Alternative Assumptions to Identify LATE in Regression Discontinuity Designs

Yingying Dong

University of California Irvine[*]

This version: August 2017

Abstract

There exist two alternative assumptions to identify local average treatment effects (LATE) in regression discontinuity (RD) designs: local independence (LI) and local smoothness (LS). Together with the usual LATE assumptions requiring existence of a first-stage and treatment monotonicity, either of these two assumptions is sufficient to identify RD LATE. I discuss the practical (and testable) implications of these alternative assumptions, and show that weakening LI by LS might be empirically relevant. However, when LI does hold, there are some practical implications one may explore. Numerical and empirical examples are briefly presented. Sharp and fuzzy empirical applications are provided in the Appendix. (Word Count: 2,283)

JEL codes: C21, C25

Keywords: Regression discontinuity design, Local independence, Local smoothness

1 Introduction

Regression discontinuity (RD) designs have been widely used in many areas of empirical research. There exist two alternative assumptions to identify local average treatment effects (LATE) in RD designs: local independence (LI) and local smoothness (LS). Together with the usual LATE assumptions requiring existence of a first-stage and treatment monotonicity, either of these two assumptions is sufficient to identify RD LATE.[1] This paper discusses these alternative assumptions, and show that weakening LI by LS might be empirically relevant. However, when LI does hold, there are some practical implications one may explore.

[*] Yingying Dong, Department of Economics, University of California Irvine, CA 92697, USA. Email: yyd@uci.edu. http://yingyingdong.com/

[1] This paper focuses on fuzzy RD designs, with sharp design following as a special case.

LI requires that individual treatment effects and potential treatment status are jointly independent of the running variable in the neighborhood of the RD cutoff (Hahn, Todd and van der Klaauw, 2001, hereafter HTK). It is a local version of the independence assumption proposed in the LATE framework (Imbens and Angrist, 1994 and Angrist, Imbens, and Rubin 1996). In contrast, LS only requires that the conditional means (or distributions) of potential outcomes and potential treatment status are smooth near the RD cutoff (see, e.g.,Frandsen, Frolich and Melly, 2012, Dong 2015, and Dong and Lewbel, 2015), which can be seen as a smooth parallel of the LATE independence assumption.

In the next Section 2, I formally present LI and LS, and briefly discuss identification of RD LATE under LS. In Section 3 I discuss their practical (and testable) implications, and illustrate these implications in empirical examples; Section 4 concludes; Sharp design and fuzzy design empirical applications are provided in the supplemental online Appendix.

2 LI, LS and RD LATE

Let y_{1i} and y_{0i} be the potential outcomes for an individual i under treatment or no treatment, respectively (Neyman 1923, Fisher 1935, Rubin 1974, 1990). Let x_i be a binary treatment indicator, so $x_i = 1$ if treated and 0 otherwise. The observed outcome can then be written as $y_i = \alpha_i + \beta_i x_i$, where $\alpha_i := y_{0i}$, and $\beta_i := y_{1i} - y_{0i}$. Define the potential treatment status as $x_i(z)$ for a given value z that z_i could take on. When z_i is a binary instrument, one of the key assumptions for identifying LATE in Imbens and Angrist (1994, Condition 1 of their Theorem 1) is that $(y_{0i}, y_{1i}, x_i(z))$ is jointly independent of z_i.

In the RD framework, z_i is the running variable, and z_0 is the RD cutoff. The following discussion applies to $z \in (z_0 - \varepsilon, z_0 + \varepsilon)$ for some small $\varepsilon > 0$. In discussing the fuzzy RD design with a variable treatment effect, HTK (Assumption A3 (i)) analogously assume the following LI assumption.

ASSUMPTION LI (HTK, 2001): $(\beta_i, x_i(z))$ is jointly independent of z_i near z_0.

Independence of $(\beta_i, x_i(z))$ from z_i near z_0 implies that the individual treatment effect β_i is independent

dent of the running variable z_i locally near z_0. That is, LI requires the RD treatment effects to be locally constant. In RD designs potential outcomes can depend on the running variable directly or indirectly through omitted covariates, so LI places a restriction on treatment effect heterogeneity. In the next section I provide empirical scenarios where this type of heterogeneity arises naturally and should be taken into account for policy evaluation.

In addition, assume $x_i = h(z_i, u_i)$, where u_i is a vector of (un)observables other than the running variable z_i. Then $x_i(z) := h(z, u_i)$. LI requires $x_i(z)$ to be independent of z_i near z_0, which further implies that u_i is independent of z_i near z_0.

LS relaxes the above restrictions. The following introduces the LS assumption. To that end, I first extend the definitions of individual types of the LATE framework (Imbens and Angrist, 1994 and Angrist, Imbens, and Rubin, 1996) to the RD setup. For an individual with $z_i = z$, let $x_{1i}(z)$ and $x_{0i}(z)$ be her potential treatment status if she is above or below the cutoff, respectively.[2] For example, if $z \geq z_0$, $x_{1i}(z)$ is her observed treatment status above the cutoff, while $x_{0i}(z)$ would be her counterfactual treatment status if she were below the cutoff. The converse holds for $z < z_0$. Given $z_i = z$, an individual's type $\psi_i(z) = A(z)$ if $x_{1i}(z) = x_{0i}(z) = 1$ (always takers), $\psi_i(z) = N(z)$ if $x_{1i}(z) = x_{0i}(z) = 0$ (never takers), $\psi_i(z) = C(z)$ if $x_{1i}(z) > x_{0i}(z)$ (compliers), and $\psi_i(z) = D(z)$ if $x_{1i}(z) < x_{0i}(z)$ (defiers). For notational convenience, whenever there is no confusion, I will suppress the argument z to use x_{0i} (x_{1i}) to denote $x_{0i}(z)$ ($x_{1i}(z)$) and similarly use A, N, C, and D to denote individual types.

