Regression Discontinuity Marginal Threshold Treatment Effects

Yingying Dong and Arthur Lewbel

University of California Irvine and Boston College*

First version July 2010, revised November 2011

Abstract

In regression discontinuity models, where the probability of treatment jumps discretely when a running variable crosses a threshold, an average treatment effect can be nonparametrically identified. We show that the derivative of this treatment effect with respect to the threshold is also nonparametrically identified and easily estimated, in both sharp and fuzzy designs. This marginal threshold treatment effect (MTTE) may be used to estimate the impact on treatment effects of small changes in the threshold. We use it to show how raising the age of Medicare eligibility would change the probability of take up of various types of health insurance.

JEL Codes: C21, C25
Keywords: regression discontinuity, sharp design, fuzzy design, treatment effects, program evaluation, threshold, running variable, forcing variable, marginal effects, health insurance, medicare.

*Yingying Dong, Department of Economics, 3151 Social Science Plaza, University of California Irvine, CA 92697-5100, USA. Email: yyd@uciedu. http://yingyingdong.com/ ; Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. (617)-552-3678, lewbel@bc.edu, http://www2.bc.edu/~lewbel/

1 Introduction

Assume we have a standard regression discontinuity model, where T is a treatment indicator, X is a so-called running or forcing variable, c is a threshold for X at which the probability of treatment changes discretely, and Y is some observed outcome that may be affected both by treatment and smoothly by X. The usual goal in these models is to estimate the effect of treatment T on the outcome Y, and the main result in this literature is that under weak conditions this treatment effect can be nonparametrically identified and estimated at the point where $X = c$. In this paper we consider the question, "how would the effect of T on Y (at $X = c$) change if c were changed a little?" We call this effect the "Marginal Threshold Treatment Effect," or MTTE.

To illustrate, consider three examples. Chay and Greenstone (2005) examine impacts of the US Clean Air Act Amendments (CAAA) of 1970, which requires that a county be given the designation of “nonattainment” status if its pollution concentrations exceed a federally determined ceiling. The CAAA imposes stringent polution abatement regulations on “nonattainment” counties. Here T indicates nonattainment status so the treatment consists of stringent pollution regulations, X is the pollution concentration measure, c is the pollution ceiling, and Y is the subsequent pollution reduction or a side effect like housing value increases. In this sharp regression discontinuity design, our MTTE would be used to address questions such as how the effectiveness of the regulations in reducing pollution would change, or how housing prices would be affected, if the pollution ceiling were marginally raised or lowered.

A simpler example is the original Thistlethwaite and Campbell (1960) regression discontinuity paper, where T is receipt of a National Merit Award, X is the test score on the National Merit Award qualifying exam, c is the exam grade required to qualify for the award, and Y is receipt of other college scholarships (and other outcomes). In this application the treatment effect is the increase in college scholarships resulting from receiving the National Merit Award, and our paper’s goal would be to evaluate how the odds of winning college scholarships would change if the National Merit Award standards were raised or lowered.

A fuzzy regression discontinuity design example is Jacob and Lefgren (2004), who consider an application in which many students are mandated to attend summer school if an exam score is below a cutoff. In this case, T is summer school attendance, X is minus test score on the exam, c is minus the required cutoff exam grade and Y is academic performance in higher grades. This design is fuzzy in part because some students obtained waivers that allowed them to avoid summer school despite failing the exam. In this case the treatment effect is the change in higher grade academic performance resulting from summer school attendence, and in this fuzzy design that treatment effect is identified for compliers, that is, the subpopulation who take the treatment when X crosses the threshold c. Our MTTE would then be the change in this treatment effect that results from a marginal change in c.

Thresholds are often set by policy, and knowing the direction and magnitude of changes in effects resulting from a change in threshold can be important in practice. Many policy debates center precisely on these types of questions, e.g., what are the effects on various health and welfare measures of changes in the legal age for drinking, smoking, or mandatory retirement? Or, in the previously described applications, what are the impacts of changing the pollution ceiling or passing test grade cutoffs?

In discussing regression discontinuity methods Hahn, Todd, and van der Klaauw (2001) note that, "A limitation of the approach is that it only identifies treatment effects locally at the point at which the probability of receiving treatment changes discontinuously... It would be of interest, for example, if the policy change being considered is a small change in the program rules, such as lowering or raising the threshold for program entry, in which case we would want to know the effect of treatment for the subpopulation affected by the change." Our MTTE addresses this issue, by showing how the effect of treatment changes given a marginal change in the threshold. Our results may also be taken as an example of a marginal policy analysis of the sort advocated by Carneiro, Heckman, and Vytlacil (2010) and Heckman (2010).

For simplicity, consider first a simple parametric treatment effect model $Y = \alpha + \beta X + \delta T + \gamma TX + e$. Suppose we have a sharp design, so T is one if and only if X exceeds a threshold c. In this parametric model the average treatment effect conditioning on $X = x$ is just $E(Y \mid T = 1, X = x) - E(Y \mid T = 0, X = x) = \delta + \gamma x$. Evaluating this expression at $x = c$ gives the treatment effect evaluated at the threshold, that is, $\delta + \gamma c$. The MTTE in this model is then the derivative of this treatment effect with respect to c, which is just the coefficient γ.

This parametric model delivers our desired slope effect, the MTTE. So would a polynomial functional form, as in chapter 6 of Angrist and Pischke (2008). But is this identification due to functional form, or can the MTTE be nonparametrically identified? In parametric models the treatment effect is identified both at $x = c$ and for values $x \neq c$ (implying identification of the MTTE) only because the functional form allows us to evaluate objects like $E(Y \mid T = 1, X = x < c)$, even though in the data we could never see any observations having both $T = 1$ and $x < c$. One might think that nothing regarding changes in c can be identified nonparametrically, because we only observe treatment at $x = c$ itself.

However, in this paper we show that, given some minimal smoothness assumptions, the effects of marginal changes in c can in fact be nonparametrically identified. We prove identification of the MTTE formally for both the sharp and fuzzy design nonparametric regression discontinuity models, and describe simple estimators for the MTTE in both designs.

Let $Y(t)$ denote the potential outcome as in Rubin (1974), meaning what Y would equal if $T = t$ for $t = 0$ and for $t = 1$, so $Y = Y(1)T + Y(0)(1 - T)$. The advantage of the regression discontinuity design is that under weak conditions it permits nonparametric identification of $\tau(c)$, the conditional average treatment effect (ATE) conditional upon $X = c$, that is, $\tau(c) = E(Y(1) - Y(0) \mid X = c)$. See, e.g., Hahn, Todd, and van der Klaauw (2001) for laying this out as well as Lee and Lemieux (2010), Imbens and Wooldridge (2009) or Imbens and Lemieux (2008) for recent surveys that discuss these conditions.

The practical usefulness of knowing $\tau(c)$, or some variant such as the effect on compliers in a fuzzy design, is a matter of debate (see, e.g. Deaton 2009, Heckman and Urzua 2010, Heckman 2010, and Imbens 2010), but at a minimum such estimands can provide useful guidance for construction of structural models if desired, or may be combined with other information to provide evidence of external validity and hence wider applicability in practice.

