $\tilde{\boldsymbol{\tau}}_{\boldsymbol{\phi_0}}$, where a tilde denotes that a quantity is hypothetical. Combined with Y and Z, $\tilde{\boldsymbol{\tau}}_{\boldsymbol{\phi_0}}$ allows researchers to recover hypothetical values of Y_C: $\tilde{Y}_{Ci\boldsymbol{\phi_0}} = Y_i - Z_i\tilde{\tau}_i,\ i \in \mathcal{W}$, where the hypothetical treatment effect is subtracted from the treated subjects' outcomes. This type of maneuver figures heavily in, for instance, Rosenbaum [2002b].

[3] One example, compatible with our approach, is based on Peters [1941], Belson [1956], and Cochran [1969]—fit the model $Y_C = f(\cdot; \theta) + \epsilon$ to the control subjects only, and extrapolate it to subjects on the other side of the cutoff, generating estimated Y_C values $\hat{Y}_C$. Then an estimate of the average treatment effect on the treated would be $Y - \hat{Y}_C$. Angrist and Rokkanen [2012] and Cattaneo et al. [2014] offer other ATE or ETT estimates for subjects in a window around the cutoff.

2.3 Outcome Analysis under Transformed Ignorability

Armed with models f and g, and assuming transformed ignorability, researchers can test a hypothesis $H_{\boldsymbol{\phi_0}}$ with the following procedure, lifted almost directly from Rosenbaum [2002b]: First, under $H_{\boldsymbol{\phi_0}}$, recover hypothetical Y_C values, setting $\tilde{Y}_{Ci\boldsymbol{\phi_0}} = Y_i - Z_i g(R_i, \boldsymbol{X}_i; \boldsymbol{\phi_0})$ for subjects $i \in \mathcal{W}$. Using $\tilde{Y}_{C\boldsymbol{\phi_0}}$, and $\boldsymbol{R}$, $i \in \mathcal{W}$, fit the model $f(R, \boldsymbol{\theta})$, estimating $\boldsymbol{\theta}$ with $\boldsymbol{hat\theta}$. Next, transform $\tilde{Y}_{Ci\boldsymbol{\phi_0}}$ into $\tilde{E}_{\boldsymbol{\phi_0}i}$, the hypothetical value of E_{Ci}. Finally, under transformed ignorability, assess the plausibility of $H_{\boldsymbol{\phi_0}}$ using randomization-based methods. The values for $\boldsymbol{\phi}$ that yield p-values greater that α forms a $1-\alpha$ confidence region for $\boldsymbol{\phi}$. Researchers can arrive at Hodges-Lehmann point estimates [Hodges Jr and Lehmann, 1963] for $\boldsymbol{\phi}$ similarly.

As an example, say researchers believe that for members of $\mathcal{W}$, Y_C is well approximated by a linear function of R, and that treatment effects are approximately constant. They may begin by testing Fisher's strict null hypothesis, $H_0 : \ Y_{Ci} = Y_{Ti} \ \forall i \in \mathcal{W}$, or equivalently, $\tau_i = 0 \ \forall i \in \mathcal{W}$ [Fisher, 1935]. Under H_0, then, $\boldsymbol{Y} = \boldsymbol{Y_C}$, and

$$\tilde{\boldsymbol{E}}_{\boldsymbol{0}} = (I - \mathcal{R}(\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}')\boldsymbol{Y} \tag{5}$$

where $\mathcal{R}$ is a matrix composed by a column of 1s of length $n = \#\mathcal{W}$ and $\boldsymbol{R}$, and I is the $n \times n$ identity matrix. That is, $\boldsymbol{E_0}$ are the residuals from a simple least squares regression of $\boldsymbol{Y}$ on $\boldsymbol{R}$ and a constant.

Researchers can then repeat this procedure for a range of values for τ_0: set $\tilde{Y}_i = Y_i - \tau_0 Z_i$, generate E_{τ_0} by replacing $\boldsymbol{Y}$ with $\tilde{\boldsymbol{Y}}$ in (5), and testing H_{τ_0} using the same statistical test as for H_0. After testing a range of H_{τ_0} hypotheses, researchers can form a $1-\alpha$ confidence interval, which would consist of the set of values for τ_0 that correspond to p-values greater that α. Similarly, setting the test statistic to its null value and solving, or picking the τ_0 corresponding to the highest p-value, yields a Hodges-Lehmann point estimate.

2.4 Randomly-Varying Treatment Effects

The strict randomization inference scheme described above requires an exact model for treatment effects. Only under such a model can Y_C, and hence E_C be reconstructed. However, a treatment effect model will rarely fit exactly. That is, even if the pattern relating τ to R or X is correctly specified, other factors may cause subjects' treatment effects to vary randomly around the model.

First, it is important to note that no treatment effect model is necessary for null hypothesis testing—the presence of treatment effect heterogeneity only affects estimation. Fortunately, in some circumstances, the method outlined above can be slightly modified so as to remain asymptotically valid under random treatment effect heterogeneity. In particular, consider the following scenario:

- $Y_{Ci} = f(R_i) + \epsilon_i$ such that $\boldsymbol{E} = HY_C$, a linear transformation matrix. This would be the case if $f(R)$ were linear in functions of R, and fit with ordinary least squares.
- True $\tau_i = g(R_i, X_i; \boldsymbol{\phi}) + \eta_i$, with $\mathbb{E}\eta_i$ and η uncorrelated with the columns of H and with Z.
- The researcher specified $\tilde{\tau}_{i\boldsymbol{\phi}} = g(R_i, X_i; \boldsymbol{\phi})$

