Jumpy or Kinky? Regression Discontinuity without the Discontinuity

Dong, Yingying

California State University Fullerton

1 August 2010

Jumpy or Kinky? Regression Discontinuity without the Discontinuity

Yingying Dong*
Department of Economics
California State University Fullerton

August 2010

Abstract

Regression Discontinuity (RD) models identify local treatment effects by associating a discrete change in the mean outcome with a corresponding discrete change in the probability of treatment at a known threshold of a running variable. This paper shows that it is possible to identify RD model treatment effects without a discontinuity. The intuition is that identification can come from a slope change (a kink) instead of a discrete level change (a jump) in the treatment probability. Formally this can be shown using L'hopital's rule. The identification results are interpreted intuitively using instrumental variable models. Estimators are proposed that can be applied in the presence or absence of a discontinuity, by exploiting either a jump or a kink.

JEL Codes: C21, C25
Keywords: Regression Discontinuity, Fuzzy design, Average treatment effect, Identification, Jump, Kink, Threshold

1 Introduction

Let T be a binary indicator for some treatment such as participation in a social program or repeating a grade (grade retention) in school, let Y be some associated outcome of interest such as employment or academic performance, and let X be a so-called running or forcing variable that affects both T and Y. For example, X could be the income level that affects eligibility for a social program, or an exam score affecting a grade retention decision. In the standard Regression Discontinuity (RD) framework, the probability of treatment given by $f(x) = E(T \mid X = x)$ changes discretely at a threshold point $x = c$. Under general conditions, this discontinuity or jump in $f(x)$, along with any observed corresponding jump in $g(x) = E(Y \mid X = x)$ at $x = c$, can be used to recover a local average treatment effect. See, e.g., Hahn, Todd, and van der Klaauw (2001), Imbens and Lemieux (2008), chapter 6 of Angrist and Pischke (2008), Imbens and Wooldridge (2009), and Lee and Lemieux (2010). The intuition is that if X and other

*Correspondence: Department of Economics, California State University Fullerton, CA 92834, USA. Email: ydong@fullerton.edu. http://business.fullerton.edu/Economics/ydong/.

The author would like to thank Arthur Lewbel for many very helpful comments and suggestions. Any errors are my own.

covariates determing Y and T are continuous at the threshold c, then untreated individuals having X just below the threshold can serve as valid counterfactuals for treated individuals just above.

In this paper I show that the RD local average treatment effect that is usually identified by a discontinuity can still be identified even if there is no discontinuity or jump in the treatment probability $f(x)$, given that there is a kink, i.e., a discrete change in slope in $f(x)$ at $x = c$. I also provide estimators for the local average treatment effect that can be used regardless of whether identification comes from a jump, a kink, or both. The identification results are further intuitively interpreted using IV models.

This paper's results could be applied in situations where the compliance rate changes less dramatically than required by the standard RD. For example, in applications where there is inertia or delay in taking up treatment, the added probability of treatment associated with crossing a threshold may rise as one gets further away from the threshold rather than jumping the moment the threshold is crossed. In this case, treatment effects based on standard RD estimators would either be weakly identified, if the jump is small, or unidentified if the jump is zero, regardless of how much the slope changes. In contrast, the estimators proposed in this paper make use of any changes in either the intercept or the slope of the treatment probability at the point $x = c$.

For example, Jacob and Lefgren (2004) examine the treatment effect of remedial eduction programs, including grade retention, on later academic performance, where the retention treatment is incurred by failing summer school tests. They note that "the probability of retention does not drop sharply (discontinuously) at the exact point of the cutoff , ...it rapidly decreases over a narrow range of values just below the cutoff." Indeed, their Figure 6 (reproduced in Figure 1 here) shows a dramatic slope change instead of a discontinuity in the retention probability at the cutoff (normalized to zero). In this case, the standard RD estimation based on a discrete change in the treatment probability is not suitable, whereas the estimators proposed in this paper can still apply.

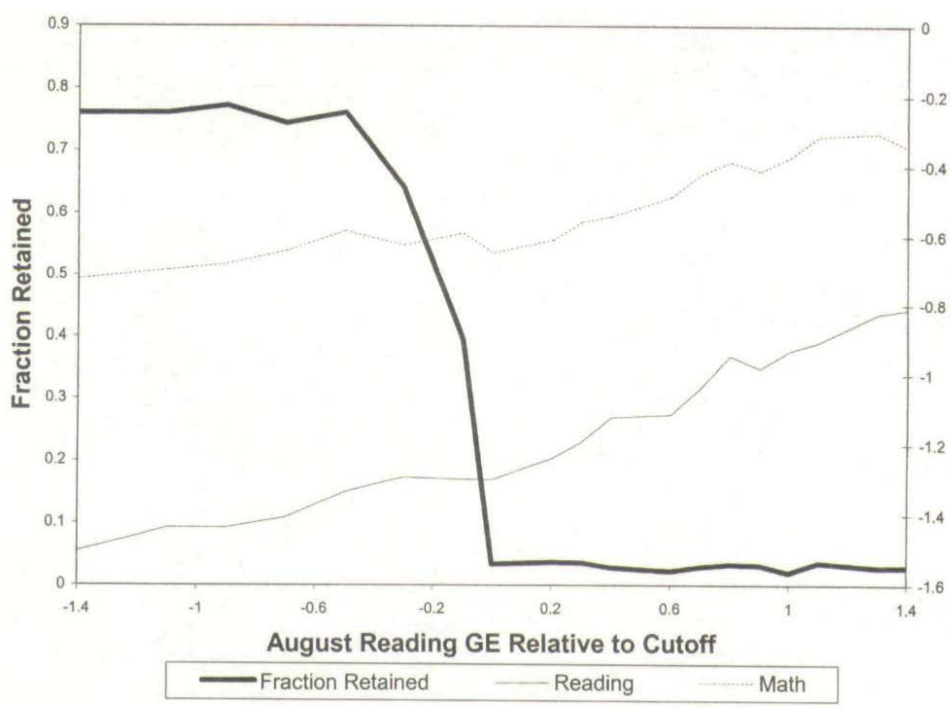

Figure 1: Retention Rate and Reading Test Scores Relative to the Cutoff

In some potential applications of RD models, there is debate about whether the probability of treatment actually jumps at a threshold. When a discrete change is small, it could be indistinguishable from a kink visually. An example is Figure 2, which reproduces Figure 4 from Card, Dobkin, and Maestas (2008), showing the probability of retirement in the US by age. It is difficult to determine whether a small jump appears in this probability at age 65, the eligibility age for full social security benefits, but there is an obvious difference in slopes above and below this threshold. The estimators proposed in this paper might then be used to identify the impact of retirement on outcomes like consumption or time use, based just on the knowledge that the propensity to retire has either a jump or a kink at 65.

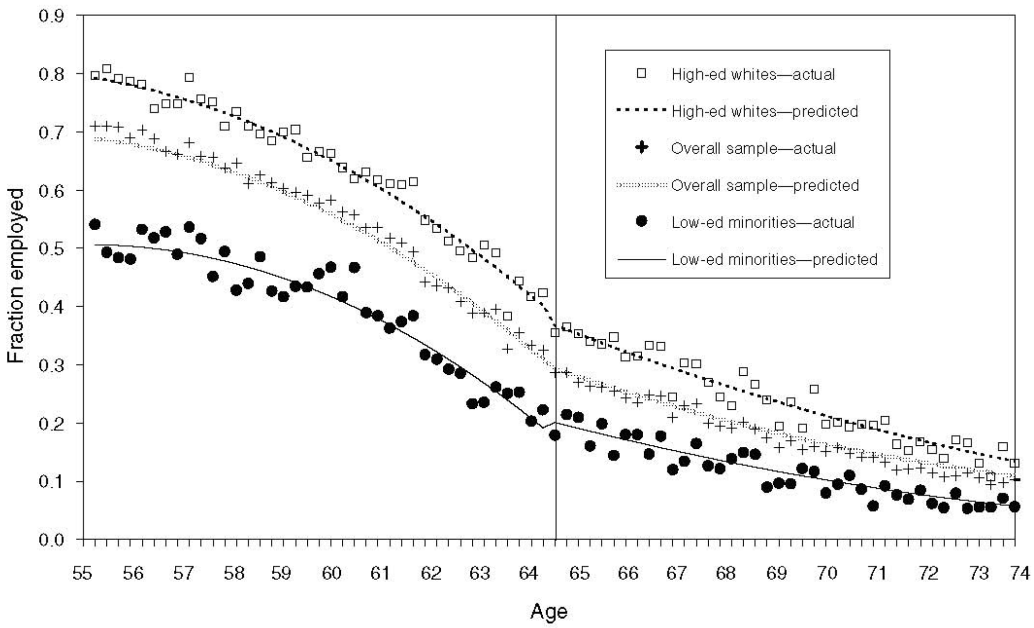

Figure 2: Employment Rates by Age and Demographic Group (1992–2003 NHIS)

