An Alternative Assumption to Identify LATE in Regression Discontinuity Designs

Yingying Dong

University of California Irvine*

This version: December 2016

Abstract

One of the key identifying assumptions for regression discontinuity (RD) designs is the local independence assumption (LI). This paper shows that it is both theoretically and empirically useful to relax LI. LATE in both sharp and fuzzy designs can be identified under alternative smoothness conditions, and the required smoothness is satisfied given a weak and empirically plausibly behavioral assumption, in the spirit of Lee (2008). A sufficient (but stronger than necessary) condition is smoothness of the conditional density of the running variable, which provides formal justification for McCrary's (2008) density test in fuzzy RD designs. (Word Count: 2,275)

JEL codes: C21, C25

Keywords: Regression discontinuity, Sharp design, Fuzzy design, LATE, Treatment effect heterogeneity, Treatment effect derivative, RD density test

1 Introduction

Regression discontinuity (RD) designs have been widely used in many areas of empirical research. One of the key identifying assumptions for RD designs is that the treatment effect 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). This local independence assumption (LI) is a local version of the independence assumption proposed in the original LATE paper by Imbens and Angrist (1994).

*I would like to thank Josh Angrist, Arthur Lewbel, Yaming Yu, Damon Clark, Marianne Bitler, Kevin Roth, David Powell and Brigham Frandsen for their helpful comments. I am also grateful to Thomas Lemieux for help with the election data. All errors are my own.

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

This paper shows that it is both theoretically and empirically useful to relax LI in RD designs. First of all, restricting the treatment effect to be independent of the running variable places an undesirable restriction on treatment effect heterogeneity, since in RD designs potential outcomes can depend on the running variable directly or indirectly through omitted covariates. This type of heterogeneity arises naturally in many empirical scenarios. Second, by not allowing the treatment effect to depend on the running variable (implying no slope or higher order derivative changes right at the RD cutoff in the sharp RD design), such an independence assumption prevents the opportunity to explore any discrete slope or derivative changes at the RD cutoff. In practice, taking into account the slope change serves as the basis for recent papers such as Calonico, Cattaneo and Titiunik (2014), Bertanha and Imbens (2014), Dong and Lewbel (2015), Yanagi (2015), Dong(2015, 2016 a, b), and Cerulli et al. (2016).

This paper discusses smoothness conditions (S1) that suffice to identify LATE in both sharp and fuzzy RD designs without LI. This paper further relates the required smoothness to a weak behavioral assumption (S2), 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, and hence local randomization and causal inference. In contrast, this paper focuses on fuzzy designs, with sharp designs following as a special case. This requires dealing with (via smoothness assumptions) the probabilities with which individuals may self-select into types such as compliers or always takers. The theorem provides precisely the same identification results as those by HTV, but under a smoothness assumption instead of LI.

This paper also discusses a testable implication of LI, given smoothness. Results in this paper provide formal support for performing McCrary's (2008) density test in fuzzy RD designs. Note that this is not the first paper that proposes smoothness conditions for RD treatment effect identification. Except for Lee (2008) on sharp designs, Frandsen, Frolich and Melly (2012) impose similar smoothness conditions to identify quantile RD treatment effects. Discussion in this paper complements these

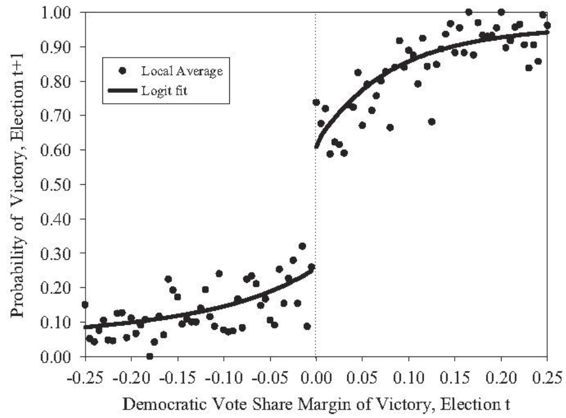

Figure 1: Probability of the Democratic Party winning election t+1 given its winning margin in election t

existing discussions.