Then let $\tau_i = \tau_{\boldsymbol{\phi}i} + \eta_i$, where $\tau_{\boldsymbol{\phi}i} = g(R_i; \boldsymbol{\phi})$, and let $\tilde{Y}_{\boldsymbol{\phi}i} = Y_i - Z_i\tau_{\boldsymbol{\phi}i}$. If there were no random variation around g, then for some value of $\boldsymbol{\phi}$, $\tilde{Y}$ would equal Y_C. $\tilde{Y}_{\boldsymbol{\phi}}$, then, is the incorrectly-recovered Y_C. Then we have the following lemma,

Lemma 1.

$$y\tilde{ch}_{\boldsymbol{\phi}} = E_C + Z'\tilde{\eta} \tag{6}$$

with $\mathbb{E}\tilde{\eta} = 0$ *and* $\mathbb{E}Z'\tilde{\eta} = 0$

Proof. We have

$$\tilde{Y} = Y - Z'\tau_0 = Y - Z'\tau + Z'\eta = Y_C + Z'\eta$$

Then

$$\begin{aligned}\check{Y}_{\tau_0} &= H\tilde{Y} \\ &= HY_C + HZ'\eta \\ &= \check{Y}_C + H\eta' Z \\ &\equiv \check{Y}_C + Z'\tilde{\eta}\end{aligned}$$

Then $\mathbb{E}\tilde{\eta} = \mathbb{E}H\mathbb{E}\eta = 0$ since $\eta \perp\!\!\!\perp R$ and $H = H(R)$. Similarly, $\mathbb{E}Z'\tilde{\eta} = \mathbb{E}Z'H\mathbb{E}\eta = 0$. □

Then, under Transformed Ignorability, a weaker null hypothesis holds, which allows for randomly varying treatment effects:

Proposition 1. *Conditional on n_T and n_C, the numbers of treated and untreated subjects, respectively, if $E_C \perp\!\!\!\perp R$ (Transformed Ignorability) then $\mathbb{E}\tilde{ych}'_{\boldsymbol{\phi}} Z/n_T = \mathbb{E}\tilde{ych}'_{\boldsymbol{\phi}}(1-Z)/n_C$, that is, the expected means of the treated and untreated groups are the same.*

Proof.

$$\mathbb{E}Z'\check{Y}_{\tau_0}/n_T = 1/n_T(\mathbb{E}Z'\check{Y}_C + \mathbb{E}Z'\tilde{\eta}) = \mathbb{E}Z'\check{Y}_C/n_T$$

and

$$\mathbb{E}\check{Y}'_{\tau_0}(1-Z)/n_C = (1/n_C)(\mathbb{E}\check{Y}'_C(1-Z) + \mathbb{E}(1-Z)Z'\tilde{\eta} = \mathbb{E}\check{Y}'_C(1-Z)/n_C$$

Finally, transformed ignorability give us $\mathbb{E}\check{Y}'_C Z/n_T = \mathbb{E}\check{Y}'_C(1-Z)/n_C$ □

The implication of Proposition 1 is that a researcher may specify a function $\tau = g(R, X; \boldsymbol{\phi})$ that is correct on average—that is, $\mathbb{E}\tau_i = g(R_i, X_i; \boldsymbol{\phi})$, even if it does not yield the exact τ_i for every i. Then, for some $\boldsymbol{\phi}$, the mean of E_{τ_0} for the treated subjects will be the same as that for the untreated subjects. A test of equality of means that asymptotically achieves its nominal level will, when inverted, yield asymptotically correct confidence intervals.

One example of such a test is, of course, the usual student's t-test. Alternatively, Chung et al. [2013] provides a method of altering popular permutation tests such that they maintain their finite-sample exactness under Fisher's strict H_0, but are asymptotically valid testing the weaker null hypothesis of equality of means.

3 Protecting Inference from Model Misspecification

In order for Transformed Ignorability to be at all useful, researchers must be able to suitably transform Y_C values into E, using a function f that relates Y_C to R. As in conventional RDD analysis [See Lee and Card, 2008], a misspecified model for relating Y_C to R can result in misleading results. The model built using the observed Y_C values on one side of the cutoff and hypothetical values on the other; to disentangle treatment effects from natural variation depends on the form of the model relating Y_C to R. Therefore, incorrect models can have a large and harmful impact on estimates and inference.

3.1 Limiting the Window of Analysis

One partial protection from misspecification is to limit the size of $\mathcal{W}$ [e.g. Imbens and Lemieux, 2008, Angrist and Pischke, 2009]. In the conventional RDD story, the target estimand is the treatment effect at the cutoff, so focusing on data near the cutoff reduces model error.

A similar intuition can apply to our limitless RDD strategy. Potential outcomes Y_C are only observed on one side of c, yet the model $Y_C = f(R) + \epsilon$ is fit to the entire sample. Hence, some amount of extrapolation from one side of the cutoff to the other is necessary. In the same vein, inference and estimation rely on specifying an approximately correct functional form f. A slightly-incorrect specification may do less harm over a shorter range of R values

than over a wider range.[4]

Often, the scientific or policy question of interest only really applies to cases near the cutoff. For instance, Thistlethwaite and Campbell [1960] studied the effect of receiving a National Merit Scholarship on students' probabilities of attending graduate school. Presumably, no one is considering awarding merit scholarships to students with particularly low PSAT scores. For these students, the hypothetical effect of merit scholarships is not relevant. The graduate school choices of students who who barely qualified, or barely missed qualifying, though, could be quite interesting. In these cases, both statistical and substantive considerations recommend restricting analysis to a window $\mathcal{W}$ around the cutoff.