ASSUMPTION LS1: $E\left[y_{xi} | \psi_i = \psi, z_i = z\right]$ for $x \in \{0, 1\}$ and $\Pr\left(\psi_i = \psi | z_i = z\right)$ for $\psi \in \{A, N, C, D\}$ are continuous in z at $z = z_0$.

[2] To see how these potential treatment status are defined, assume that the treatment x_i is determined by $x_i = h(z_i, u_i)$, where u_i is a vector of (un)observables. Let $d_i = 1(z_i \geq z_0)$ be the indicator for being right above the cutoff z_0. Given that d_i is binary and is a deterministic function of z_i, without loss of generality, one can write $x_i = h_1(z_i, u_i) d_i + h_0(z_i, u_i)(1 - d_i)$, where the functions $h_1(z_i, u_i)$ and $h_0(z_i, u_i)$ describe the treatment assignment above and below the cutoff, respectively. When $h_{di}(z, u_i)$, $d = 0, 1$ is viewed as a random variable, I use $x_{di}(z)$ to denote $h_{di}(z, u_i)$.

LS1 assumes that the conditional means of potential outcomes for each type of individuals and the probabilities of different types are continuous at the RD cutoff. For sharp designs, everyone is a complier. LS1 then reduces to the assumption that $E\left[y_{xi}|z_i = z\right]$, $x = 0, 1$, are continuous at $z = z_0$.[3]

LS1 can be justified by a weak and empirically plausible behavioral assumption, in the spirit of Lee (2008). For the special case of sharp RD designs, Lee (2008) provides behavioral assumptions that lead to continuity of the conditional density (conditional on an individual's 'identity') of the running variable, which further implies local randomization and hence causal inference. For fuzzy designs, one needs to take into account (via smoothness assumptions) the probabilities with which individuals may self-select into types such as compliers.

Define the random vector $w_i := (y_{0i}, y_{1i}, \psi_i)$ with support $\mathcal{W}$. Denote the conditional density of the running variable z_i conditional on w_i as $f_{z|w}(\cdot|\cdot)$ and the unconditional density as $f_z(\cdot)$.

ASSUMPTION LS2: $f_{z|\mathbf{w}}(z|\mathbf{w})$ is continuous in a neighborhood of $z = z_0$ for all $w \in \mathcal{W}$. $f_z(\cdot)$ is continuous and strictly positive in a neighborhood of $z = z_0$.

LS2 imposes smoothness on the conditional and unconditional densities of the running variable. In particular, it asserts that for each 'individual' defined by w_i, the conditional density of the running variable z_i conditional on w_i is smooth. By Bayes' Rule, $f_{\mathbf{w}|z}(\mathbf{w}|z) = f_{z|\mathbf{w}}(z|\mathbf{w}) f_{\mathbf{w}}(\mathbf{w})/f_z(z)$, so LS2 implies that $f_{\mathbf{w}|z}(\mathbf{w}|z)$ is continuous in z at $z = z_0$, where $f_{\mathbf{w}|z}(\mathbf{w}|z)$ denotes the (possibly mixed) joint density of $\mathbf{w}_i$ conditional on $z_i = z$. Other than smoothness, LS2 imposes no restrictions on $y_{1i} - y_{0i}$, so treatment effects can be arbitrarily heterogeneous and be correlated with z_i. LS2 allows for self-selection into treatment and into different types. In particular, there can be endogenous selection into compliers, as long as the probability of being a complier is smooth at the cutoff.

LS2 implies LS1, which provides theoretical ground for performing McCrary's (2008) density test

[3] To identify quantile treatment effects (QTE) in RD designs, Frandsen, Frolich and Melly (2012) assume that the conditional distributions of potential outcomes, conditional on individual types are smooth.

in fuzzy RD designs. LS2 is stronger than necessary. For example, some observations may be missing above or below the cutoff so that the density of z_i has a discontinuity. However, LS1 can still hold, as long as the observations are missing at random. LS2 is appealing considering its plausible behavioral interpretation and testable implications. Note that one cannot test continuity of the conditional density, but only the unconditional density of the running variable. Therefore, passing the RD density test is neither necessary nor sufficient for validity of RD designs.

Define $y^{+} := \lim_{\varepsilon \to 0} E\left[y_i | z_i = z + \varepsilon\right]$, $y^{-} := \lim_{\varepsilon \to 0} E\left[y_i | z_i = z - \varepsilon\right]$, and similarly x^{+} and x^{-}.

THEOREM: Assume that $\Pr\left(\psi_i = D\right) = 0$ in the neighborhood of z_0, and $x^{+} \neq x^{-}$. Then under LS1, $E\left[y_{1i} - y_{0i} | z_i = z_0, \psi_i = C\right] = \frac{y^{+} - y^{-}}{x^{+} - x^{-}}$.

