Jumpy or Kinky? Regression Discontinuity without the Discontinuity

Yingying Dong*
Department of Economics
University of California Irvine

First draft August 2010, revised November 2011

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 the RD model treatment effect without a discontinuity. In particular, identification can come from a slope change (a kink) instead of a discrete level change (a jump) in the treatment probability. The intuition is based on L'hopital's rule. The identification results can also be interpreted using instrumental variables models. Estimators are proposed that can be applied in the presence or absence of a discontinuity, by exploiting either a jump, or a kink, or both. The proposed estimators are applied to investigate the "retirement-consumption puzzle." In particular, I estimate the impact of retirement on household food consumption by exploiting changes in the retirement probability at 62, the early retirement age in the US.

JEL Codes: C21, C25
Keywords: Regression discontinuity, Fuzzy design, Local average treatment effect, Identification, Jump, Kink, Threshold, Retirement, Consumption

*Correspondence: Department of Economics, 3151 Social Science Plaza, University of California Irvine, CA 92697-5100, USA. Email: yyd@uciedu. http://yingyingdong.com/.

The author would like to thank Arthur Lewbel for many very helpful comments and suggestions, and thank Peter Gottschalk and Devlin Hanson for providing the data. Any errors are my own.

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 age or 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 the mean outcome $g(x) = E(Y \mid X = x)$ 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 all other observed and unobserved covariates determining Y and T are continuous at the threshold c, then individuals having X just below the threshold will be comparable to those having X just above, and hence may provide valid counterfactuals. In particular, any difference in their mean outcomes can be attributed to their treatment probability change.

In this paper I show that the RD local average treatment effects that are usually identified by discontinuities can still be identified even if there is no discontinuity or jump, given that there is a kink, i.e., a discrete change in slope in $f(x)$ at $x = c$. I also provide estimators that can be used regardless of whether identification comes from a jump, a kink, or both. Just as the standard RD treatment effect estimator is numerically equivalent to an instrumental variable (IV) estimator (Hahn, Todd, van der Klaauw, 2001), I show that the proposed estimators are numerically equivalent to IV estimators.

This paper's results can be applied in situations where the compliance rate changes less dramatically than required by standard RD models. For example, if the benefits or incentives

for taking up treatment depend on one's distance from the threshold, or an administrator's discretion or incentive to assign treatment depends on one's distance to the threshold, then 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. This would cause slopes to change at the threshold. When parametric models are employed in a RD design in the existing literature, information is implicitly exploited on first derivatives by allowing for different slopes on either side of the discontinuity threshold, so the pure kink case can be taken as an extreme case where the jump at the threshold is essentially zero. [1] 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, 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 $x = c$.

Jacob and Lefgren (2004) examine the effect of remedial education programs, including grade retention, on later academic performance, where the treatment, grade retention, 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).[2] 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.

In some potential applications of RD models, there is debate about whether the probability

[1] In some cases, policy rules could directly generate kinks if the provided benefits (punishment) for taking up (not taking up) the treatment depends on the distance from the cutoff.

[2] Because of the uncertainty regarding both performance and the grading metric in their case, they note particularly that it is unlikely that a student would have the incentive or ability to marginally change her test score near the cutoff. This rules out the possibility that the observed kink is due to test takers' endogenous behavior. Also, test scores in this case are grade equivalents (GEs), which are typically reported up to the tenths place, and are not really continuous. However, the discreteness of the running variable tends to enlarge the discontinuity gap in this case, so using a more refined measure of the test score is not likely to yield a significant jump either.

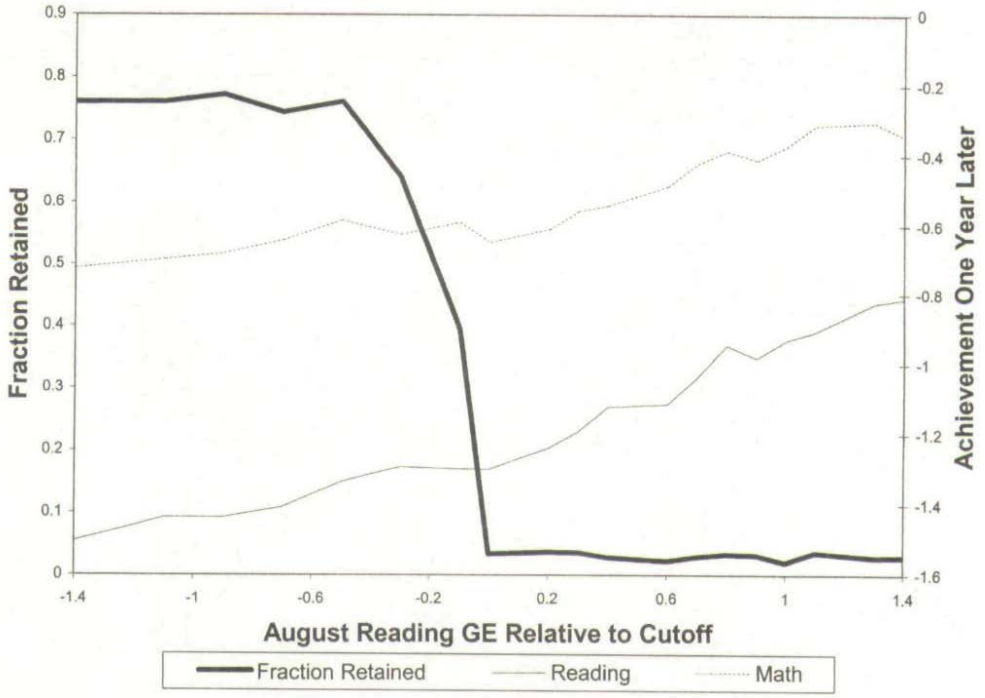

Figure 1: Retention Rate and Reading Test Scores Relative to the Cutoff

of treatment actually jumps at a threshold. When a discrete change is small, it could be indistinguishable from a kink. An example is Figure 2, which reproduces Figure 4 from Card, Dobkin, and Maestas (2008), showing the employment rates in the US by age. It is difficult to determine whether a small jump appears 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 employment on outcomes like heath conditions among the close to retirement age people, based just on the knowledge that the propensity to work has either a jump or a kink at 65. Since age (in quarters) is used as a running variable, individuals are not likely to manipulate their age to sort near the cutoff, and so the observed kink should not be caused by endogenous sorting.

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 in estimating the average (unconditional) treatment effect, though they

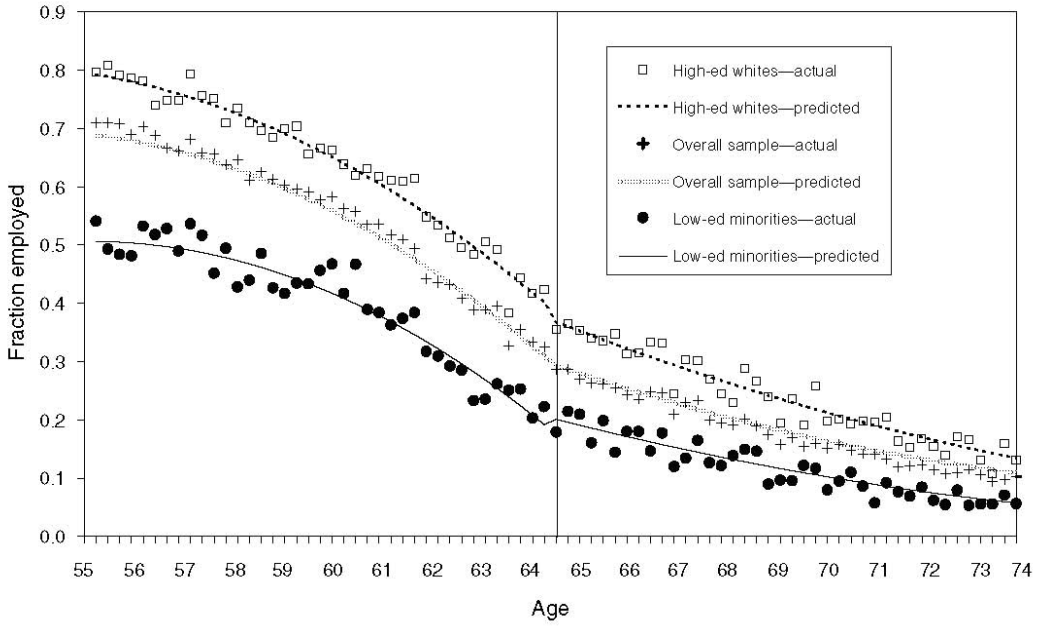

Figure 2: Employment Rates by Age and Demographic Group (1992–2003 NHIS)

may be useful for efficiency or for testing the validity of RD 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.[3] In applications, one could either partial out these covariates prior to analysis, or include them in the models as additional regressors.

I apply the proposed estimators to estimate the impact of retirement on household food consumption at 62, the early retirement age in the US. Graph analyses show that food consumption and retirement may have a jump and/or a kink at 62, so estimators based on either a jump, or a kink, or both are performed. It's shown that using either one or both sharp changes in the retirement probability at 62 yields very similar estimates and that the results are robust to different estimation windows (bandwidths) and weightings (kernel functions). Food consumption is estimated to drop by about 15% to 23% when household heads retire at 62.