3.2 Covariate Balance Placebo Tests

If researchers have access to covariates that are informative about the outcome of interest, these can provide guidance on the width of the window $\mathcal{W}$, and the appropriateness of f, the function modeling Y_C's relationship to R. In particular, they may serve as placebo tests, since, by definition, the treatment could not have affected them. To do so, the researcher would fit the same functional form f to each available X, and transform each X into $\check{X}$. If f models X's relationship to R approximately well, $\check{X}$ will be independent of Z. Conversely, if f fails to model X well, then the residuals on one side of the cutoff are likely to have a different distribution, or location, than those on the other side, in which case $\check{X}$ will not be independent of Z. The dependence of $\check{X}$ on Z can be tested, in the same way as the dependence between E and Z.

Checking covariate balance, as a placebo test, is not uncommon in RDD literature; see, for instance, Cattaneo et al. [2014], which does not recommend transforming X into $\check{X}$ before checking balance, and Lee and Lemieux [2010] which does.

[4]This need not be the case: if the true relationship between R and Y_C fluctuates, but the researcher's f is monotonic, some wider intervals may yield superior results to some smaller intervals.

If there are several covariates available, the hope is that for at least one of them, f will fail at least as badly as with Y_C. That is, for some covariate X_k, f will model X_k the same or worse than Y_C. Then, if Y_C's departure from f is large enough to cause a false positive result, then it will also cause a false positive result in the placebo test for X_k.

More formally, before estimating effects, using the observed outcome values Y, the researcher conducts a placebo test with X. That is, the researcher fits misspecified model f to X and extracts residuals $\check{X}$, then tests if $\check{X} \perp\!\!\!\perp Z$, rejecting at level α_X. If the placebo test fails to reject—that is, its p-value p_X is greater than α_X—the researcher proceeds to analyze the RDD in Y_C, using model f and window $\mathcal{W}$. Then the inference in Y is conditional on the inference in X. The true type-I error rate is

$$Pr(p_Y < \alpha_Y \text{ and } p_X > \alpha_X | H_0, \beta_2, \gamma_2) = \\ Pr(p_Y < \alpha_Y | H_0, \beta_2, p_X > \alpha_X) Pr(p_X > \alpha_X | \gamma_2)$$

where p_Y and α_Y are the p-value and level of the outcome analysis. The type-I error rate, then, is a function not only of the distribution of p_Y, but also of p_X, and their dependence. These, of course, are unknown. Nevertheless, the constraint of balance in $\check{X}$ serves to substantially lower the probability of a type-I error in the case of serious model misspecification.

To get a sense of how these joint tests might operate, and hence an appropriate choice for α_X, we will make a number of simplifying assumptions. First, assume that the test statistics that give rise to p_Y and p_X, T_Y and T_X, say, are normally distributed with unit standard deviation. This would be the case, asymptotically, if the test statistic were a difference-in means or a properly standardized Wilcoxon test, among others. Next, assume that T_Y and T_X are equal in distribution. Again, the hope is that at least one covariate will depart further from f than Y will, in such a way that its test statistic will be stochastically greater than T_Y; in this case, the assumption of equality in distribution is conservative.

Finally, assume that $T_X \perp\!\!\!\perp T_Y$, and hence $p_X \perp\!\!\!\perp p_Y$. This assumption will generally be false, but it is likely to be conservative as well. If T_X and T_Y are positively correlated, then $Pr(p_Y < \alpha_Y | H_0, \beta_2, p_X > \alpha_X) \leq Pr(p_Y < \alpha_Y | H_0, \beta_2)$, so the true type I error rate α is likely to be less than the rate that we would calculate ignoring the dependence between T_X and T_Y.

In this scenario, imagine the researcher decided in advance to only proceed with outcome analysis if the p-value from a covariate balance test $p > \alpha_X$. Further, he would reject H_0 at a level of α_Y. If H_0 were true, the test might not achieve its true level due to misspecification of f. Say f were misspecified to the extent that the expected values of T_Y and T_X are both μ. Then the true type-I error rate is

$$\alpha = (1 - \Phi(z_{\alpha_X/2} - \mu) - \Phi(-z_{\alpha_X/2} - \mu))(\Phi(z_{\alpha_Y/2} - \mu) + 1 - \Phi(z_{\alpha_Y/2} - \mu)) \tag{7}$$

where Φ is the standard normal CDF, and z_p is the p^{th} quantile of the standard normal distribution.

This calculus may, indeed, provide some guidance on α_X, and hence on the width of $\mathcal{W}$ or the choice of f. For a given μ, there is a unique α_X that will produce the desired type-I error rate. That is, given μ, α_Y, and α, one can solve equation (7) for α_X. Of course, μ is unknown: it depends both on the sample size and on the true, unknown, distribution of Y_C on both sides of the cutoff. However, as the magnitude of μ increases, beyond a certain point the true α tends towards 0. This is because $\Phi(z_{\alpha_X/2} - \mu)$ tends towards 1, and $\Phi(-z_{\alpha_X/2} - \mu)$ tends towards 0, so $(1 - \Phi(z_{\alpha_X/2} - \mu) - \Phi(-z_{\alpha_X/2} - \mu))$ tends to 0, while the second half of the expression, $(\Phi(z_{\alpha_Y/2} - \mu) + 1 - \Phi(z_{\alpha_Y/2} - \mu))$, is bounded by 1. Hence, in terms of the true α, there is a worst-case scenario, which corresponds to a conservative value for α_X.

