JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **30**: 422–446 (2015)
Published online 11 March 2014 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2369

REGRESSION DISCONTINUITY APPLICATIONS WITH ROUNDING ERRORS IN THE RUNNING VARIABLE

YINGYING DONG*
Department of Economics, University of California Irvine, CA, USA

SUMMARY

Many empirical applications of regression discontinuity (RD) models use a running variable that is rounded and hence discrete, e.g. age in years, or birth weight in ounces. This paper shows that standard RD estimation using a rounded discrete running variable leads to inconsistent estimates of treatment effects, even when the true functional form relating the outcome and the running variable is known and is correctly specified. This paper provides simple formulas to correct for this discretization bias. The proposed approach does not require instrumental variables, but instead uses information regarding the distribution of rounding errors, which is easily obtained and often close to uniform. Bounds can be obtained without knowing the distribution of the rounding error. The proposed approach is applied to estimate the effect of Medicare on insurance coverage in the USA, and to investigate the retirement-consumption puzzle in China, utilizing the Chinese mandatory retirement policy. Copyright © 2014 John Wiley & Sons, Ltd.

Received 2 August 2012; Revised 2 August 2013

Supporting information may be found in the online version of this article.

1. INTRODUCTION

Regression discontinuity (RD) models identify local average treatment effects by associating a discrete change in the treatment probability with a corresponding discrete change in the mean outcome when a continuous running variable crosses a known threshold. The RD treatment effect is given by the ratio of these two discrete changes. In the case of a sharp RD design, where the treatment probability changes by 1 at the threshold, the RD treatment effect is simply the difference in the mean outcome of interest just above and just below the threshold.

Many empirical applications of regression discontinuity methods use a running variable that is rounded and hence involves rounding or discretization errors. Examples of rounded and hence discrete running variables include age in years, birth weight in ounces, an integer-valued test score, calendar year or quarter, etc. Rounding issues exist frequently due to data limitation in survey datasets. For example, many survey datasets only report an individual's age in years at the time of the survey. in this case the reported age is thus an individual's actual age rounded down to the nearest integer. Similar to age in years, calendar year or similarly quarter involves rounding down to the nearest integer, whereas an integer test score or birth weight in ounces typically involves ordinary rounding, i.e. the true birth weight or test score is rounded either up or down to the nearest integer.

Barreca *et al.* (2010, 2011) note that birth weight tends to heap at ounce or 100 g multiples, possibly due to the limited resolutions of scales. Their Monte Carlo simulations show that standard RD analysis arrives at biased estimates when the running variable displays heaps. They therefore recommend dropping observations at the heaping points, and only use the remaining continuous birth weight data

* Correspondence to: Yingying Dong, Department of Economics, 3151 Social Science Plaza, University of California Irvine, Irvine, CA 92697-5100, USA. E-mail: yyd@uci.edu

as a way to reduce the bias. However, as they have noted, data heaps may comprise a large share of the data, and one may be interested in knowing the treatment effect for the heaped type, which tends to be associated with different quality of hospitals (and hence different treatment) as well as different family backgrounds from the continuous type. This paper's approach might be applied to heaped birth weight data to correctly estimate the RD treatment effect.

In addition, many survey datasets report income in brackets. One way to use bracket data is to include a separate dummy variable for each bracket in a regression. This is essentially a functional form assumption that is made not for economic plausibility but just to accommodate the limitations of data reporting. More commonly, interval censoring is ignored, and income is just imputed as the midpoint of each bracket. This imputation causes rounding errors, resulting in estimation bias. As long as the true model depends on the underlying continuous income measure rather than interval dummies, this paper's approach may be applied to deal with the rounding error bias resulting from interval censoring.

RD models crucially rely on the continuity of the running variable for identification. In particular, identification is achieved by shrinking the bandwidth to zero in the limit and hence essentially compares outcomes for observations 'just above' and 'just below' the treatment threshold. As noted by Lee and Card (2008), when the observed running variable X is rounded and hence is discrete, it is impossible to shrink the bandwidth to zero to compare units just above and just below the threshold (since there are no observations 'just below' the threshold regardless of the sample size). Identification in this case has to rely on extrapolation based on functional form. Essentially, the RD treatment effect is not nonparametrically identified in this case. Standard practice is therefore to estimate parametric regressions of Y on reported age X (generally low-order polynomial models; see, for example, Card and Shore-Sheppard, 2004; DiNardo and Lee, 2004; Lee, 2008) above and below the RD treatment threshold.

Although using a rounded discrete running variable has been common in RD applications due to the limitation of many survey datasets, very few studies directly address this issue. This paper contributes to the growing RD literature by providing an easy and practical way to test and correct the rounding bias caused by using a rounded discrete running variable. This paper uses age in years as a leading motivating example and focuses on the case of rounding down first. The results are then extended to cases involving non-integer cutoffs or other common forms of rounding, such as ordinary rounding or rounding up to the nearest integer. Unlike the case of rounding down, these other cases may involve misclassification of falling on either side of the threshold.

I first show that the standard method for dealing with a rounded discrete running variable leads to inconsistent estimates of the RD treatment effect, even if the true functional form relating the outcome to the running variable is known and is correctly specified. This inconsistency exists for similar reasons that measurement errors in regressors yield biased and inconsistent estimates of regression coefficients, even when the functional form of the regression is correctly specified. In this case the observed rounded running variable can be taken as a mismeasure of the true exact running variable.

I next provide a formula for the size of the bias in the standard RD model estimators that use a rounded running variable based on rounding down, and describe the restrictive conditions under which this bias will be zero. For example, a sufficient condition for the rounding or discretization bias to be zero is that the slope and higher derivatives of the outcome as a function of age do not change at the cutoff. The bias can be either positive or negative, depending on how the slope and higher derivatives change at the threshold, so with a rounded running variable the data may reveal a larger or smaller discontinuity than the true discontinuity.

I also show how to correct this rounding bias and thereby obtain consistent estimates of the true RD treatment effect. These corrections are very simple to implement in practice, and it is also simple to obtain standard errors for the bias-corrected estimator of the treatment effect. Knowing exactly when the rounding error matters might be useful in practice. For example, if the correction does not

DOI: 10.1002/jae

make much difference, one can be assured that obtaining a more refined measure of the running variable would not significantly change one's estimates. These bias formulas and resulting corrections are not the same as the corrections for standard measurement error models.[1] The t statistic on the bias-corrected estimator provides a valid test for the presence of a true discontinuity, and hence a non-zero true treatment effect.

A convenient feature of the proposed bias correction is that it does not require an instrument. It instead assumes that one has some information regarding the distribution of the rounding error within the discretized variable cell. In the case of age in years, the required information is moments of the distribution of ages within a year, i.e. the distribution of birthdays for the underlying population the data are drawn from, which is readily available from census data. Alternatively, in some applications this distribution can be well approximated by a uniform distribution. I also briefly discuss how to construct bounds on the true treatment effect based on this paper's identification results. These bounds do not rely on having any information regarding the distribution of the rounding error.