The rest of the paper is organized as follows. Section 2 reviews the related literature.

[3]Conditional on Z is necessary if treatment effects vary across covariate values and if one is interested in estimating conditional treatment effects.

Section 3 provides the main identification results. Section 4 gives an instrumental variables interpretation of the identification results. Section 5 discusses some extensions, including possible identification based on higher order derivatives. Section 6 provides associated estimators. Section 7 presents an empirical application, and Section 8 concludes.

2 Literature Review

This section reviews two strands of literature, the standard RD literature and the recent regression kink design literature. This paper is directly built upon the standard RD literature, which is originated in Thistlethwaite and Campbell (1960). Important theoretical research on RD analysis includes Hahn, Todd, and van der Klaauw (2001), Porter (2003), Lee (2008), Lee and Card (2008), McCrary (2008), Imbens and Kalyanaraman (2009). Recent comprehensive reviews of the RD literature include Imbens and Lemieux (2008), van der Klaauw (2008), and Lee and Lemieux (2010).

In a seminal paper, Hahn, Todd, and van der Klaauw (2001) formally show the RD identification in the treatment effect framework and provide assumptions required to identify causal effects. They also propose local linear estimators for nonparametric estimation of the RD treatment effect. Porter (2003) proposes alternative nonparametric estimators and discusses optimal convergence rates. Lee (2008) establishes weak behavioral conditions under which causal inferences from RD analysis can be credible. In particular, Lee (2008) shows that when agents have only imprecise control over the running variable and hence the running variable, along with other covariates, is continuous at the cutoff (due to the random component), RD analysis can still deliver valid inferences. The author proposes to test this assumption by examining whether baseline covariates are continuous at the threshold of the running variable. McCrary (2008) develops a formal density test to test the manipulation of the running variable. Lee and Card (2008) consider the case when the running variable is discrete. They interpret

deviations of the chosen approximating regression function from the true regression function as random specification errors and discuss the impact of this on inference. In particular, they propose how to make the standard errors correct and do possibly more efficient estimation in this case. Imbens and Kalyanaraman (2009) discuss the optimal bandwidth choice for RD models.

A recently developed literature considers a regression kink design. The concept of regression kink design (RKD) is first introduced by Nielsen, Sorensen, and Taber (2009) in their study of financial aid effect on college enrollment. They also propose an associated estimand for their RKD in their application. The estimand takes the form of the ratio of the slope change or kink in the conditional mean of outcome and the kink (the subsidy rate change) in the magnitude of a continuous treatment (the amount of subsidy) as a function of the running variable (family income). Other empirical studies that use kinks to identify effects of continuous endogenous regressors include Guryan (2003) and Simonsen et al. (2009).

Building upon Nielsen, Sorensen, and Taber (2009), the paper by Card, Lee, and Pei (2009) considers nonparametric identification of the average marginal effect of a continuous endogenous treatment variable in a generalized nonseparable model when the treatment of interest is a known, deterministic but kinked function of an observed continuous assignment variable. They characterize a broad class of models for which a RKD provides valid inferences regarding the underlying marginal effects. Under suitable conditions they show that the RKD estimand identifies the "treatment on the treated" parameter.

The models in Nielsen, Sorensen, and Taber (2009) and Card, Lee, and Pei (2009) can be taken as sharp designs in the sense that everyone is a complier and obeys the same treatment assignment rule. The treatment is continuous and is assumed to be a known deterministic function of the running variable. The fundamental identification problem is then to separate the effect of the possibly endogenous running variable from that of treatment in a general nonparametric model, because the latter is completely determined by the former. The identification is

based on the magnitude of the treatment as a kinked function of the running variable.

The goal of this paper, however, is to estimate the same treatment effect parameter of interest as in the standard RD literature, but under more general conditions. In particular, this paper considers a fuzzy design where the treatment is binary, and the functional form for the treatment is unknown. As in the standard RD literature, the estimated effect is a local average effect for those who take up the treatment when crossing a threshold value of the running variable. The identification is based on a kink and/or a jump.

The purely kink-based estimand (Theorem 1) in this paper superficially resembles the RKD estimand as in Nielsen, Sorensen, and Taber (2009) and Card, Lee, and Pei (2009). A key difference is that the RKD estimand depends on the derivative of the treatment variable, which would be infeasible when treatment is binary, while the estimand here depends on the derivative of the expected value of a binary treatment, i.e., the treatment probability. This paper also discusses generalizations that work regardless of whether the treatment probability has a jump, a kink, or both. In addition, this paper shows the identification results (and the proposed estimators) can be intuitively interpreted using IV models. This extends the known result that the standard RD estimator is numerically equivalent to an IV estimator although the IV validity assumption does not hold, as noted by Hahn, Todd, and van der Klaauw (2001).

3 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 is $X = x$. In the standard RD model one would expect both $f(x)$ and $g(x)$ to have a jump (discontinuity)

at the fixed threshold $x = c$.

Let T^* be a dummy for crossing the 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 $T = T^*$ when assigned $X = x$ for all x in some neighborhood of c, so a complier is an individual who takes up treatment if and only if he crosses the threshold. Let D^* be a binary indicator for compliers, i.e., $D^* = 1$ if an individual is a complier and 0 otherwise. $E(D^* \mid X = x)$ then equals the compliance rate among all individuals having $X = x$ for x in a neighborhood of c. We do not observe D^* and so do not know who are compliers. Assumption A1 below and Lemma 1 later make it clear that by conditioning on compliers, one does not have to impose additional conditions like the conditional mean independence or alternative assumptions as imposed by Hahn, Todd, and van der Klaauw (2001, Theorem 2) for identification of the RD model treatment effect in a general setup.

The standard RD model requires $E(D^* \mid X = c) \neq 0$, which would result in $f(x)$ having 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 $E(Y(t) \mid X = x, D^* = 1)$ for $t = 0, 1$, $E(Y \mid X = x, D^* = 0)$, and $E(T \mid X = x, D^* = 0)$, as well as $E(D^* \mid X = x)$, 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.

Assumption A1 says that for compliers the conditional mean potential outcomes $E(Y(t) \mid X = x, D^* = 1)$ for $t = 0, 1$ are smooth, for noncompliers the conditional mean outcome $E(Y \mid X = x, D^* = 0)$ and treatment probability $E(T \mid X = x, D^* = 0)$ are smooth, and the treatment probability change or the compliance rate $E(D^* \mid X = x)$ is also smooth. All this required smoothness is in the sense of continuous differentiability. Intuitively, these

assumptions rule out individuals' sorting behavior, i.e., it is assumed that individuals can not precisely manipulate the running variable to be just above or below the threshold and hence take or avoid the treatment (more discussion on this can be found in Lee 2008). These assumptions also rule out defiers (or defiers, i.e., individuals who have $T = 1 - T^*$ for all x in the neighborhood of c), guaranteeing that any jumps or kinks in the outcome or the treatment probability are due only to compliers. Below provides further discussion.

Assumption A1 differs from the standard RD assumptions in requiring more smoothness. For example, standard RD models require only continuity of the conditional mean potential outcomes for identification rather than continuous differentiability. This paper requires 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, standard asymptotic properties of kernel or local linear regressions require continuous differentiability. Similar continuous differentiability of conditional potential outcomes in the running variable X is also used by Dong and Lewbel (2010) to identify the treatment effect change given a marginal change in the threshold.

Assumption A1 imposes smoothness on the conditional mean potential outcomes particularly for compliers ($D^* = 1$).[4] One could instead impose smoothness without conditioning on $D^* = 1$ by having either a constant treatment effect or a local conditional independence of treatment assumption, i.e., having potential outcomes conditional on $X = x$ be independent of treatment in a neighborhood of $x = c$, as in Hahn, Todd, and van der Klaauw (2001).

For noncompliers ($D^* = 0$), Assumption A1 requires their conditional mean outcome $E(Y \mid X = x, D^* = 0)$ to be smooth at the cutoff, so the observed outcome difference when crossing the threshold are from compliers. By assuming smoothness of $E(T \mid X = x, D^* =$

[4]Intuitively, the smoothness of potential outcomes along with the definition of compliers means that among compliers those just below the cutoff would provide valid counterfactuals for those just above. Note that this still allows for self-selection into the group of compliers.

0), Assumption A1 rules out a positive probability of defiers.

One way to interpret the smoothness of $E(Y \mid X = x, D^* = 0)$ is to assume that there exists a small neighborhood of c where noncompliers consist of always-takers and/or never-takers. Then a sufficient condition for the smoothness of the conditional mean outcome for noncompliers would be continuous differentiability of their conditional mean potential outcomes, i.e., $E(Y(t) \mid X = x, D^* = 0)$ for $t = 0, 1$. This is due to the fact that for both always-takers and never-takers, treatment status does not change and hence is smooth when crossing the threshold, i.e., treatment is always one for always-takers, and zero for never-takers, and that $E(T \mid X = x, D^* = 0)$ and $E(D^* \mid X = x)$ are assumed to be smooth at the cutoff.