Figure 3.2 illustrates this point. Setting $\alpha = \alpha_Y = 0.05$, we numerically solved (7) for α_X at a range of values for μ, using the `rootSolve` package in `R` [Soetaert and Herman,

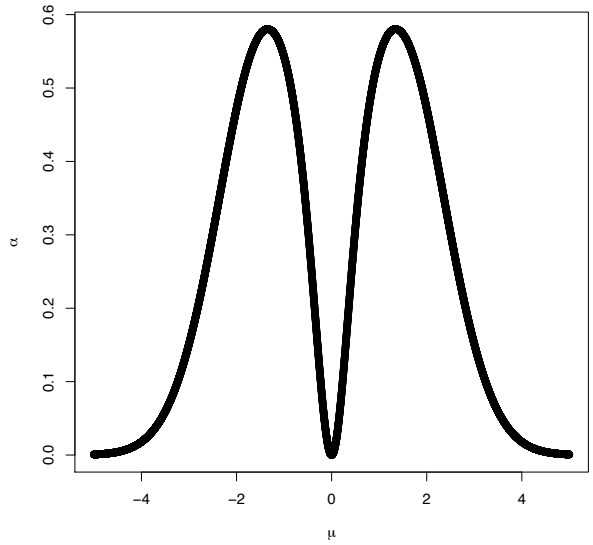


Figure 1: The value of α_X necessary to achieve $\alpha = 0.05$ with $\alpha_Y = 0.05$ at various values of μ

2008, Soetaert, 2009, R Development Core Team, 2011]. The maximal α_X in the figure is approximately 0.6, which will serve as a conservative choice for α_X.

3.3 Recommendations for Practice

The estimation and inference for RDDs, under our framework, takes place entirely for the set of subjects $i \in \mathcal{W} = \{i : c - b \leq R_i \leq c + b\}$ for a bandwidth $b > 0$. The first motivation for a choice of b should be substantive: for which subjects does an effect estimate make sense? Which subjects could, conceivably, be candidates for treatment? For instance, in the LSO dataset, it is hardly reasonable to ask what the effect of academic probation would be on straight-A students. In the absence of a clear guideline of this sort, a first choice for $\mathcal{W}$ can be the smallest window that contains substantively meaningful variation in R: this identifies an interpretable group of subjects.

In addition to $\mathcal{W}$, researchers must pick f, a model for Y_C. Gelman and Imbens [2014]

recently argued against using polynomials of order higher than two. As a first pass, we endorse that recommendation: higher-order polynomials are hard to interpret, and it is hard to choose the order in a non-arbitrary fashion. For that reason, researchers may start by modeling Y_C as a linear or quadratic function of R. Visual inspection of a scatterplot of Y against R, and possibly other regression diagnostic tools, can, of course, inform this choice.

After choosing $\mathcal{W}$ and f, researchers will then transform, and test balance, on a set of covariates that may be informative about Y_C. As above, a conservative value of $\alpha_X = 0.6$ is reasonable, although other considerations may suggest other values for α_X (for instance, some confidence in f and $\mathcal{W}$). If the balance test rejects, researchers can add more flexibility to f, perhaps by adding polynomial terms, or shrinking $\mathcal{W}$.

Cattaneo et al. [2014] and Angrist and Rokkanen [2012] both describe a similar data-driven window selection approach: use covariates to successively test model specification for windows $\mathcal{W} = \{i : c - b \leq R_i \leq c + b\}$ for increasing bandwidths b. Then, choose the larges b whose covariate-balance p-value exceeds a pre-specified level. This approach easily extends to our method, testing balance of transformed covariates at a range of values for b. It is not without its problems, however: p-values at neighboring b values will typically be highly correlated, since they share a large amount of data. Moreover, multiple-comparison problems can confuse this process, causing idiosyncratically low or high p-values. The practical consequences of these problems, and whether they may be overcome, is an open statistical question.

4 Similarities and Differences with Existing RDD Strategies

4.1 The Standard Method

At the highest level, the standard approach to estimating RDDs is identical to what we present here. Analysts specify and fit a model for Y_C and another model for τ, The difference between these fitted models, at $R = c$ is the average treatment effect for subjects at $R = c$. For instance, if the analyst specifies a linear models of Y_C and a constant model for τ, then the procedure reduces to a linear regression:

$$Y_i = \alpha + \beta R_i + \tau Z_i + \epsilon_i. \tag{8}$$

The standard approach to RDDs differs in two important ways from the method we describe here. First, the source of identification: in the standard approach, identification comes from the continuity of the expected values of Y_T and Y_C, as R varies—this allows researchers to estimate their values at the cutoff and estimate the treatment effect there. In contrast, our method relies more heavily on the local randomization assumption: that, after transforming Y_C values, treatment assignment may be modeled as random. For this reason, a discrete R may pose a problem for the conventional approach; our method avoids this issue. See, however, Lee and Card [2008] which explicitly addresses discrete R.

The estimand standard approach is a limit of average treatment effects in ever-tightening windows. This can be difficult to interpret. In some scenarios, it may be possible to interpret as the average treatment effect for subjects for whom $R = c$; however, this will often differ from the researcher's estimand of interest. Our estimand is simply the function $\tau = g(R; X)$—the function describing the treatment effect for subjects in the window $\mathcal{W}$.

Perhaps surprisingly, the standard estimator coincides with our estimator in one par-

ticular setting. If analysts use (8) to estimate τ, and, in a different analysis, model Y_C as linear, model τ as constant, and use the difference in means between treatment and control subjects as a test statistic, the two estimates will be identical. We formalize this in the following proposition:

Proposition 2. *If, in $\mathcal{W}$, a researcher uses OLS to model $Y_C = f(R) + \epsilon = \beta R + \epsilon$ and $\tau = \tau_0$, and takes as the test statistic $\boldsymbol{E'Z}$, then the Hodges-Lehmann estimate of τ_0 will be equal to the OLS estimate of τ from the regression (8), fit using the data in $\mathcal{W}$.*

The proof to this proposition is basically the same as an argument from Rosenbaum [2002b, p. 290]