All proofs are in the online supplemental Appendix. Assuming no defiers and existence of a discontinuity, LS1 yields the standard RD identification result that was first established by HTK.

3 Discussion

Either LI or LS can be used to identify RD LATE. They also have readily testable implications.

Assume $y_i = g\left(x_i, z_i, v_i\right)$, where v_i is a vector of other (un)observable covariates other than the running variable z_i. Consider a local linear approximation of it above and below the cutoff (assuming a uniform kernel): $y_i = \alpha_0 + \alpha_1 \left(z_i - z_0\right) + \tau_0 x_i + \tau_1 \left(z_i - z_0\right) x_i + e_i$. $\tau_0 + \tau_1 z_i$ is then the treatment effect. LI requires the treatment effect to be independent of z_i near z_0, and hence $\tau_1 = 0$. In a sharp design, $\tau_1 = 0$ means no slope change at the cutoff. In fact, having treatment effects to be independent of the running variable in sharp designs implies no slope or any higher order derivative changes right at the RD cutoff.

Consider the sharp RD design of Lee (2008). x_i is an indicator for the Democratic Party being the incumbent party. z_i is the Democratic Party's winning margin, and $z_0 = 0$. y_i is whether a Democrat won the next election. In this case, LI requires that the incumbent party's electoral advantage does not

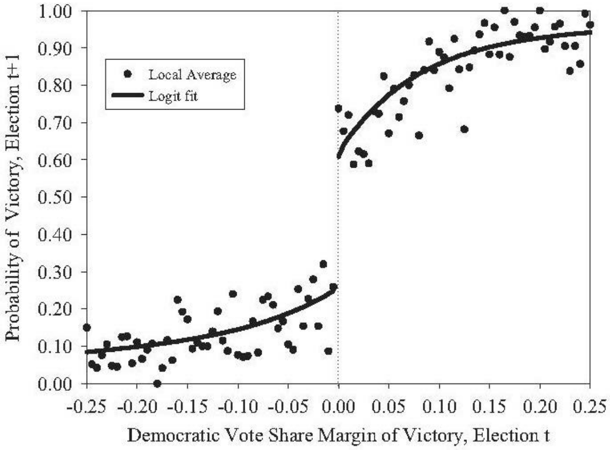

Figure 1: Probability of the Democratic Party winning election t+1 against its winning margin in election t

depend on its winning margin. However, such a dependency may exist directly or indirectly, due to, e.g., omitted (un)observables.

Figure 1 (reproduced from Figure 5-(a) of Lee 2008) shows the probability of a Democrat winning election $t+1$ given its winning margin in election t. The slope gets steeper right above the threshold, implying that the larger the incumbent party's share is in the previous election, the greater their chance of winning the next election, i.e., the incumbency advantage depends on the winning margin.

Consider another RD design of Goodman (2008) estimating the effect of the Adams Scholarship program on college choices. The treatment x_i is eligibility for the Adams Scholarship. It is determined by a student's standard test score z_i exceeding a certain threshold. The Adams Scholarship program provides qualified students tuition wavers at in-state public colleges in Massachusetts, United States.

Figure 2 shows the probability of choosing a four-year public college conditional on the number of grade points to the eligibility threshold. The probability of choosing a four-year public college jumps up at the eligibility threshold, but then declines quickly once further above the threshold. The dramatic downward slope change at the threshold suggests that a student's response to an Adams Scholarship likely depends on her test score, and thereby invalidates LI. Goodman (2008) shows that marginally qualified react much more strongly to the price change than students with test scores further above the threshold.[4]

[4] The dramatic downward slope change is induced neither by manipulation or by missing covariates.

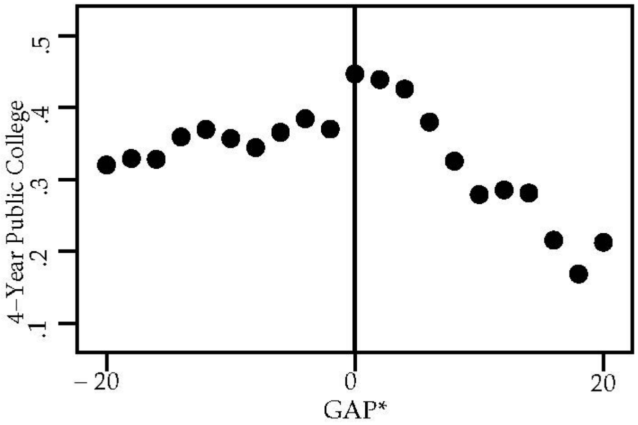

Figure 2: Probability of choosing a 4-year public college against the grade points from the eligibility threshold

Students trade college quality with prices. Better qualified students may be admitted to private colleges of much higher quality, and hence would face a large quality drop had they accepted the Adams Scholarship. In contrast, the quality difference is smaller or non-existent for marginal winners.