The main result in the literature on sharp design regression discontinuity is that, given some mild regularity conditions, $\tau(c)$ is identified as $\lim_{x\downarrow c} E(Y \mid X = x) - \lim_{x\uparrow c} E(Y \mid X = x)$. For fuzzy designs, this expression is divided by a similar difference in conditional expectations of T, corresponding to the change in treatment probabilities at the threshold. The required regularity conditions include continuity of $E(Y(t) \mid X = x)$ at $x = c$. In practice, local linear (or higher order local polynomial) regressions are used for estimation, for technical reasons as discussed by Hahn, Todd, and van der Klaauw (2001) and Porter (2003), and the asymptotic theory for local linear or polynomial estimation (see Fan and Gijbels 1996) requires not just continuity but continuous differentiability. Parametric models likewise consist of specifications like polynomials that are continuously differentiable in X given T. In most regression discontinuity applications it would be difficult to construct a convincing economic argument as to why $E(Y(t) \mid X = x)$ would be continuous as required without also being differentiable, and this is reflected in the fact that empirical applications of regression discontinuity models all use parametric or nonparametric estimators that assume continuous differentiability.

What we show in this paper is that, under the same conditions (continuous differentiability) that are always assumed to estimate $\tau(c)$ in empirical applications, one can also nonparametrically identify the derivative of $\tau(c)$. Formally define this Marginal Threshold Treatment Effect (MTTE) $\tau'(c)$ to be:

$$\tau'(c) = \frac{\partial \tau(c)}{\partial c} = \frac{\partial E(Y(1) - Y(0) \mid X = c)}{\partial c}.$$

In the fuzzy design case we might also condition on an individual being a complier.

The usual intuition underlying regression discontinuity models is that, without untestable functional form assumptions, nothing can be identified about treatment effects at points other than $X = c$, because nothing can be observed about $Y(0)$ at $X > c$ and nothing can be observed about $Y(1)$ at $X < c$. However, we show that just as smoothness permits identification of the treatment effect itself at c, so too will smoothness permit identification of how treatment effects change when c marginally changes.

Essentially, the logic is this. We start with the regression discontinuity based estimate of $\tau(c)$, so to estimate $\tau'(c) = \lim_{\varepsilon\to 0}[\tau(c) - \tau(c-\varepsilon)]/\varepsilon$ we would like in addition an estimate of $\tau(c-\varepsilon)$ for some tiny positive ε. The obstacle to identifying $\tau(c-\varepsilon)$ is that it depends in part on $E(Y(1) \mid X = x)$ for $x = c - \varepsilon < c$ and we do not observe individuals who have both $X < c$ and $Y = Y(1)$. A similar problem would arise if we *tried to estimate* $\tau'(c)$ using $\tau'(c) = \lim_{\varepsilon\to 0}[\tau(c+\varepsilon) - \tau(c)]/\varepsilon$, since in this case we would not observe individuals having both $Y = Y(0)$ and $X > c$. To overcome these obstacles, observe that differentiability of $E(Y(1) \mid X = x)$ implies that this function becomes arbitrarily close to a line in an arbitrarily small neighborhood around c. We can identify this line using data where X is greater than but arbitrarily close to c, and then extrapolate the line an arbitrarily small distance to $X = c - \varepsilon$, thereby identifying $\tau(c-\varepsilon)$, and hence identifying $\tau'(c)$. This line and its extrapolation is approximate, but the approximation error goes to zero as ε goes to zero.

Once we have obtained an estimate of $\tau'(c)$, we can use this MTTE to provide estimates of the approximate effect of small discrete changes in the threshold c, exactly the way that, e.g., price elasticity estimates are used to approximate the effects of small changes in prices such as those arising from a marginal change in a sales tax or value added tax. For example, if the government raised or lowered the ceiling threshold pollution level (at which stringent polution abatement regulations were imposed) a small amount from c to some other level c_{new}, then the Taylor expansion $\tau(c_{new}) \approx \tau(c) + \tau'(c)(c_{new} - c)$ could be used to approximate $\tau(c_{new})$.

As in all reduced form analyses, the policy relevance of our estimand $\tau'(c)$ or $\tau(c_{new})$ will depend on stability assumptions. We are identifying and estimating features of the function $E(Y(1) - Y(0) \mid X)$, and so to interpret $\tau(c_{new})$ as the conditional ATE that would be observed if the threshold were changed to c_{new}, one would need to assume that the function $E(Y(1) - Y(0) \mid X)$ (evaluated in the neighborhood of $X = c$) would not itself change if the threshold changed marginally. This policy invariance assumption (see, e.g., Heckman 2010) should be at least a reasonable approximation in most RD applications, because RD already assumes X cannot be precisely manipulated by individuals to cross the threshold, and because we are only considering marginal changes in c.

A couple of other papers exist that appeal to derivative conditions for identification in RD analyses. Dong (2010) uses changes in the derivative of conditional expectations at the threshold to identify treatment effects in applications where there is a kink (i.e., a change in slope) but no actual discontinuity at the threshold. Perhaps the closest result to ours is a few paragraphs in a survey article by Dinardo and Lee (2011), in which they informally propose using a Taylor expansion at the threshold to identify an average treatment effect on the treated (ATT) parameter. In contrast, we use a similar expansion to estimate a different object, that is, we consider the impact of changing the threshold, and we provide results for both fuzzy and sharp designs.

For simplicity we give assumptions and results first without consideration of covariates other than the running variable X. We later discuss how additional covariates Z could be included in the regressions. In addition, we show how our method can be extended to estimation of higher derivatives of $\tau(c)$. We also provide an empirical illustration of our results, showing how estimates reported by Card, Dobkin, and Maestas (2008) can be used to estimate how the probability of take up of various types of insurance would change if the age of medicare eligibility were marginally raised or lowered.

2 The Marginal Threshold Treatment Effect

We present our results for sharp designs first, and later consider the extension to fuzzy designs.

ASSUMPTION A1: For each unit (individual) i we observe Y_i, T_i, X_i where T_i is a binary treatment indicator, X_i is a running variable, and $Y_i = Y_i(1)\, T_i + Y_i(0)\,(1 - T_i)$ for potential outcomes $Y_i(1)$ and $Y_i(0)$.

For ease of notation we will drop the i subscript when refering to the random variables $Y(1)$, $Y(0)$, Y, T, and X.

ASSUMPTION A2 (sharp design): $T = I(X \geq c)$ for some known constant threshold c. The support of X includes a neighborhood of c. $E(Y(1) \mid X = x)$ and $E(Y(0) \mid X = x)$ are continuously differentiable in x in a neighborhood of $x = c$.

The Rubin (1974) unconfoundedness assumption for treatment estimation that $Y(1), Y(0) \perp T \mid X$ holds trivially given Assumption A2, because T is a deterministic function of X. Continuity of $E(Y(t) \mid X = x)$ for $t = 0, 1$ coupled with discontinuity of $E(T \mid X = x)$ at the point $x = c$ takes the place of the usual common support assumption that, along with unconfoundedness, is used for identifying average treatment effects.

Standard regression discontinuity identification only requires continuity, not differentiability, of

$E(Y(t) \mid X = c)$ as in Assumption A2, and only requires a continuous density for X, not a differentiable density. However, virtually all empirical implementations of regression discontinuity models satisfy these stronger smoothness conditions. In particular, parametric models generally assume polynomials or other differentiable functions for these expectations, while most nonparametric estimators, including local linear regressions, impose continuous differentiability of both regression functions and densities in their list of technical assumptions required by asymptotic theory. It would be difficult to construct an economic argument for why the expected value of potential outcome functions $Y(t)$ with respect to X should be continuous in X but not be smooth enough to satisfy Assumptions A2.