Proof. Let $\mathcal{R}$ be the matrix formed by joining a column of ones to $\boldsymbol{R}$. Then let $H = \mathcal{R}(\mathcal{R}^T\mathcal{R})^{-1}\mathcal{R}')$. Under $H_{\tau_0} : Y_{Ti} - Y_{Ci} = \tau_0$, $E_\tau = (I - \mathcal{H})(Y - Z\tau)$ and the test statistic is $Z^T E_\tau = Z^T(I - \mathcal{H})(Y - Z\tau)$. When $\tau = \tau_0$, the expected value of the test statistic is $\mathbb{E}Z^T E = \mathbb{E}Z\mathbb{E}E = 0$ since $\mathcal{R}$ contains a constant term and the model is fit with OLS. The Hodges-Lehmann estimate of τ, solves the equation

$$Z^T E_\tau = 0 \tag{9}$$

$$Z^T(I - \mathcal{H})(Y - Z\tau) = 0 \tag{10}$$

which is solved when $\tau = \frac{Z^T(I-\mathcal{H})Y}{Z^T(I-\mathcal{H})Z}$, which is equal to the OLS estimate of the coefficient of Z from the regression of Y on a constant, R, and Z. □

That is, one of the simplest conventional RDD estimates can be re-interpreted as a Hodges-Lehmann estimate of a constant treatment effect under transformed ignorability. In this case, rather than provide a new method, we reinterpret the conventional method.

4.2 Local-Randomization-Based Inference (Cattaneo et al. 2014)

Cattaneo et al. [2014] developed a randomization-based approach to RDDs. Their approach was to limit the data to a small window around the cutoff, analogous to our $\mathcal{W}$, and argue that within $\mathcal{W}$, subjects are for practical purpose, randomized into treatment and control groups, so $Z \perp\!\!\!\perp Y$. As we mentioned above, this corresponds to the case where one "models" the relationship between Y_C and R with a constant function—that is, assumes no relationship between Y_C and R—in $\mathcal{W}$, in which case transformed ignorability is equivalent to standard ignorability (2).

We argued above that in many instances the relationship between R and Y_C is both strong and important, so ignoring it, even in a small window $\mathcal{W}$, can yield misleading results. That being said, the randomization-based approach in Cattaneo et al. [2014] has some advantages over the more general approach here. To wit, Cattaneo et al. [2014]'s set-up allows researchers to estimate a broader class of estimands. For instance, if one models the data in $\mathcal{W}$ as arising from a randomized experiment, one may estimate quantile treatment effects or displacement effects [Rosenbaum, 2001]. Estimating these in the framework we have developed here requires a non-trivial extension.

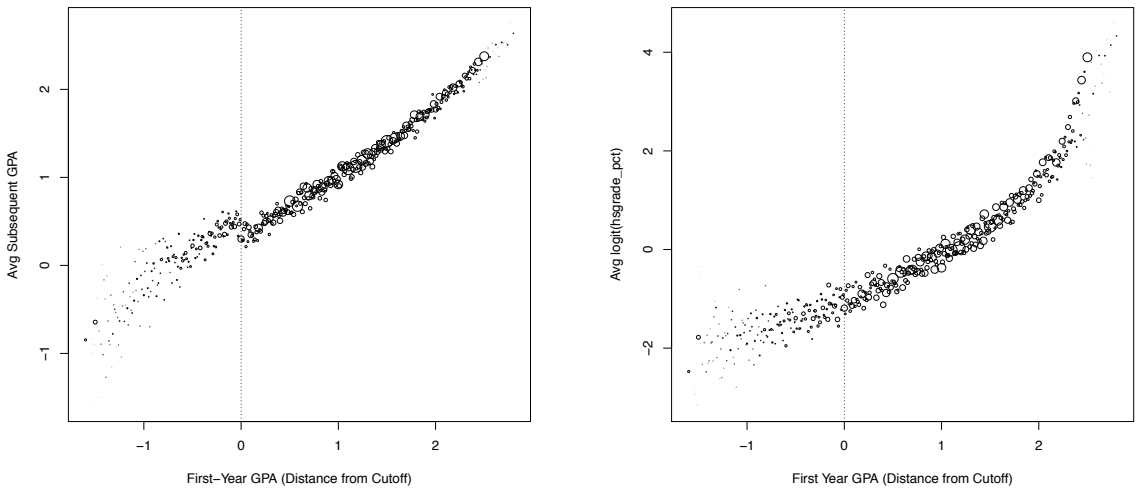
4.3 Conditional Ignorability Assumption (Angrist and Rokkanen 2012)

Angrist and Rokkanen [2012] addressed the question of estimating treatment effects away from the cutoff with a new assumption, called the "Conditional Ignorability Assumption," or CIA. The assumption states that, conditional on covariates X, treatment assignment is mean-independent of the running variable. This approach shares three important similarities with ours. First, it explicitly formulates causal identification in terms of an ignorability assumption. Second, it is interested in effects for a sample of subjects which is not asymp-

totically vanishing. Finally, it uses covariates to justify causal identification away from the cutoff.

In its details, though, it is quite different—its identification assumption is different from ours, as is the method it proposes. Unlike our approach, CIA assumes mean independence, which is sufficient for unbiased estimation but not inference. Therefore, for estimation, CIA may be weaker than Transformed Ignorability, but for inference it requires some stronger assumptions. Additionally, covariates are explicit in the definition of CIA, but not in Transformed Ignorablility. Both assumptions seek to make R ignorable—Transformed Ignorability does so with modeling, while CIA does so with covariates.

5 Example: The Effect of Academic Probation



(a) (b)

Figure 2: (a) The RDD from LSO. The first-year GPAs were shifted so that the cutoff is at zero—that is, each campus's cutoff was subtracted from its students first-year GPAs. Subsequent GPA was averaged according to first-year GPA. (b) Students log-transformed high-school GPAs, also averaged by first-year college GPAs.