To motivate the discussion, consider an outcome model (or a local linear approximation of it) $y_i = a + b(z_i - z_0) + \tau x_i + \tau_1 (z_i - z_0) x_i + e_i$, where y_i is the observed outcome, x_i is a binary treatment indicator, so $x_i = 1$ if treated and 0 otherwise, and z_i is the running variable with z_0 being the RD threshold. The treatment effect is then $\tau + \tau_1 z_i$. LI requires the treatment effect to be independent of z_i near z_0, so $\tau_1 = 0$. In a sharp design, $\tau_1 = 0$ means no slope change at the 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 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 (formal estimation of this is provided in the online supplemental Appendix).

Consider another RD design of Goodman (2008) estimating the effect of the Adams Scholarship

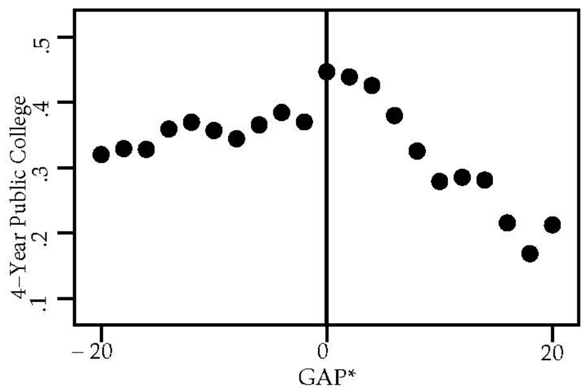

Figure 2: Probability of choosing a 4-year public college given the grade points from the eligibility threshold

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 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. Indeed, Goodman (2008) shows that qualified students with test scores near the eligibility threshold react much more strongly to the price change than students with test scores further above the threshold.[1] 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 and attended a Massachusetts public college. In contrast, the quality difference is smaller or non-existent for marginal winners.

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

2 Identification Given Smoothness

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). The observed outcome can then be written as $y_i = \alpha_i + \beta_i x_i$, where x_i is the treatment indicator, $\alpha_i := y_{0i}$, 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 an instrument, one of the key assumptions for identifying LATE in Imbens and Angrist (1994, Condition 1 of their Theorem 1) is that the triplet $(y_{0i}, y_{1i}, x_i(z))$ is jointly independent of z_i.

In the RD framework, z_i is the running variable, and z_0 is the RD cutoff. In discussing the fuzzy RD design with a variable treatment effect, HTV (Theorem 2) analogously assume that $(\beta_i, x_i(z))$ is jointly independent of z_i in a neighborhood of z_0.

Let an individual's treatment status below and above the RD cutoff be generated by the random functions $x_{0i}(z)$ and $x_{1i}(z)$, respectively.[2] For notational convenience, I simply use x_{0i} (x_{1i}) to denote $x_{0i}(z)$ ($x_{1i}(z)$). One can then define four types of individuals following Angrist, Imbens, and Rubin (1996): an individual's type $\psi_i = A$ if $x_{1i} = x_{0i} = 1$ (always takers), $\psi_i = N$ if $x_{1i} = x_{0i} = 0$ (never takers), $\psi_i = C$ if $x_{1i} > x_{0i}$ (compliers), and $\psi_i = D$ if $x_{1i} < x_{0i}$ (defiers). Note that individual types can implicitly be functions of z.

ASSUMPTION S1 (Smoothness): $E\left[y_{ti}|\psi_i = \psi, z_i = z\right]$ for $t \in \{0, 1\}$ and $\Pr\left(\psi_i = \psi | z_i = z\right)$ for $\psi \in \{A, N, C\}$ are continuous in z at $z = z_0$.

S1 replaces LI. S1 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. S1 nests the sharp design assumption as a special case. For sharp designs, everyone is a complier, so S1 reduces to the

[2] Assume $x_i = h(z_i, u_i)$ for unobservables u_i, which can be a vector. 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 $d_i = 1(z_i \geq z_0)$. The function $h_d(z_i, u_i)$ for $d = 0, 1$ describes the treatment assignment below or above the cutoff. Define then $x_{id}(z) \equiv x_{id}(z, u_i)$ for $d = 0, 1$.

assumption that $E\left[y_{ti}|z_i = z\right]$, $t = 0, 1$ are continuous at $z = z_0$.

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 S2 (Stronger Smoothness): $f_{z|w}(\cdot|\cdot)$ 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 $z = z_0$.

S2 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.[3] 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 S2 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, S2 imposes no restrictions on $y_{1i} - y_{0i}$, so treatment effects can be arbitrarily heterogeneous and be correlated with z_i. S2 allows for self-selection into treatment based on idiosyncratic gains and selection into different types. For example, there can be endogenous selection into compliers, as long as the probability of being a complier is smooth at the cutoff.