Dong (2010) also exploits differentiability of $Y(1) - Y(0)$, but in that paper the derivative is used to help identify and estimate the average treatment effect itself, under more general conditions than usual for regression discontinuity models (specifically, in fuzzy designs when both the conditional mean of Y and the probability of treatment, as a functions of X, may have a kink or change in slope at c instead of a discontinuous jump).

Define $\tau(c)$ to be the average treatment effect (ATE) conditional upon $X = c$, that is

$$\tau(c) = E(Y(1) - Y(0) \mid X = c)$$

and define

$$g(x) = E(Y \mid X = x).$$

Given Assumptions A1 and A2, the main result in this literature is that $\tau(c)$ is identified by

$$\tau(c) = \lim_{x \downarrow c} g(x) - \lim_{x \uparrow c} g(x). \tag{1}$$

and can be consistently estimated by replacing the conditional expectations $g(x)$ for $x > c$ and for $x < c$ with either nonparametric regressions (assuming X is continuously distributed with a sufficiently smooth density function) or by parametric regression estimators.

It will be convenient later to use the notation $h_+(x) = \lim_{\varepsilon \downarrow 0} h(x + \varepsilon)$ and $h_-(x) = \lim_{\varepsilon \uparrow 0} h(x + \varepsilon)$ for any function h, so we can rewrite equation (1) as

$$\tau(c) = g_+(c) - g_-(c). \tag{2}$$

To show identification of $\tau'(c) = \partial \tau(c) / \partial c$ we require one-sided derivatives. The right and left derivatives of a function $h(x)$ at the point x, which we will denote as $h'_+(x)$ and $h'_-(x)$

respectively, are defined by

$$h'_{+}(x) = \lim_{\varepsilon \downarrow 0} \frac{h(x+\varepsilon) - h(x)}{\varepsilon} \quad \text{and} \quad h'_{-}(x) = \lim_{\varepsilon \uparrow 0} \frac{h(x+\varepsilon) - h(x)}{\varepsilon}$$

A property of right and left derivatives is that if a function $h(x)$ is differentiable at a point x, then $h'_{+}(x) = h'_{-}(x) = \partial h(x) / \partial x = h'(x)$.

THEOREM 1: If Assumptions A1 and A2 hold then

$$\frac{\partial E(Y(0) \mid X = c)}{\partial c} = g'_{-}(c) \quad \text{and} \quad \frac{\partial E(Y(1) \mid X = c)}{\partial c} = g'_{+}(c) \tag{3}$$

so the marginal threshold treatment effect MTTE is given by

$$\tau'(c) = g'_{+}(c) - g'_{-}(c). \tag{4}$$

Proofs are in the appendix. Given identification of the threshold derivatives in Theorem 1, we can use a Taylor expansion to obtain an approximate estimate of the effect of a discrete change in the threshold. For example, an estimate of what the treatment effect $\tau(c_{new})$ would be if the threshold were changed a small amount from c to c_{new} is

$$\tau(c_{new}) \approx \tau(c) + (c_{new} - c)\,\tau'(c). \tag{5}$$

To provide some intuition for Theorem 1, suppose for the moment that potential outcomes were linear in X, so for $t = 0$ and $t = 1$ we would have $Y(t) = a_t + b_t(X - c) + e_t$. Then $\tau(c) = a_1 - a_0$ and $\tau'(c) = b_1 - b_0$, which shows that in a linear model the MTTE is constant and the same for all possible thresholds c. Here b_1 is identified and can be estimated as the coefficient of X in a linear least squares regression of Y on X using observations having

$X > c$ and similarly b_0 is the coefficient of X in a linear regression using observations having $X < c$, so the MTTE $b_1 - b_0$ is easily identified and estimated in this case. In this linear model, we can identify how the treatment effect $\tau(c) = a_1 - a_0$ would change in response to any size change in the threshold, since in this model if the threshold were changed from c to any c_{new}, the treatment effect would change by exactly $(b_1 - b_0)(c_{new} - c)$, so $\tau(c_{new}) = a_1 - a_0 + (b_1 - b_0)(c_{new} - c)$.

Return now to the nonparametric case where all we know about $Y(t)$ is that it is continuously differentiable at c. This smoothness means that $Y(t)$ is approximately linear in the neighborhood of c, and so the above linear model logic applies just using data in the neighborhood of c. This is the logic that underlies local linear regression.

For estimation, one could use either parametric or local polynomial regressions to estimate $g(x)$ separately above and below the threshold using observations having $X > c$ and with observations having $X < c$. These models directly provide consistent estimates of $g'(x)$ for $x > c$ and for $x < c$, and which when evaluated at $x = c$ will equal estimators of $g'_+(c)$ and $g'_-(c)$. Taking the difference between these two derivative estimates then provides a consistent estimator of the threshold effect $\tau'(c)$.

In particular, suppose we estimate $Y = a_1 + b_1(X - c) + e_1$ by linear least squares regression using just observations having $c < X < c + \varepsilon$ for some small positive ε, and estimate $Y = a_0 + b_0(X - c) + e_0$ using just observations having $c - \varepsilon < X < c$. These regressions will be special cases of nonparametric local linear estimators (with a uniform kernel function). Assuming that $\varepsilon \to 0$ as the sample size goes to infinity, the resulting estimated nonparametric average treatment effect and threshold treatment effect will just be given by $\widehat{\tau}(c) = \widehat{a}_1 - \widehat{a}_0$ and $\widehat{\tau}'(c) = \widehat{b}_1 - \widehat{b}_0$.

Even more simply, these two regressions are equivalent to estimating $Y = a_0 + b_0(X - c) + AT + B(X - c)T + e$ by weighted least squares (or ordinary least squares if the variances of e_0 and e_1 are the same), using observations having $c - \varepsilon < X < c + \varepsilon$, in which case the average treatment effect and threshold treatment effect will be given by $\widehat{\tau}(c) = \widehat{A}$ and $\widehat{\tau'}(c) = \widehat{B}$.

Regression discontinuity models are *often estimated with an interaction term* $(X - c)\, T$, and we have shown that the coefficient B of this term corresponds to the MTTE. This term is generally included in RD model estimators as a control to improve precision of the estimated treatment effect $\widehat{\tau}$. Parametrically, inclusion of the interaction term allows for *locally nonconstant treatment effects,* while nonparametrically inclusion of this term corresponds to local linear estimation vs ordinary kernel regression, which reduces biases associated with estimation of τ at the boundary. $\widehat{B}$ is an estimate of how the treatment effect varies with X, but Theorem 1 shows that $\widehat{B}$ is also the response of the treatment effect to a change in the treatment threshold c.

Regression discontinuity models are often estimated with the interaction term $(X - c)\, T^*$, allowing the slopes of the conditional mean function $E(Y \mid X)$ to be different on either side of the threshold. In the sharp RD design, $T = T^*$, and so $(X - c)\, T^*$ is the same as $(X - c)\, T$. We have shown that the coefficient B of this term corresponds to the MTTE. This term is generally included in RD model estimators as a control to improve precision of the estimated treatment effect $\widehat{\tau}$, as nonparametrically inclusion of this term corresponds to local linear estimation vs ordinary kernel regression, which reduces biases associated with estimation of τ at the boundary. Parametrically, inclusion of the interaction term allows for treatment effects to vary with the running variable. $\widehat{B}$ is then an estimate of how the treatment effect varies with X. Under the policy invariance assumption that the function $E(Y(1) - Y(0) \mid X)$ (in the neighborhood of $X = c$) itself does not change if the threshold changed marginally, this is the same as the response of the treatment effect to a change in the treatment threshold c, or the MTTE.