Lindo et al. [2010]—LSO—attempted to estimate the effect of academic probation (AP)

on college students at an unnamed Canadian university. One of the outcomes that LSO measures is $nextGPA$, students' subsequent GPAs, either for the summer or fall term after students' first years. Figure 2 (a) displays $nextGPA$ as a function of students' first-year GPAs. Their causal question of interest is whether AP causes a change in $nextGPA$: do students on AP tend to have higher (or lower) subsequent GPAs? Recall that AP is determined almost exclusively[5] by first-year cumulative GPA, which in this case is the running variable R. That is, students with GPAs below the cutoff are "treated" with AP, and students above the cutoff are in the control group. The university in question has three campuses, two of which have cutoffs of 1.5; the other has a cutoff of 1.6. To combine data from the three schools, LSO centered each student's first-year GPA at the appropriate c, so R_i is a student i's first year GPA, minus the cutoff at his college. Then, $Z = \mathbb{1}_{[R \leq 0]}$

A relevant region in which to estimate a treatment effect is within 0.3 grade-points of the cutoff c. Conventionally, 0.3 represents the difference in grade points between a C, say, and a C-, or any other grade half-step. As a first pass at $\mathcal{W}$, then, we will look at $\mathcal{W}_1 = \{i : R_i \in [-0.3, 0.3]$. A first pass at f will be $f_1(R_i) = \alpha + \beta R_i + \epsilon_i$, a linear function. Simplicity recommends a linear approximation, and an examination of a scatterplot of Y versus R suggests that it will be a close approximation in $\mathcal{W}$.

To test f_1 and $\mathcal{W}_1$, we will conduct placebo tests, examining balance of transformed covariates. LSO exploits seven pre-treatment variables to test the RDD assumptions in the AP case: high-school GPA (available, for some reason, on a scale from 0–100), the total number of credits attempted in students' first year and students' age at college entry, which are continuous, and dummy variables for campus (the university has three campuses), whether

[5]There are 48 cases, out of a total of 44,362, where students were not put on AP despite GPAs below the cutoff, and three cases in which students were put on AP despite having GPAs above the cutoff. In this paper, as in LSO, we will follow the "intent to treat" principle and estimate the effect of treatment *assignment*—that is, having a GPA below the cutoff—instead of actual treatment (AP). Since the number of cases in which these two disagree is such a small proportion of the total, we anticipate that this will have a minimal impact on our estimates.

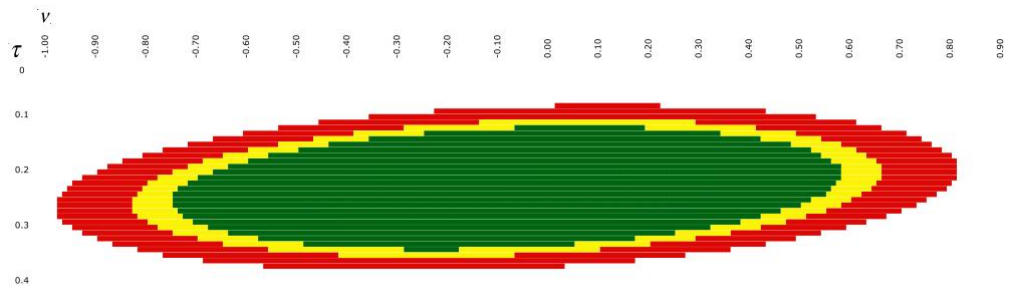


Figure 3: Confidence regions for ν and τ from model (11). The green region is a 90% confidence region, the yellow is 95% and the red is 99%.

students' first language is English, and whether students were born in North America. Figure 2 (b) displays the logit of students' high-school GPAs, which had been on a 1–100 scale suggests that the curvature in relationship of (transformed) high-school GPA with R is greater than $nextGPA$, suggesting that high-school GPA may be a good candidate for a conservative placebo test. Hansen and Bowers [2008] suggested an omnibus statistic for testing covariate balance; their routine, implemented as `xBalance` in the `R` package `RItools` [Bowers et al., 2010, **?**], yields a p-value of p =0.03, well above the 0.6 threshhold.

With f_1 and $\mathcal{W}_1$ in hand, we can test Fisher's $H_0 : Y_{Ci} = Y_{Ti}$ for all subjects i. Using the same testing procedure as the covariate placebo test, `xBalance`, the p-value is 0.0092, which rejects H_0 at level $\alpha = 0.05$.

To estimate the magnitude of the effect, we need, additionally, a model g for treatment effects. Examination of Figure 2 suggests that the slope relating R to $nextGPA$ does not shift from one side of c to the other, so a constant effect model may be sufficient. Inverting the hypothesis test for a range of constant effects yields a 95% confidence interval of 0.05–0.38, with a Hodges-Lehmann point estimate of 0.21.

It may be, though, that the effect of AP varies with first-year GPAs. Students with higher GPAs may be more motivated than their lower-GPA colleagues, and therefore respond more

heartily to the threats that accompany AP. This suggests a more nuanced model for τ:

$$\tau_i = g(R_i; \boldsymbol{\phi}) = \tau_0 + \nu R_i \tag{11}$$

Where the treatment effect τ is decomposed into τ_0, a constant effect for all students, and ν, an effect that varies linearly with R. That is, $\boldsymbol{\phi} = \{\tau_0, \nu\}$. Test statistics that focus only on the relationship between hypothetical values for Y_C and Z will perform poorly when testing a hypothesis that involves both Z and R, as in (11). This is especially true for test statistics that are sensitive to location shifts, such as the Wilcoxon test or the `xBalance` procedure that we have used so far: for each hypothetical τ_0, a hypothetical ν is available so that transformed, hypothetical values for Y_C, that is, $E_{\tau_0,\nu}$, have the same locations in both treatment and control groups. An appropriate test statistic for (11) is sensitive to differences in both the slope and the intercept of the $E_{\tau_0,\nu}$-R regression lines between the treatment and control groups. One such statistic is the omnibus F-statistic from the regression of $E_{\tau_0,\nu}$ on R, Z, and $R : Z$. Inverting the permutation test of this statistic yields the confidence region in Figure 3. Marginally, the inference for τ_0 is roughly the same as the constant effects model above, but the data are, somewhat surprisingly, uninformative about ν.