One may formally assess validity of LI, given smoothness. LI implies that any derivatives of the RD treatment effects evaluated at the RD cutoff are zero. I focus on the first derivative $\frac{\partial}{\partial z}E\left[\beta_i|z_i = z, \psi_i = C\right]|_{z=z_0}$. Under minimal further smoothness assumptions, i.e., assuming continuous differentiability instead of just continuity of the conditional means and probabilities in LS1, Dong and Lewbel (2015) show that $E\left[\beta_i|z_i = z, \psi_i = C\right]|_{z=z_0}$, referred to as treatment effect derivative (TED), can be nonparametrically identified and estimated. One can then test significance of the estimated TED to evaluate LI. Consider again the local linear approximation $y_i = \alpha_0 + \alpha_1 (z_i - z_0) + \tau_0 x_i + \tau_1 (z_i - z_0) x_i + e_i$. TED is captured by τ_1. One can estimate τ_1 by a local two stage least squares (2SLS) estimator for a given bandwidth, using $d_i := 1(z_i - z_0 \geq 0)$ and $d_i (z_i - z_0)$ as excluded IVs for x_i.[5] For sharp designs, $x_i = d_i$, this reduces to estimating the slope change at the RD cutoff.

TED in the RD design of Lee (2008) is estimated to be between 1.143 and 1.349, statistically significant at the 5% or 1% level (provided in the online supplemental Appendix). That is, given a 1 percentage point increase in the Democrats' winning margin, the probability for a Democrat to win the next election

[5] Any other desirable kernel functions than a uniform kernel can be applied by estimating kernel weighted local 2SLS.

increases by 1.143% to 1.349%. In another fuzzy design empirical application, I investigate how placement on academic probation affects the dropout probability in college. The negative effect of probation is shown to increase significantly as a student's GPA moves marginally below the probation threshold. In both cases, empirical evidence suggests that smoothness is plausible, even though there is a great treatment effect heterogeneity with respect to the running variable or how far one is from the cutoff.

Additionally, LI implies that (un)observable covariates that determine potential treatment status, u_i, are independent of the running variable near z_0, so one can test independence of covariate means from the running variable near z_0 to check for LI. This is similar to testing smoothness of the conditional means of covariates to evaluate the smoothness conditions. This kind of tests are essentially falsification tests.

LI might be restrictive in some empirical scenarios. However, when LI holds, there are practical implications one may explore. First, the estimated RD local effects have stronger external validity. Second, for sharp designs when LI holds, one does not need to under-smooth in order to shrink the bias to zero to have correct inference. In practice, the robust bias-corrected inference of Calonico, Cattaneo, and Tituinik (2014) is proposed by taking into account the slope change so that inference remains valid even for "too large" bandwidth choices. LI implies no slope changes at the cutoff in sharp designs, so one can enjoy robust estimation that is less sensitive to bandwidth choice and bias correction.

4 Conclusion

Either local independence (LI) or a smooth parallel of it, local smoothness (LS), can be used to identify RD LATE. Both have readily testable implications. This paper discusses these two alternative assumptions and note that LS is closely related to a weak and empirically plausible behaviorial assumption, in the spirit of Lee (2008). However, when LI holds, there are some practical implications one may explore. The discussion also provides theoretical ground for McCrary's (2008) density test in fuzzy RD designs.

References

[1] Angrist, J., G. Imbens and D. Rubin (1996), “Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91(434), 444-455.

[2] Dong, Y. (2015), “Regression Discontinuity Applications with Rounding Errors in the Running Variable," Journal of Applied Econometrics, 30, 422-446.

[3] Dong Y. and A. Lewbel (2015), "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," The Review of Economics and Statistics, 97(5), 1081-1092.

[4] Fisher, R. (1935): "Design of Experiments," Oliver and Boyd, Edinburgh.

[5] Frandsen B., M. Frolich, and B. Melly (2012): "Quantile treatment effects in the regression discontinuity design," Journal of Econometrics, 168, 382-395.

[6] Goodman, J. (2008): "Who Merits Financial Aid?: Massachusetts’ Adams Scholarship," Journal of Public Economics 92(10) 2121-2131.

[7] Hahn, J., P. Todd, W. van der Klaauw (2001): "Identification and estimation of treatment effects with a regression-discontinuity design," Econometrica 69(1), 201-209.

[8] Imbens, G. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62(2), 467-475.

[9] Lee, D.S. (2008): "Randomized Experiments from Non-random Selection in U.S. House Elections," Journal of Econometrics, 142(2), 675-97.

[10] McCrary, J. (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test,” Journal of Econometrics, 142(2), 698-714.

[11] Neyman, J. 1923: "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," Translated in Statistical Science, 5(4), 465-480, 1990.

[12] Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," Journal of Educational Psychology, 66, 688-701.

[13] Rubin, D. (1990): "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," Statistical Science, 5, 472-480.

Supplemental Appendix to "Alternative Assumptions to Identify LATE in Regression Discontinuity Designs"

Yingying Dong

University of California Irvine[*]

August 2017

Abstract

This document provides two empirical applications for the discussion in the author's paper "Alternative Assumptions to Identify LATE in Regression Discontinuity Designs." Also provided are proofs to the implication of the assumption LS2 and the Theorem in the paper.

JEL codes: C21, C25
Keywords:Regression discontinuity design, Local independence, Local smoothness

1 Empirical Applications

This section provides two empirical applications. One is for the sharp design, and the other is for the fuzzy design. The goal is to illustrate evaluating LI vs. LS in empirical settings.