For simplicity, this paper will mostly not deal with covariates other than the running variable X in the analysis. The standard RD argument applies that covariates are generally not needed for consistency, though they may be useful for efficiency or for testing continuity assumptions. However, if desired, additional covariates Z could be included in the analysis by letting all the assumptions hold conditional upon the values Z may take on. In applications, one could either partial out these covariates prior to analysis, or include them in the models as additional regressors.

The rest of the paper is organized as follows. Section 2 provides the main identification results. Section 3 gives an instrumental variables interpretation of the identification. Section 4 discusses some extensions, including possible identification based on higher order derivatives. Section 5 provides associated estimators, and Section 6 concludes.

2 RD Treatment Effects without A Discontinuity

I will use Rubin's (1974) potential outcome notation. Let $Y(1)$ and $Y(0)$ denote an individual's potential outcomes from being treated or not, respectively. The observed outcome can then be written as $Y = Y(1)T + Y(0)(1-T)$. As in the introduction, define $g(x) = E(Y \mid X = x)$ and $f(x) = E(T \mid X = x)$, so $g(x)$ and $f(x)$ are the expected outcome and expected probability of treatment when the running or forcing variable $X = x$. In the standard RD model one would expect both $f(x)$ and $g(x)$ have a jump (discontinuity) at $x = c$.

Let T^* be a dummy for crossing the fixed threshold c, i.e., $T^* = I(X \geq c)$, so T^* is one for individuals who have X above the threshold and zero otherwise. An individual is defined to be a complier if he has $T = T^*$. Let $D^* = I(T = T^*)$ be a binary indicator for compliers, i.e., those whose treatment is determined or induced just by crossing the fixed threshold. $E(D^* \mid X = x)$ then equals the compliance rate (percentage of compliers) at $X = x$, i.e., the change in the treatment probability at $X = x$. We do not observe D^* and so do not know who are compliers.

The standard RD model requires $E(D^* \mid X = c) \neq 0$, i.e., a discontinuity at c. The sharp design RD model is the special case where $E(D^* \mid X = c) = 1$ so everyone is a complier.

ASSUMPTION A1: Assume that for each unit (individual) i we observe Y_i, T_i, and X_i. The threshold c is a known constant. The conditional means $E(Y(t) \mid X = x, D^* = d)$ for $t = 0, 1$, and $d = 0, 1$, as well as $E(T \mid X = x, D^* = 0)$ and $E(D^* \mid X = x)$ are continuously differentiable for all x in a neighborhood of $x = c$.

For ease of notation I will drop the i subscript when referring to the random variables Y, T, and X.

It follows from Assumption 1 that $E(Y(t) \mid X = x)$ for $t = 0, 1$ are continuously differentiable in the neighborhood of $x = c$. Assumption A1 differs from the standard RD assumptions in requiring more smoothness. Specifically, standard RD models require only continuity of the conditional mean potential outcomes in x rather than continuous differentiability for identification. This paper requires this additional smoothness to rule out not only jumps but also kinks (formally defined below) caused by factors other than changes in the treatment probability at the threshold $x = c$. In practice, estimators of standard RD models generally impose at least as much smoothness as Assumption A1. For example, asymptotic theories of kernel or local linear regressions require differentiability.

Assumption A1 also imposes smoothness on the conditional mean potential outcomes separately for compliers ($D^* = 1$) and non-compliers ($D^* = 0$). One could instead impose smoothness without conditioning on D^* by having either a constant treatment effect or a local conditional independence of treatment assumption (i.e., conditional on the running variable $X = x$, potential outcomes are independent of treatment in a neighborhood of $x = c$) as in Hahn, Todd, and van der Klaauw (2001). Assumption A1 also rules out a positive probability of deniers (individuals who have $T = 1 - T^*$) by assuming smoothness of $E(T \mid X = x, D^* = 0)$. If one is willing to assume constant treatment effects, then Assumption A1 could be extended to allowing for deniers as follows. Let d^* be a binary indicator for deniers, so $d^* = I(T = 1 - T^*)$. Then one can replace smoothness of $E(Y(t) \mid X = x, D^* = d)$ in Assumption A1 with smoothness of $E(Y(t) \mid X = x)$ for $d = 0, 1$ and $t = 0, 1$, replace $E(D^* \mid X = x)$ with $E(D^* - d^* \mid X = x)$, and replace $E(T \mid X = x, D^* = 0)$ with $E(T \mid X = x, D^* = d^* = 0)$.

Results in this paper require one-sided limits and one-sided derivatives. For any function $h(x)$, define (when they exist) $h_+(x)$ and $h_-(x)$ as the right-sided and left-sided limits, and define $h'_+(x)$ and $h'_-(x)$ as the right-sided and left-sided derivatives, respectively. Also let $h'(x) = \partial h(x) / \partial x$. A standard result is that if $h(x)$ is differentiable, then $h'_+(x)$, $h'_-(x)$, and $h'(x)$ exist and $h'_+(x) = h'_-(x) = h'(x)$. With these notations, a discontinuity at $x = c$ means $f_+(c) - f_-(c) \neq 0$, and the treatment effect that is estimated by standard RD models can be written as $(g_+(c) - g_-(c)) / (f_+(c) - f_-(c))$.

LEMMA 1: If Assumption A1 holds then

$$g_+(c) - g_-(c) = \tau(c) E\left(D^* \mid X = c\right) \tag{1}$$

and

$$f_+(c) - f_-(c) = E\left(D^* \mid X = c\right), \tag{2}$$

where

$$\tau(c) = E\left(Y(1) - Y(0) \mid X = c, D^* = 1\right). \tag{3}$$

Proofs are in the Appendix. Lemma 1 shows that Assumption A1 suffices to reproduce the standard result in the RD literature. In particular, it follows immediately from Lemma 1 that if there is a discontinuity, meaning that $f_+(c) - f_-(c) \neq 0$ then

$$\tau(c) = \frac{g_+(c) - g_-(c)}{f_+(c) - f_-(c)}. \tag{4}$$

That is, the standard RD treatment effect estimator estimates $\tau(c) = E(Y(1) - Y(0) \mid X = c, D^* = 1)$, the average treatment effect for the compliers ($D^* = 1$) at the threshold c, as discussed in, e.g., Hahn, Todd, and van der Klaauw (2001) and Imbens and Lemieux (2008).

If one alternatively assumes constant treatment effects and allows for deniers, then the above conclusion would hold by replacing equation (2) with $f_+(c) - f_-(c) = E(D^* - d^* \mid X = c)$ and replacing equation (3) with $\tau(c) = E(Y(1) - Y(0) \mid X = c)$.

I now consider identifying the RD model treatment effect under alternative assumptions. In particular, I consider: What if there is no jump in the treatment probability? Can we still identify the RD model treatment effect when there is no discontinuity? Formally define a jump and a kink as follows.

DEFINITION: At the point x, a jump in the function $f(x)$ (or simply a jump) is defined as $f_+(x) - f_-(x) \neq 0$ and a kink in the function $f(x)$ (or simply a kink) is defined as $f'_+(x) - f'_-(x) \neq 0$.