Higher order terms like $(X - c)^2$ and $(X - c)^2\, T$ can be added to the regression without changing the above analysis. Nonparametrically adding these terms will correspond to local quadratic regression, which, as shown by Fan and Gijbels (1996), will generally have smaller asymptotic bias (as a function of the bandwidth) for estimation of slopes than local linear estimation.

3 Sharp Design Extensions: Covariates and Higher Order Derivatives

It may sometimes be desirable to include covariates in RD models, e.g. to assess how treatment effects vary across subpopulations. Let Z denote a vector of covariates, which is added to the list of observables in Assumption A1. Then Theorem 1 still holds replacing $E\left(Y\left(t\right) \mid X=x\right)$ with $E\left(Y\left(t\right) \mid X=x, Z=z\right)$ for $t=0$ and $t=1$ everywhere (including in the proof), which also implies replacing $\tau'\left(c\right)$, $\tau\left(c\right)$, and $g\left(c\right)$ with $\tau'\left(c,z\right)$, $\tau\left(c,z\right)$, and $g\left(c,z\right)$ respectively. In practice functions of Z, possibly interacted with functions of X and T, can just be included as additional regressors in the regression models discussed at the end of the previous section.

Theorem 1 can also be extended to identify and estimate higher order derivatives. For example, if we replace the continuous differentiability in Assumption A2 with the assumption that $E\left(Y\left(t\right) \mid X=x\right)$ for $t=0$ and $t=1$ are continuously twice differentiable for all x in the neighborhood of c, then by twice applying the proof of Theorem 1 we obtain $\partial^2\tau\left(c\right)/\partial^2 c = g''_+\left(c\right) - g''_-\left(c\right)$.

If sufficient data are available to precisely estimate these higher order derivatives in the neighborhood of $x=c$, then these could be used to further refine estimates of the effects of small discrete changes in c, e.g. for c_{new} close to c, a second order Taylor expansion gives

$$\tau\left(c_{new}\right) \approx \tau\left(c\right) + \left(c_{new}-c\right)\tau'\left(c\right) + \frac{\left(c_{new}-c\right)^2}{2}\tau''\left(c\right). \qquad (6)$$

These higher order derivatives would be estimated using local polynomials of degree two or more, corresponding to the inclusions of terms like $(X - c)^2 \, T$ in the RD regressions.

4 Fuzzy Designs

We now extend Theorem 1 to fuzzy designs, analogous to Hahn, Todd, and van der Klaauw (2001). Let T continue to indicate whether one is treated, but now let $T^* = I\,(X \geq c)$, so T would be the same as T^* for all individuals if the design were sharp. An individual is defined to be a complier if he has $T = T^*$. Let $D^* = 1$ if an individual is a complier and zero otherwise, so $D^* = I\,(T = T^*)$. As before define $g\,(x) = E\,(Y \mid X = x)$ and now also define $f\,(x) = E\,(T \mid X = x)$, so $f\,(x)$ is the probability of treatment given $X = x$.

For the fuzzy design we replace Assumption A2 with the following.

ASSUMPTION A3 (Fuzzy design): Assume the threshold c is a known constant. The support of X includes a neighborhood of c. $E\,(D^* \mid X = x) > 0$ for all x in a neighborhood of c. $E\,(Y\,(1) \mid X = x)$, $E\,(Y\,(0) \mid X = x)$, $E\,((1 - D^*)\,Y \mid X = x)$ and $E\,((1 - D^*)\,T \mid X = x)$ are continuously differentiable for all x in a neighborhood of c.

Assumption A3 as stated rules out deniers (also known in the literature as defiers), that is, individuals having $T = 1 - T^*$, because their presence would violate the assumed smoothness of $E\,((1 - D^*)\,T \mid X = x)$. It would be possible to allow for deniers by placing restrictions on the treatment effects. In particular, Assumption A3 permits the local avearge treatment effect to vary with x, but suppose instead the effect of treatment were assumed to be constant across individuals having x in a neighborhood of c. Then, letting $d^* = I\,(T = 1 - T^*)$ be the indicator of deniers, our results will still hold if, in addition to assuming this local constant treatment effect, we also add to Assumption A3 the condition that $E\,(D^* - d^* \mid X = x) \neq 0$, and replace $1 - D^*$ with $(1 - D^*)\,(1 - d^*)$ everywhere it appears in Assumption A3.

In addition to $E\,(Y\,(t) \mid X)$ being smooth in X as in Assumption A1, Assumption A3 also requires that the conditional means of $(1 - D^*)\,Y$ and $(1 - D^*)\,T$ be smooth in X, that is, the outcomes and the treatments of noncompliers do not have jumps or kinks at the threshold $x = c$. This smoothness may be derived from more primitive assumptions, e.g., monotonicity assumptions on treatment T coupled with smoothness of $E\left(Y\left(t\right) \mid X\right)$, or assuming equivalences among expected outcomes across compliers and noncompliers. Constructing alternative primitive conditions that suffice for regression discontinuity estimation is an active area of research (see, e.g., Battistin, Brugiavini, Rettore, and Weber (2009) and Lee and Lemieux (2010) for recent examples) which we will not pursue further here.

The standard fuzzy design treatment effect estimator as in Hahn, Todd, and van der Klaauw (2001) is

$$\widetilde{\tau}\left(c\right) = \frac{g_+\left(c\right) - g_-\left(c\right)}{f_+\left(c\right) - f_-\left(c\right)} \tag{7}$$

which given Assumptions A1 and A3 can be shown to equal the local average treatment effect for compliers defined as

$$\widetilde{\tau}\left(c\right) = E\left(Y\left(1\right) - Y\left(0\right) \mid X = c, D^* = 1\right) \tag{8}$$

as described in Imbens and Lemieux (2008), among others. This known result is consistent with our specific assumptions, the proof of which is provided in the Appendix as Lemma 1.

Now consider estimation of the fuzzy marginal threshold treatment effect $\widetilde{\tau}'\left(c\right)$, defined by

$$\widetilde{\tau}'\left(c\right) = \frac{\partial \widetilde{\tau}\left(c\right)}{\partial c} = \frac{\partial E\left(Y\left(1\right) - Y\left(0\right) \mid X = c, D^* = 1\right)}{\partial c}. \tag{9}$$

THEOREM 2: If Assumptions A1 and A3 hold then the fuzzy marginal threshold treatment effect MTTE is given by

$$\widetilde{\tau}'\left(c\right) = \frac{g'_+\left(c\right) - g'_-\left(c\right) - \left[f'_+\left(c\right) - f'_-\left(c\right)\right]\widetilde{\tau}\left(c\right)}{f_+\left(c\right) - f_-\left(c\right)} \tag{10}$$

As before, the MTTE can be used to approximate the effect of treatment on compliers if the threshold is changed a small amount from c to c_{new}, since by a Taylor expansion

$$\widetilde{\tau}\left(c_{new}\right) \approx \widetilde{\tau}\left(c\right) + \left(c_{new} - c\right) \widetilde{\tau}'\left(c\right). \tag{11}$$

One must be careful in interpreting this change in treatment effects, since $\widetilde{\tau}\left(c\right)$ is the average treatment effect over individuals who are compliers when the threshold is c, while $\widetilde{\tau}\left(c_{new}\right)$ is the average treatment effect over individuals who are compliers at the new threshold c_{new}. So for example, if the eligibility threshold for some social welfare or assistance program were changed from c to c_{new}, individuals who were compliers when the threshold was c might no longer be compliers at c_{new}, and vice versa.