It is worth emphasizing that, although this paper's results are discussed in the context of RD models, they can be readily used in correcting for the bias in the estimated coefficients in ordinary regressions when the true model depends on continuous regressors, but one has only available rounded, coarsened or interval censored variables.

Two empirical applications are provided. The first looks at the effect of Medicare eligibility on medical insurance coverage in the USA. For this application, a finer (monthly versus yearly) measure of age is available. Estimates based on these monthly age data provide a benchmark. I show that the proposed methodology works well and produces estimates that are consistent with having and using data where age is more accurately measured. In particular, both the benchmark estimates and the bias-corrected estimates imply that rounding leads to an overestimate of the impact of Medicare eligibility on the medical insurance coverage rate. The bias-corrected estimates are similar to those estimates using monthly age data and depart further from the uncorrected estimates based on age in years. A bounds calculation confirms that rounding in this application leads to an overestimate of the treatment effect, whatever the rounding error distribution may be. Interestingly, this is in opposite direction of the usual attenuation bias that results from classical measurement error.

The second application investigates the retirement-consumption puzzle in China. The puzzle (see, for example, Banks *et al.*, 1998) refers to the empirical finding that, in many datasets of developed Western countries, consumption (typically food consumption) drops significantly at retirement, which is inconsistent with consumption smoothing by the standard life cycle model. In China, the official retirement age for male workers is 60. This mandatory retirement rule yields a significant jump in the retirement rate at age 60, which helps identify the true causal impact of retirement on various outcomes of interest.

Since age is recorded in years in the available Chinese dataset, accurately estimating the retirement impact using RD models requires correcting for the rounding bias. I show that the slope of the food consumption profile changes substantially at the threshold age 60, resulting in a relatively large rounding bias. I find that there is a significant drop in food consumption at the retirement of male household heads, but the drop is not as large as one would estimate from a standard RD analysis that ignores the rounding bias.

The rest of the paper proceeds as follows. Section 2 briefly reviews the literature. Section 3 provides the main identification result for the sharp design RD based on rounded and hence discrete data. Also provided is a numerically simple correction to the standard discrete data RD estimation. Section 4 describes how to estimate the true RD treatment effect with rounded data. Section 5 extends the approach to fuzzy design RD models. Sections 6 and 7 present two empirical applications. Sections 8

[1] Classical measurement errors will lead to attenuation bias in estimated regression coefficients. With rounding errors, the treatment effect can either be overestimated or underestimated.

DOI: 10.1002/jae

discusses some extensions to the basic setup. Short concluding remarks are provided in Section 9. Proof of the main result is provided in the Appendix.

2. LITERATURE REVIEW

RD methods have gained great popularity in the treatment effect evaluation literature in recent years. Recent surveys include Imbens and Lemieux (2008) and Lee and Lemieux (2010). Age is commonly used as a running variable in RD applications. Examples include Behaghel *et al.* (2008), Card *et al.* (2008, 2009), Carpenter and Dobkin (2009), Chen and van der Klaauw (2008), De Giorgi (2005), Edmonds (2004), Edmonds *et al.* (2005), Ferreira (2010), Lalive *et al.* (2006), Lalive (2007, 2008), Lee and McCrary (2005), Lemieux and Milligan (2008) and Leuven and Oosterbeek (2004). This paper's main results may also be similarly applied to RD applications with calender years as a running variable. For example, Oreopoulos (2006) uses birth year as a running variable to estimate returns to education in the UK.

One paper that specifically considers discreteness of an RD running variable is Lee and Card (2008). They model deviations of the true regression function from a given approximating function, i.e. the chosen parametric (e.g. polynomial) function for the discrete running variable, as random specification errors. They then discuss the impact of these random specification errors on inference.

As noted by Lee and Card (2008), when the random specification errors are not identical for the regressions above and below the cutoff, the true parameter of interest, i.e. the true RD treatment effect, is not the same as the simple difference of the expected values of these two regressions at the cutoff. In particular, the true treatment effect equals the latter plus the expected difference between the two random specification errors at the cutoff.

If one interprets their specification errors as rounding errors due to discretization of the running variable, then it can be shown that the expected difference between these two rounding errors is not necessarily zero. In this paper, I show that under general conditions this bias term can be identified and estimated, using moments of the distribution of the rounding error within the discretized cell.

A few existing papers deal with RD models in which the running variable is mismeasured with a classical measurement error. Unlike rounding errors classical measurement errors are assumed to be independent of the unobserved true running variable. With a continuous running variable, as required by standard RD models, rounding errors cannot be classical.[2] In particular, Pei (2011) discusses identification of the running variable distribution and the RD treatment effect under the assumption that the true underlying running variable and the classical measurement error are discrete and bounded; note that here the true running variable is continuous.

Battistin *et al.* (2009) deal with measurement error in the running variable in their empirical application of an RD model. Their RD design exploits the Italian pension eligibility rule and is used to investigate the retirement-consumption puzzle in Italy. The observed running variable (distance to pension eligibility, constructed based on age and pension contribution years) is measured with error. They show that, assuming a classical measurement error and assuming that the measurement error is orthogonal to the treatment and the potential outcomes conditioning the true running variable, one can estimate the RD treatment effect by the standard procedure ignoring the measurement error. In this case, the numerator and the denominator of the fuzzy design RD treatment effect estimator are scaled by the same factor, i.e. the fraction of individuals who correctly report their running variable among all individuals near the eligibility threshold.

[2] Let $Z = Z^* + u$, where Z is the observed mismeasured variable, Z^* is the true variable and u is the measurement error; in the case of a classical measurement error, u is assumed to be independent of Z^*. However, the rounding errors take on the form $X^* = X + e$, where X^* is the unobserved true continuous running variable, and X is the observed rounded and hence discrete variable. By construction, the rounding error e cannot be independent of the true variable X^*.

DOI: 10.1002/jae

Hullegie and Klein (2010) also deal with the measurement error problem in the running variable in an empirical application of an RD model. They estimate the effect of private insurance on health care utilization and health in Germany. Private insurance eligibility is determined by income above a certain threshold. Due to the potential measurement error in income, no discontinuity is observed in private insurance probability at the qualifying income threshold. They assume a normally distributed classical measurement error, which is independent of private insurance status and potential outcomes, and then develop an estimator based on the distributional assumptions. For the effect of measurement error in the general average treatment effect framework, one can refer to Battistin and Chesher (2011) and references therein.