Smoothness of $E(T \mid X = x, D^* = 0)$, i.e., no defiers, means that $E(D^* \mid X = x)$ equals the change in the treatment probability at $X = x$. The smoothness of $E(D^* \mid X = x)$ then guarantees that its ordinary derivative exists and that its one-sided derivatives equal the ordinary derivative at $x = c$.

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 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

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

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

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

That is, the standard RD treatment effect estimator estimates $\tau\left(c\right) = E(Y\left(1\right) - Y\left(0\right) \mid X = c,$ $D^* = 1)$, 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).

Note that if one is willing to assume locally constant treatment effects, then Assumption A1 could be extended to allow for defiers as follows. Let d^* be a binary indicator for defiers, so $d^* = 1$ for individuals who have $T = 1 - T^*$ when assigned $X = x$ for all x in the neighborhood of c. Then in addition to assuming smoothness of $E(Y(t) \mid X = x, D^* = 1)$, one also needs to similarly assume smoothness of $E(Y(t) \mid X = x, d^* = 1)$. Furthermore, one needs to replace $E(Y \mid X = x, D^* = 0)$ and $E(T \mid X = x, D^* = 0)$ in Assumption A1 with $E(Y \mid X = x, D^* = d^* = 0)$ and $E(T \mid X = x, D^* = d^* = 0)$, respectively, and replace $E(D^* \mid X = x)$ with $E(D^* - d^* \mid X = x)$. In this case, Lemma 1 would hold by replacing equation (2) with $f_+(c) - f_-(c) = E(D^* - d^* \mid X = c)$ and replacing equation (3) with $\tau\left(c\right) = E\left(Y\left(1\right) - Y\left(0\right) \mid X = c\right)$.

I now consider identifying this 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 reasoning 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 this 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 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 a standard RD model individuals just below the cutoff and those just above are comparable and so their mean outcome difference can be attributed to the treatment probability change. Here, individuals just below the kink point and those just above are also comparable, so one can use the slope change of their mean outcome and the associated slope change of their treatment probability to identify the local average treatment effect at the cutoff.

Just as jumps 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, unusual jumps and kinks 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 in this case, the identified $\tau(c)$ in Theorem 1 would not be interpretable purely as a causal effect. To address this concern, one can easily extend the standard RD validity tests, e.g., tests on the smoothness of the density of the running variable and the smoothness of the conditional means of covariates, to this paper's case.

A formal test of continuity of density can be found in McCrary (2008). More informally, one could draw a histogram of X based on a fixed number of bins on each side of the cutoff. The overall shape of the distribution can provide a sense whether there is an unusual jump or kink in the density of X at the cutoff. Alternatively, on each side of the cutoff one could do a linear regression of the number of observations in each bin on the mid-point value of each bin and examine if there is a significant intercept or slope change.

For other base-line covariates, analogous to the test suggested by Lee (2008) and Lee and Lemieux (2010), one could do a parallel RD analysis by replacing the outcome variable Y with these covariates and examine the significance of the coefficients on T, for which the equation includes both a jump T^* and a kink $(X-c)T^*$. Finding a significant effect of treatment on these pre-determined baseline covariates would suggest unusual jumps or kinks of these covariates at the cutoff. Alternatively, one could do local linear regressions of these covariates at each side of the threshold to examine if there is an intercept or slope change in those variables at the threshold.[5]

Combining Lemma 1 with Theorem 1 gives the following Corollary.

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

[5]The latter, when using a uniform kernel, visually corresponds to using a fixed number of bins on each side of $X = c$ and graphing the mean value of each covariate in each bin against the mid-point of those bins.

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. If there is either a jump, or a kink, or both along with $\tau'(c) = 0$, 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 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. $\tau'(c) = 0$ means that the treatment effect does not vary linearly with the running variable X, as in the case where the treatment effect is locally constant. The proof of Theorem 2 shows that given $\tau'(c) = 0$, both $g_+(c) - g_-(c) / \left(f_+(c) - f_-(c)\right)$ and $g'_+(c) - g'_-(c) / \left(f'_+(c) - f'_-(c)\right)$ can be valid estimands for the treatment effect $\tau(c)$. Intuitively, when the treatment effect varies linearly with X, the kink based estimator would not converge to the local average treatment effect as the jump based estimator does. 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.

Another use of Theorem 2 is to construct a simple test regarding locally constant treatment effects when the treatment probability has both a jump and a kink. Define τ_1 and τ_2 by

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

If the treatment effect is locally constant, then $\tau'(c) = 0$, and by Theorem 2 one will have both $\tau_1 = \tau_2 = \tau(c)$, so one could test $\tau'(c) = 0$ by testing whether the difference between the two corresponding estimates $\widehat{\tau}_1$ and $\widehat{\tau}_2$ is significant. Failing this test indicates that the treatment effect is not locally constant. 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 $\widehat{\tau}_1 - \widehat{\tau}_2$.

The weight ω could be chosen to maximize efficiency, i.e., choosing the value of ω that minimizes the estimated standard error of the corresponding estimate of $\tau(c)$. The following Section 4 provides a two stage least squares estimator (2SLS) that uses weights based on a measure of the relative strength of the two possible sources of identification, the jump and kink.[6]

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 2 versus Theorem 2 is that asymptotically Corollary 2 sets $\tau(c)$ equal to the standard RD estimator 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 $\tau(c)$.

[6]The result is not surprising. It is in fact a generic feature of 2SLS that when there exit more than one instrumental variables, 2SLS uses efficient weights in combining these instrumental variables (see, e.g., Davidson and MacKinnon, 1993).

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

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

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. In Section 4 I will show that the weights in the local 2SLS estimator, a special case of the proposed estimator here, have this property. So Corollary 2 justifies on a formal ground that local 2SLS estimators utilizing both the jump and kink as IV's are valid estimators when one is not sure whether there is a jump, a kink or both and they also do not impose constant treatment effects.

Estimators based on the above theorems and corollary will be discussed in more detail later. 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 $g'_-(c)$. Replacing Y with T in the local linear regressions above and below the threshold 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 $\tau(c)$.

4 Instrumental Variables Interpretation

This section provides an instrumental variables interpretation for the identification results of the previous section. This extends the known result by Hahn, Todd and van der Klaauw (2001) that the standard RD fuzzy design estimator based on discontinuities is numerically

equivalent to an IV estimator. I show that when there is either a jump, or a kink, or both, the RD treatment effect estimators are numerically equivalent to IV estimators. I will also show how instrumental variable methods can be used to construct simple estimators based on these results.

Suppose that for $c-\varepsilon \leq X \leq c+\varepsilon$ for some small positive ε, one has the outcome model

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

where α, β, and τ are coefficients, and 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. 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 given ε, even though the IV zero correlation assumption is violated. Continuity of potential outcomes (essentially continuity of X and e in this case) at the threshold and having the bandwidth $\varepsilon \to 0$ as the sample size $n \to \infty$ establish the consistency of the standard RD estimator.

The above model can be taken as a nonparametric regression function having X and hence T^* become independent of e (i.e., randomly determined) as ε gets arbitrarily close to zero. For compliers, treatment is entirely determined by T^* and so is independent of e (randomly assigned) in the arbitrarily small neighborhood of c, i.e., $T \perp e \mid D^* = 1, X = x$ for $c-\varepsilon \leq x \leq c+\varepsilon$ as $\varepsilon \to 0$. The local randomness of T^* assignment will hold if individuals, in particular compliers who have x close to c, can not precisely manipulate the running variable X, and hence they will be randomly distributed just above versus just below the threshold (see details regarding this assignment mechanism in Lee, 2008).

For example, let T be a grade retention treatment, X be negative test score, and c be the negative threshold score. T^* then indicates whether one fails the test or not. Y could be later academic performance, and one component of e could be ability, which in general is

correlated with test score X and hence T^*. Marginal students may try to be just below the threshold and hence avoid the treatment; however, depending on whether or not they are lucky on the test day, they will score randomly below or above the threshold, which implies a local independence (randomization) of X and hence T^* from e.

Since strictly speaking equation (8) holds only in the limit as $\varepsilon \to 0$, the model does not place any functional restrictions on the function $\tau(c)$. For example, if the true model contains higher order terms like $(X-c)^2$ or any interaction terms like $(X-c)T$ and $(X-c)^2T$, those terms would converge to zero as $\varepsilon \to 0$. Similarly, if the true model has other covariates in it, or the treatment effect in the true model depends on covariates that are omitted, the misspecification of equation (8) may cause e to be correlated with X for $\varepsilon > 0$; however, in the limit as $\varepsilon \to 0$ this correlation will go away, as long as the omitted covariates are smooth at $X = c$. Therefore, τ in the above equation would still consistently identify the local average (unconditional) treatment effect even when there are omitted variables or when there are unknown forms of treatment effect heterogeneity, as long as the omitted component is smooth around the cutoff.

Note that e could still be correlated with T in the limit, due to the existence of noncompliers whose treatment is not determined by T^*. Correlation of e with T means that there could be (self-)selection into treatment based on factors other than X that could affect Y. 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.

If the treatment probability $f(x)$ has a jump at $x = c$, $f(x)$ will be correlated with $T^* = I(X \geq c)$, and then T^* can be a valid instrument for T asymptotically in equation (8). One could then estimate equation (8) using 2SLS, with instruments $X-c$ and T^*.