THEOREM 1: Let Assumption A1 hold. Assume there is a kink but no jump at $x = c$. Then

$$\tau(c) = \frac{g'_+(c) - g'_-(c)}{f'_+(c) - f'_-(c)}. \tag{5}$$

First note that Assumption A1 suffices to guarantee that the one-sided derivatives $g'_+(x)$, $g'_-(x)$, $f'_+(x)$, and $f'_-(x)$ exist at $x = c$. Theorem 1 says that if there is no jump in $f(x)$, then the treatment effect will equal the ratio of the kinks in $g(x)$ and $f(x)$ at $x = c$ instead of the ratio of the jumps. The intuition is that if $f(x)$ does not have a jump, then both the denominator and the numerator of the standard RD estimator given by equation (4) will equal zero as x goes to c. In that case, by L'hopital's rule, that ratio will equal the ratio of derivatives of the numerator and denominator, given that these derivatives exist.

Theorem 1 requires that the slope of the treatment probability changes at the threshold, which provides additional identification. So unlike in the standard RD model where the treatment effect $\tau(c)$ is identified off a jump in the treatment probability, here $\tau(c)$ is identified off a kink.

In practice, just as a jump in the density of X or conditional means of other baseline covariates at the threshold would cast doubt on the validity of the smoothness assumption in standard RD models, a kink in the density of X or conditional means of other baseline covariates at the threshold would cast doubt on the validity of the smoothness assumption in A1 and hence the identified $\tau(c)$ in Theorem 1 may not be interpreted as a causal effect. To address this concern, one can search for any unusual jumps or kinks in covariates using methods suggested by Imbens and Lemieux (2008) and Lee and Lemieux (2010).

In particular, one could draw a histogram of X based on a fixed number of bins at each side of the cutoff. The overall shape of the distribution can provide a sense whether there is an ususal jump or kink at the cutoff. A more formal test of a discontinuity in the density can be found in McCrary (2008). For other covariates, since their conditional means conditional on X need to be continuously differentiable in the neighborhood of c, one could parametrically regress these covariates on T^* and $(X - c)T^*$ or nonparametrically do local linear regressions of these covariates on $(X - c)$ to examine if there is an intercept or slope change in those variables at the threshold. The latter, when using a uniform kernel, corresponds to using a fixed number of bins at each side of $X = c$ and graphing the mean value of each covariate in each bin against the mid-point of those bins.

Combining Lemma 1 with Theorem 1 gives the following Corollary.

ASSUMPTION A2: Assume there is either a jump or a kink (or both) at $x = c$.

COROLLARY 1: Let Assumptions A1 and A2 hold. Assume that the one-sided limits and one-sided derivatives of $f(x)$ and $g(x)$ at $x = c$ are identified from the data. Then $\tau(c)$ is identified.

Given identification, in the following I provide results that are more directly useful for estimation. In each of the remaining theorems and corollaries, estimators are obtained by replacing functions g and f with corresponding estimates $\widehat{g}$ and $\widehat{f}$.

THEOREM 2: Assume A1 and A2. If either there is no jump or if $\tau'(c) = 0$, then

$$\tau(c) = \frac{g_+(c) - g_-(c) + w\left(g'_+(c) - g'_-(c)\right)}{f_+(c) - f_-(c) + w\left(f'_+(c) - f'_-(c)\right)} \tag{6}$$

for any $w \neq -\left(f_+(c) - f_-(c)\right) / \left(f'_+(c) - f'_-(c)\right)$.

Theorem 2 uses a weight w to combine both the standard RD estimator (4) and the new kink based estimator (5). When there is no jump, i.e., $f_+(c) - f_-(c) = 0$, then equation (6) will reduce to equation (5). In practice, if one is sure that there is no jump, then it will generally be preferable to base estimation directly on equation (5) rather than equation (6), because in that case equation (6) will entail estimation of the terms $f_+(c) - f_-(c)$ and $g_+(c) - g_-(c)$, which are known to equal zero if there is no jump.

When there may be both a jump and a kink, Theorem 2 shows that information in both the jump and the kink can be used to estimate the treatment effect $\tau(c)$, given $\tau'(c) = 0$, as in the case where the treatment effect is locally constant. However, note that $\tau'(c) = 0$ is a strictly weaker condition than assuming a locally constant treatment effect, because the latter would imply that all derivatives of $\tau(c)$ were zero, not just the first derivative $\tau'(c)$. I will discuss the interpretation of this restriction in more detail in Section 3, and provide an extension to Theorem 2 in Section 4. This extension will permit $\tau'(c)$ to be non-zero. Note also that when it is not clear whether there is a jump, a kink, or both, the above estimator can be used as long as $\tau'(c) = 0$ holds, which might be appealing empirically.

The weight w could be chosen to maximize efficiency, i.e., choosing the value of w that minimizes the estimated standard error of the corresponding estimate of $\tau(c)$. Later in Section 3 I will provide a two stage least squares estimator that uses weights based on a measure of the relative strength of the two possible sources of identification, the jump and kink.

In addition, if in practice one knows that there is both a jump and a kink, then the results in Theorem 2 can be used to construct a simple test for locally constant treatment effects. Define τ_1 and τ_2 by

$$\begin{aligned} \tau_1 &= \left(g_+(c) - g_-(c)\right) / \left(f_+(c) - f_-(c)\right) \\ \tau_2 &= \left(g'_+(c) - g'_-(c)\right) / \left(f'_+(c) - f'_-(c)\right). \end{aligned}$$

If the treatment effect is locally constant, then $\tau'(c) = 0$. By Theorem 2 one will have both $\tau_1 = \tau_2 = \tau(c)$, so one could test a local constant treatment effect by testing whether the difference between the two corresponding estimates $\widehat{\tau}_1$ and $\widehat{\tau}_2$ is significant. For parametric RD models, this amounts to a simple t test with the test statistic $(\widehat{\tau}_1 - \widehat{\tau}_2) / \sigma_{(\widehat{\tau}_1 - \widehat{\tau}_2)}$, where the denominator is the standard error of the difference $\widehat{\tau}_1 - \widehat{\tau}_2$.

Theorem 2 requires knowing either that there is no jump or that $\tau'(c) = 0$. The following Corollary provides a weighted estimator that requires neither. The disadvantage of this Corollary versus Theorem 2 is that asymptotically Corollary 2 sets $\tau(c)$ equal to τ_1 when there is a jump, regardless whether there is a kink, whereas when $\tau'(c) = 0$ Theorem 2 can exploit information from both the jump and the kink to estimate $\tau(c)$.

COROLLARY 2: Assume A1 and A2 hold. Given any sequence of nonzero weights w_n such that $\lim_{n\to\infty} w_n = 0$, then

$$\tau(c) = \lim_{n\to\infty} \frac{g_+(c) - g_-(c) + w_n \left(g'_+(c) - g'_-(c)\right)}{f_+(c) - f_-(c) + w_n \left(f'_+(c) - f'_-(c)\right)}. \tag{7}$$

The notable feature of Corollary 2 versus Theorem 2 is that it can be applied to construct estimators for $\tau(c)$ that do not require the user to know whether an observed break at $X = c$ is a jump or a kink, or to know if the treatment effect is locally constant or not. Later I will show that the weights in the local linear 2SLS estimator, which is a special case of the proposed estimator here, have this property.

I will discuss estimators based on the above theorems and corollary in more detail later, but for now observe that one could directly construct nonparametric estimators of $g_+(c)$ and $g'_+(c)$ as the intercept and slope of a local linear regression of Y on $X - c$ just using data having $X \geqslant c$. Doing the same with data having $X < c$ will give estimators of $g_-(c)$ and $g'_-(c)$, and replacing Y with T will give estimates of $f_+(c)$, $f'_+(c)$, $f_-(c)$ and $f'_-(c)$. These could then be substituted into equations (6) or (7) to obtain consistent estimates of $\tau(c)$.

3 Instrumental Variable Interpretation

In this section I provide an instrumental variable interpretation for the identification results of the previous section. I will also show how instrumental variable methods can be used to construct simple estimators based on these results.