Let $p\left(c\right)$ denote the fraction of the population that are compliers when the threshold equals c. The proofs of Lemma 1 and Theorem 2 show that $p\left(c\right) = f_{+}\left(c\right) - f_{-}\left(c\right)$ (this could also be obtained by applying the sharp design estimator using T in place of Y and T^{*} in place of T) and that $p'\left(c\right) = f'_{+}\left(c\right) - f'_{-}\left(c\right)$, so Equation (10) can be written as the sum of two terms, $\left[g'_{+}\left(c\right) - g'_{-}\left(c\right)\right]/p\left(c\right)$ and $p'\left(c\right)\widetilde{\tau}\left(c\right)/p\left(c\right)$. The first of these terms is essentially the MTTE given the probability of compliance $p\left(c\right)$, while the second term, which is proportional to $p'\left(c\right)$, accounts for the effect on the MTTE of changes in the probability of compliance that occur when c marginally changes.

Applying the Taylor expansion again we can approximate the proportion of the population who would be compliers at a new threshold c_{new} by $p\left(c_{new}\right) \approx p\left(c\right) + \left(c_{new} - c\right) p'\left(c\right)$. So even though the set of compliers can change in unknown ways when the threshold changes, we can approximate both $p\left(c_{new}\right)$, the fraction of the population who would be compliers at the new threshold c_{new}, and the treatment effect $\widetilde{\tau}\left(c_{new}\right)$ on those compliers. For example, in the above social welfare case, given a proposed change in the eligibility threshold, we could approximately estimate both the probability of compliance at the new threshold and the corresponding average treatment effect at the new threshold. The smaller the proposed change in threshold, the better will be the quality of these approximations.

One way to do estimation in this fuzzy design case would be to estimate $Y_i = a_+ + (X_i - c)\, b_+ + e_{+i}$ and $T_i = r_+ + (X_i - c)\, s_+ + u_{+i}$ by ordinary least squares using just observations having $c < X_i < c + \varepsilon$, and estimate $Y_i = a_- + (X_i - c)\, b_- + e_{-i}$ and $T_i = r_- + (X_i - c)\, s_- + u_{-i}$ using just observations having $c - \varepsilon < X_i < c$, where ε is some small positive constant. Here e and u are error terms and a, b, r, and s are constant regression coefficients, with subscripts + and - denoting whether they are estimated using data above or below the threshold, respectively. With these estimates the fuzzy design treatment effect and fuzzy design marginal threshold treatment effect estimators are then given by

$$\widehat{\widehat{\tau}}\,(c) = \frac{\widehat{a}_+ - \widehat{a}_-}{\widehat{r}_+ - \widehat{r}_-} \quad \text{and} \quad \widehat{\widehat{\tau}}'\,(c) = \frac{\widehat{b}_+ - \widehat{b}_- - (\widehat{s}_+ - \widehat{s}_-)\, \widehat{\widehat{\tau}}\,(c)}{\widehat{r}_+ - \widehat{r}_-}. \tag{12}$$

These estimators are equivalent to nonparametric local linear based estimation using a uniform kernel. The next section provides additional estimation results.

5 Instrumental Variables Estimation

Fuzzy design models are often expressed and estimated in the form of instrumental variables models. Here we show the relationship between these IV model coefficients and the MTTE. Consider the model

$$Y_i = \alpha + X_i \beta + T_i \delta + X_i T_i \gamma + e_i \tag{13}$$

for data having $c - \varepsilon \le X_i \le c + \varepsilon$ where it is assumed that

$$E\left(e_i \mid X_i = x, T_i^* = t, D^*,\ c - \varepsilon \le x \le c + \varepsilon\right) = 0 \tag{14}$$

either for some fixed $\varepsilon > 0$, or just in the limit as $\varepsilon \to 0$. It follows from these equations that the coefficients α, β, δ, and γ can be estimated by applying linear instrumental variables (or equivalently two stage least squares) estimation to equation (13), using X_i, T_i^*, and $X_i T_i^*$ as instruments.

There are two ways to interpret this model. If equation (13) is assumed to hold for some constant ε, then this model corresponds to imposing the parametric functional form of equation (13), in which the treatment effect is assumed to be linear in a neighborhood of c.

Alternatively, if $\varepsilon \to 0$ as a sample size $n \to \infty$, then the linear regressions in each stage of the two stage least squares are like local linear estimators (with a uniform kernel) of arbitrary smooth nonparametric specifications of Y and T as functions of X and T^*. In this case equation (14) only needs to hold in the limit as $\varepsilon \to 0$, meaning that there is local randomization of who lies above versus below the threshold c among individuals having X arbitrarily close to c. This will occur if, e.g., individuals do not have perfect control over X, such as in the test score cases where among individuals of identical skill or education levels, there is some random variation in the exact score that each achieves on the test.

It follows from equations (13) and (14) that

$$E\left(Y_i \mid X_i = x, T_i = t, D^* = 1\right) = \alpha + x\beta + \delta t + xt\gamma \quad \text{for } c - \varepsilon \leq x \leq c + \varepsilon$$

(because compliers have $D^* = 1$ and $T_i = T_i^*$) so in particular at $x = c$ the average treatment effect for the compliers is

$$\widetilde{\tau}\left(c\right) = E\left(Y_i \mid X_i = c, T_i = 1, D^* = 1\right) - E\left(Y_i \mid X_i = c, T_i = 0, D^* = 1\right) = \delta + c\gamma$$

and the MTTE in this model is therefore just

$$\widetilde{\tau}'\left(c\right) = \partial\widetilde{\tau}\left(c\right)/\partial c = \gamma\,.$$

The MTTE exactly equals the coefficient of the interaction term $X_i T_i$ in this model.

An equivalent way to write equations (13) that is more convenient empirically is

$$Y_i = \widetilde{\alpha} + (X_i - c)\,\widetilde{\beta} + T_i\widetilde{\delta} + (X_i - c)\,T_i\widetilde{\gamma} + e_i \tag{15}$$

which can be estimated as above using $X_i - c$, T_i^*, and $(X_i - c)\,T_i^*$ as instruments, and has $\widetilde{\tau}\,(c) = \widetilde{\delta}$ and $\widetilde{\tau}'\,(c) = \widetilde{\gamma} = \gamma$. Note here that some of the coefficients of equation (15), though not $\widetilde{\gamma}$, are implicitly functions of c. In particular, $\widetilde{\delta} = \delta + c\gamma$.

As in the sharp design case, higher order terms like $(X_i - c)^2$ and $(X_i - c)^2\,T_i$ (the latter now instrumented by $(X_i - c)\,T_i^{*2}$) can be added to the regression without changing the above analysis, and do so may reduce nonparametric bias in the slope coefficient estimates as in Fan and Gijbels (1996). Other covariates can also easily be added as additional regressors, possibly interacted with T and X. In this case both the average treatment effect and the threshold treatment effect could depend on covariates. Alternatively, with some restrictions on how covariates appear in the model, one could partial covariates out by first regressing Y_i on covariates both above and below the threshold, and then use the residuals from those regressions in place of Y_i in the estimation of treatment and threshold effects. See the estimation section of the Appendix for more details.