Similar to how a jump in $f(x)$ at $x = c$ implies that T^* can be used as an instrument,

a kink in $f(x)$ at the threshold implies that the interaction term $(X-c)\,T^*$ could also be an 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 reduced-form treatment as

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

for $c-\varepsilon \leq X \leq c+\varepsilon$ for some arbitrarily small ε, where r, s, p, and q, are the coefficients of this equation.

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 $U = \tau V + e$.

Given equations (9) and (10), one has

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

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

Equations (9) and (10) with $c-\varepsilon \leq X \leq c+\varepsilon$ are numerically identical to local linear regressions of T and Y respectively on X, using a uniform kernel and bandwidth ε. 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), equations (11) and (12) would hold regardless of the true functional forms of Y and T.

Let y, t, t^*, and z be Y, T, T^*, and $(X-c)\,T^*$ after partialling out $(X-c)$, respectively,

i.e., they are the residuals from local linear regressions of Y, T, T^*, and $(X-c)\,T^*$ on a constant and $(X-c)$. 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 in this case 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

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

This shows that, the IV estimator in equation (13) is numerically 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 an 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 an instrument for T. The resulting estimator for τ, given by equation (5), equals the ratio of the coefficients for $T^*(X - c)$ in the reduced-form Y and T equations, which confirms that the slope change of the treatment probability provides identification.

Note that the local 2SLS estimator that has a variable bandwidth $\varepsilon \to 0$ as the sample size $n \to \infty$ has the property specified in Corollary 2, i.e., 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

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

which by Theorem 1 is still equal to the local treatment effect parameter $\tau\,(c)$.

5 Extensions

The previously described estimand 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 $\tau'(c) = 0$. As mentioned, 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 linearly with the running variable X, while still 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 that 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

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

for any weights $w \neq -p/\left(2q^2 - rp\right)$. Similar estimands 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 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 corresponding estimator does not allow the treatment effect to vary quadratically with $(X-c)$, because in this case the estimator using the second derivative changes $(2qC - Dp)/\left(2q^2 - rp\right)$ would not be a valid estimator for the local treatment effect at 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 any 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 considering other covariates except for X. It is worth emphasizing that if one is interested in estimating the average treatment effect, or the unconditional treatment effect, covariates are not necessary for consistency, but may be useful to increase efficiency or for robustness check. If desired, one can directly include covariates in the treatment and outcome equations, or 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.

If in a particular application, one believes that treatment effects vary with other covariates,

and one is interested in estimating the conditional treatment effect, then covariates are necessary. In this case, additional covariates Z can 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, allowing Z to be interacted with T and X, as additional regressors in the local polynomial or IV regressions. Or more generally, one can estimate the treatment effect conditional on a specific value of Z, say z, i.e., estimate $\tau(c|Z=z)$.

6 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 proposed estimators are simple functions of these local linear or polynomial regression coefficients. For example, 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 regressions of Y and T on a constant and $(X-c)$ 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 4.

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., Ludwig and Miller (2007), Imbens and Lemieux (2008), Imbens and Kalyanaraman (2009) or Lee and Lemieux (2010) and references therein. Just as Hahn, Todd, and van der Klaauw (2001) and Porter (2003) recommend using local linear or local polynomial estimation to reduce boundary bias in the estimated constant terms of these

regressions, it might be advisable to use local quadratic or higher-order polynomial 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.

7 Empirical Application

This section applies the results in the previous sections to estimate the effect of retirement on food consumption using changes in the retirement probability at age 62, the early retirement age in the US. The existing literature generally reports a greater change in retirement rates around 62 than at the full retirement age 65 in US, which is confirmed later in this paper's sample.[7]

Many empirical studies document a significant decrease in consumption at retirement. The estimated drops range from about 10% to more than 40% (Ameriks et al. 2007). The finding that consumption drops at retirement is referred to as the "retirement-consumption puzzle," because a systematic fall in consumption is inconsistent with the life cycle/permanent income hypothesis (LCPIH), which holds that rational people smooth consumption over their life-cycle and so consumption should not fall when the future date of retirement is anticipated.

To the extent that retirement can be affected by a negative income or health shock such

[7]Starting from 62, individuals in the US are eligible for social security retirement benefits, which is documented to cause an increased probability of retirement. If one retires earlier than the normal (or full) retirement age (NRA), typically 65 for the sample of individuals used in this paper, their social security benefits will be reduced by a certain percentage for each month they retire earlier than the NRA, but the percentage schedule is claimed to be set so that the expected values of life-time benefits are about the same regardless when one chooses to retire. If this is true, then it may explain why there are no obvious sharp changes in retirement probabilities at age 65.

as a job loss or a disability (so that retirement is endogenous to consumption), the observed consumption fall does not necessarily contradict the LCPIH. This section estimates the size of the drop in household food consumption due to the household head's retirement, exploiting changes in retirement probabilities when workers turn 62 and hence first become eligible for social security retirement benefits. In this case, an RD model essentially compares individuals who just turn 62 with those who are just under, and identifies the retirement effect for individuals who retire because they qualify for social security retirement benefits. Given that the early retirement age 62 is fully anticipated, the estimated effect is then the causal impact of retiring at 62. I examine food consumption because food is a nondurable good and so one would expect immediate changes after retirement if any.

Food consumption here is measured by the total expenditure on food consumed at home, delivered to door, and eaten out per week. Y is then defined as the logarithm of food consumption adjusted for family size and composition using an equivalence scale. I use the equivalence scale that was recommended to the US Census by the National Resource Council's Panel on Poverty and Family Assistance (see Citro and Michaels 1995).[8][9] Since multiple years' data are used, food expenditures are adjusted for inflation. T is the retirement treatment. X is household head's age, and the cutoff c is 62. Retirement T is defined as the household head's self-reported retirement status, which equals one when the household head is retired and zero otherwise. The sample does not include non-labor-force participants, such as students, the disabled, and homemakers.

The data are from the 1994 to 2007 US Panel Study of Income Dynamics (PSID). To

[8]This equivalence scale assigns a value of 1 to each adult and of 0.5 to each child in a household and raises the sum of these assigned values to the power of 0.7.

[9]Instead of dividing food expenditure by family size, I divide it by an equivalence scale to account for economies of scale to consumption (for food, these can take the form of reduced waste and other gains from joint food preparation). In theory, consistency of the RD estimator when the bandwidth shrinks to zero does not depend on the choice of equivalence scale, because the RD model assumes that family size and other related variables are distributed smoothly around the cutoff, as is confirmed in Figure 5 later. An equivalence scale adjustment (especially for observations some distance from the cutoff) is mainly intended to improve efficiency and reduce finite sample bias.

avoid measurement and behavioral issues associated with the use of food stamps, the sample is restricted to households who did not use food stamps to buy food. Detailed information on food expenditures and food stamp usage is available on a consistent basis in the PSID starting from 1994. There are ten waves of data in this sample period (data are collected every two years since 1997). Years 2001 and 2003 are arguably recession years, when individuals may have different retirement and consumption behavior, so the analysis here focuses on the sample excluding these two years of data. However the main results are reported both with and without using these two years of data. The differences between including and excluding these recession years are found to be modest. To reduce the impact of outliers, households in the top 1% of food consumption are trimmed out of the sample. Three windows consisting of 6, 8, or 10 years at each side of the cutoff, age 62, are used for estimation, yielding final sample sizes of 6,278, 8,565, and 11,048 observations, respectively.

Figures 3 and 4 show changes in retirement rates and food consumption at 62, based on a ten-year window at each side of the cutoff. The scatter plots in these figures show sample averages by age. Also shown are fitted quadratic regression lines above and below the cutoff age 62.[10] As one can see, the retirement rate has plausibly a small jump and may also have a small kink (slope change) at 62, whereas the food consumption has a more obvious kink than a jump at the cutoff. In particular, the age profile of food consumption jumps around a relatively flat line before age 62, but then steeply declines afterwards.[11] Since the data appear to be noisier at younger ages, a specification considered later incorporates variation in within cell sampling variances. Note that no particular jumps or kinks are present in the retirement rate at 65.

Based on these figures, in the following I will estimate the retirement effect at 62 exploiting

[10] These scatter plots are equivalent to histograms or bin average graphs that are recommended in the standard RD literature (see, e.g., Imbens and Lemieux 2008 and Lee and Lemieux 2010), since the reported age is in years.

[11] Questions regarding food consumption changed somewhat after 1997, and overall the food consumption data are noisier for early waves of PSID, which may contribute to the observed larger variances at younger ages.