Suppose that for some constants α, β, and τ one has the outcome model

$$Y = \alpha + \beta(X - c) + \tau T + e \tag{8}$$

for $c - \varepsilon \leq X \leq c + \varepsilon$, where the error e may be correlated with the treatment indicator T. In general, e might also be correlated with X and hence T^* for strictly positive ε, which would violate the ordinary IV assumption. Hahn, Todd, and van der Klaauw (2001) show that the standard fuzzy design RD estimator given by equation (4) is numerically equivalent to the IV estimator of τ in equation (8), using $(X - c)$ and T^* as instruments for any fixed ε, even the IV assumption is violated. Instead, continuity of potential outcomes at the threshold and having the bandwidth $\varepsilon \to 0$ as the sample size $n \to \infty$ establishes the consistency of the standard RD estimator.

One could take the above model as that for a small positive ε neighborhood of $x = c$, X and hence T^* are independent of e as $\varepsilon \to 0$. Then conditional on compliers, those whose treatment is a deterministic function of T^*, treatment is independent of e, i.e., $T \perp e \mid D^* = 1, X = x$ for $c - \varepsilon \leq x \leq c + \varepsilon$ as $\varepsilon \to 0$.[1] This in practice could hold if individuals, in particular compliers, can not precisely manipulate the

[1]Note that in this case T for all individuals would still depend on e, because for noncompliers, i.e., always takers and never takers, their treatment is not a deterministic function of T^*.

running variable X, and hence they will be randomly above or below the threshold. That is, treatment will be randomly assigned among compliers. (See details regarding this assignment mechanism in Lemieux, 2008). For example, let T be a grade retention treatment and X be negative test score, so c is then the negative threshold score and T^* indicates whether one fails the test or not. Y could then be later academic performance, and e could be ability, which in general is correlated with test score X and hence T^*. Individuals may try to be just below the threshold score and hence avoid the treatment; however, depending on whether or not they are lucky on the test day, they will be randomly below or above the threshold, which implies a local independence of X (and hence T^*) from e.

Also, equation (8) can be taken as a parametric functional form that holds for some small fixed ε, which then requires independence of X (and hence T^*) from e for $c - \varepsilon \leq X \leq c + \varepsilon$. In this case, equation (8) assumes a constant local average treatment effect. One could relax the constant local average treatment effect assumption by adding interaction terms like $(X - c)T$ and $(X - c)^2 T$, etc., to the model; however, in this case one would still be imposing some parametric functional form on $\tau(c)$. In the extension section I will consider this type of generalization of equation (8).

Alternatively, one could take equation (8) as one that only holds in the limit as $\varepsilon \to 0$. In this case the model would not be placing any functional restrictions on the function $\tau(c)$, since if the true model contained any interaction terms like $(X - c)T$ and $(X - c)^2 T$ that were omitted from equation (8), those terms would converge to zero as $\varepsilon \to 0$. Similarly, in this case even if misspecification of equation (8) caused e to be correlated with any function of X when $\varepsilon > 0$, that correlation would go to zero as $\varepsilon \to 0$, because in the limit X would no longer be a random variable, but would just converge to the constant c. Note however that e could still be correlated with T in this limit.

Correlation of e with T means that the assignment to treatment could be related to the outcome through factors other than X. For example, if the treatment is grade retention and the decision of who to retain is based both on whether test score X is below a threshold and on teachers' judgments of who would benefit the most from being retained, then that judgement criterion could induce a correlation between T and e.

Now consider $T^* = I(X \geq c)$. If the treatment probability $f(x)$ has a jump at $x = c$, which means it will depend on T^*, then T^* will be a valid instrument for T in equation (8). One could then estimate equation (8) using two stage least squares, with instruments $X - c$ and T^*.

Similar to how a jump in $f(x)$ at $x = c$ implies that T^* is a valid instrument, a kink in $f(x)$ at the threshold implies that the interaction term $(X - c)T^*$ would also be a valid instrument for T. So if there is no jump but a kink in the treatment probability, one would still be able to use this kink, the slope change in the treatment probability, to identify the RD model treatment effect.

To include either T^* or $(X - c)T^*$ or both as possible instruments for T, write the first stage linear projection of a linear two stage least squares regression as

$$T = r + s(X - c) + pT^* + q(X - c)T^* + V \tag{9}$$

for $c - \varepsilon \leq X \leq c + \varepsilon$, where r, p, and q, are the coefficients of this linear projection and the error V is by construction uncorrelated with $(X - c)$, T^*, and $(X - c)T^*$. Substituting equation (9) into equation (8) yields the reduced form Y equation

$$Y = A_1 + A_2(X - c) + BT^* + C(X - c)T^* + U, \tag{10}$$

where $A_1 = \alpha + \tau r$, $A_2 = \beta + \tau s$, $B = p\tau$, $C = q\tau$, and $U = \tau V + e$.

As with equation (8), one may interpret equations (9) and (10) in one of two possible ways. If ε is a fixed constant then these linear regressions are parametric reduced form equations that imply that $f(x)$ and $g(x)$ have the functional forms $f(x) = r + s(X - c) + pT^* + q(X - c)T^*$ and $g(x) = Y = A_1 + A_2(X - c) + BT^* + C(X - c)T^*$, which in turn would imply that

$$f_+(c) - f_-(c) = p, \quad f'_+(c) - f'_-(c) = q, \tag{11}$$

$$g_+(c) - g_-(c) = B, \quad g'_+(c) - g'_-(c) = C. \tag{12}$$

Alternatively, if $\varepsilon \to 0$ then equations (9) and (10) are numerically identical to local linear regressions of T and Y respectively on X, using a uniform kernel, and estimated separately using data where X is above and below the threshold. In this case it will no longer be true that the functional forms for f and g are linear in $(X-c)$, T^*, and $(X-c)T^*$ but equations (11) and (12) will still hold since the coefficients in local linear regressions equal conditional means and derivatives of conditional means regardless of their true functional forms (as long as they are sufficiently smooth).

Let the residuals from linear regressions of Y, T, T^*, and $(X-c)\,T^*$ on a constant and $(X-c)$ be y, t, t^*, and z, respectively. Then the first and second stage regression equations can be rewritten as

$$\begin{aligned} t &= pt^* + qz + \upsilon, \\ y &= \tau t + e, \end{aligned}$$

and the reduced form for y as

$$y = Bt^* + Cz + U.$$

The IV estimator is then

$$\begin{aligned} \tau &= \frac{cov\,(y, pt^* + qz)}{cov\,(t, pt^* + qz)} = \frac{cov\,(Bt^* + Cz, pt^* + qz)}{cov\,(t, pt^* + qz)} \\ &= \frac{var\,(t^*)\,Bp + cov\,(t^*, z)\,(Bq + Cp) + var\,(z)\,Cq}{cov\,(t, t^*)\,p + cov\,(t, z)\,q} \\ &= \frac{\left[var\,(t^*)\,p + cov\,(t^*, z)\,q\right] B + \left[cov\,(t^*, z)\,p + var\,(z)\,q\right] C}{cov\,(t, t^*)\,p + cov\,(t, z)\,q} \\ &= \frac{cov\,(t, t^*)\,B + cov\,(t, z)\,C}{cov\,(t, t^*)\,p + cov\,(t, z)\,q} \end{aligned}$$

which is the same as

$$\tau = \frac{w_1 B + w_2 C}{w_1 p + w_2 q}$$

where the weights are given by $w_1 = cov\,(t, t^*)$ and $w_2 = cov\,(t, z)$, so the relative weight reflects the relative strength of the two IVs, T^* and $(X-c)T^*$. Plugging in B, C, p, and q, gives

$$\tau = \frac{w_1 B + w_2 C}{w_1 p + w_2 q} = \frac{w_1\,(g_+(c) - g_-(c)) + w_2\left(g'_+(c) - g'_-(c)\right)}{w_1\,(f_+(c) - f_-(c)) + w_2\left(f'_+(c) - f'_-(c)\right)}. \tag{13}$$

This shows that, asymptotically, the IV estimator in equation (13) is equivalent to the special case of the estimator in Theorem 2 where $w = w_2/w_1$.