6 Empirical Illustration

Card, Dobkin, and Maestas (2008) employ a sharp design regression discontinuity model to evaluate the impact of reaching age 65 on a variety of outcomes relating to health insurance coverage. The almost universal eligibility of medicare coverage at age 65 in the US is assumed to produce the required discontinuity in eligibility status. In this model X is age, c is 65, $T = T^* = I(X \geq c)$, and outcomes Y considered include various types of health insurance. Some people are eligible for and possess medicare coverage before age 65, and not everyone takes up medicare afterwards, so modeling the impact of medicare coverage itself would require a fuzzy design. However, in this application the treatment is sharply defined as reaching the age of near universal eligibility.

Table 1: Treatment and Marginal Threshold Treatment Effects of
Age 65 Universal Medicare Eligibility on Insurance Coverage.

	Medicare	*Any*	*Private*	*2+ Forms*	*Managed*
Percent at age 63-64	12.3	87.9	71.8	10.8	59.4
Age effect	1.5 (0.2)	-0.2 (0.2)	-1.3 (2.8)	1.1 (0.2)	-2.7 (0.4)
Treatment effect τ (65)	59.7 (4.1)	9.5 (0.6)	-2.9 (1.1)	44.1 (2.8)	-28.4 (2.1)
MTTE τ' (65)	3.3 (1.6)	0.8 (0.2)	1.2 (0.5)	2.7 (1.2)	0.8 (0.9)
Approximate τ (66)	63.0	10.3	-1.7	46.8	-27.6

The first three rows of Table 1 reproduce data from Table 1 in Card, Dobkin, and Maestas (2008). The outcomes Y listed across the top of Table 1 are various types of insurance coverage, specifically, Medicare coverage, any insurance, private coverage, two or more types of insurance coverage, and managed care. For each of these outcomes the first row of the table reports the percentage of people possessing that type of coverage at ages 63-64 and the second row gives the coefficient of age in the model, showing the estimated change in percentage covered that results from each year of aging. The third row is the estimated sharp design regression discontinuity treatment effect τ (65), corresponding to the increase in percentage of people possessing insurance coverage that results from crossing the age 65 threshold.

The fourth row of Table 1 is taken from Table 4 in a supplemental online appendix to Card, Dobkin, and Maestas (2008). There these authors report the coefficients of other regressors in their model, including the coefficient of $I(X \geq 65)(X - 65)$ which by our Theorem 1 corresponds in the sharp design to the marginal threshold treatment effect $\tau'(65)$. In the fifth row of the Table we provide estimates of τ (66) based on equation (5), showing how the treatment effect would differ if the age of eligibility were 66 instead of 65. Where available, standard errors are provided in parentheses.

To interpret the results in Table 1, consider the first column on medicare coverage. The standard analysis of these estimates says that each year a person ages increases the chance that he has medicare coverage by 1.5 percentage points. At age 64 the chance of having medicare coverage is 12.3% and crossing the age 65 universal eligibility threshold increases this coverage probability by 59.7 percentage points.

What we have shown is that, in this parametric model with a sharp regression discontinuity design, the estimated MTTE is 3.3, which means that if the threshold age of universal eligibility were raised marginally, say from 65 to 66, then the treatment effect (where treatment is crossing the age of universal eligibility) would increase by 3.3, from 59.7 to 63.0. Similarly, if the threshold age were lowered marginally from 65 to 64, the treatment effect would decrease by 3.3, from 59.7 to 56.4.

Card, Dobkin, and Maestas also reported squared age effects and cross products with treatment, so a second order Taylor expansion refinement would also be possible, though in this application the estimated second order effects are small.

Every MTTE estimate in Table 1 is positive, showing that if the age of medicare eligibility were raised, the impact of the eligibility age on all types of insurance coverage rates would increase. However, as a policy prescription this gain in coverage rates for individuals at the threshold age would have to be weighed against the individuals between age 65 and the new eligibility age who postpone obtaining coverage until they became Medicare eligible.

7 Conclusions

We have proved nonparametric identification of the marginal threshold treatment effect (MTTE), defined as the marginal change in a local treatment effect resulting from a change in the regression discontinuity threshold. We also provided simple estimators of the MTTE, and discussed its usefulness for policy analysis.

One concern regarding our results is that policy changes of interest may be larger than marginal. Given a parametric model for the outcome Y as a function of the threshold c, one could estimate the effect of any size change in c. But the effects of nonmarginal changes in c are then identified only by functional form. Functional restrictions could instead be used to extrapolate the impacts of our nonparametric estimates. For example, if treatment effects are linear, then the approximate formula for evaluating a marginal policy change, $\widetilde{\tau}\left(c_{new}\right) \approx \widetilde{\tau}\left(c\right) + \left(c_{new} - c\right)\widetilde{\tau}'\left(c\right)$, becomes exact and so can be applied to larger changes in c. Similarly, if treatment effects are quadratic then equation (6) becomes exact. These assumptions would still be less restrictive than the requirement that one have a complete, correctly specified parametric model.

As in all reduced form analyses, the policy relevance of the MTTE will depend on its external validity. our MTTE is a feature of the functions $E\left(Y\left(1\right) - Y\left(0\right) \mid X\right)$ or $E\left(Y\left(1\right) - Y\left(0\right) \mid X, D^{*} = 1\right)$, and so to interpret $\tau\left(c_{new}\right)$ as the conditional ATE that would be observed if the threshold were changed to c_{new}, one would need to assume that these functions, (evaluated in the neighborhood of $X = c$), would not themselves change if the threshold changed. This is a policy invariance assumption, as discussed in, e.g., Heckman (2010). We feel this invariance will be at least a reasonable approximation in most RD applications, because RD already assumes X cannot be precisely manipulated by individuals to cross the threshold, and because we are only considering marginal changes in c.

8 Appendix A: Estimation

Here we provide more details regarding parametric and nonparametric threshold treatment effect estimation. The treatment model estimators themselves that we provide here are not new; they are equivalent to estimators summarized in surveys such as Imbens and Wooldridge (2009) and Lee and Lemieux (2010). What is new here is just the application of these estimators to the construction of threshold treatment effect estimators.

For parametric models, assume that for observations i having $X_i \geq c$, so $T_i^* = 1$, the outcome Y_i has the functional form $Y_i = G(X_i, \theta_+) + e_i$ while for $X_i < c$ we have $Y_i = G(X_i, \theta_-) + e_i$, where G is known and $E(e_i \mid X) = 0$. The parameter vectors θ_+ and θ_- can then be estimated by the least squares regression

$$\widehat{\theta}_-, \widehat{\theta}_+ = \arg \min_{\theta_-, \theta_+} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - T_i^* G(X_i, \theta_+) - \left(1 - T_i^*\right) G(X_i, \theta_-) \right]^2 \omega_i \qquad (16)$$

where $\omega_i = 1$ for ordinary least squares, while values $\omega_i \neq 1$ would correspond to weighted least squares, which might be used to increase efficiency if e_i has some heteroskedasticity of known form. In particular, ω_i could vary with T_i^*, which would correspond to doing ordinary least squares separately on data with $T_i^* = 0$ and $T_i^* = 1$. Then in the sharp design $T_i = T_i^*$, $\widehat{\tau}(c) = G\left(c, \widehat{\theta}_+\right) - G\left(c, \widehat{\theta}_-\right)$ and the estimator $\widehat{\tau'}(c)$ is given by the ordinary derivatives

$$\widehat{\tau'}(c) = \frac{\partial G\left(c, \widehat{\theta}_+\right)}{\partial c} - \frac{\partial G\left(c, \widehat{\theta}_-\right)}{\partial c}.$$

Note that in the sharp design $G(x, \theta_+) = E(Y(1) \mid X = x)$ and $G(x, \theta_-) = E(Y(0) \mid X = x)$. Ordinary derivative formulas can be used in $\widehat{\tau'}(c)$ because these potential outcome functions are differentiable at c, and so have left and right derivatives that equal ordinary derivatives at c.