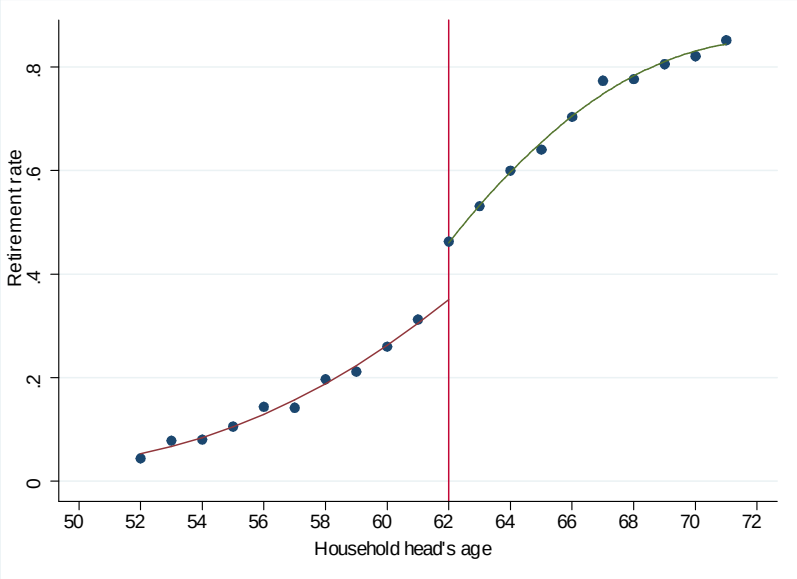

Figure 3: Retirement rate by age for window [-10, +10]

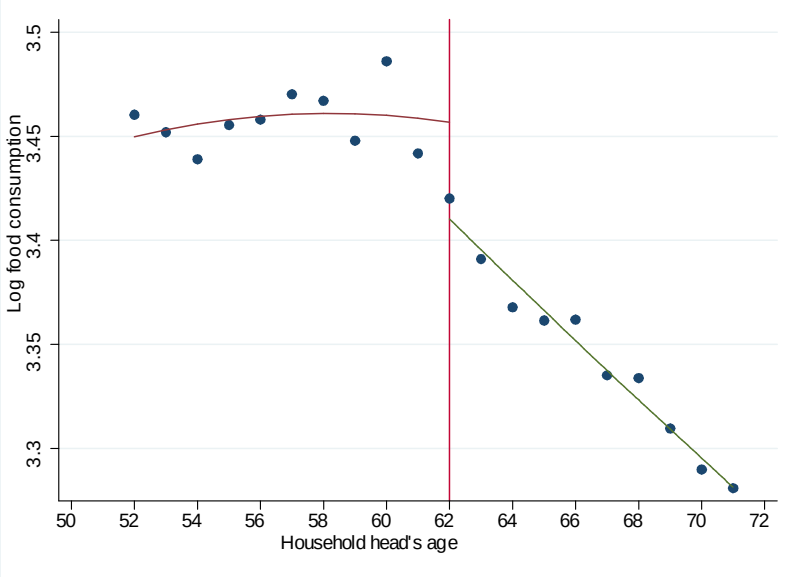

Figure 4: Household food consumption by head's age for window [-10, +10]

either the jump, or the kink, or both for identification. The jump based estimator is given by equation (16). The estimator using both sources of identification is given by equation (18). Both estimators can be easily implemented using local 2SLS with either only the jump or both the jump and the kink as instruments, based on the different windows stated above and with different weights, corresponding to nonparametric estimation with varying bandwidths and kernel functions. In particular, according to Corollary 2, the estimator in the last two columns is valid regardless of whether there is a jump, a kink, or both.

Now consider estimating the kink based estimator given by equation (17), using only the kink information. In this case as shown by the figures there might be jumps in the conditional means of retirement and log food consumption. One way to consistently estimate the purely kink based estimator is the method described in Section 6, based on separately estimating local linear or quadratic regressions above and below the threshold and plugging the resulting consistent estimates of all the slopes into equation (17). Alternatively, one can remove the possible jumps from the retirement and log food consumption functions by partialling out the jumps first, i.e., regressing log food consumption and retirement on a constant and the crossing threshold dummy, and then performing 2SLS using the residuals as dependent variables and the kink as an instrument. More detail will be provided later.

Because the running variable age is discrete, similar to Battistin et al. (2009) and Lemieux and Milligan (2008), I adopt specifications based on age-cell means. In particular, the outcome model is specified as

$$Y_{\widetilde{X}} = \beta_0 + \beta_1 \widetilde{X} + \beta_2 \widetilde{X}^2 + \beta_3 R_{\widetilde{X}} + e, \tag{19}$$

where $\widetilde{X} = X - 62$, and represents the distance to the cutoff. $Y_{\widetilde{X}}$ is the average logged food consumption in each age cell and $R_{\widetilde{X}}$ is the empirical probability of retirement, i.e., the observed fraction of household heads in the cell who are retired. Both are indexed by $\widetilde{X}$ to emphasize that they are defined as sample averages by age. As noted by Lemieux and Milligan

(2008), the corresponding regression estimates based on micro-data are identical to weighted estimates of equation (19) if the weight used is the number of observations by age, while weighting only affects the efficiency, but not the consistency of least squares estimation.[12]

The sample size in each age cell for the ten-year windows below and above the cutoff age 62 (covering 52 to 71) ranges from a minimum of 479 to a maximum of 868 observations. Table 1 shows the number of observations at each age.

Table 1 Number of observations at each age

$\widetilde{X}$	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
No. of observations	868	831	763	725	678	655	609	572	552	578
$\widetilde{X}$	0	1	2	3	4	5	6	7	8	9
No. of observations	527	520	530	517	576	538	488	483	489	479

Ideally observations of the running variable would be continuous, not discretized by year as in Table 1. However, as discussed by Lee and Card (2008), given a discretely observed running variable, if the deviations of the specified approximating function from the true regression function can be taken to be random specification errors, then the point estimates will still be consistent, though calculation of standard errors for micro-data regressions would then need to take into account the clustered nature of these specification errors at the cell level. They show that under certain conditions, robust standard errors from this cell level regression are valid.

Corresponding to the food consumption equation (19), retirement is specified as

$$R_{\widetilde{X}} = \gamma_0 + \gamma_1 \widetilde{X} + \gamma_2 \widetilde{X}^2 + \gamma_3 1(\widetilde{X} \geqslant 0) + \gamma_4 \widetilde{X} * 1(\widetilde{X} \geqslant 0) + u, \tag{20}$$

[12] An alternative specification would be to use year specific cell means and then include in the model year dummies. However, since in this case cross-year variations are captured by those dummies, it would be similar to using averages over all years. A small disadvantage of using year specific cell means is having to estimate year specific effects.

where as before $1(\cdot)$ denotes an indicator function that equals one if its argument is true and zero otherwise. Based on the identification theorems provided earlier, either γ_3 or γ_4 could be set to zero, depending on whether the source of identification is a jump or a kink. Allowing both γ_3 and γ_4 to be nonzero permits identification based on either a jump or a kink, or both.

Log food consumption and retirement are specified in equations (19) and (20) as second order polynomial regressions in $\widetilde{X}$. Given asymptotically shrinking windows (bandwidths), these may be interpreted as nonparametric local quadratic regressions, as recommended by Porter (2003). Although in theory one could include higher order polynomial terms, higher order terms are asymptotically unnecessary for consistency, and empirically cause numerical multicollinearity issues, given the relatively small number of age cell means used for estimation here.

For a given degree of polynomial, shrinking the window width generally reduces bias (at the cost of increasing variance by reducing the effective sample size). So to reduce bias, one can shrink window width. For example, although Figure 4 based on a ten-year window looks like a quadratic form might not be a good fit for log food consumption below the cutoff, when looking at a narrower window, say 8 or 6 years from the cutoff, the scatter plot would be better fitted by simple quadratics. This is confirmed empirically below, where quadratic specifications are shown to provide good fit and yield estimates that are robust across specifications, including varying window widths and kernel weights.

Using either only the jump or both the jump and the kink, equation (19) is estimated by a weighted 2SLS, with the first stage given by equation (20), where γ_4 is set to zero for the former. Using only the kink, γ_3 is set to zero in equation (20), and $Y_{\widetilde{X}}$ and $R_{\widetilde{X}}$ in these two equations are replaced by residualized $Y_{\widetilde{X}}$ and $R_{\widetilde{X}}$, respectively, i.e., residuals from regressions of $Y_{\widetilde{X}}$ and $R_{\widetilde{X}}$ on one and $1(\widetilde{X} \geqslant 0)$. They are then estimated similarly by a weighted 2SLS.[13]

[13] In practice, for the kink based estimator, one may restrict the second derivatives to be the same at each side of the threshold so that the (residualized) retirement treatment equation does not include an interaction term between $\widetilde{X}^2$ and $1(\widetilde{X} \geqslant 0)$. The resulting 2SLS estimator would then correspond directly to the estimator given by equation (17). Alternatively, one can include this interaction term to allow the second derivatives

In all the weighted 2SLS, each observation is weighted by $1/(1+|\widetilde{X}|)$, so the observation at the cutoff having $\widetilde{X}=0$ is weighted by one, whereas those further away are weighted by values less than one. This weighting gives the greatest influence to observations that are most informative about the treatment effect, that is, the observations that are closest to the cutoff. This weighting also makes each stage of the 2SLS equivalent to a local polynomial regression at the cutoff point, with the weights corresponding to the kernel function.

Besides the above kernel weighting, I also try weighting each observations additionally by the inverse sample standard deviation of the dependent variable $Y_{\widetilde{X}}$, log food consumption within each age cell. This weighting scheme takes into account differences in the sampling variances of log food consumption at different ages, as indicated by Figure 4. This weighting was used by Lemieux and Milligan (2008) for RD estimation of the disincentive effects of social assistance. As shown below, the results are not sensitive to the different choices of weighting.