In addition, the biases associated with the use of interval data have been examined in the context of standard regression models. Heitjan and Rubin (1991) present a general statistical model of data coarsening, and establish that when data are coarsened at random coarsening can be ignored in drawing Bayesian and likelihood inferences. Empirical studies confronting interval regressor data often set observations equal to the midpoint of the interval, which is similar to ordinary rounding. Another common practice is to use a set of dummy variables indicating falling in a particular interval. Hsiao (1983) critiques these approaches as applied to linear regression models. Alternatively, discretization or rounding leads to a loss of identification. Manski and Tamer (2002) provide bounds on model parameters in regressions with one of the regressors being interval valued. Tsiatis (2006) reviews estimation of semiparametric models with missing, censored and coarsened data.

3. IDENTIFICATION

Begin with the standard RD setup. I consider sharp design RD first, and later provide the straightforward extension to fuzzy designs. Let T be the indicator of treatment, so an individual has $T = 1$ if treated and $T = 0$ otherwise. Let c denote the cutoff or threshold age for treatment, which is assumed to be an integer. Let X^* be an individual's exact age at the time of the survey minus c, so X^* is the underlying continuous running variable.

Follow Rubin (1974) to let $Y(1)$ and $Y(0)$ denote an individual's potential outcomes of interest from being treated or not, respectively. The observed outcome Y can then be written as $Y = Y(0) + [Y(1) - Y(0)]T$. An individual's potential outcome can depend on X^*. Define conditional means of the potential outcomes conditioning on the true age as $g_t(X^*) = E(Y(t) \mid X^*)$ for $t = 0, 1$. The conditional mean of the observed outcome is then $E(Y \mid X^*, T) = g_0(X^*) + [g_1(X^*) - g_0(X^*)]T$.

Define the dummy indicating crossing the threshold as $T^* = I(X^* \geq 0)$, where $I(\cdot)$ is the indicator function that equals one if its argument is true and zero otherwise. Sharp design RD means that $T = T^*$, so for now an individual is assumed to be treated if and only if his true age equals or exceeds the cutoff c. It follows that when $X^* \geq 0$, $g_1(X^*) = E(Y \mid X^*, T^* = 1) = E(Y \mid X^*)$ and when $X^* < 0$, $g_0(X^*) = E(Y \mid X^*, T^* = 0) = E(Y \mid X^*)$. By assuming continuity of $g_0(X^*)$ and $g_1(X^*)$ at the threshold point $X^* = 0$, the standard sharp design RD local average treatment effect is given by $\tau = g_1(0) - g_0(0)$.

Outcomes could also depend on other covariates, which are suppressed for now. All the statements and theorems in this paper can be assumed to hold conditioning on the values of other covariates, though it should be noted that a generic virtue of the RD approach is that inclusion of other covariates generally only affects efficiency but not consistency of estimated RD treatment effects.

Let $\widehat{g}_1(X^*)$ be a consistent estimator of $g_1(X^*)$ for $X^* \geq 0$, obtained by regressing Y on X^* either nonparametrically, or by a correctly specified parametric model, using observations of data having $X^* \geq 0$. Similarly, let $\widehat{g}_0(X^*)$ be a consistent estimator of $g_0(X^*)$ for $X^* < 0$. The sharp design RD treatment effect τ is consistently estimated by $\widehat{\tau} = \widehat{g}_1(0) - \widehat{g}_0(0)$.

Now suppose one does not observe continuous X^*. Instead, one observes X, defined as X^* rounded down to the nearest integer, i.e. X is an individual's reported age in years minus c at the time of the

DOI: 10.1002/jae

survey. This X is what is usually available in survey datasets, since surveys typically report individuals' age in (integer) years up to their most recent birthday at the time of the survey. The same analysis can be applied for finer or coarser reporting of age; e.g. if age is recorded in months, then the same analysis can apply by taking the units of X and X^* as measured by month rather than by year.

Let $h_1(X) = E(Y \mid X, T^* = 1)$ when $X \geq 0$ and that $h_0(X) = E(Y \mid X, T^* = 0)$ when $X < 0$, so $h_t(X)$ is the discrete data analog of $g_t(X^*)$ for $t = 0, 1$. By construction, $E(Y \mid X, T) = E(Y \mid X, T^*) = h_0(X) + [h_1(X) - h_0(X)]T^*$. Let $\widehat{h}_1(X)$ be a consistent estimator of $h_1(X)$ for $X \geq 0$, obtained by regressing Y on X in a correctly specified parametric model using observations of data having $X \geq 0$. Similarly let $\widehat{h}_0(X)$ be a consistent estimator of $h_0(X)$ for $X < 0$, obtained by regressing Y on X in a correctly specified parametric model using observations of data having $X < 0$.

Suppose one were to ignore the fact that the reported data are rounded, and attempted to estimate the RD treatment effect as $\widehat{h}_1(0) - \widehat{h}_0(0)$ instead of $\widehat{g}_1(0) - \widehat{g}_0(0)$. Refer to this as the naive discrete data RD treatment effect estimator and denote it by $\widehat{\tau}'$. Denote the probability limit of $\widehat{\tau}'$ as τ', which is referred to as the naive discrete data RD treatment effect, so $\tau' = h_1(0) - h_0(0)$.

If the discrete data RD treatment effect τ' equals the true RD treatment effect τ, i.e. $\tau' = \tau$, then the naive discrete data estimator $\widehat{\tau}'$ is a consistent estimator of the true RD treatment effect τ, otherwise $\widehat{\tau}'$ will be an inconsistent estimator of the true RD treatment effect. Define the bias in the discrete data RD treatment effect as $\tau' - \tau$.

Access to only rounded age data will make local nonparametric estimation in the neighborhood of the threshold impossible, because one does not observe data anywhere in the neighborhood of zero except at zero itself. It will therefore be necessary to assume that one knows the parametric form of the underlying model $g_t(X^*)$ for $t = 0, 1$.

Let $e = X^* - X$, so e is the measurement error in the reported rounded age, and has $0 \leq e < 1$. Let $\mu_k = E\left(e^k\right)$ be the kth non-central moment of the rounding error e.

I will follow the usual practice in the literature of specifying these models as polynomials with possibly unknown but finite degrees. Polynomials are by far the most commonly used models in empirical practice. Also, any sufficiently smooth (i.e. analytic) function can be approximated arbitrarily well by a polynomial. Given the parametric model for the true continuous age, I then derive the corresponding specifications for the rounded age. It will follow from the assumptions below (as proved in Corollary 1) that the correct specifications for $h_0(X)$ and $h_1(X)$ will also be polynomials. When estimating these models, standard covariate selection tests can be employed to determine the degree of these polynomials.

The following assumptions allow one to identify and consistently estimate the true RD local treatment effect τ using what one can identify from rounded data. Equivalently, these assumptions will permit one to quantify and correct for the bias in the discrete data RD treatment effect τ'.