In the above IV estimator, if $q = 0$ and $p \neq 0$, meaning there is a jump, but no kink, then $C = 0$ and $w_2 = 0$, and hence τ equals equation (4), which is the standard fuzzy design RD treatment effect estimator. Identification comes from T^* being a valid instrument for T in this case.

If $p = 0$ and $q \neq 0$, meaning there is no jump, but a kink, then $B = 0$ and $w_1 = 0$, and hence the IV estimator reduces to (5), which is the estimator proposed in Theorem 1. In this case T^* drops out of both the instrument equation (9) and the reduced form Y equation (10) but $(X-c)T^*$ appears in both, providing a valid instrument for T. The resulting estimator for τ, given by equation (5), equals the ratio of the coefficient for $T^*(X-c)$ in the reduced-form Y equation and that in the T equation, which confirms that the slope change of the treatment probability provides identification.

Note that the local linear 2SLS estimator that has a variable bandwidth ($\varepsilon \to 0$ as the sample size $n \to \infty$) has the property specified in Corollary 2. Asymptotically the local 2SLS puts a zero weight on the slope change if there is a discrete jump. As the sample size $n \to \infty$, the bandwidth used in the local regressions shrinks to zero (using observations closer and closer to the threshold), so $X - c$ and hence $(X - c)T^*$ goes to zero, which makes z go to zero. It follows that $w_2 = cov(t, z)$, and hence w_2/w_1 goes to zero. So with the local 2SLS if there is a jump, i.e., $p = f_+(c) - f_-(c) \neq 0$, the 2SLS weight $w_2/w_1 = w_n \to 0$ as $n \to \infty$. Alternatively, if the treatment probability does not have a jump, i.e., $p = f_+(c) - f_-(c) = 0$ and hence $B = g_+(c) - g_-(c) = 0$, then the weights are asymptotically irrelevant, since in that case one has

$$\frac{w_1 B + w_2 C}{w_1 p + w_2 q} = \frac{w_2 C}{w_2 q} = \frac{C}{q} = \frac{g'_+(c) - g'_-(c)}{f'_+(c) - f'_-(c)}.$$

4 Extensions

The previously described estimator in Theorem 2 uses either a jump, or a kink, or both, but asymptotically if there is a jump, then the only case in which the kink information is used is when $\tau'(c) = 0$ in Theorem 2. Having $\tau'(c) = 0$ means that the treatment effect does not vary linearly with X. For example, in the true parametric form, Y cannot be a function of $(X - c)T$.

This section provides an extension of Theorem 2 to allow $\tau'(c) \neq 0$, so the treatment effect can vary with the running variable X, while exploiting information in both a jump and a kink. For example, if the treatment is grade retention, the running variable is test score, and the outcome is later academic performance. Then $\tau'(c) \neq 0$ would mean the effect of repeating a grade on later performance depends on the pre-treatment test score, and in this case one still could use both jump and kink information to estimate the treatment effect.

For convenience of notation, formally define $B(c) = g_+(c) - g_-(c)$, $C(c) = g'_+(c) - g'_-(c)$, $p(c) = f_+(c) - f_-(c)$, and $q(c) = f'_+(c) - f'_-(c)$. Further define $D(c) = g''_+(c) - g''_-(c)$ and $r(c) = f''_+(c) - f''_-(c)$. So $B(c)$, $C(c)$, $D(c)$, $p(c)$, $q(c)$, and $r(c)$ are the intercept (level), slope, and second derivative changes in the outcome functions and the treatment probability, respectively. The proof of Theorem 1 shows that $B'(c) = C(c)$ and $p'(c) = q(c)$. Similarly it follows that $B''(c) = D(c)$ and $p''(c) = r(c)$. Whenever possible, I will drop the argument (c), and simply use B, C, p, q, D, and r, but note that all these parameters are in general functions of c.

THEOREM 3: Assume A1, A2, and further assume that the conditional means specified in A1 are continuously twice differentiable. If either there is no jump or $\tau''(c) = 0$, then

$$\tau(c) = \frac{B + w(2qC - Dp)}{p + w(2q^2 - rp)} \tag{14}$$

for any weights $w \neq -p/(2q^2 - rp)$. Similar estimators can be constructed if the d-th derivative $\tau^{(d)}(c) = 0$, as is the case if the treatment effect is up to a polynomial of degree $d - 1$ in $(X - c)$, for any positive integer d.

The conditional means in Assumption A1 are twice differentiable, which guarantees that all the involved derivatives in B, C, D, p, q, and r exist. They can be estimated by regression coefficients if one does local quadratic regressions using a uniform kernel at each side neighborhood of the cutoff c.

Analogous to Theorem 2, the assumption for Theorem 3 that $\tau''(c) = 0$ will hold if the treatment effect is locally linear or locally constant. However, while a locally linear or constant treatment effect is sufficient for $\tau''(c) = 0$, it is stronger than necessary, because it implies that all derivatives higher than the first are zero, instead of just the second derivative being zero. With the assumption $\tau''(c) = 0$, the estimator does not allow the treatment effect to be quadratic in $(X-c)$. So for example, in the parametric form, Y cannot be a function of $T(X-c)^2$, but can be a function of T or $T(X-c)$ or both.

Similar to the estimator in Theorem 2, when there is no jump, i.e., $p=0$ and $B=0$, then the above estimator reduces to C/q, which is the estimator in Theorem 1. So when one is sure there is no jump, it is more efficient to use the estimator in Theorem 1. Otherwise if one assumes that the treatment effect is locally linear or locally constant, then this estimator works regardless of whether there is a jump, a kink, or both, and exploits the identification information in both when both are present.

Construction of the estimator when $\tau^{(d)}(c)=0$ for some finite d is briefly discussed in the Appendix. In this case, the treatment effect can be an arbitrarily high-order (e.g., up to the $(d-1)$-th order) polynomial of $(X-c)$, as long as the order is finite.

From Theorem 3, one has the following corollary.

COROLLARY 3: Assume A1 and A2 hold. Given any sequence of nonzero weights ω_n such that $\lim_{n\to\infty}\omega_n=0$, then

$$\tau(c)=\frac{B+\omega_n(2qC-Dp)}{p+\omega_n\left(2q^2-rp\right)} \tag{15}$$

Compared with the estimator in Corollary 2, when the treatment effect is locally linear instead of locally constant, the above estimator uses this information, while the estimator in Corollary 2 does not. In particular, for a local linear treatment effect model, given a kink ($q\neq 0$), $(2qC-Dp)/\left(2q^2-rp\right)$ would be a valid estimator for the treatment effect $\tau(c)$ regardless whether there is a jump or not, while C/q is not unless there is no jump ($p=0$).

Note that the above estimator exploits possible higher order derivative changes for identification. For example, in the absence of both a jump and a kink, the above estimator reduces to D/r. Similar to C/q identifying the RD model treatment effect in the absence of a jump, applying L'hopital's rule to C/q gives D/r as a valid estimator when there is neither a jump nor a kink, but a second derivative change. However, a possible disadvantage of using Corollary 3 for estimation instead of Corollary 2 is that Corollary 3 requires estimation of higher order derivatives (second instead of first), which in practice might be very imprecisely estimated.

So far, all the estimators have been discussed without other covariates except for X. Sometimes it may be desirable to include covariates in a RD model, for example, treatment effects could vary with the other covariates. In that case, additional covariates Z could be included by letting all the assumptions hold conditional upon the values Z may take on. The RD treatment effect estimators are then all conditional on the specific value of Z. For estimation, one could directly include Z, possibly interacted with T and X, as additional regressors in the local polynomial or IV regressions. Then by averaging the estimates over all possible values of Z one could get unconditional RD treatment effects. Alternatively, one could partial covariates out by first regressing Y and T on covariates both above and below the threshold, and then use the residuals from those regressions in place of Y and T in estimation.

5 Estimation

In this section I describe how to implement the proposed RD estimators. The estimation methods provided here are not new. All that is new is their application to the Theorems in this paper.