Table 2 presents the main estimation results. Each estimate in Table 2, when multiplied by 100, equals the estimated average percentage change in food consumption at retirement, when retirement is caused by reaching the age at which one qualifies for social security benefits. The first two columns present jump based estimates. The middle two columns present kink based estimates as discussed above. The last two columns present estimates using both the jump and kink for identification. By Corollary 2 and the properties of the 2SLS weights discussed in Section 4, this estimator is valid regardless whether there is a kink, a jump, or both. For each of these three estimators, Column (1) uses the inverse distance weighting, and Column (2) uses both the inverse distance and the inverse sampling standard deviation weighting as discussed above.

to be different. Either way in existence of kinks, the resulting kink estimator converges to the same limiting value (analogous to the way either local linear or local quadratic regressions will consistently estimate the same regression slope). In most cases examined in this paper, the results from including or excluding the interaction between $\widetilde{X}$ and $1(\widetilde{X}\geq 0)$ are not very different, though including this interaction yields estimates that are slightly more stable across different windows widths.

Table 2 Estimated retirement effects on food consumption at age 62-(I)

	Jump		Kink		Both jump and kink	
	(1)	(2)	(1)	(2)	(1)	(2)
[-6,+6]	-0.191	-0.171	-0.221	-0.226	-0.181	-0.181
	(0.089)**	(0.086)**	(0.045)***	(0.043)***	(0.090)**	(0.085)**
[-8,+8]	-0.211	-0.198	-0.194	-0.200	-0.226	-0.215
	(0.089)**	(0.085)**	(0.035)***	(0.033)***	(0.087)***	(0.084)***
[-10,+10]	-0.206	-0.194	-0.211	-0.216	-0.225	-0.215
	(0.080)***	(0.077)**	(0.027)***	(0.026)***	(0.077)***	(0.075)***

Note: estimates are based on the 1994 - 2007 PSID data, with data from the recession years 2001 and 2003 omitted. (1) uses the inverse distance weighting; (2) adds the inverse sampling standard deviation weighting. Using weight (1), the first stage F statistics range from 35.64 to 12579.95. Using weight (2), the first stage F statistics range from 75.45 to 15304.00. For all specifications, the instrumental variables are (jointly) significant at the 1% level in the first stage regression of the 2SLS. Robust standard errors are in parentheses; * significant at the 10% level, ** significant at the 5% level, *** significant at the 1% level.

For all the specifications, the first stage regression of the 2SLS is highly significant with instrumental variables that are (jointly) significant at the 1% level. The estimates remain similar regardless of whether identification is based on the jump, the kink, or both. For example, when using a six-year window (covering age 56 to 67) and inverse distance weighting, the estimated food consumption drops based on the jump, the kink, or both, are 20.6%, 21.1%, and 21.5%, respectively. Note that Theorem 2 shows that when there is a jump, the kink based estimator given by equation (17), the ratio of two kinks, is valid only when the derivative of the treatment effect with respect to the cutoff age 62 is zero, i.e., the treatment effect does not vary linearly with age. The fact that these three estimators all yield similar results therefore suggests that this derivative may be zero, as would be the case if the retirement treatment effect is locally constant.

The results are also robust to different weightings. For example, by the kink based estimator and the inverse distance weighting (marked by (1) in Table 1) the estimated food consumption drops are 21.1%, 19.4%, and 21.1% for the 6, 8, and 10 years windows, respectively, in contrast to 22.6%, 20.0%, and 21.6% based on the alternative weighting. In all specifications,

estimates based on the two different ways of weighting are within one standard error of each other. Furthermore, the results are not very sensitive to different window widths. For example, using only the jump for identification and the inverse distance weighting, the estimated drops in food consumption for the 6, 8, and 10 years windows are 19.1%, 21.1%, and 20.6%, respectively. Results based on alternative equivalence scales are reasonably close. For example, when using another commonly used equivalence scale, the OECD equivalence scale or the Oxford scale, the estimated retirement effects are between 15% and 22% (Appendix B Table A1).[14]

Table 3 Estimated retirement effects on food consumption at age 62-(II)

	Jump		Kink		Both jump and kink	
	(1)	(2)	(1)	(2)	(1)	(2)
[-6,+6]	-0.226	-0.228	-0.226	-0.227	-0.215	-0.219
	(0.088)**	(0.083)***	(0.036)***	(0.035)***	(0.083)**	(0.079)***
[-8,+8]	-0.220	-0.241	-0.210	-0.212	-0.239	-0.224
	(0.080)***	(0.081)***	(0.027)***	(0.026)***	(0.085)***	(0.076)***
[-10,+10]	-0.240	-0.242	-0.211	-0.211	-0.217	-0.222
	(0.079)***	(0.076)***	(0.021)***	(0.021)***	(0.079)***	(0.070)***

Note: estimates are based on the 1994 - 2007 PSID data, including data from the recession years 2001 and 2003. (1) uses the inverse distance weighting, (2) adds the inverse sampling standard deviation weighting. Using weight (1), the first stage F statistics range from 96.89 to 5273.45. Using weight (2), the first stage F statistics range from 116.82 to 6094.19. For all specifications, the instrumental variables are (jointly) significant at the 1% level in the first stage regression of the 2SLS. Robust standard errors are in parentheses. * significant at the 10% level; ** significant at the 5% level;*** significant at the 1% level.

Estimates in this paper are largely consistent with what is documented in the literature. For example, Ameriks et al. (2007) find that a typical U.S. household experiences roughly a 20% fall in consumption at retirement. Bernheim et al. (2001) estimate an average 10%–20% downward shift in the consumption profile around the time of retirement based on the 1978-1990 PSID. Hurd and Rohwedder (2005) estimate the decline at 15%–20% using data

[14]The OECD equivalence scale assigns a value of 1 to the first household member, of 0.7 to each additional adult and of 0.5 to each child. The equivalence scale then equals the sum of these values across all household members.

from the Health and Retirement Study (HRS) and from a supplemental survey to the HRS, the Consumption and Activities Mail Survey (CAMS). More recently, using panel data from 1980 to 2000 Consumer Expenditure Survey (CEX), Aguila, Attanasio, and Meghir (2010) find that food expenditure declines at retirement, but not nondurable expenditure.

For another robustness check, I re-estimate the model using all years' data from 1994 to 2007, including those from the recession years 2001 and 2003. The results are presented in Table 3. The estimated consumption drops are comparable to and are only slightly larger on average than the estimates reported in Table 2 where the recession year's data are omitted. It is plausible that food consumption drops more at retirement during recessions than in other time periods.

Table 4 Estimated retirement effect on food consumption based on micro-data

	Sample (I)				Sample (II)			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
[-6,+6]	-0.258	-0.198	-0.273	0.006	-0.253	-0.223	-0.255	-0.010
	(0.304)	(0.199)	(0.297)	(0.023)	(0.333)	(0.195)	(0.322)	(0.022)
[-8,+8]	-0.261	-0.185	-0.282	0.003	-0.249	-0.214	-0.244	-0.011
	(0.265)	(0.177)	(0.246)	(0.023)	(0.285)	(0.173)	(0.261)	(0.022)
[-10,+10]	-0.231	-0.210	-0.251	0.002	-0.233	-0.216	-0.221	-0.012
	(0.202)	(0.165)	(0.209)	(0.123)	(0.250)	(0.061)	(0.221)	(0.022)

Note: Sample (I) omits years 2001 and 2003 data; Sample (II) includes years 2001 and 2003 data. (a) Jump based 2SLS; (b) Kink based (partialling out jumps) 2SLS; (c) Jump and kink based 2SLS; (d) OLS. All specifications are weighted by the inverse distance weight. Robust standard errors are in parentheses.

Table 4 reports 2SLS and OLS regression results based on individual household micro-data instead of cell means, using data either omitting the recession years (Sample (I) in Table 4) and not omitting (Sample (II)). Since these are not cell mean data, the estimates are weighted only by the inverse distance weighting, not by within cell standard deviations. The specifications are the same as equations (19) and (20), except that year dummies are added as additional covariates in these two equations. Note that use of age-cell means averages over, and hence

smooths out, variation across years, which is analogous to including year dummies in the micro-data regressions. The difference is that estimating the year specific effects tends to increase the standard errors (see, e.g., Lee and Lemieux, 2010). Not surprisingly, for all three window widths (6, 8, or 10 years from the cutoff), the point estimates based on the micro-data 2SLS remain reasonably close to the cell-mean based estimates, while the standard errors are much larger. These increased standard errors cause the estimated retirement effects to become statistically insignificant at conventional significance levels.

Estimates based on OLS are in striking contrast to those by IV 2SLS. OLS yields effects that are mostly small and insignificant and also vary in signs depending on the samples used. In particular, when using the sample that leaves out recession years, OLS estimates of retirement effects have implausible positive signs.