1.1 Sharp Design

First I revisit the sharp RD model of Lee (2008) investigating the incumbency advantage in the US house election. The treatment in this case is an indicator for the Democratic Party being the incumbent party. The running variable is the Democratic Party's winning margin in election t. The outcome is whether a democratic candidate won in election $t+1$.

[*] Yingying Dong, Department of Economics, University of California Irvine, CA 92697, USA. Email: yyd@uci.edu. http://yingyingdong.com/

The analysis draws on the same data as those used in Lee (2008) and Lee and Lemieux (2010).[1] The sample consists of 6,558 elections from 1946 to 1998. Following Lee and Lemieux (2010), I use local linear regressions to estimate the local causal effect of being an incumbent party. Analogous to using local linear regressions to estimate means at a boundary point, local quadratic regressions may be appropriate for estimating slopes. See, e.g., discussion in Porter (2003) and Calonico, Cattaneo, and Titiunik (2014). I therefore use local quadratic regressions to estimate the derivative of the RD treatment effect (corresponding to the slope change) at the RD cutoff.

Table 1 Sharp RD Estimates of the Incumbency Advantage

	CCT	IK	CV	CCT_u	IK_u	CV_u
RD LATE	0.387	0.402	0.411	0.364	0.386	0.414
	(0.050)***	(0.046)***	(0.039)***	(0.051)***	(0.046)***	(0.040)***
Bandwidth	0.160	0.147	0.202	0.118	0.116	0.153
N	1,850	1,725	2,291	1,405	1,375	1,784
Polynomial order	1	1	1	1	1	1

Note: This table uses data from Lee (2008); All RD LATE estimates are based on bias-corrected robust inference proposed by Calonico, Cattaneo and Titiunik (CCT, 2014), using local linear regressions; using local linear regressions; CCT and IK refer to the optimal bandwidths proposed by CCT and Imbens and Kalyanaraman (2014), respectively; CV refers to the cross validation bandwidth proposed by Ludwig and Miller (2007); CCT, IK, and CV use the triangular kernel, and CCT_u, IK_u and CV_u use the Uniform kernel; Standard errors are in parentheses; Standard errors are in parentheses; * significant at the 10% level, ** significant at the 5% level, ***significant at the 1% level.

For kernel choices, I adopt both the boundary optimal triangular kernel (Fan and Gijbels, 1996), and the uniform kernel, which is frequently used for convenience. Three different bandwidth estimators are used to choose the optimal bandwidth for the local linear or local quadratic regressions. These are the plug-in estimator proposed by Calonico, Cattaneo and Titiunik (2014), the plug-in estimator proposed by Imbens and Kalyaranaman (2014), and the cross-validation estimator proposed by Ludwig and Miller (2007).

Table 1 and Table 2 report estimates of the treatment effect and the treatment effect derivative,

[1] Caughey and Sekhon (2011) show possible manipulation in this case. However, Lee and Lemieux (2014) notice that this can be explained by the sampling differences between Caughey and Sekhon (2011) and Lee (2008).

Table 2 The Derivative of the Incumbency Advantage

	CCT	IK	CV	CCT_u	IK_u	CV_u
TED	1.349	1.240	1.209	1.391	1.143	1.169
	(0.599)**	(0.442)***	(0.411)***	(0.560)**	(0.434)***	(0.450)***
Bandwidth	0.353	0.432	0.452	0.304	0.361	0.352
N	3,796	4,410	4,570	3,289	3,866	3,785
Polynomial order	2	2	2	2	2	2

Note: This table uses data from Lee (2008); All estimates are based on local quadratic regressions; CCT and IK refer to the optimal bandwidths proposed by Calonico, Cattaneo, and Titiunik (2014) and Imbens and Kalyanaraman (2014), respectively; CV refers to the cross validation bandwidth proposed by Ludwig and Miller (2007); CCT, IK, and CV use the triangular kernel, and CCT_u, IK_u, and CV_u uses the uniform kernel; Bandwidth and sample size N refer to those of the outcome equation; Bootstrapped Standard errors based on 500 simulations are in parentheses; * significant at the 10% level, ** significant at the 5% level, ***significant at the 1% level.

respectively. The treatment effect derivative in this case measures how the incumbency advantage depends on the incumbent party's winning margin, corresponding to the slope change at the cutoff in Figure 1.

The estimated incumbency effects and their derivatives are largely robust to different bandwidth choices. Consistent with estimates in Lee (2008) and Lee and Lemieux (2010), the average incumbency effect is estimated to be between 0.364 and 0.414, meaning that when the Democratic Party is the incumbent party, it increases their probability of winning the next election by 36.4% to 41.4%. The estimated treatment effect derivative is between 1.143 and 1.349, so given a 1 percentage point increase in the Democrats' winning margin, the probability for their candidates to win the next election increases by 1.143% to 1.349%. The estimated incumbency effects and their derivatives are all statistically significant. Therefore, LI is not likely to hold in this case. In particular, the incumbency advantage depends on the incumbent party's winning margin.