5.1 Robustness Checks

The choices we made for b and for f could have been made differently; they were motivated, respectively, by substantive background and a desire for simplicity, and were validated with covariate placebo tests. That being the case, it may be wise to examine the robustness of our analysis to different choices for b and f.

A reasonable alternative for f allows Y_C to vary with R as a quadratic polynomial, so $Y_{Ci} = f(R_i) = \beta_0 + \beta_1 R_i + \beta_2 R_i^2$. Using this quadratic model for f leaves our inference, in this case, virtually unchanged: a p-value of 0 for H_0, a confidence interval of 0.04–0.39, and

a Hodges-Lehmann point estimate of 0.22.

	b	p-value	0.95 CI	HL Estimate
main	0.3	0.0092	(0.05,0.38)	0.21
DataDriven	1.03	1.2e-10	(0.17,0.33)	0.25
IK	1.25	5.8e-16	(0.2,0.33)	0.26

Table 1: Null Hypothesis p-values, 95% confidence intervals, and Hodges-Lehmann point estimates for a variety of bandwidths in the LSO data.

To examine robustness to choices of bandwidth b and hence $\mathcal{W}$, we tried two alternative bandwidths. The results are summarized in Table 1. First, a data-driven bandwidth, using a procedure similar to what was suggested in Cattaneo et al. [2014]: setting b to the largest value for which a covariate placebo test yields a p-value above $\alpha_X = 0.6$. The largest b for which this is true turns out to be $b = 1.03$; at this bandwidth, the p-value for a covariate placebo test is 0. This yields a p-value for H_0 of 1.2e-10, a confidence interval of 0.17–0.33 and a point estimate of 0.25. Finally, Imbens and Kalyanaraman [2012] suggests a procedure for choosing a bandwidth that minimizes mean-squared-error for the conventional local-linear RDD analysis. We implemented this method, specifying a "rectangular" kernel (weighting equally all observations within the window, and assigning all others a weight of zero) using the `rdd` package in `R` [R Development Core Team, 2011, Dimmery, 2013]. In the LSO case, that procedure yields a bandwidth of 1.25, with a covariate balance p-value of 0. The Imbens and Kalyanaraman [2012] bandwidth yields an H_0 p-value of 5.8e-16, a 95% confidence interval of 0.2–0.33, and a point estimate of 0.26. These results suggest an insensitivity to the choice of b.

5.2 LSO Results from Other Methods

How does our method compare with others in the LSO analysis? To see, we use the `rdd` package in `R` [Dimmery, 2013] to implement the latest limit-based RDD analysis. This used a local linear regression, with the bandwidth recommended by Imbens and Kalyanaraman

	b	p-value	0.95 CI	HL Estimate
main	0.3	0.0092	(0.05,0.38)	0.21
Conventional	0.79	7.5e-15	(0.17,0.29)	0.23
CFT	0.16	3.2e-05	(0.07,0.18)	0.12

Table 2: Null Hypothesis p-values, 95% confidence intervals, and point estimates for our main analysis, compared with the analysis in Cattaneo, et al. (2014), and the conventional local-linear estimate with the Imbens and Kalyanaraman (2012) bandwidth.

[2012] (this time with the recommended "triangular" kernel), to estimate the effect of AP at the cutoff. Next, we implemented the randomization-based routine recommended in Cattaneo et al. [2014]. First, to choose a bandwidth, we used the `xBalance` function to examine covariate balance—without transformation—at a range of bandwidths near the cutoff.[6] The highest bandwidth that corresponded to a p-value greater than $\alpha_X = 0.15$, which they recommend, was $b = 0.15$. Then, we used `xBalance` as a difference-in-means test to test the strict null hypothesis within the window, and we inverted it for a 95% confidence interval and a Hodges-Lehmann point estimate.

Our method gives roughly the same estimate as the conventional, local linear approach, though with a wider confidence interval. The Cattaneo et al. [2014] method gives a slightly different answer; in fact, the estimates from our method and from the conventional method are outside the randomization-based method's confidence interval.This may be because Cattaneo et al. [2014] ignores the relationship between $nextGPA$ and R. The general trend—higher $nextGPA$ for higher R—is in the opposite direction as the effect of AP—students with lower R are treated, and experience a positive effect. These two factors may partially cancel each other out, leading to a point estimate that is biased toward zero. If this is the case, it illustrates the importance of explicitly modelling R in an RDD analysis.

[6]Cattaneo et al. [2014] recommends testing balance separately for each covariate, and choosing the minimum p-value at each possible bandwidth. We found this to be overly conservative—it rejected the null hypothesis at every possible bandwidth, due, possibly, to multiple comparisons.

5.3 Other Issues in the LSO Dataset

Our analysis of the LSO dataset is intended as an illustration of our novel RDD analysis method; as a study of the effect of AP, it is incomplete. In particular, there are two serious statistical issues that we ignored, for the sake of brevity. We mention them briefly here. First, 4.39% of the subjects in the LSO dataset were missing a GPA either for their first-semester, their subsequent semester, or both. Our analysis implicitly treated these subjects as missing completely at random; however, this assumption may not be true.