To evaluate the plausibility of the RD assumptions, I examine whether the baseline covariates have any unusual jumps or kinks at the cutoff. The RD modeling assumption implies that individuals do not have precise manipulation of the running variable. If this is true, then there should be no sharp changes at the cutoff age 62 in variables that are determined prior to the treatment. Otherwise it would cast doubt on the validity of the smoothness assumption of potential outcomes as specified in assumption A1. I test smoothness of the conditional means of a battery of baseline covariates, conditional on household head's age. The covariates tested include household head's gender, white/non-white, Hispanic/non-Hispanic, marital status, wife's age, education (in years), and family size. As examples, figures 7 and 8 show the average values of wife's age and family size by head's age. One can see that they both change smoothly with household head's age and there are no unusual jumps or kinks at 62.

Formally, this imprecise manipulation assumption can be tested by the method proposed by Lee (2008) and Lee and Lemieux (2010). First choose covariates that are known to be unaffected by retirement but are correlated with food consumption, and then test the null hypothesis of a zero average effect on these pseudo outcomes by conducting parallel RD analyses,

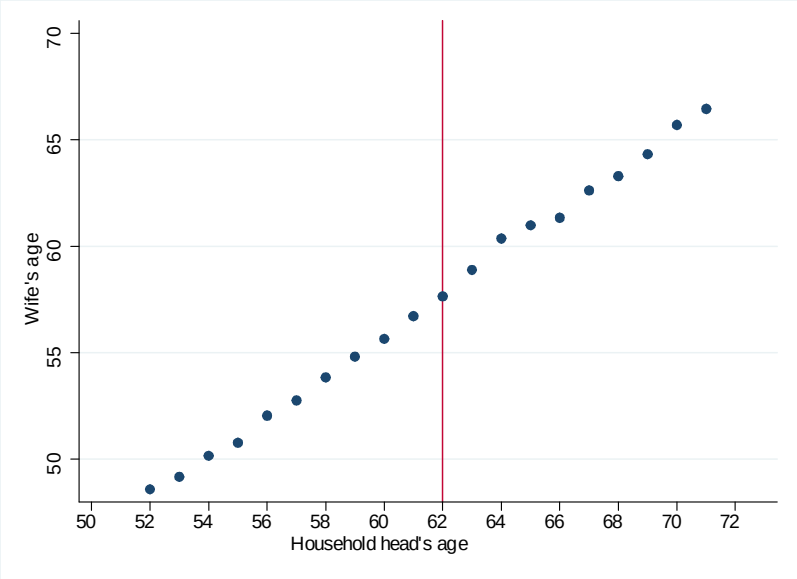

Figure 5: Wife's age by household head's age

i.e., replacing the dependent variable in the second stage of 2SLS with these covariates and performing similar 2SLS estimation. The test results are reported in Appendix B Table A2. For all the covariates listed above, the coefficients of retirement are not statistically significant at any conventional levels and hence confirm the imprecise manipulation assumption.

I next test the smoothness of the density of age at the cutoff, though intuitively it is unlikely that individuals could manipulate age to qualify for social security retirement benefits. Following the general idea of the density test in McCrary (2008), I estimate a local regression the same as the retirement equation except that the retirement rate $R_{\widetilde{X}}$ is now replaced by the fraction of observations at each age. I then test the significance of the estimated coefficients of $1(\widetilde{X} \geqslant 0)$ and $\widetilde{X} * 1(\widetilde{X} \geqslant 0)$ to determine whether there is a significant jump or kink at the cutoff. This regression is intuitively equivalent to graphing the fraction of each age (or a histogram) and inspecting if there are jumps or kinks at the cutoff. As shown in Table A2 in Appendix B, the estimated coefficients are not statistically significant at any conventional levels, thereby indicating no significant jumps or kinks in the density of age at the cutoff, which further confirms the validity of the RD model here.

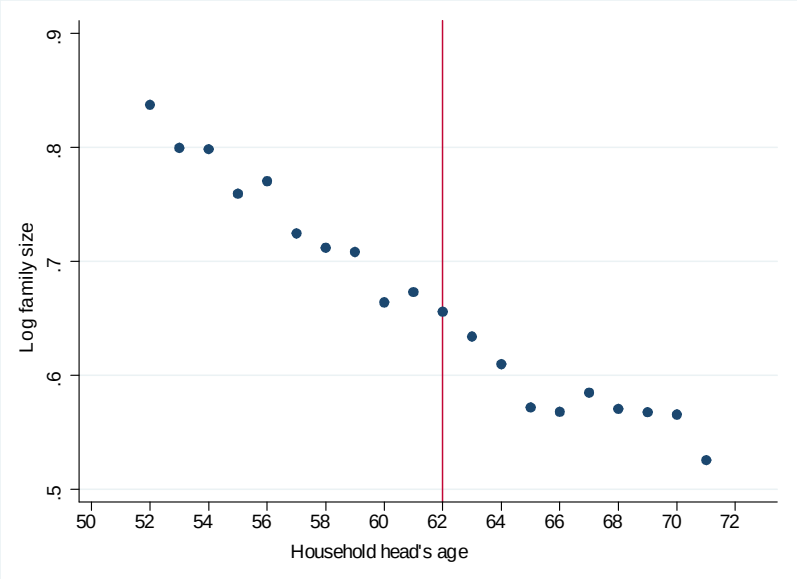

Figure 6: Family size by household head's age

8 Conclusions

Regression discontinuity models identify local average treatment effects by associating a discrete change (a jump) in the mean outcome with a corresponding jump in the treatment probability at a fixed threshold value of the running variable. Lack of discontinuity would make the standard RD estimator infeasible. However, this paper shows that it is possible to identify the standard RD model treatment effect under more general conditions, i.e., from a slope change (a kink) rather than, or in addition to, a jump in the probability of treatment.

Mathematically, the intuition for identification off a kink in the absence of a jump is based on L'hopital's rule. Behaviorally, in this case individuals just below the kink point and those just above are comparable except for the different rate of treatment probability changes. This slope change along with any observed slope change in the mean outcome can be used to estimate the RD model treatment effect. Note that the RD identification here can not handle dynamic effects like anticipation effects (e.g., individuals change behaviors in anticipation of the treatment) or delayed treatment effects (e.g. treatment effects show up gradually with

time) when the running variable is age or time, but the same is true for standard RD models.

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 or the slope of the treatment probability at a threshold of the running variable.

The identification results in this paper can be intuitively interpreted using IV models. Just as the standard fuzzy design RD estimator is numerically equivalent to an IV estimator, I show that the proposed estimators are numerically equivalent to IV estimators. In particular, a kink in the treatment probability provides an additional instrument that one can use to identify the RD treatment effect. It is known that a jump in the treatment probability at the threshold implies that the binary indicator for crossing the threshold can be used as an instrument. Similarly, a kink at the threshold implies that the interaction term between this binary indicator and the running variable can also be an instrument. So if there is no jump but a kink in the treatment probability, one would still be able to use this kink at the threshold to identify the same local average treatment effect as would be identified by a jump, if the jump were to exist. 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 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.

The identification results are applied to estimate the retirement impact on household food consumption at the early retirement age 62 using the PSID data in the US. Graphical analyses show that there might be a jump, a kink, or both in food consumption and retirement probabilities at 62. Estimators based on either a jump, or a kink, or both are performed. I show that all three yield very similar estimates and that the results are robust to different window widths and weightings. Food consumption is estimated to drop by about 15% to 23% when household heads retire. The estimates are largely consistent with what is documented in the literature.

Given this paper's 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).

9 Appendix A: Proofs

First note that for any x such that $c-\varepsilon \leq x \leq c+\varepsilon$, given Assumption A1, $E\left(Y\left(1-D^*\right) \mid X=x\right)$ and $E\left(T\left(1-D^*\right) \mid X=x\right)$ are continuously differentiable in the neighborhood of $x=c$, since $E\left(Y \mid X=x, D^*=0\right)$, $E\left(T \mid X=x, D^*=0\right)$ and $E\left(D^* \mid X=x\right)$ are assumed to be continuously differentiable. The following proof will use these results.

PROOF of LEMMA 1:

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

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

For any $x > c$, by the definition of compliers $D^* = 1$, and continuity of $E(Y(1) \mid X = x, D^* = 1)$ and $E(Y(1 - D^*) \mid X = x)$, one has

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

Similarly, for any $x < c$, by the definition of compliers $D^* = 1$ and continuity of $E(Y(0) \mid X = x, D^* = 1)$ and $E(Y(1 - D^*) \mid X = x)$, one has

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

Therefore,

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

$E(Y(1) - Y(0) \mid X = c, D^* = 1)$ 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 $T = T D^* + T(1 - D^*)$ and the definition of compliers, 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

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

which is equation (2).