LEMMA: S2 implies S1.

All proofs are in the online supplemental Appendix. Compared with S1, S2 is stronger than necessary. For example, some observations may be missing above or below the cutoff so that the density has a discontinuity. However, S1 can still hold, as long as the observations are missing at random. S2 is more 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 density test is neither necessary nor sufficient for validity of RD designs.

[3]The independence assumption imposed by HTV may be seen to hold in the limit as z approaches to z_0.

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 $\Pr\left(\psi_i = D\right) = 0$ and $x^{+} \neq x^{-}$. Then given S1, LATE for compliers at $z_i = z_0$ is identified and is given by $E\left[y_{1i} - y_{0i} | z_i = z_0, \psi_i = C\right] = \frac{y^{+} - y^{-}}{x^{+} - x^{-}}$.

That is, given monotonicity $\Pr\left(\psi_i = D\right) = 0$ and existence of a discontinuity $x^{+} \neq x^{-}$, smoothness S1 suffices to obtain the standard RD identification results established by HTV.

3 Smoothness vs. Local Independence

LI requires that treatment effects are locally constant, while the alternative smoothness assumption implies only that treatment effects be smooth near the RD cutoff. Below I briefly discuss the theoretical importance of relaxing LI.

One example is Calonico, Cattaneo and Tituinik (2014), who propose robust bias-corrected confidence intervals that are not sensitive to "too large" bandwidth choices. When LI is plausible, one does not need to under-smooth in order to shrink the bias to zero and hence to have correct inference. The robust biased-corrected inference proposed by Calonico, Cattaneo and Titiunik (2014) will not be necessary. Intuitively, for a sharp design RD model, the bias from the local linear regression above or below depends on the derivative of the conditional mean of the outcome right above and below the RD cutoff. Under the local independence assumption, the two derivatives will be the same and hence the incurred biases from above and below will be canceled out.[4]

Another example is Dong and Lewbel (2014). They investigate external validity of RD LATE, and propose a non-parametric approach of extrapolating the RD LATE away from the RD cutoff, and identifying how the RD LATE would change when the threshold is marginally changed with additional

[4]For fuzzy design RD models, the bias depends on the derivative differences in both the conditional means of treatment and the conditional means of the outcome from above and blow the RD cutoff.

assumptions. Under LI, the external validity away from the cutoff is guaranteed and so one can directly apply the standard RD LATE to points near but not at the RD cutoff. See also Dong (2016 a) for using slope changes at the RD cutoff to identify a limit form of RD LATE.

A third example is Bertanha and Imbens (2014), who also discuss evaluating the external validity of the RD LATE. They employ smoothness conditions (citing this paper) since under LI, external validity of the RD LATE to points other than the RD cutoff holds automatically.

One may assess the validity of LI, given smoothness. LI requires that individuals' treatment effects and types do not depend on the running variable near the RD cutoff, i.e., $\partial E\left[y_{1i} - y_{0i} | z_i = z, \psi_i = C\right] / \partial z|_{z=z_0} = 0$. Under minimal further smoothness assumptions, in particular assuming continuous differentiability instead of just continuity of the conditional means and probabilities in S2, one can nonparametrically identify and estimate $\partial E\left[y_{1i} - y_{0i} | z_i = z, \psi_i = C\right] / \partial z|_{z=z_0}$ (Dong and Lewbel, 2015). In a local linear approximation $y_i = a + b\,(z_i - z_0) + \tau x_i + \tau_1 (z_i - z_0)\, x_i + e_i$, this amounts to estimating the coefficient of $x_i\,(z_i - z_0)$ using $d_i := 1\,(z_i - z_0 \geq 0)$ and $d_i\,(z_i - z_0)$ as excluded IVs. For sharp designs, $x_i = d_i$, this reduces to estimating the slope change at the RD cutoff. One can then evaluate LI by testing significance of this estimated derivative.

In the online supplemental Appendix, I show that the estimated treatment effect derivative in the RD design of Lee (2008) is between 1.143 and 1.349, statistically significant at the 5% or 1% level. That is, given a 1 percentage point increase in the Democrats' winning margin, the probability for a Democrat to win the next election increases by 1.143% to 1.349%. I also provide another fuzzy design RD application, investigating how placement on academic probation affects the dropout probability in college. I show that the discouragement effect of probation increases significantly as a student's GPA moves marginally below the probation threshold. In both applications, evidence strongly support that the smoothness conditions plausibly hold, even though LI does not.