Table 3 reports the estimated jumps or kinks at the RD cutoff in the conditional mean of an important covariate, the Democratic vote share from the previous election. Also reported are the estimated jumps and kinks in the empirical density of the running variable at the RD cutoff. I only report estimates by the local linear or quadratic regressions with triangular kernels and bandwidths

Table 3 Smoothness of the Covariate Mean and Density of the Running Variable

	Estimates	Bandwidth	No of obs.	Polynomial order
	Previous Election Vote Share			
Jump	-0.001 (0.015)	0.190	2,170	1
Kink	-0.150 (0.337)	0.239	2,663	2
	Density of the Winner Margin in Election t			
Jump	0.138 (0.161)	0.205	82	1
Kink	2.400 (3.629)	0.212	84	2

Note: Standard errors are in parentheses; All estimates use the CCT optimal bandwidth and the triangular kernel.

chosen by the Clonico, Cattaneo and Tituinik's (2014) plug-in estimator. Estimates using uniform kernels and other bandwidths are similar and are therefore suppressed to save space. None of the estimated jumps and kinks are statistically significant. These results support the plausibility of the smoothness assumption, and hence validity of the RD design, even though LI likely does not hold by the results in Table 2.

1.2 Fuzzy Design

I next conider a fuzzy RD design based on the probation rule in college and evaluate the impact of academic probation on the dropout probability. I evaluate how the discouragement effect of academic probation depends on the running variable, a student's pre-treatment GPA.

Nearly all colleges and universities in the US adopt academic probation to motivate students to stay above a certain performance standard. Typically students are placed on academic probation if their GPAs fall below a pre-determined threshold. Students on academic probation face the real threat of being suspended if their performance continues to fall below.

Let Y be the binary indicator for dropout, which is 1 if a student drop out of college and 0 otherwise. The running variable R is the first semester GPA. The treatment T is the indicator for ever being placed on academic probation. I use confidential data from a large Texas university collected under the Texas Higher Education Opportunity Project (THEOP). The actual probation status is not

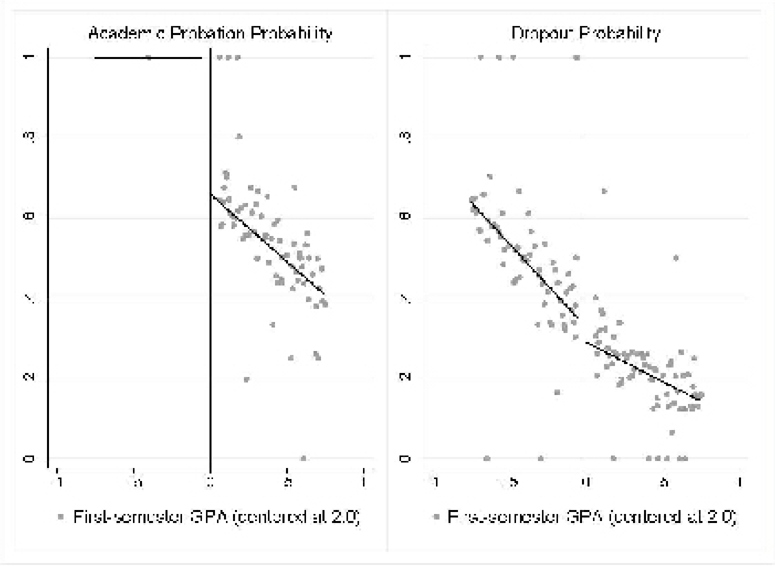

Figure 1: Academic Probation and Dropout Rates against First-semester GPA

observed in the data. Here I define treatment to be 1 as long as a student's cumulative or semester GPA is below the school-wide cutoff 2.0, i.e., when a student is considered as 'scholastically deficient.'[2] The data used here represent the entire population of the first-time freshmen cohorts between 1992 and 2002. The total sample size is 64,310.

Figure 3 presents the probation probability and the dropout rate conditional on the first semester GPA. In the left graph, for those whose first semester GPAs fall below the probation threshold, the probation probability is 1 by construction; for those whose first semester GPAs fall marginally above, there is still an over 60% chance for them to be placed on probation later. Note that a dramatic slope change is present at the RD cutoff, indicating that the fraction of 'compliers' depends on the running variable. In the right graph, the dropout rate also shows a discernible slope change the RD cutoff, in

[2] An undergraduate at this university is considered as 'scholastically deficient' if his or her semester or cumulative GPA falls below 2.0. In practice, when a student is considered as scholastically deficient, he or she may only be given an academic warning. However, a quick survey administered to the relevant academic deans shows that students are generally placed on probation in this case.

addition to a small level change.

Table 4 Fuzzy RD Estimates of the Impact of Academic Probation on Dropout Rates

	CCT	IK	CV	CCT_u	IK_u	CV_u
1st-stage discontinuity	-0.343	-0.345	-0.352	-0.336	-0.341	-0.378
	(0.010)***	(0.010)***	(0.016)***	(0.010)***	(0.010)***	(0.023)***
RD-LATE	0.068	0.108	0.108	0.073	0.128	0.120
	(0.053)	(0.084)	(0.085)	(0.049)	(0.084)	(0.106)
Bandwidth	0.869	0.723	0.681	0.762	0.568	0.487
N	31,396	25,149	23,623	26,780	19,413	15,763
Polynomial order	1	1	1	1	1	1

Note: This table uses the Administration and Transcript Data from a Public University in Texas; All RD LATE estimates are based on bias-corrected robust inference proposed by Calonico, Cattaneo and Titiunik (CCT, 2014), using local linear regressions; CCT and IK refer to the optimal bandwidths proposed by CCT and Imbens and Kalyanaraman (2014), respectively; CV refers to the cross validation bandwidth proposed by Ludwig and Miller (2007); CCT, IK, and CV use the triangular kernel, and CCT_u, IK_u and CV_u use the Uniform kernel; Standard errors are in parentheses; * significant at the 10% level, ** significant at the 5% level, ***significant at the 1% level.