Assumption 1. $T = I(X^* \geq 0)$.

Assumption 2. $g_0(X^*)$ and $g_1(X^*)$ are continuous at $X^* = 0$.

Assumption 3. The conditional mean functions $g_0(X^*)$ for $X^* < 0$ and $g_1(X^*)$ for $X^* \geq 0$ are polynomials of possibly unknown degree J.

Assumption 4. $h_0(X)$ is identified for all $-(J+1) \leq X < 0$, and $h_1(X)$ is identified for all $0 \leq X \leq J$.

Assumption 5. $I(X \geq 0) = I(X^* \geq 0)$.

Assumption 6. For all integers $k \leq J$, $E\left(e^k \mid X\right) = E\left(e^k\right) = \mu_k$, and these J moments are identified.

DOI: 10.1002/jae

Assumption 1 is the standard sharp design RD identifying assumption that treatment occurs if and only if age exceeds the threshold c, and hence when X^* crosses zero. Assumption 2 is the standard identifying assumption of an RD design that the conditional means of the potential outcomes are continuous at $X^* = 0$, so the discontinuity in the observed conditional mean of Y at the threshold can be attributed to the treatment.

Assumption 3 is a functional form restriction. As noted previously, unlike the standard RD, local nonparametric estimation is not possible with rounded discrete data, so estimation requires some assumed functional form. Note that Assumption 3 is an assumption only about observables, not counterfactuals, and so could be tested (using, for example, a validation sample where true age X^* is reported). Assumption 3 can be easily extended to allow $g_0(X^*)$ and $g_1(X^*)$ to be polynomials of different degrees. For simplicity, I assume that they both are polynomials of degree J. One can take J to be the maximum of the degree of these two polynomials.[3]

Assumptions 4, 5 and 6 are the only assumptions that are imposed on the observed age X (as opposed to the true age X^*). Assumption 4 says that one can identify the mean of Y in each observed age cell X. So, for example, $h_1(X)$ is just the mean of Y across everyone with reported age X above the threshold.

Assumption 5 says that the crossing threshold indicator T^*, when defined in terms of X rather than X^* is not mismeasured. This holds automatically for the type of rounding considered here, where the observed age is the true age rounded down to the nearest integer and the cutoff c is also an integer. This assumption may not hold for other types of rounding such as ordinary rounding or rounding up to the nearest integer, which I will discuss separately in the extension Section 8.

Assumption 6 says that the moments of the rounding error e do not depend upon X, and that these moments are identified. Later bounds will be derived in case these moments are not identified. In the case of age, these are essentially moments of the distribution of birthdays within a year (among individuals in the population where the survey sample is drawn), so Assumption 6 will hold if birthdays are uniformly distributed. In this case $\mu_k = \int_0^1 e^k de = 1/(k+1)$ is known. Note that, regardless of the distribution of e, all moments μ_k are finite, because e is bounded between zero and one by construction.

There exists evidence of small but statistically significant seasonal departures from uniformity in the distribution of births within a year (see, for example, Beresford, 1980; Murphy, 1996). However, this seasonal variation appears to have very little impact on the lower-order moments μ_k. For example, Murphy (1996) provides birthdays of 480,040 life insurance applicants. The first four empirical moments μ_1 to μ_4 in his data are 0.506, 0.339, 0.254 and 0.203, which are numerically quite close to the corresponding moments of a true uniform distribution, 0.500, 0.333, 0.250 and 0.200.[4]

Small departures from uniformity could also arise among very old populations, where those with birthdays earlier in the year may be slightly underrepresented due to the higher mortality risk. However, even without assuming a uniform distribution, these distributions may be estimated using data from other sources such as census data, so it is not restrictive to assume that moments from these distributions are identified.

Recall that $\tau = g_1(0) - g_0(0)$ is the true local RD treatment effect.

[3] Intuitively, the difference between the 'true' treatment effect and the 'naive' discrete data treatment effect depends on how the outcome function varies with e in the discretization cell for observations close to the cutoff.

[4] Data from other sources show similar empirical moments. For example, the first four empirical moments of the birth date distribution in the NLSY97 data are 0.501, 0.336, 0.252 and 0.201, and in the Italian anagraphic records data (2001–2011) are 0.507, 0.339, 0.255 and 0.203.

DOI: 10.1002/jae

Theorem 1. Let Assumptions 1–6 hold. Then τ can be identified even if X^* is not observed.

Theorem 1 says that the above assumptions are sufficient to identify the true local RD treatment effect τ. The only data these assumptions require are age cell means $E(Y \mid X = x)$ and moments $\mu_k = E(e^k)$. Given just these data one can consistently estimate τ.

Corollary 1 below describes the bias in the discrete-data RD treatment effect τ', and provides a general method for constructing a consistent estimator for τ using rounded age based on Theorem 1.

Given Assumptions 1 and 3, one can write the true data model as

$$Y = \sum_{j=0}^{J} a_j X^{*j} + \sum_{j=0}^{J} b_j X^{*j} T^* + \varepsilon^* \tag{1}$$

where $\varepsilon^* = Y^* - E(Y \mid X^*, T)$, $\sum_{j=0}^{J} a_j X^{*j} = g0(X^*)$ and $\sum_{j=0}^{J} b_j X^{*j} = g_1(X^*) - g0(X^*)$. Define $A = (a_0, a_1, \ldots, a_J)'$ and $B = (b_0, b_1, \ldots, b_J)'$. The true RD treatment effect is then $\tau = b_0$.

Corollary 1 below shows that the rounded data model has the same functional form as equation (1) but with different coefficients, so

$$Y = \sum_{j=0}^{J} d_j X^j + \sum_{j=0}^{J} c_j X^j T^* + \varepsilon \tag{2}$$

where $\varepsilon = Y - E(Y \mid X, T)$, $\sum_{j=0}^{J} d_j X^j = h_0(X)$, and $\sum_{j=0}^{J} c_j X^j = h_1(X) - h_0(X)$. Define $D = (d_0, d_1, \ldots, d_J)'$ and $C = (c_0, c_1, \ldots, c_J)'$. The naive discrete data RD treatment effect is then given by $\tau' = c_0$. Given the discrete data regression (2), the identification and estimation problem is to recover b_0, and more generally all the coefficients A and B in equation (1), from the coefficients D and C in equation (2).

Let $\binom{j}{k}$ denote the binomial coefficient $\frac{j!}{k!\,(j-k)!}$. Define the $J+1$ by $J+1$ matrix M as the upper triangular matrix that has the element $\binom{j}{k}\mu_{j-k}$ in row $k+1$ and column $j+1$ for all j, k satisfying $J \geq j \geq k \geq 0$, with all elements below the diagonal being zero. For the special case in which e is uniformly distributed in the range zero to one, the element of M in row $k+1$ and column $j+1$ will be $\binom{j}{k}\mu_{j-k} = \binom{j}{k}\frac{1}{j-k+1} = \frac{j!}{k!(j-k+1)!}$.