4 Conclusion

Given a discontinuity in the treatment probability and monotonicity, identification of LATE in RD designs can be established under just smoothness conditions without local independence (LI). This paper shows that it is both theoretically and empirically useful to relax LI in RD designs. This paper discusses a weak behavioral assumption, in the spirit of Lee (2008), that leads to the required smoothness. The discussion formally justifies using McCrary's (2008) density test to evaluate validity of fuzzy design RD models. I also show that LI can be tested, given smoothness.

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] Bertanha, M. and G. W. Imbens (2014), "External Validity in Fuzzy Regression Discontinuity Designs," NBER Working Paper No. 20773.

[3] Cerulli, G., Y. Dong, A. Lewbel, and A. Poulsen, (2016) "Testing Stability of Regression Discontinuity Models," Forthcoming, Advances in Econometrics.

[4] 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.

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

[6] Dong, Y. (2016 a), "Jump or Kink? Regression Probability Jump and Kink Design for Treatment Effect Evaluation," Working Paper.

[7] Dong, Y. (2016 b), “Regression Discontinuity Designs with Sample Selection,” Working Paper.

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

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

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

[11] 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.

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

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

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

[15] 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.

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

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

[18] Yanagi (2015): "Regression Discontinuity Designs with Nonclassical Measurement Error," working paper, Department of Economics, Hitotsubashi University.

Supplemental Appendix to "An Alternative Assumption to Identify LATE in Regression Discontinuity Designs"

Yingying Dong
University of California Irvine*

November 2016

Abstract

This document provides two empirical applications for the discussion in the author's paper "An Alternative Assumption to Identify LATE in Regression Discontinuity Designs." Also provided are proofs to the Lemma and Theorem in the paper.

JEL codes: C21, C25
Keywords: Regression discontinuity, Sharp design, Fuzzy design, LATE, Treatment effect heterogeneity, Treatment effect derivative, RD density test

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 evaluate the plausibility of the two alternative assumptions, i.e., independence vs. smoothness, in RD designs. I provide evidence that the smoothness conditions plausibly hold in both cases, while the independence assumption does not, given smoothness.

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

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

party. The running variable is the Democratic Party's winning margin in election t. The outcome is whether a democratic candidate won in election $t+1$.

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 Kalyanaraman (2014), and the cross-validation estimator proposed by

[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).

Ludwig and Miller (2007).

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.

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

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.

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 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. More importantly, I evaluate how the discouragement effect of academic probation depends on the running variable, a student's pre-treatment GPA, which would not be allowed for under LI.

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 oth-

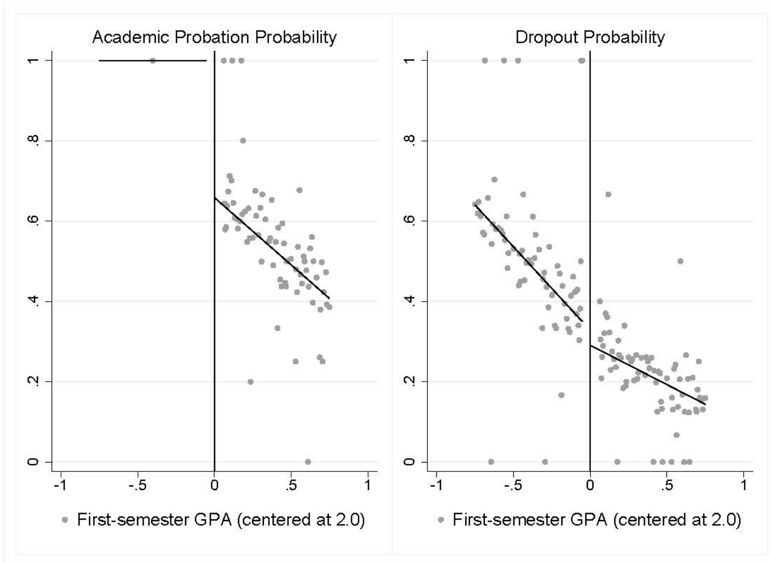

Figure 1: Academic Probation and Dropout Rates against First-semester GPA

erwise. 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 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,

[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.

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 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.

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.

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 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 Appendix

Proof of Lemma: 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 A1a 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 A1a, monotonicity, and the definitions of individual types, we have

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

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

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.