PROOF of THEOREM 1:

For $t = 0, 1$, define the function $G_t(x)$ by

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

Taking the ordinary derivative of this function gives

$$G_t'(x) = \frac{\partial E(Y(t) \mid X = x, D^* = 1) E(D^* \mid X = x)}{\partial x} + \frac{\partial E(Y(1 - D^*) \mid X = x)}{\partial x}.$$

This derivative $G_t'(x)$ exists and is continuous at $x = c$ because $E(Y(t) \mid X = x, D^* = 1)$, $E(D^* \mid X = x)$, and $E(Y(1 - D^*) \mid X = x)$ are all continuously differentiable at $x = c$. It follows that

$$\begin{aligned} G_1'(c) - G_0'(c) &= \frac{\partial E(Y(1) - Y(0) \mid X = c, D^* = 1) E(D^* \mid X = c)}{\partial c} \\ &= \tau'(c) E\left(D^* \mid X = c\right) + \tau(c) \frac{\partial E(D^* \mid X = c)}{\partial c}. \end{aligned}$$

By the proof of Lemma 1 for x in the neighborhood of c we have $g(x) = G_1(x)$ for $x \geqslant c$, and by continuity of $G_t'(x)$ we have $g_+'(c) = G_1'(c)$. In the same way based on $x \leqslant c$ we get $g_-'(c) = G_0'(c)$, and so

$$g_+'(c) - g_-'(c) = \tau'(c) E\left(D^* \mid X = c\right) + \tau(c) \frac{\partial E(D^* \mid X = c)}{\partial c}. \tag{21}$$

Similarly, we have

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

and

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

Given the continuous differentiability of $E\left(T\left(1 - D^*\right) \mid X = x\right)$ and $E\left(D^* \mid X = x\right)$ at the point $x = c$, the ordinary derivatives of the right-hand side in the above two equations exist. So analogous to the above analysis we obtain

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

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

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

and so

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

Note that the above can be alternatively shown by L'hopital's rule. To see this, let $B(x) = G_1(x) - G_0(x)$ for all x in the neighborhood of c. Similarly, let $p(x) = E\left(D^* \mid X = x\right)$. By Assumption A1, both $B(x)$ and $p(x)$ are continuously differentiable.

From the above, one has $B(c) = G_1(c) - G_0(c) = g_+(c) - g_-(c)$ and $p(c) = f_+(c) - f_+(c)$, as well as $B'(c) = G'_1(c) - G'_0(c) = g'_+(c) - g'_-(c)$ and $p'(c) = f'_+(c) - f'_-(c)$. Lemma 1 shows $B(c) = \tau(c)\, p(c)$, so given $p(c) = 0$, one has $B(c) = \tau(c)\, p(c) = 0$.

Then

$$\begin{aligned}\tau(c) &= \frac{\lim_{x\to c} B(x)}{\lim_{x\to c} p(x)} = \frac{\lim_{x\to c} B'(x)}{\lim_{x\to c} p'(x)} \\ &= \frac{B'(c)}{p'(c)} = \frac{g'_+(c) - g'_-(c)}{f'_+(c) - f'_-(c)},\end{aligned}$$

where the second equality follows from L'hopital's rule, and the third from the continuous differentiability of $B(x)$ and $p(x)$.

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 $B(c)$, $p(c)$, $B'(c)$, and $p'(c)$ as in the proof of Theorem 1. If there is no jump, i.e., $p(c) = 0$ and hence $B(c) = \tau(c)\, p(c) = 0$, by Assumption A2, there is a kink. Then Theorem 1 gives

$$\tau(c) = \frac{B'(c)}{p'(c)} = \frac{B(c) + wB'(c)}{p(c) + wp'(c)}.$$

Now consider the case where $\tau'(c) = 0$. By equations (21) and (22), if $\tau'(c) = 0$ then $B'(c) = \tau(c)\, p'(c)$, and in addition it has already been shown that $B(c) = \tau(c)\, p(c)$ with equations (1) and (2). Taking a weighted sum of these two equations gives $B(c) + wB'(c) = \tau(c)\,(p(c) + wp'(c))$. Then

$$\tau(c) = \frac{B(c) + wB'(c)}{p(c) + wp'(c)}.$$

The denominator of this equation is nonzero, since by Assumption A2 either $p(c)$ or $p'(c)$ is

nonzero.

PROOF of COROLLARY 2:

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

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

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

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

where the last equality follows from Theorem 1. Since this equality holds for all n, it must hold in the limit as $n \to \infty$.

PROOF of THEOREM 3:

Use the notation in the proof of Theorem 1, and for simplicity replace $B(c)$ and $p(c)$ with B and p, respectively. From Lemma 1,

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

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 in the text $B' = C$, $p\prime = q$, $B'' = D$, and $p'' = r$. If there is no jump, $p = 0$. By Assumption 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 solving for $\tau(c)$ from equations (24) and (25) gives

$$\tau(c) = \frac{2qC - Dp}{2q^2 - rp}.$$

Also if there is a jump, the standard RD estimator applies, $\tau(c) = \frac{B}{p}$. By the rule of fraction, 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 taking derivatives on both sides of $B = \tau(c)\, p$, 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, so it holds in the limit as $n \to \infty$.

10 Appendix B: Tables

Table A1 Estimated retirement effects on food consumption at age 62 - (III)

	Jump		Kink		Both jump and kink	
	(1)	(2)	(1)	(2)	(1)	(2)
[-6,+6]	-0.186	-0.176	-0.175	-0.180	-0.193	-0.184
	(0.097)*	(0.093)*	(0.048)***	(0.046)***	(0.096)**	(0.091)**
[-8,+8]	-0.212	-0.200	-0.150	-0.155	-0.223	-0.212
	(0.095)**	(0.091)**	(0.037)***	(0.036)***	(0.092)**	(0.088)**
[-10,+10]	-0.205	-0.194	-0.166	-0.170	-0.219	-0.210
	(0.085)**	(0.082)**	(0.029)***	(0.028)***	(0.081)***	(0.079)***

Note: estimates are based on the 1994 - 2007 PSID data, not including the recession years 2001 and 2003. Food consumption is scaled by an alternative OECD equivalence scale. (1) uses the inverse distance weighting; (2) uses both the inverse distance and inverse sampling standard deviation weighting. Using weight (1), the first stage F statistics range from 63.44 to 12579.95. Using weight (2), the first stage F statistics range from 74.03 to 15122.38. For all specifications, the instrumental variables are (jointly) significant at 1% level in the first stage regression of the 2SLS. Robust standard errors are in the parentheses. * significant at the 10% level; ** significant at the 5% level; *** significant at the 1% level.

Table A2 The smoothness of conditional means of covariates and density of age

Retirement effects on base-line covariate:	
Male	0.089 (0.066)
White	-0.060 (0.058)
Married	0.112 (0.069)
Wife's age	1.197 (1.000)
Hispanic	-0.031 (0.037)
Education	1.041(0.705)
Family size	-0.014 (0.061)
Density of household head's age:	
Jump	-0.000 (0.002)
Kink	0.001 (0.001)

Note: Robust standard errors are in the parentheses; All the estimates are not statistically significant at the conventional significance levels; Estimation is based on a 10 years window with the inverse density weighting and inverse sampling standard deviation weighting.

References

[1] Aguila, E., O. Attanasio, and C. Meghir (2010), "Changes in consumption at retirement: evidence from panel data," forthcoming, The Review of Economics and Statistics.

[2] Ameriks, J., A. Caplin, and J. Leahy (2007), "Retirement Consumption: Insights from a Survey," The Review of Economics and Statistics 89, 265-274.

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

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

[5] Bernheim, B. D., J. Skinner, and S. Weinberg (2001), "What Accounts for the Variation in Retirement Wealth Among U.S. Households?" American Economic Review 91, 1–26.

[6] 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.

[7] Card, D., D. S. Lee and Z. Pei (2009), "Quasi-Experimental Identification and Estimation in the Regression Kink Design," Princeton University, WP # 553.

[8] 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.

[9] Citro, C. F. and R. T. Michaels (1995), Measuring Poverty: A New Approach, C. F. Citro and R. T. Michaels (Eds.) National Academy Press, 161-162.

[10] Davidson, R. and J. G. MacKinnon (1993), Estimation and Inference in Econometrics, Oxford University Press.

[11] Dong, Y. and A. Lewbel (2010), "Regression Discontinuity Marginal Threshold Treatment Effects," Boston College working paper number 759.

[12] Guryan, J. (2003), "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts," National Bureau of Economic Research, WP8269.

[13] 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.

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

[15] Hurd, M., and S. Rohwedder (2005), "The Retirement-Consumption Puzzle: Anticipated and Actual Declines in Spending at Retiremen," RAND Labor and Population working paper series WR-242.

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

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

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

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

[20] Lee, D. S. (2008), "Randomized Experiments from Non-random Selection in U.S. House Elections," Journal of Econometrics, 142 (2), 675–697.

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

[22] Lemieux, T. and K. Milligan (2008), "Incentive effects of social assistance: A regression discontinuity approach," Journal of Econometrics 142, 807–828.

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

[24] Mariger, R. (1987), "A Life-Cycle Consumption Model with Liquidity Constraints: Theory and Empirical Results," Econometrica 55, 533–557.

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

[26] Nielsen, H. S., T. Sorensen, and C. R. Taber (2009), "Estimating the Effect of Student Aid on College Enrollment: Evidence from a Government Grant Policy Reform," American Economic Journal: Economic Policy," forthcoming.

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

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

[29] Simonsen, M., L. Skipper, and N. Skipper (2009), "Price Sensitivity of Demand of Prescription Drugs: Exploiting a Kinked Reimbursement Scheme," University of Aarhus Economics Working Papers no. 2010-03.

[30] Thistlethwaite, D. L. and D. T. Campbell (1960), "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment," Journal of Educational Psychology 51, 309–317.