One convenient way to implement the proposed RD estimators is to do local linear or polynomial regressions using a uniform kernel. The coefficients in the local linear or polynomial regressions provide parameters one needs to construct estimates for the proposed estimators. For example, for a constant treatment effect model, one could estimate $g_+(X) = E(Y \mid X, T^* = 1) = B_+ + (X - c)\,C_+$ and $f_+(X) = E(T \mid X, T^* = 1) = p_+ + (X - c)\,q_+$ by ordinary least squares using observations right above the threshold c, and estimate $g_-(X) = E(Y \mid X, T^* = 0) = B_- + (X - c)\,C_-$ and $f_-(X) = E(T \mid X, T^* = 0) = p_- + (X - c)\,q_-$ using observations right below the threshold. Here B, C, p, and q are constant regression coefficients, and the subscripts $+$ and $-$ denote whether they are estimated using data from above or below the threshold. With these estimates the standard RD treatment effect estimator given a jump can be estimated by

$$\widehat{\tau}(c) = \frac{\widehat{B}_+ - \widehat{B}_-}{\widehat{p}_+ - \widehat{p}_-}. \tag{16}$$

This estimator can also be implemented as the estimated coefficient of T using IV estimation, regressing Y on a constant, $X - c$, and T, using $(X - c)$ and T^* as instrumental variables.

The RD treatment effect estimator given a kink but no jump at the threshold c (the estimator in Theorem 1) can be estimated by

$$\widehat{\tau}(c) = \frac{\widehat{C}_+ - \widehat{C}_-}{\widehat{q}_+ - \widehat{q}_-}. \tag{17}$$

Equivalently, one could take $\widehat{\tau}(c)$ to be the estimated coefficient of T in an IV estimation, regressing Y on a constant, $X - c$, and T, using $(X - c)$ and $(X - c)T^*$ as instrumental variables.

The RD treatment effect estimator proposed in Theorem 2 can be implemented as

$$\widehat{\tau}(c) = \frac{\widehat{B}_+ - \widehat{B}_- + \widehat{w}\left(\widehat{C}_+ - \widehat{C}_-\right)}{\widehat{p}_+ - \widehat{p}_- + \widehat{w}\left(\widehat{q}_+ - \widehat{q}_-\right)}. \tag{18}$$

where the weight $\widehat{w}$ can be chosen to minimize the bootstrapped standard error for $\widehat{\tau}(c)$. Alternatively, equation (18) could be estimated by a 2SLS regression of Y on a constant, $X - c$, T, and $(X - c)T$, using as instruments $(X - c)$, T^*, and $(X - c)T^*$. The resulting estimated weights will then be as described in Section 3.

For all the estimators in the above, one could use the Delta method to calculate standard errors. Alternatively, parametric IV estimation provides standard errors directly along with the point estimate of the local average treatment effect.

These estimators can be interpreted as a special case of nonparametric local linear based estimation, using a uniform kernel. The bandwidth might be chosen using cross validation or other methods as described in, e.g., Imbens and Kalyanaraman (2009) or Lee and Lemieux (2010) and references therein. Just as Porter (2003) recommends using local linear estimation to reduce boundary bias in the estimated constant terms of these regressions, it might be advisable to use local quadratic rather than local linear estimation for reducing boundary bias in the derivative estimates.

To apply the estimator proposed in Theorem 3, where the treatment effect is allowed to vary with X, one would need to estimate local quadratic or higher-order polynomial regressions to obtain the second or higher-order derivatives involved in those estimators. Similarly, IV estimation can be implemented using the higher-order interaction terms as additional instruments. Since these extensions are straightforward, I do not explicitly give their formulas here.

6 Conclusions

Regression discontinuity models identify local treatment effects by associating a discrete change (a jump) in the mean outcome with a corresponding jump in the probability of treatment at a fixed threshold value of the running variable. Lack of discontinuity would make the standard RD estimator inapplicable. However, this paper shows that it is possible to identify the RD model treatment effect from a slope change (a kink) rather than, or in addition to, a jump in the probability of treatment. The intuition for identification off a kink in the absence of a jump is based on L'hopital's rule.

I propose extensions of the usual RD estimator that can be used regardless of whether the source of identification is a jump or a kink. This is empirically appealing because in some potential applications of RD models, it is hard to determine whether the probability of treatment actually jumps or just have a kink at the threshold. In these cases, treatment effects based on standard RD estimators would either be weakly identified, if the jump is small, or unidentified if the jump is zero, regardless of how much the slope changes. In contrast, this paper's estimators make use of any changes in either the intercept (a jump) or the slope (a kink) of the treatment probability at a threshold of the running variable.

The identification results in this paper can be intuitively interpreted using IV models. I show that a kink in the treatment probability provides an additional valid instrument that one can use to identify the same RD treatment effect as a jump. Specifically, similar to a jump in the treatment probability at the threshold implying that the binary indicator for crossing the threshold is a valid instrument, a kink at the threshold implies that the interaction term between this binary indicator and the running variable could also be a valid instrument for the treatment. So if there is no jump but a kink in the treatment probability, one would still be able to use this kink, or the slope change in the treatment probability, to identify the RD model treatment effect. I also show that in some cases (e.g., when the treatment effect is locally constant in the neighborhood of the threshold), one can use the information in both the intercept change and the slope change, i.e., both the jump and a kink to estimate the RD treatment effect.

All of the proposed estimators can be computed using just the estimated coefficients from the same local linear or polynomial regressions that are typically used to estimate standard RD models, so no new estimation methods are required. As usual, one can alternatively do IV or 2SLS estimation using observations in the neighborhood of the threshold to obtain not only point estimates of the local average treatment effect but also parametric standard errors, with an added advantage in this paper's context that 2SLS provides the type of weights that some of the proposed estimators require.

Given our results, it would be useful to explore identification and estimation of other treatment related parameters in the presence of kinks instead of jumps, such as the marginal policy effects of Carneiro, Heckman, and Vytlacil (2010) and Heckman (2010).

7 Appendix: Proofs

First note that for any x such that $c-\varepsilon \leq x \leq c+\varepsilon$, given Assumption A1, one has that $E(Y(t)D^* \mid X=x)$, $E(Y(t)(1-D^*) \mid X=x)$, for $t=0,1$, and hence $E(Y(1-D^*) \mid X=x)$ and $E(T(1-D^*) \mid X=x)$ are all continuously differentiable in the neighborhood of $x=c$. The last two equations hold because $Y=Y(0)+(Y(1)-Y(0))T$ and because $E(T \mid X=x, D^*=0)$ and $E(D^* \mid X=x)$ are assumed to be continuously differentiable. The following proof will use these results.

PROOF of LEMMA 1:

Consider the conditional mean of Y in a RD model for a fixed threshold c,

$$\begin{aligned} E(Y \mid X=x) &= E\left(YD^* + Y\left(1-D^*\right) \mid X=x\right) \\ &= E\left(YD^* \mid X=x\right) + E\left(Y\left(1-D^*\right) \mid X=x\right) \\ &= E\left(Y \mid X=x, D^*=1\right) E\left(D^* \mid X=x\right) + E\left(Y\left(1-D^*\right) \mid X=x\right) \end{aligned}$$

For any $x > c$, by the definition of compliers $D^*=I(T=T^*)$, one has

$$E\left(Y \mid X=x, D^*=1\right) = E\left(Y(1) \mid X=x, D^*=1\right). \tag{19}$$

By continuity of $E(Y(1) \mid X=x, D^*=1)$, equation (19) also holds in the limit as $x \downarrow c$. Therefore one has

$$
\begin{aligned}
g_+(c) &= \lim_{x \downarrow c} E(Y \mid X=x) \\
&= E\left(Y(1) \mid X=c, D^*=1\right) E\left(D^* \mid X=c\right) + E\left(Y\left(1-D^*\right) \mid X=c\right),
\end{aligned}
$$

where the first equality follows from the definition of $g_+(c)$. Similarly, for any $x < c$, one has

$$
E\left(Y \mid X=x, D^*=1\right) = E\left(Y(0) \mid X=x, D^*=1\right). \tag{20}
$$

By continuity of $E(Y(0) \mid X=x, D^*=1)$, equation (20) also holds in the limit as $x \uparrow c$. Then one has

$$
\begin{aligned}
g_-(c) &= \lim_{x \uparrow c} E(Y \mid X=x) \\
&= E\left(Y(0) \mid X=c, D^*=1\right) E\left(D^* \mid X=c\right) + E\left(Y\left(1-D^*\right) \mid X=c\right).
\end{aligned}
$$

Further continuity of $E(Y(1-D^*) \mid X=x)$ at $x=c$ implies

$$
g_+(c) - g_-(c) = E\left(Y(1) - Y(0) \mid X=c, D^*=1\right) E\left(D^* \mid X=c\right).
$$

$E(Y(1) - Y(0) \mid X=c, D^*=1)$ is denoted as $\tau(c)$, so one has

$$
g_+(c) - g_-(c) = \tau(c) E\left(D^* \mid X=c\right),
$$

which is equation (1).