Nonparametric local polynomial estimation is given by these same formulas, with the functions G being low order polynomials, and weights ω_i defined by

$$\omega_i = \frac{1}{h} K \left(\frac{c - X_i}{h} \right) \tag{17}$$

where K is an ordinary kernel function (e.g., a normal or other symmetric probability density function) and h is a bandwidth parameter that goes to zero as n goes to infinity. For example, local linear estimation takes $\theta_+ = (a_+, b_+)$, $G(x, \theta_+) = a_+ - (x - c) b_+$, and similarly for θ_-, so

$$\widehat{a}_-, \widehat{b}_- = \arg \min_{a_-, b_-} \frac{1}{n} \sum_{i=1}^{n} \left[\left(1 - T_i^*\right) \left(Y_i - \left(a_- + \left(X_i - c\right) b_-\right)\right) \right]^2 \frac{1}{h} K \left(\frac{c - X_i}{h} \right), \tag{18}$$

and

$$\widehat{a}_+, \widehat{b}_+ = \arg \min_{a_+, b_+} \frac{1}{n} \sum_{i=1}^{n} \left[T_i^* \left(Y_i - \left(a_+ + \left(X_i - c\right) b_+\right)\right) \right]^2 \frac{1}{h} K \left(\frac{c - X_i}{h} \right). \tag{19}$$

Then, in the sharp design, the estimated average treatment effect and threshold treatment effect are given by

$$\widehat{\tau}(c) = \widehat{a}_+ - \widehat{a}_- \quad \text{and} \quad \widehat{\tau}'(c) = \widehat{b}_+ - \widehat{b}_-$$

Details regarding the use of local polynomial estimators for regression discontinuity estimation are provided in, e.g., Hahn, Todd, and van der Klaauw (2001), Porter (2003), Imbens and Lemeiux (2008). Bandwidth choice is discussed in Ludwig and Miller (2007) and Imbens and Kalyanaraman (2009), among others. As discussed there, local linear or polynomial estimation is preferable in terms of finite sample properties to local constant (i.e., ordinary kernel regression) estimation for estimation of the levels of functions in the neighborhood of a boundary. By the same logic, local quadratic or higher order polynomial regression might be preferable to local linear regression for estimation of derivatives at the boundary, as we require for $\widehat{\tau}'(c)$.

It should be noted that estimation of the threshold treatment effect $\widehat{\tau}'(c)$ is more demanding in terms of data requirements than estimation of the treatment effect itself, $\widehat{\tau}(c)$, since more data in the neighborhood of c is required to accurately estimate a slope than an intercept. In nonparametric estimation, this shows up in the form of slower optimal rates of convergence for estimation of the derivatives of a conditional mean than for estimation of the conditional mean itself.

If the design is fuzzy then define $T_i^* = I(X_i \geq c)$. Analogous to equation (16) let

$$\widehat{\psi}_-, \widehat{\psi}_+ = \arg \min_{\psi_-, \psi_+} \frac{1}{n} \sum_{i=1}^{n} \left[T_i - T_i^* F\left(X_i, \psi_+\right) - \left(1 - T_i^*\right) F\left(X_i, \psi_-\right) \right]^2 \omega_i \tag{20}$$

where $F\left(X_i, \psi_+\right)$ is a model for the conditional mean of T_i (i.e., a propensity score) given $X_i \geq c$ and $F\left(X_i, \psi_-\right)$ is the model for $X_i < c$. Here F either corresponds to parametric models, or is a polynomial when the weights ω_i are given by equation (17) and we have local polynomial estimation of the conditional mean of T_i. We then obtain the fuzzy design estimators

$$\widehat{\widehat{\tau}}(c) = \frac{G\left(c, \widehat{\theta}_+\right) - G\left(c, \widehat{\theta}_-\right)}{F\left(c, \widehat{\psi}_+\right) - F\left(c, \widehat{\psi}_-\right)} \tag{21}$$

and

$$\widehat{\widehat{\tau}'}(c) = \frac{\frac{\partial G\left(c, \widehat{\theta}_+\right)}{\partial c} - \frac{\partial G\left(c, \widehat{\theta}_-\right)}{\partial c} - \left(\frac{\partial F\left(c, \widehat{\psi}_+\right)}{\partial c} - \frac{\partial F\left(c, \widehat{\psi}_-\right)}{\partial c} \right) \widehat{\widehat{\tau}}(c)}{F\left(c, \widehat{\psi}_+\right) - F\left(c, \widehat{\psi}_-\right)}. \tag{22}$$

In the particular example of local linear estimation, we have equations (18) and (19) along with

$$\widehat{r}_-, \widehat{s}_- = \arg \min_{r_-, s_-} \frac{1}{n} \sum_{i=1}^{n} \left[I(X_i < c)(T_i - (r_- - (X_i - c) s_-)) \right]^2 \frac{1}{h} K\left(\frac{c - X_i}{h} \right), \tag{23}$$

and

$$\widehat{r}_+, \widehat{s}_+ = \arg \min_{r_+, s_+} \frac{1}{n} \sum_{i=1}^{n} \left[I(X_i \geq c)(T_i - (r_+ - (X_i - c) s_+)) \right]^2 \frac{1}{h} K\left(\frac{c - X_i}{h} \right). \tag{24}$$

Then the above fuzzy design treatment effect and fuzzy design threshold treatment effect estimators become

$$\widehat{\widetilde{\tau}}\left(c\right) = \frac{\widehat{a}_{+} - \widehat{a}_{-}}{\widehat{r}_{+} - \widehat{r}_{-}} \tag{25}$$

and

$$\widehat{\widetilde{\tau}}'\left(c\right) = \frac{\widehat{b}_{+} - \widehat{b}_{-} - \left(\widehat{s}_{+} - \widehat{s}_{-}\right)\widehat{\widetilde{\tau}}\left(c\right)}{\widehat{r}_{+} - \widehat{r}_{-}}. \tag{26}$$

In all of these models, one could straightforwardly add covariates Z_i if desired. For example, parameters like a_+ and a_- could be replaced with linear functions like $Z_i' A_+$ and $Z_i' A_+$. Treatment and threshold effects would then be obtained conditional on $Z = z$ for given values of z, or these conditional effects could be averaged across Z to obtain unconditional average effects.

9 Appendix B: Proofs

PROOF of Theorem 1: Let $h_t\left(x\right) = E\left(Y\left(t\right) \mid X = x\right)$ for $t = 0, 1$. For any $x > c$ we have $h_1\left(x\right) = g\left(x\right)$ so these functions must have the same one sided derivatives $h_{1+}'\left(x\right) = g_+'\left(x\right)$ for any $x > c$. By assumption $h_t\left(x\right)$ is differentiable for x in a neighborhood of c, so $h_{1+}'\left(x\right) = h_1'\left(x\right)$, and continuity of the derivatives $h_1'\left(x\right) = g_+'\left(x\right)$ for all $x > c$ in some neighborhood of c implies that $g_+'\left(c\right) = h_1'\left(c\right)$. The same argument based on $x < c$ shows that $g_-'\left(c\right) = h_0'\left(c\right)$, so equation (3) holds.

LEMMA 1: If Assumptions A1 and A3 hold, then $\widetilde{\tau}\left(c\right)$ given by equation (8) satisfies equation (7).