Corollary 1. Let Assumptions 1–6 hold. Then:

(i) Equation (2) holds, with D identified as the vector of coefficients of the polynomial $h_0(x)$ that goes through the points $x = -1, -2, \ldots, -J, -(J+1)$, and C is identified by $C = C_1 - D$, where C_1 is the vector of coefficients of the polynomial $h_1(x)$ that goes through the points $x = 0, 1, \ldots, J$.

(ii) The coefficients in the true underlying model A and B are identified by $A = M^{-1}D$ and $B = M^{-1}C$.

(iii) The true treatment effect is $\tau = b_0$, the naive discrete data treatment effect is $\tau' = c_0$, and the difference or the bias is $\tau' - \tau = \sum_{j=1}^{J} b_j \mu_j$.

Corollary 1 shows that the rounded data model is itself a polynomial, and that the matrix M connects the polynomial coefficients D and C in the rounded data model to the coefficients A and B in the true model by $MA = D$ and $MB = C$.

To illustrate the potential size of the rounding bias, and hence the size of the proposed correction, consider the case where the e distribution is uniform so $\mu_k = 1/(k+1)$. It then follows immediately from Corollary 1 that the bias in the discrete data estimator is

DOI: 10.1002/jae

$$\tau' - \tau = \frac{1}{2}b_1 + \frac{1}{3}b_2 + \ldots + \frac{1}{J+1}b_J$$

As noted earlier, the coefficient b_j for $j = 1, 2, \ldots, J$ is the change in the jth derivatives of $E(Y \mid X^*)$ at the threshold.

This bias formula shows that if the slope and higher derivatives of the conditional mean outcome do not change at the threshold, meaning that the treatment effect is locally constant, then the bias from rounding will be zero. Otherwise, the larger the changes in slope and the higher derivatives are at the threshold, the larger the rounding bias tends to be. Note that the rounding bias can result in either an overestimate or an underestimate of the true RD treatment effect, depending on the changes in the polynomial coefficients crossing the threshold. In particular, what appears to be a discontinuity in the rounded discrete data may not exist with continuous data. The next section describes how to estimate the true treatment effect from rounded discrete data, and test the significance of a true discontinuity.

The analysis in this section assumes that the threshold c is an integer, which means that by observing X instead of X^* one can still determine T^*, whether one is above or below the threshold, without error. In particular, X is non-negative if and only if X^* is non-negative, so $T^* = I(X^* \geq 0) = I(X \geq 0)$. For example, in Card *et al.* (2008) the treatment threshold is defined to be age 65 (the age of near-universal Medicare eligibility), so their data correctly sorts individuals into those who are above 65 and hence are eligible for Medicare from those who are not, even though they only observe age in years. Note that when the threshold is not an integer, $I(X^* \geq 0) = I(X \geq 0)$ may not hold. Cases like this are discussed in the extension Section 8.

These results can be extended immediately to applications involving other types of discretization or rounding, as long as they maintain this property of no mismeasurement in the crossing threshold indicator. Extensions to rounding involving mismeasurement of the crossing threshold status at the cutoff, i.e. $I(X^* \geq 0) \neq I(X \geq 0)$, are discussed in the extension Section 8.

4. ESTIMATION

This section describes how to apply Theorem 1 and Corollary 1 to estimate the true treatment effect τ with rounded data of the running variable. For simplicity, I first present the estimator for the case where the polynomial regressions are fourth (or lower)-order polynomials, which should cover most actual empirical applications. I then describe the general estimation method for any order polynomials.

Given n observations $\{X_i, T_i^*, Y_i\}$ for $i = 1, 2, \ldots, n$, the first step is to estimate the following equation:

$$Y_i = d_0 + d_1 X_i + d_2 X_i^2 + d_3 X_i^3 + d_4 X_i^4 + \left(c_0 + c_1 X_i + c_2 X_i^2 + c_3 X_i^3 + c_4 X_i^4\right) T_i^* + \varepsilon_i \quad (3)$$

which is equation (2) with $J = 4$. Assuming all the assumptions hold conditioning on covariates, one can add other covariates to the model if desired.

The naive discrete data treatment effect τ' will just be c_0 in this regression (3). However, by applying Theorem 1 and Corollary 1 (detailed derivations are provided as supporting information in the supplemental online Appendix), if the distribution of ages within a year (for individuals in the population from which the data are drawn) is uniform, then the true treatment effect τ is given by

$$\tau = c_0 - \frac{1}{2}c_1 + \frac{1}{6}c_2 - \frac{1}{30}c_4 \quad (4)$$

More generally, the true treatment effect is

$$\tau = c_0 - \mu_1 c_1 + \left(2\mu_1^2 - \mu_2\right) c_2 + \left(-6\mu_1^3 + 6\mu_2\mu_1 - \mu_3\right) c_3 + \left(24\mu_1^4 - 36\mu_1^2\mu_2 + 8\mu_3\mu_1 + 6\mu_2^2 - \mu_4\right) c_4 \quad (5)$$

DOI: 10.1002/jae

where $\mu_j = E\left(e^j\right)$ for $j = 1, \ldots, 4$, and the distribution of e is the distribution of ages within a year, on a scale of zero to one, where zero means a birth at the beginning of the first day of the year, and one means a birth at the end of the last day. For lower-order polynomial models these same formulas can be used, by just setting the higher-order coefficients equal to zero.

Since the bias-corrected treatment effect is a linear combination of regression coefficients, standard methods, such as bootstrapping and the delta method, can be used to obtain standard errors. More generally, the corrected treatment effect is estimated based on equation (5), where the moments are estimated, then one also needs to take into account the estimation error of the moments (details are provided in the online Appendix).

Equation (4) or more generally (5) can be used to test for the presence of rounding bias, by applying an ordinary t-test to the hypothesis that the bias $\tau' - \tau = 0$. For example, when e has a uniform distribution and the polynomial is degree $J \leq 4$, by equation (4) one would just need to test if $c_1/2 - c_2/6 + c_4/30$ equals zero. A sufficient condition for no rounding bias when the polynomial $J \leq 4$ is that c_1 , c_2, c_3 and c_4 equal zero, so even if we did not know moments of the distribution of e, one could still do a standard F-test of the joint hypothesis that the regression coefficients equal zero, i.e. $c_1 = c_2 = c_3 = c_4 = 0$.

More generally, for higher-order polynomials, one may estimate the regression in equation (2) by ordinary least squares and apply the bias correction in Corollary 1 (ii).