Similarly, given $T = TD^* + T(1-D^*)$, one has

$$
\begin{aligned}
E(T \mid X=x) &= E\left(T \mid X=x, D^*=1\right) E\left(D^* \mid X=x\right) + E\left(T\left(1-D^*\right) \mid X=x\right) \\
&= E\left(T^* \mid X=x, D^*=1\right) E\left(D^* \mid X=x\right) + E\left(T\left(1-D^*\right) \mid X=x\right).
\end{aligned}
$$

$E(T^* \mid X=x, D^*=1) = 1$ for all $x \geq c$. Also $E(T^* \mid X=x, D^*=1) = 0$ for all $x < c$, so it must hold in the limit as $x \uparrow c$. By continuity of $E(T(1-D^*) \mid X=x)$ and $E(D^* \mid X=x)$, one has

$$
f_+(c) - f_-(c) = E\left(D^* \mid X=c\right),
$$

which is equation (2).

PROOF of THEOREM 1:

Since Assumption 1 holds for all x in the the neighborhood of $x=c$, similar to the proof of Lemma 1, for some small positive $\varepsilon_1 < \varepsilon$, one has

$$
\begin{aligned}
g_+(c+\varepsilon_1) &= \lim_{x \downarrow c+\varepsilon_1} E(Y \mid X=x) \\
&= E\left(Y(1) \mid X=c+\varepsilon_1, D^*=1\right) E\left(D^* \mid X=c+\varepsilon_1\right) + E\left(Y\left(1-D^*\right) \mid X=c+\varepsilon_1\right)
\end{aligned}
$$

and

$$
\begin{aligned}
g_-(c-\varepsilon_1) &= \lim_{x \uparrow c-\varepsilon_1} E(Y \mid X=x) \\
&= E\left(Y(0) \mid X=c-\varepsilon_1, D^*=1\right) E\left(D^* \mid X=c-\varepsilon_1\right) + E\left(Y\left(1-D^*\right) \mid X=c-\varepsilon_1\right)
\end{aligned}
$$

Given continuous differentiability of $E(Y(t) \mid X = x_0, D^* = 1)$ for $t = 0, 1$ and $E(Y(1 - D^*) \mid X = x_0)$, the one sided derivatives $g'_+(c + \varepsilon_1)$ and $g'_-(c - \varepsilon_1)$ exist and are given by, respectively,

$$g'_+(c+\varepsilon_1) = \frac{\partial (E(Y(1) \mid X = x, D^* = 1) E(D^* \mid X = x))}{\partial x} \mid x = c + \varepsilon_1 + \frac{\partial E(Y(1 - D^*) \mid X = x)}{\partial x} \mid x = c + \varepsilon_1$$

and

$$g'_-(c-\varepsilon_1) = \frac{\partial (E(Y(0) \mid X = x, D^* = 1) E(D^* \mid x))}{\partial x} \mid x = c - \varepsilon_1 + \frac{\partial E(Y(1 - D^*) \mid X = c + \varepsilon_1)}{\partial x} \mid x = c - \varepsilon_1$$

Note that continuous differentiability of the involved conditional means also implies that $g'_+(c+\varepsilon_1)$ and $g'_-(c-\varepsilon_1)$ are continuous. It follows that that $g'_+(c)$ and $g'_-(c)$ exist and are given by $\lim_{\varepsilon_1 \to 0} g'_+(c+\varepsilon_1)$ and $\lim_{\varepsilon_1 \to 0} g'_+(c-\varepsilon_1)$, respectively. Therefore,

$$\begin{aligned} g'_+(c) - g'_-(c) &= \frac{\partial (E(Y(1) - Y(0) \mid X = x, D^* = 1) E(D^* \mid X = x))}{\partial x} \mid x = c \\ &= \tau'(c) E\left(D^* \mid X = c\right) + \tau(c) \frac{\partial E(D^* \mid X = x)}{\partial x} \mid x = c. \end{aligned} \tag{21}$$

This equality holds as $\partial E(Y(1 - D^*) \mid X = x) / \partial x \mid x = c$ cancels out.

Similarly, one has

$$\begin{aligned} f_+(c+\varepsilon_1) &= \lim_{x \downarrow x_{c+\varepsilon_1}} E(T \mid X = x) \\ &= E\left(D^* \mid X = c + \varepsilon_1\right) + E\left(T\left(1 - D^*\right) \mid X = c + \varepsilon_1\right) \end{aligned}$$

and

$$f_-(c-\varepsilon_1) = \lim_{x \uparrow c - \varepsilon_1} E(T \mid X = x) = E\left(T\left(1 - D^*\right) \mid X = c - \varepsilon_1\right).$$

Given the continuous differentiability of $E(T(1 - D^*) \mid X = x)$ and $E(D^* \mid X = x)$, the one sided derivatives $f'_+(c+\varepsilon_1)$ and $f'_-(c-\varepsilon_1)$ exist and are continous in the neighborhood of c. So analogous to the above analysis, one has

$$f'_+(c) - f'_-(c) = \frac{\partial E(D^* \mid X = x)}{\partial x} \mid x = c. \tag{22}$$

This equality holds as $\partial E(T(1 - D^*) \mid X = x) / \partial x \mid x = c$ cancels out.

Given equations (21) and (22) and the assumption $E(D^* \mid X = c) = f_+(c) - f_-(c) = 0$, one has

$$g'_+(c) - g'_-(c) = \tau(c)\left(f'_+(c) - f'_-(c)\right), \tag{23}$$

That is,

$$\tau\left(c\right)=\frac{g_{+}^{\prime}\left(c\right)-g_{-}^{\prime}\left(c\right)}{f_{+}^{\prime}\left(c\right)-f_{-}^{\prime}(c)}.$$

Note that the above essentially applies L'hopital's rule. To see this, let $B\left(x\right)=g_{+}\left(x\right)-g_{-}\left(x\right)$ and $p\left(x\right)=f_{+}\left(x\right)-f_{-}\left(x\right)$. From the above analysis, $B\left(x\right)$ and $p\left(x\right)$ are continuously differentiable at $x=c$. Given $p\left(c\right)=0$, one has $B\left(c\right)=\tau\left(c\right)p\left(c\right)=0$. L'hopital's rule gives

$$\begin{aligned}\tau\left(c\right) &= \frac{\lim_{x\rightarrow c}B\left(x\right)}{\lim_{x\rightarrow c}p\left(x\right)}=\frac{\lim_{x\rightarrow c}B^{\prime}\left(x\right)}{\lim_{x\rightarrow c}p^{\prime}\left(x\right)}\\ &= \frac{B^{\prime}(c)}{p^{\prime}(c)}=\frac{g_{+}^{\prime}\left(c\right)-g_{-}^{\prime}\left(c\right)}{f_{+}^{\prime}\left(c\right)-f_{-}^{\prime}\left(c\right)}.\end{aligned}$$

PROOF of COROLLARY 1:

If there is a jump, i.e, the identified difference $f_{+}(c)-f_{-}(c)$ is nonzero, then $\tau(c)$ is identified by equation (4). Alternatively, if there is no jump ($f_{+}(c)-f_{-}(c)=0$), then by Assumption A2 there must be a kink. So by Theorem 1 $\tau(c)$ is identified by equation (5).