More importantly, perhaps, are the results of a McCrary density test [McCrary, 2008], which yielded a p-value of 6.8e-08, indicating that some subjects may have manipulated their first-year GPAs to avoid AP. In an analysis not shown here, but available upon request, we found suggestive evidence that a number of students may have dropped a course in order to achieve a GPA just above the cutoff. In fact, if we remove the students at the cutoff who took only four (the mode is five), the McCrary p-value increases to 0.34. Removing these students has negligible effect on our estimates or inferences. The justification for this maneuver is debatable, and doing so here would distract from our central point; suffice to say that, from a substantive perspective, the LSO result needs further investigation.

6 Conclusion

This paper presents a novel interpretation and modeling approach to regression discontinuity designs. The new approach has some advantages over the conventional approach, including natural interpretation and statistical inference when the running variable is discrete or the the sample size is small, identifying assumptions that speak to the link between RDDs and randomized experiments, and a role for covariates in validating inference.

The approach presented here may have some weaknesses as well. Firstly, it requires a set of covariates in the data whose relationships to the outcome of interest are similar to

the running variable R's. Not only are such covariates not always available, but even in a rich dataset it may be hard to assess the covariates' usefulness. Secondly, there may be some scenarios in which the conventional RDD assumptions are more plausible than those presented here. Finally, in the study of academic probation that we discuss here, our method yielded a wider confidence than the conventional method. It is unclear if this will generally be the case.

Though starting from a similar point, this paper's approach differs in some important ways from the approach in Cattaneo et al. [2014]. That paper suggests assuming that subjects in a window close to the cutoff are randomized, with equal probabilities, to various treatment conditions. In particular, within the window of analysis, the potential outcomes are not related to treatment assignment—and, therefore, the running variable R. In contrast, this paper allows a relationship between potential outcomes and R in the window of analysis. The question of which set of assumptions is more plausible will depend on the substantive question of interest. Cattaneo et al. [2014]'s approach allows more flexibility in the choice of estimands, whereas the approach here is limited to estimating mean-based estimands. This paper's approach will hopefully be attractive to researchers who are drawn to randomization-based arguments, but are reluctant to abandon the familiar trappings of the conventional RDD approach.

Recently, the RDD methodology literature has begun to address the case of multiple running variables [Papay et al., 2011, Reardon and Robinson, 2012]. The method we present here extends to that case in a straightforward way, using multivariate modeling techniques to disentangle outcomes from the running variables and joint permutation tests for inference.

Finally, this paper highlights the need for future work on choosing a window of analysis based on a sequence of specification tests.

References

Joshua Angrist and Miikka Rokkanen. Wanna get away? rd identification away from the cutoff. Technical report, National Bureau of Economic Research, 2012.

Joshua D Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press, 2009.

William A. Belson. A technique for studying the effects of a television broadcast. 5(3): 195–202, November 1956.

R.A. Berk and D. Rauma. Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, pages 21–27, 1983.

J. Bowers, M. Fredrickson, and B. Hansen. Ritools: Randomization inference tools. *R package version 0.1-11. URL: http://www. jakebowers. org/RItools. html*, 2010.

M. D. Cattaneo, B. Frandsen, and R. Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. Technical report, University of Michigan, 2014. URL `http://www-personal.umich.edu/ titiunik/papers/RD-Randominf.pdf`.

EunYi Chung, Joseph P Romano, et al. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 2013.

W. G. Cochran. The use of covariance in observational studies. 18:270–275, 1969.

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

D.R. Cox. *The Planning of Experiments*. John Wiley, 1958.

S.L. DesJardins and B.P. McCall. 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. *Unpublished Manuscript*, 2008.

Drew Dimmery. *rdd: Regression Discontinuity Estimation*, 2013. URL `http://CRAN.R-project.org/package=rdd`. R package version 0.54.

R. A. Fisher. *Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. Technical report, National Bureau of Economic Research, 2014.

J. Hahn, P. Todd, and W. Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

Ben B. Hansen. The prognostic analogue of the propensity score, 2008. doi: 10.1093/biomet/asn004.

Ben B. Hansen and Jake Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236, 2008.

JL Hodges Jr and E.L. Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, pages 598–611, 1963.

Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959, 2012.

G.W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

David S. Lee. Randomized experiments from non-random selection in u.s. house elections. Technical report, Department of Economics, UC Berkeley, 2005. 40 pages; posted at Berkeley Econ website.

David S Lee. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697, 2008.

D.S. Lee and D. Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

D.S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355, 2010.

J.M. Lindo, N.J. Sanders, and P. Oreopoulos. Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2 (2):95–117, 2010.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008.

P. Oreopoulos. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *The American Economic Review*, pages 152–175, 2006.

John P Papay, John B Willett, and Richard J Murnane. Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161 (2):203–207, 2011.

Charles C. Peters. A method of matching groups for experiment with no loss of population. *Journal of Educational Research*, 34:606–612, 1941.

E.J.G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Sean F Reardon and Joseph P Robinson. Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, 5(1):83–104, 2012.

Paul R. Rosenbaum. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(1):219–231, 2001.

Paul R. Rosenbaum. *Observational Studies*. Springer-Verlag, second edition, 2002a.

P.R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3), 2002b.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688, 1974.

D.B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

Karline Soetaert. rootsolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. *R package version*, 1, 2009.

Karline Soetaert and Peter MJ Herman. *A practical guide to ecological modelling: using R as a simulation platform*. Springer, 2008.

J. Splawa-Neyman, DM Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

D.L. Thistlethwaite and D.T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.

Frank Wilcoxon, SK Katti, and Roberta A Wilcox. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259, 1970.

Vivian C Wong, Thomas D Cook, W Steven Barnett, and Kwanghee Jung. An effectiveness-based evaluation of five state pre-kindergarten programs using regression-discontinuity. *Journal of Policy Analysis and Management*, pages 872–884, 2007.