Bounds can be constructed directly based on Theorem 1 and Corollary 1 when the distribution of the rounding error is unknown. In particular, rounding errors e lie between 0 and 1, which implies that non-central moments $\mu_j = E(e^j)$ satisfy $1 > \mu_1 \geq \mu_2 \geq \ldots \geq \mu_j \geq 0$. For example, in a quadratic model, without knowing anything about the distribution of rounding errors (other than its support), lower and upper bounds on the treatment effect τ are given by the minimum and maximum of $c_0 - \mu_1 c_1 + \left(2\mu_1^2 - \mu_2\right) c_2$ over the set $1 > \mu_1 \geq \mu_2 \geq 0$. Given estimates of each c_j, one can easily search over this set of values of μ_1 and μ_2 to obtain bounds. In the case of a linear model, the bounds are given by c_0 and $c_0 - c_1$, so the sign of the bias is entirely determined by the sign of c_1, the slope change at the cutoff (further discussion can be found in the online Appendix).

5. FUZZY DESIGNS

Continue to let T be the indicator of whether one is treated or not, and let $T^* = I\left(X^* \geq 0\right)$ be the crossing threshold indicator. Unlike the sharp design, the fuzzy design RD no longer assumes $T = T^*$.

Under well-known standard conditions, the fuzzy design RD local treatment effect τ_f is given by

$$\tau_f = \frac{\tau_Y}{\tau_T} \tag{6}$$

where $\tau_Y = E\left(Y \mid X^* = 0, T^* = 1\right) - E\left(Y \mid X^* = 0, T^* = 0\right)$, which is the size of the jump or discontinuity in the mean outcome at the threshold, and $\tau_T = E\left(T \mid X^* = 0, T^* = 1\right) - E\left(T \mid X^* = 0, T^* = 0\right)$, which is the size of the jump in the treatment probability at the threshold, and $0 < \tau_T < 1$.

Both the numerator and the denominator may involve rounding bias. The estimator of the previous sections can therefore be immediately extended to estimation of fuzzy design treatment effects. First apply the exact same estimator as in the sharp design case to obtain a consistent estimator $\widehat{\tau}_Y$ of τ_Y. Then replace Y with T and apply the exact same estimator again to obtain a consistent estimator $\widehat{\tau}_T$ of τ_T. The fuzzy design treatment effect estimator is then $\widehat{\tau}_Y/\widehat{\tau}_T$.

To illustrate the fuzzy design estimator, consider the case where the functional forms for $E\left(Y \mid X^*, T^* = t\right)$ and $E\left(T \mid X^*, T^* = t\right)$ are fourth-order polynomials, so $J = 4$.[5] Then one can

[5] The orders of the polynomials in these two equations need not be the same, and these fourth-order polynomial formulas cover lower-order polynomials as special cases by setting higher-order coefficients equal to zero.

DOI: 10.1002/jae

use least squares to estimate the coefficients in the outcome regression specified in equation (3) and a similarly specified treatment regression:

$$T_i = r_0 + r_1 X_i + r_2 X_i^2 + r_3 X_i^3 + r_4 X_i^4 + \left(s_0 + s_1 X + s_2 X_i^2 + s_3 X_i^3 + s_4 X_i^4\right) T_i^* + \epsilon_i \quad (7)$$

Then, assuming a uniform distribution for e, the fuzzy design treatment effect will be given by

$$\tau_f = \frac{c_0 - \frac{1}{2}c_1 + \frac{1}{6}c_2 - \frac{1}{30}c_4}{s_0 - \frac{1}{2}s_1 + \frac{1}{6}s_2 - \frac{1}{30}s_4} \quad (8)$$

This is in contrast to the incorrect discrete data treatment effect $\tau'_f = \tau'_Y/\tau'_T = c_0/s_0$. Therefore the size of the bias in a fuzzy design RD depends on the rounding bias in the numerator and that in the denominator. If the change in slopes above and below the threshold in the numerator is opposite in sign compared with that in the denominator, then the impact of rounding error on the ratio will be magnified. The second empirical application in Section 7 illustrates this point. Similar to the case of sharp design, standard errors can be obtained either by bootstrapping or by the delta method, after jointly estimating the reduced form outcome and treatment equations.

Bounds can also be similarly constructed as in the case of sharp design. For example, in the case of linear models, it can be shown that the bounds for the correct treatment effect are given by $\frac{c_0}{s_0}$ and $\frac{c_0 - c_1}{s_0 - s_1}$ when $\mu_1 = 0$ and $\mu_1 = 1$, respectively.

6. MEDICARE AND INSURANCE COVERAGE

This section provides an empirical application examining the impact of qualifying for Medicare at age 65 on the health insurance coverage rate in the USA. Others have applied RD analyses to Medicare qualification (see, for example, Card *et al.*, 2008). I use this example because the presence of a discontinuity in the insurance rate due to the treatment (Medicare eligibility) is uncontroversial, and because the available data can be used to verify the accuracy of the proposed method for correcting rounding bias.

The data used are from the US Health and Retirement Study (HRS). The HRS is a national panel survey of individuals over age 50 and their spouses in the USA. It has extensive information on health insurance, health, employment, demographics, etc. Data have been collected every 2 years since 1992. Nine waves of data have been released so far. The HRS is suitable because it focuses on the elderly, covering an age period particularly relevant to the empirical analysis here. More importantly, the HRS has available both age in months and age in years, so one can compare the bias-corrected estimates based on yearly age data with estimates based on monthly age data, and thereby empirically evaluate how well the proposed correction works. This paper uses all waves of data. After observations with missing values deleted, the final samples have 60,290–135,582 observations, depending on the age ranges examined.

Let the outcome Y be the dummy indicating whether one has any health insurance, the cutoff c be age 65 and X be the reported age minus 65. The (sharp design) treatment $T = T^*$ then corresponds to crossing age 65 and thereby becoming eligible for Medicare.

Figures 1 and 2 show the age profiles of health insurance rates, i.e. the age cell means of Y against age in years and in months, respectively. These figures clearly show a jump in insurance coverage at age 65. To model this treatment effect, I fit second-, third- and fourth-order polynomials to annual data. The quadratic model appears to underfit the model, while the fourth-order polynomial tends to overfit especially for the narrower age ranges considered (both graphically and in terms of statistical significance of higher-order terms as well as the adjusted $\overline{R}^2$ of the regressions). I therefore focus on the third-order polynomial as the preferred model (i.e. equation (3) with c_4 and d_4 set to zero), though