PROOF of Lemma 1: Define $G_t\left(x\right)$ for $t = 0$ and $t = 1$ by

$$G_t\left(x\right) = E\left(Y\left(t\right) \mid X = x, D^* = 1\right) E\left(D^* \mid X = c\right) + E\left(Y\left(1 - D^*\right) \mid X = c\right)$$

First consider $g(x) = E(Y \mid X = x)$ in the fuzzy design.

$$\begin{aligned} g(x) &= E\left(YD^* + Y\left(1 - D^*\right) \mid X = x\right) = E\left(YD^* \mid X = x\right) + E\left(Y\left(1 - D^*\right) \mid X = x\right) \\ &= \Sigma_{d=0}^{1} E\left(Yd \mid X = x, D^* = d\right) \Pr\left(D^* = d \mid X = x\right) + E\left(Y\left(1 - D^*\right) \mid X = x\right) \\ &= E\left(Y \mid X = x, D^* = 1\right) E\left(D^* \mid X = x\right) + E\left(Y\left(1 - D^*\right) \mid X = x\right) \end{aligned}$$

so

$$g(x) = G_1(x) \quad \text{for } x > c \quad \text{and} \quad g(x) = G_0(x) \quad \text{for } x < c \tag{27}$$

By Assumption A3, $G_1(x)$ and $G_0(x)$ are continuous for x in the neighborhood of c, and therefore the equalities in equation (27), which hold on open sets of x, extend to the boundary c of those sets, that is,

$$\lim_{x \downarrow c} E(Y \mid X = x) = g_+(c) = G_1(c)$$

and

$$\lim_{x \uparrow c} E(Y \mid X = x) = g_-(c) = G_0(c).$$

Next, the assumed continuity of $E((1 - D^*)Y \mid X = x)$ at $x = c$ then makes $E((1 - D^*)Y \mid X = x)$ be the same whether $x \downarrow c$ or $x \uparrow c$, so

$$g_+(c) - g_-(c) = G_1(c) - G_0(c) = E\left(Y(1) - Y(0) \mid X = c, D^* = 1\right) E\left(D^* \mid X = c\right). \tag{28}$$

Now consider $f(x) = E(T \mid X = x)$. We have

$$\begin{aligned} f(x) &= E\left(TD^* + T\left(1 - D^*\right) \mid X = x\right) = E\left(TD^* \mid X = x\right) + E\left(T\left(1 - D^*\right) \mid X = x\right) \\ &= \Sigma_{d=0}^{1} E\left(Td \mid X = x, D^* = d\right) \Pr\left(D^* = d \mid X = x\right) + E\left(T\left(1 - D^*\right) \mid X = x\right) \\ &= E\left(T^* \mid X = x, D^* = 1\right) E\left(D^* \mid X = x\right) + E\left(T\left(1 - D^*\right) \mid X = x\right) \end{aligned}$$

so

$$f(x) = E\left(D^* \mid X = x\right) + E\left(T\left(1 - D^*\right) \mid X = x\right) \text{ for } x > c \tag{29}$$

and

$$f(x) = E\left(T\left(1 - D^*\right) \mid X = x\right) \text{ for } x < c \tag{30}$$

so by the assumed continuity of $E(D^* \mid X = x)$ and $E(T(1 - D^*) \mid X = x)$ for x in the neighborhood of c,

$$f_+(c) - f_-(c) = \lim_{x \downarrow c} E(T \mid X = x) - \lim_{x \uparrow c} E(T \mid X = x) = E\left(D^* \mid X = c\right). \tag{31}$$

Substituting equations (28) and (31) into equation (7) then yields equation (8).

PROOF of Theorem 2: By equation (27), for $x > c$ we have $g(x) = G_1(x)$. Taking one sided derivatives of both sides for x in a neighborhood of c with $x > c$ gives $g'_+(x) = G'_{1+}(x) = G'_1(x)$, where the second equality holds because $G_1(x)$ is differentiable and so has one sided derivatives equal to ordinary derivatives. It follows from Assumption A3 that the derivative $G'_1(x)$ is continuous for x in the neighborhood of c, and therefore the equality $g'_+(x) = G'_1(x)$ that holds for x in the range $c < x < c + \varepsilon$ for some ε extends to the lower boundary of this interval, making $g'_+(c) = G'_1(c)$. The same logic starting from equation (27) for $x < c$ shows that $g'_-(c) = G'_0(c)$ and therefore

$$\begin{aligned} g'_+(c) - g'_-(c) &= \frac{\partial\left[G_1(c) - G_0(c)\right]}{\partial c} = \frac{\partial\left[E\left(D^* \mid X = x\right)\widetilde{\tau}(c)\right]}{\partial c} \\ &= \frac{\partial E\left(D^* \mid X = c\right)}{\partial c}\widetilde{\tau}(c) + \left[f_+(c) - f_-(c)\right]\frac{\partial \widetilde{\tau}(c)}{\partial c} \end{aligned} \tag{32}$$

where the second equality holds by equations (31) and (7) from Lemma 1, and the last equality is from the derivative product rule and equation (31).

By equation (29) and Assumption A3 $f(x)$ is differentiable in x for $x > c$, so

$$f'_{+}(x) = f'(x) = \frac{\partial E(D^* \mid X = x)}{\partial x} + \frac{\partial E(T(1 - D^*) \mid X = x)}{\partial x} \text{ for } x > c$$

and these derivatives are continuous so this equation for $x > c$ also holds at $x = c$. In the same way by equation (30) and Assumption A3 we get

$$f'_{-}(x) = \frac{\partial E(T(1 - D^*) \mid X = x)}{\partial x} \text{ for } x < c$$

which also holds at $x = c$, and putting together the last two equations at $x = c$ with differentiability of $E(T(1 - D^*) \mid X = x)$ at $x = c$ gives

$$f'_{+}(c) - f'_{-}(c) = \frac{\partial E(D^* \mid X = c)}{\partial c} \tag{33}$$

Substituting equation (33) into equation (32) and solving the result for $\partial\widetilde{\tau}(c)/\partial c$ gives equation (10) which proves the theorem.

References

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

[2] Battistin, E., A. Brugiavini, E. Rettore and G. Weber (2009) "The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach," American Economic Association, 99, 2209-2226.

[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] Chay, K. Y. and M. Greenstone (2005), "Does air quality matter? Evidence from the housing market," Journal of Political Economy, 113, 376–424.

[6] Deaton A., (2009), "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development," NBER Working Paper No. 14690.

[7] Dinardo, J. and D. S. Lee (2011), "Program Evaluation and Research Designs," in Handbook of Labor Economics, Ashenfelter and Card, eds., vol. 4a, Chap. 5, 463-536.

[8] Dong, Y. (2010), "Jumpy or Kinky? Regression Discontinuity Without the Discontinuity," Unpublished Manuscript.

[9] Fan, J. and Gijbels, I. (1996), Local Polynomial Modelling and its Applications. London: Chapman and Hall.

[10] 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.

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

[12] Heckman, J., J. and S. Urzua, (2010), “Comparing IV With Structural Models: What Simple IV Can and Cannot Identify," Journal of Econometrics, 156, 27-37

[13] Imbens, G. W. (2010), "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)" Journal of Economic Literature, 48, 399–423.

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

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

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

[17] 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

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

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

[20] Porter, J. R. (2003) “Estimation in the Regression Discontinuity Model,” Unpublished Manuscript.

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