Table 4 reports estimates of RD LATE. Placement on academic probation is shown to have a small, positive, yet insignificant impact on the college dropout rate right at the first-semester probation threshold. However, in Table 5, I report estimates of the treatment effect derivative. These derivative estimates are largely robust to the choices of bandwidths and kernel functions and are statistically significant at the 1% level. In particular, the estimated treatment effect derivative is -0.568 by the CCT bias-corrected estimation with a triangular kernel. This implies that the discouragement effect of placement on academic probation increases significantly as a student's GPA moves marginally below the cutoff. In particular, when a student's first-semester GPA decreases by 0.1, the probability of dropping out of college once on probation increases by 5.68%. This is large in magnitude, compared with the change of 7-13% in the dropout rate at the cutoff. These results stronly suggest that the probation effect depends on how far one is from the probation threshold and hence LI is violated in this case.

To check for smoothness in this case, I estimate the jumps and kinks in the conditional means of covariates and those in the density of the running variable at the probation threshold. Covariates

Table 5 The derivative of the Impact of Academic Probation on Dropout Rates

	CCT	IK	CV	CCT_u	IK_u	CV_u
1st-stage derivative	-0.371	-0.269	-0.336	-0.400	-0.298	-0.342
	(0.022)***	(0.077)***	(0.033)***	(0.018)***	(0.090)***	(0.045)***
TED	-0.568	-0.584	-0.555	-0.754	-0.639	-0.591
	(0.150)***	(0.376)	(0.144)***	(0.103)***	(0.422)	(0.147)***
Bandwidth	1.604	0.868	1.648	1.807	0.726	1.358
N	54,151	31,396	54,595	59,306	25,185	47,846
Polynomial order	2	2	2	2	2	2

Note: This table uses the Administration and Transcript Data from a Public University in Texas; All estimates are based on local quadratic regressions; CCT and IK refer to the optimal bandwidths proposed by Calonico, Cattaneo and Titiunik (2014) and Imbens and Kalyanaraman (2014), respectively; CV refers to the cross validation bandwidth proposed by Ludwig and Miller (2007); CCT, IK, and CV use the triangular kernel, and CCT_u, IK_u, and CV_u uses the uniform kernel; Bandwidth and sample size N refer to those of the outcome equation; Bootstrapped Standard errors based on 500 simulations are in parentheses; * significant at the 10% level, ** significant at the 5% level, ***significant at the 1% level.

investigated include dummies for male, Black, and Hispanic, as well as an indicator for being ranked among the top 25% of the high school class. These results are reported in Table 6. None of the estimated jumps and kinks are statistically significant, indicating that the smoothness conditions are plausible and hence the RD design is still valid, even though LI does not hold.

Table 6 Smoothness of Covariate Means and Density of 1st-semester GPA

	Jump		Kink	
Male	-0.007	(0.019)	-0.084	(0.220)
Black	-0.004	(0.010)	0.035	(0.075)
Hispanic	0.010	(0.013)	0.102	(0.172)
Top 25% of HS Class	0.013	(0.018)	0.085	(0.175)
Density of 1st-semester GPA	0.381	(0.369)	-0.266	(1.646)

Note: CCT bias-corrected intercept and slope change estimates are reported; Robust standard errors are in parentheses. The bin size used to generate the empirical density is .006.

2 Proofs

Proof that LS2 implies LS1: For simplicity, assume that y_{0i} and y_{1i} are continuous, though analogous analysis can be done when y_{0i} and y_{1i} are discrete. The following discussion applies to $z_i = z \in (z_0 - \varepsilon, z_0 + \varepsilon)$ for some small $\varepsilon > 0$. Let $f_{\cdot}(\cdot)$ and $f_{\cdot|\cdot}(\cdot|\cdot)$ denote the unconditional and conditional probability density or mass functions, respectively. In particular, let $f_{\mathbf{w}|z}(\mathbf{w}|z)$ denote the mixed joint density of $\mathbf{w}_i$ conditional on $z_i = z$, i.e., $f_{\mathbf{w}|z}(\mathbf{w}|z) = f_{y_0,y_1,z|\psi}(y_0, y_1, z|\psi_i = \psi)\Pr(\psi_i = \psi)/f_z(z)$.

Assumption LS2 states that $f_{z|\mathbf{w}}(z|\mathbf{w})$ is continuous in z, and $f_z(z)$ is continuous and strictly positive at $z = z_0$. By Bayes' Rule, $f_{\mathbf{w}|z}(\mathbf{w}|z) = f_{z|\mathbf{w}}(z|\mathbf{w}) f_{\mathbf{w}}(\mathbf{w})/f_z(z)$, so $f_{\mathbf{w}|z}(\mathbf{w}|z)$ is continuous in z at $z = z_0$. By definition $\mathbf{w}_i := (y_{0i}, y_{1i}, \psi_i)$, then probability of each type of individual $\Pr(\psi_i = \psi|z_i = z) = \int_{\Omega_1}\int_{\Omega_0} f_{\mathbf{w}|z}(\mathbf{w}|z)dy_0dy_1$ for $\psi_i = \psi \in \{A, N, C\}$ is continuous in z at $z = z_0$, where Ω_t is the conditional support of y_{ti} for $t = 0, 1$ conditional on $z_i = z$.