DOI: 10.1002/jae

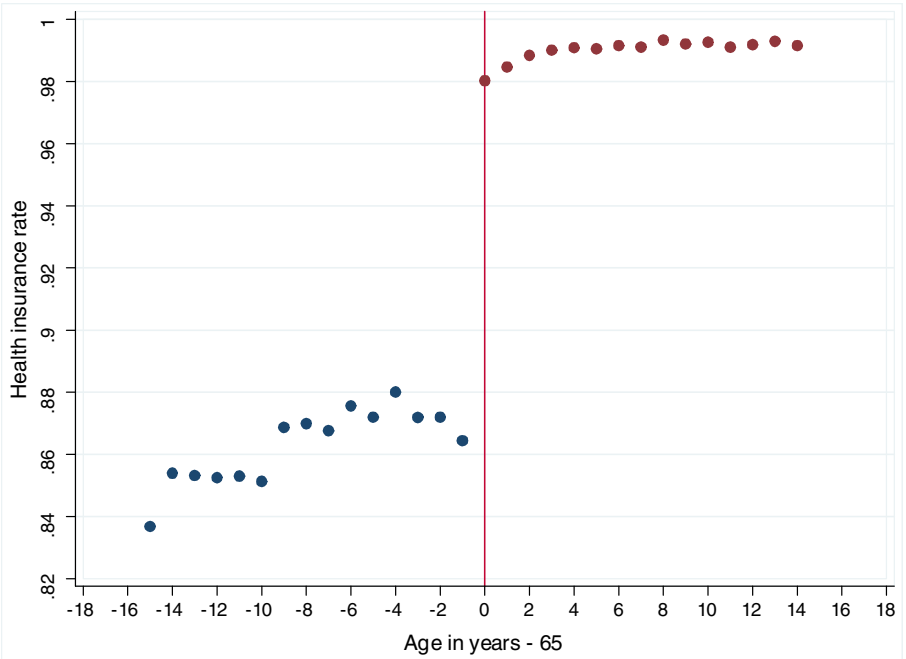

Figure 1. The age (in years) profile of health insurance coverage rates, HRS 1992–2008

estimates using both third- and fourth-order polynomials are reported for comparison. Attempts to include terms of degree five or more are completely insignificant and including these higher-order terms does not improve overall fit of the model.

In practice, there is a trade-off regarding what age range of data around the threshold to include in the model. A wider age range provides more observations, thereby adding to the precision with which the model coefficients can be estimated. However, the further the included age are from the threshold, the more likely it is that the correct specification for these distant observations will differ from the correct specification near the threshold, risking specification errors. I consider four ranges of data, specifically, 6, 9, 12 and 15 years below and above the threshold, corresponding to age ranges 59–70, 56–73, 53–76 and 50–79. Note that the smallest window width here is less than half the largest window width.

Another specification issue is inclusion of covariates. To assess the impact of covariates, I estimate models that include year of survey dummies with or without additional demographic characteristics such as gender, race (white/non-white), ethnicity (Hispanic/non-Hispanic) and education levels. Three education levels represent less than high school (the default), high school or GED (General Educational Development), and college or above.

The results are reported in Table I. For each specification, the top panel in Table I presents the naive discrete data estimates corresponding to $\tau' = c_0$, and the bottom panel represents the bias-corrected estimates corresponding to τ in equation (4). Estimates based on monthly age data are reported in the middle panel. Clustered standard errors are reported, taking into account the panel structure of the data. All of the reported estimates are statistically significant at the 1% level.

For the preferred third-order polynomial, the results are similar across different specifications, i.e. controlling for different covariates or using different ranges of data. Note that the monthly data estimates are systematically smaller than the yearly data estimates, which implies that rounding of age by years in these data results in an overestimate of the impact of the Medicare program. In contrast,

DOI: 10.1002/jae

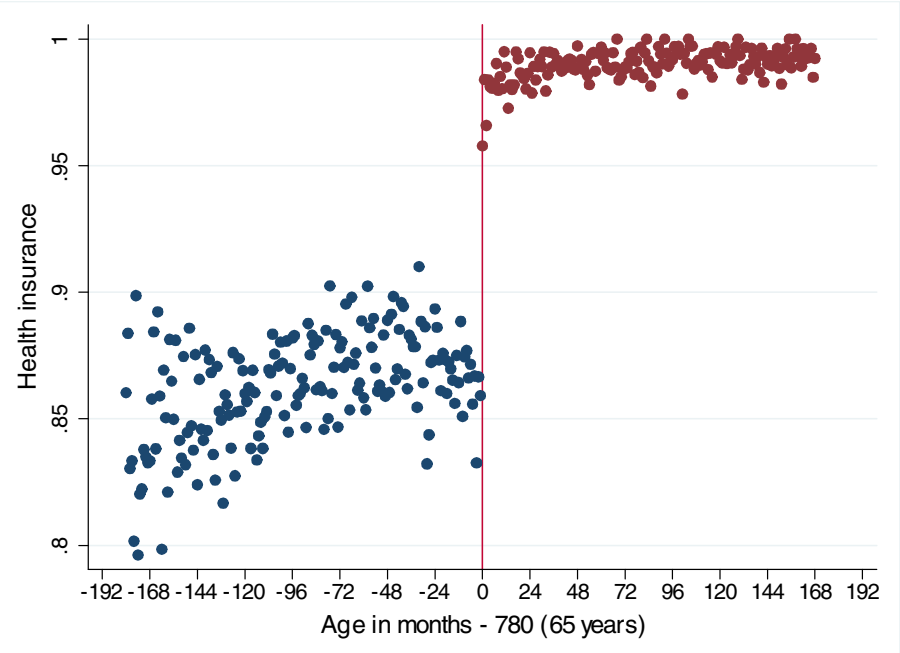

Figure 2. The age (in months) profile of health insurance coverage rates, HRS 1992–2008

the bias-corrected annual data estimates are all close to, and slightly smaller than, the monthly data estimates. Thus going from annual to monthly data appears to correct most but not all of the bias associated with rounding. This is what one would expect if the proposed model and bias-correcting methodology are valid.

When using the annual age data, the discrete data treatment effect τ' is estimated to be in the range of 0.124–0.128, which means that health insurance coverage rate increases by 12.4–12.8% due to individuals qualifying for Medicare. In contrast, the estimates based on age in months are on average about 5% lower, in the range of 0.118–0.121. The bias-corrected estimates are in a similar range of 0.117–0.119, averaging about 6% lower than estimates using yearly age data.

Table A1 in the online Appendix reports the estimated biases and their standard errors. These estimates range mostly from 0.005 to 0.010, representing 4–9% upward biases. For the third-order polynomials, most of these estimates are significant at the 1% level. For the fourth-order polynomials, the estimates are significant only when using the longer 15 years window. This is not surprising because the fourth-order polynomial is imprecisely estimated with shorter windows.

Statistically, the bias due to rounding is relatively small in percentage terms in this application. This is not surprising given the fact that the slope of the health insurance profile changes very little at the threshold age (as clearly shown by Figures 1 and 2, and by the coefficient estimates). As a result, the leading term c_1 in the correction expression $\tau - \tau' = -\frac{1}{2}c_1 + \frac{1}{6}c_2 - \frac{1}{30}c_4$ is quite small. Given that the correction does not make much difference, one can expect that using exact birth date information would not significantly improve the estimates in this case.