PROOF of THEOREM 2:

For convenience I will continue to use the $B\left(c\right)$, $p\left(c\right)$, $B^{\prime}(c)$, and $p^{\prime}(c)$ as in the proof of Theorem 1. If there is no jump, by Assumption A2, there is a kink. Then Theorem 1 gives

$$\tau\left(c\right)=\frac{B^{\prime}\left(c\right)}{p^{\prime}\left(c\right)}=\frac{B\left(c\right)+wB^{\prime}\left(c\right)}{p\left(c\right)+wp^{\prime}\left(c\right)}$$

where the second equality follows from $p(c)=0$ and hence $B(c)=\tau\left(c\right)p\left(c\right)=0$.

Now consider the case where $\tau^{\prime}\left(c\right)=0$. By equations (21) and (22), if $\tau^{\prime}\left(c\right)=0$ then $B^{\prime}\left(c\right)=\tau\left(c\right)p^{\prime}\left(c\right)$, and in addition it is already showed that $B\left(c\right)=\tau\left(c\right)p\left(c\right)$ with equations (1) and (2). So regardless of whether there is a jump, a kink or both, if $\tau^{\prime}\left(c\right)=0$ then taking a weighted sum of these two equations gives $B\left(c\right)+wB^{\prime}\left(c\right)=\tau\left(c\right)\left(p\left(c\right)+wp^{\prime}\left(c\right)\right)$. Then

$$\tau\left(c\right)=\frac{B\left(c\right)+wB^{\prime}\left(c\right)}{p\left(c\right)+wp^{\prime}\left(c\right)}.$$

The denominator of this equation is nonzero since by Assumption A2 either $p\left(c\right)$ or $p^{\prime}\left(c\right)$ is nonzero.

PROOF of COROLLARY 2:

Suppose first that there is a jump, $f_{+}\left(c\right)-f_{-}\left(c\right)\neq 0$, then

$$\lim_{n\rightarrow\infty}\frac{g_{+}\left(c\right)-g_{-}\left(c\right)+w_{n}\left(g_{+}^{\prime}\left(c\right)-g_{-}^{\prime}\left(c\right)\right)}{f_{+}\left(c\right)-f_{-}\left(c\right)+w_{n}\left(f_{+}^{\prime}\left(c\right)-f_{-}^{\prime}\left(c\right)\right)}=\frac{g_{+}\left(c\right)-g_{-}\left(c\right)}{f_{+}\left(c\right)-f_{-}\left(c\right)}=\tau(c).$$

Alternatively, suppose there is no jump, $f_{+}\left(c\right)-f_{-}\left(c\right)=0$, then $g_{+}\left(c\right)-g_{-}\left(c\right)=\tau(c)\left(f_{+}\left(c\right)-f_{-}\left(c\right)\right)=0$. So

$$\frac{g_{+}\left(c\right)-g_{-}\left(c\right)+w_{n}\left(g_{+}^{\prime}\left(c\right)-g_{-}^{\prime}\left(c\right)\right)}{f_{+}\left(c\right)-f_{-}\left(c\right)+w_{n}\left(f_{+}^{\prime}\left(c\right)-f_{-}^{\prime}\left(c\right)\right)}=\frac{w_{n}\left(g_{+}^{\prime}\left(c\right)-g_{-}^{\prime}\left(c\right)\right)}{w_{n}\left(f_{+}^{\prime}\left(c\right)-f_{-}^{\prime}\left(c\right)\right)}=\tau(c),$$

where the last equality follows from Theorem 1. This equality holds for all n, and so it must also hold in the limit as $n\rightarrow\infty$.

PROOF of THEOREM 3:

Use the notation in the proof of Theorem 1, and rewrite equation (2) as

$$B = \tau(c)\, p$$

Twice differentiability gives

$$B' = \tau'(c)\, p + \tau(c)\, p' \tag{24}$$

$$B'' = \tau''(c)\, p + 2\tau(c)' p' + \tau(c) p'' \tag{25}$$

Recall by notation $B' = C$, $p\prime = q$, $B'' = D$, and $p'' = r$. If there is no jump, then $p = 0$. By A2, there is a kink, so $p' = q \neq 0$. From Theorem 1, one has

$$\tau(c) = \frac{B'}{p'} = \frac{C}{q} = \frac{B + w(2qC - Dp)}{p + w\left(2q^2 - rp\right)}$$

for any $w \neq -p/\left(2q^2 - rp\right)$. The last equality follows from $p = 0$ and hence $B = \tau(c)\, p = 0$.

If $\tau''(c) = 0$, then $p\tau''(c) = 0$. Then regardless if there is a jump, a kink or both, one has

$$\tau(c) = \frac{B + w(2qC - Dp)}{p + w\left(2q^2 - rp\right)}.$$

First, if there is no jump, which, given A2, means there is a kink, then by Theorem 1, $\tau(c) = \frac{C}{q} = \frac{2qC - Dp}{2q^2 - rp} = \frac{B + w(2qC - Dp)}{p + w\left(2q^2 - rp\right)}$, where the last equality follows from $p = 0$ and $B = 0$. Second, if there is a jump, the standard RD estimator applies, so $\tau(c) = \frac{B}{p}$. Also by assumption, $\tau''(c) = 0$ in this case, so solving for $\tau(c)$ from equations (24) and (25) gives $\tau(c) = \frac{2qC - Dp}{2q^2 - rp}$. By the rule of fraction, $\tau(c) = \frac{B + w(2qC - Dp)}{p + w\left(2q^2 - rp\right)}$.

So either there is no jump or $\tau''(c) = 0$, one has

$$\tau(c) = \frac{B + w(2qC - Dp)}{p + w\left(2q^2 - rp\right)}$$

for any $w \neq -p/\left(2q^2 - rp\right)$.

The same procedure can be applied to cases where the d-th derivative $\tau^{(d)}(c) = 0$ for any finite positive integer d. Keep doing derivatives on both sides of equation (4), until the d-th derivative. With the system of d equations and $\tau^{(d)}(c) = 0$, one can back out $\tau(c)$, as the system of equations are recursive in nature.

PROOF of COROLLARY 3:

Similar to Corollary 2, suppose first that there is a jump, $p \neq 0$, then

$$\lim_{n \to \infty} \frac{B + \omega_n (2qC - Dp)}{p + \omega_n \left(2q^2 - rp\right)} = \frac{B}{p} = \tau(c).$$

Alternatively, suppose there is no jump, $p = 0$, and hence $B = \tau(c) p = 0$, so

$$\frac{B + \omega_n (2qC - Dp)}{p + \omega_n \left(2q^2 - rp\right)} = \frac{\omega_n (2qC - Dp)}{\omega_n \left(2q^2 - rp\right)} = \frac{C}{q} = \tau(c),$$

where the last equality follows from Theorem 1. It holds for all n, and so it holds in the limit as $n \to \infty$.

References

[1] Angrist, J. D. and J.-S. Pischke (2008) Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press.

[2] Hahn, J., P. E. Todd, and W. van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica, 69, 201–09.

[3] Card, D., C. Dobkin, and N. Maestas, (2008), "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare," American Economic Review, 98, 2242–2258.

[4] Carneiro, P., J. J. Heckman, and E. Vytlacil, (2010), "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," Econometrica, 78, 377–394.

[5] Heckman, J. J. (2010), "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy," Journal of Economic Literature, 48, 356-398.

[6] Imbens, G. W. and K. Kalyanaraman (2009), "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," NBER working paper number 14726.

[7] Imbens, G. W. and T. Lemieux (2008), "Regression Discontinuity Designs: A Guide to Practice," Journal of Econometrics, 142, 615–35.

[8] Imbens, G. W. and J. M. Wooldridge (2009), "Recent Developments in the Econometrics of Program Evaluation," Journal of Economic Literature 47, 5–86.

[9] Jacob, B. A., and L. Lefgren, (2004) "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," Review of Economics and Statistics, 86, 226–244.

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

[11] Porter, J. R. (2003) "Estimation in the Regression Discontinuity Model," Unpublished Manuscript.

[12] Rubin, D. B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66, 688–701.