By Bayes' Rule, $f_{y_0,y_1|\psi,z}(y_0, y_1|\psi_i = \psi, z_i = z) = f_{\mathbf{w}|z}(\mathbf{w}|z)/Pr(\psi_i = \psi|z_i = z)$ for $\psi_i = \psi \in \{A, N, C\}$. Both $f_{\mathbf{w}|z}(\mathbf{w}|z)$ and $\Pr(\psi_i = \psi|z_i = z)$ are continuous in z at $z = z_0$, so $f_{y_0,y_1|\psi,z}(y_0, y_1|\psi_i = \psi, z_i = z)$ for $\psi_i = \psi \in \{A, N, C\}$ is continuous in z at $z = z_0$. It follows that type-specific conditional means of potential outcome $E(y_{ti}|\psi_i = \psi, z_i = z)$ for $t = 0, 1$,and $\psi_i = \psi \in \{A, N, C\}$ are continuous in z at $z = z_0$.

Proof of Theorem: Given LS1, monotonicity, and the definitions of individual types, we have

$$
\begin{aligned}
& y^+ - y^- = \lim_{\varepsilon \to 0} E\left[\alpha_i + \beta_i x_i | z_i = z_0 + \varepsilon\right] - \lim_{\varepsilon \to 0} E\left[\alpha_i + \beta_i x_i | z_i = z_0 - \varepsilon\right] \\
= \; & \lim_{\varepsilon \to 0} E\left[\beta_i x_i | z_i = z_0 + \varepsilon\right] - \lim_{\varepsilon \to 0} E\left[\beta_i x_i | z_i = z_0 - \varepsilon\right] \\
= \; & \lim_{\varepsilon \to 0} E\left[\beta_i | z_i = z_0 + \varepsilon, x_{1i} = 1\right] \Pr\left[x_{1i} = 1 | z_i = z_0 + \varepsilon\right] \\
& - \lim_{\varepsilon \to 0} E\left[\beta_i | z_i = z_0 - \varepsilon, x_{0i} = 1\right] \Pr\left[x_{0i} = 1 | z_i = z_0 - \varepsilon\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\alpha_i|z_i = z_0\right] + E\left[\beta_i|z_i = z_0, \psi_i = C\right]\Pr\left[\psi_i = C|z_i = z_0\right] \\
&\quad +E\left[\beta_i|z_i = z_0, \psi_i = A\right]\Pr\left[\psi_i = A|z_i = z_0\right] \\
&\quad -E\left[\alpha_i|z_i = z_0\right] + E\left[\beta_i|z_i = z_0, \psi_i = A\right]\Pr\left[\psi_i = A|z_i = z_0\right] \\
&= E\left[\beta_i|z_i = z_0, \psi_i = C\right]\Pr\left[\psi_i = C|z_i = z_0\right].
\end{aligned}
$$

In addition,

$$
\begin{aligned}
x^+ - x^- &= \lim_{\varepsilon\to 0} E\left[x_i|z_i = z_0+\varepsilon\right] - \lim_{\varepsilon\to 0} E\left[x_i|z_i = z_0-\varepsilon\right] \\
&= E\left[x_{1i}|z_i = z_0\right] - E\left[x_{0i}|z_i = z_0\right] \\
&= E\left[x_{1i} > x_{0i}|z_i = z_0\right] \\
&= \Pr\left[\psi_i = C|z_i = z_0\right].
\end{aligned}
$$

By A2, $x^+ - x^- \neq 0$, then

$$
E\left[y_{1i} - y_{0i}|z_i = z_0, \psi_i = C\right] = \frac{y^+ - y^-}{x^+ - x^-}.
$$

References

[1] Calonico, S., M.D. Cattaneo and R.Titiunik, (2014), “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs,” Econometrica, 82(6), 2295–2326.

[2] Caughey, D. and J. S. Sekhon (2011): "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008,” Political Analysis, 19, 385-408.

[3] Imbens W. G. and Kalyanaraman, K. (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," Review of Economic Studies,79 (3): 933-959.

[4] Lee, D.S. (2008): "Randomized Experiments from Non-random Selection in U.S. House Elections," Journal of Econometrics, 142(2), 675-97.

[5] Lee, D.S., T. Lemieux (2010): "Regression Discontinuity Designs in Economics," Journal of Economic Literature, 48, 281-355.

[6] Lee, D.S., T. Lemieux (2014): "Regression Discontinuity Designs in Social Sciences," in Regression Analysis and Causal Inference, H. Best and C. Wolf (eds.), Sage.

[7] Ludwig J., and D. L. Miller (2007): "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," The Quarterly Journal of Economics, 122, 159-208.

[8] Porter, J. (2003): "Estimation in the Regression Discontinuity Model," unpublished manuscript, Department of Economics, University of Wisconsin, Madison.