Economically, failing to correct for the rounding bias results in an overestimate of insurance coverage of 0.5–1.0% of the relevant population. The current population of the USA that is over age 65 and hence qualifies for Medicare is approximately 38 million (according to the US Census), so even half of 1% of this total is a large number of people.

DOI: 10.1002/jae

Table I. Estimated increases in health insurance coverage rate at the Medicare eligibility age 65

	Third-order polynomial			Fourth-order polynomial		
	(1)	(2)	(3)	(1)	(2)	(3)
Naive estimates using age in years						
$[-6,+6)$	0.128	0.125	0.124	0.107	0.106	0.107
	(0.015)***	(0.015)***	(0.015)***	(0.035)***	(0.034)***	(0.035)***
$[-9,+9)$	0.128	0.128	0.127	0.124	0.119	0.121
	(0.009)***	(0.009)***	(0.009)***	(0.016)***	(0.016)***	(0.014)***
$[-12,+12)$	0.126	0.127	0.126	0.120	0.119	0.119
	(0.007)***	(0.007)***	(0.007)***	(0.011)***	(0.010)***	(0.011)***
$[-15,+15)$	0.124	0.126	0.126	0.129	0.128	0.129
	(0.006)***	(0.006)***	(0.006)***	(0.009)***	(0.009)***	(0.009)***
Naive estimates using age in months						
$[-6,+6)$	0.119	0.118	0.119	0.112	0.113	0.113
	(0.008)***	(0.008)***	(0.008)***	(0.011)***	(0.011)***	(0.011)***
$[-9,+9)$	0.119	0.120	0.120	0.116	0.115	0.115
	(0.007)***	(0.006)***	(0.007)***	(0.008)***	(0.008)***	(0.009)***
$[-12,+12)$	0.119	0.120	0.121	0.116	0.117	0.116
	(0.006)***	(0.006)***	(0.006)***	(0.007)***	(0.007)***	(0.007)***
$[-15,+15)$	0.119	0.121	0.120	0.119	0.120	0.120
	(0.005)***	(0.005)***	(0.005)***	(0.006)***	(0.006)***	(0.006)***
Bias-corrected estimates using age in years						
$[-6,+6)$	0.118	0.117	0.117	0.112	0.113	0.113
	(0.009)***	(0.009)***	(0.009)***	(0.014)***	(0.014)***	(0.014)***
$[-9,+9)$	0.117	0.117	0.118	0.116	0.115	0.116
	(0.006)***	(0.006)***	(0.006)***	(0.009)***	(0.009)***	(0.009)***
$[-12,+12)$	0.118	0.119	0.119	0.114	0.113	0.114
	(0.006)***	(0.006)***	(0.006)***	(0.007)***	(0.007)***	(0.007)***
$[-15,+15)$	0.117	0.118	0.119	0.119	0.120	0.120
	(0.005)***	(0.005)***	(0.005)***	(0.006)***	(0.006)***	(0.006)***

Note: Estimates are based on HRS 1992–2008; (1) does not control for covariates; (2) controls for year dummies; (3) controls for year dummies and additional demographic variables. Bottom panel, bias-corrected estimates, applying formula in equation (4). Robust clustered standard errors are calculated by the delta method; *significant at 10% level; **significant at 5% level; ***significant at 1% level.

Table II. Bounds for the bias-corrected estimates of health insurance rate increase at 65

	(1)			(2)			(3)		
$[-6,+6)$	0.128	(0.109,	0.128]	0.125	(0.111,	0.125]	0.124	(0.111,	0.124]
$[-9,+9)$	0.128	(0.108,	0.128]	0.127	(0.109,	0.127]	0.127	(0.109,	0.127]
$[-12,+12)$	0.126	(0.111,	0.126]	0.127	(0.113,	0.127]	0.126	(0.111,	0.126]
$[-15,+15)$	0.124	(0.111,	0.124]	0.126	(0.112,	0.126]	0.126	(0.112,	0.126]

Note: Estimates are based on HRS 1992–2008; all estimates utilize third-order polynomials; bounds are provided in parentheses next to the naive estimates; (1) does not control for covariates; (2) controls for year dummies; (3) controls for year dummies and additional demographic variables.

Table II reports the bounds on the bias-corrected estimates for the preferred third-order polynomial. Since the slope change dominates other higher-order derivative changes, not surprisingly these bounds are narrow, ranging from 0.013 to 0.02 in width. Also, the naive uncorrected estimates are the upper bounds of the correct estimates, implying an overestimation of the true effect of Medicare eligibility by using rounded age in years.

DOI: 10.1002/jae

7. THE RETIREMENT-CONSUMPTION PUZZLE IN CHINA

This section applies the proposed approach to investigating consumption changes around retirement in China. Standard life cycle models suggest that rational people smooth consumption over the life cycle, so consumption should not change at retirement when retirement is expected. However, many empirical studies find that consumption (typically food consumption) drops significantly at retirement. This finding is referred to as the 'retirement-consumption puzzle'.

Evidence of this puzzle has been mostly obtained from developed Western countries, including the UK (Banks *et al.*, 1998), the USA (Bernheim *et al.*, 2001; Aguila *et al.*, 2011; Ameriks *et al.*, 2007; Haider and Stephens, 2007; Hurd and Rohwedder, 2008), Canada (Robb and Burbridge, 1989), Germany (Schwerdt, 2005) and Italy (Battistin *et al.*, 2009; Borella *et al.*, 2011). Evidence from developing countries is scanty.

Most analyses of the retirement-consumption puzzle depend on structural models. One exception is Battistin *et al.* (2009), who estimate RD models that exploit pension eligibility rules in Italy.

This section conducts an RD analysis of the retirement-consumption puzzle in China, taking advantage of the Chinese mandatory retirement rule. Since age is reported in years in the dataset used here, I apply the proposed approach to correct the associated rounding bias. The Chinese case is interesting due to its unique social and cultural environment, which differs in many ways from developed Western countries.

In China, the official retirement age is 60 for male workers, 55 for white-collar female workers and 50 for blue-collar female workers, with some exceptions applying to certain occupations and to disabled workers.[6] These mandatory retirement ages have not changed ever since the retirement system was founded in the 1950s. Compared with pension eligibility rules, the mandatory retirement policy in China induces a larger change in the retirement probability and hence helps more precisely identify the causal impact of retirement on consumption.

The analysis here focuses on male workers, because female workers' labor supply is more complicated and their mandatory retirement age depends on the types of their work. I look at food consumption, because so far the evidence of the 'retirement-consumption puzzle' is mostly about food consumption declines. Similarly, I find that in the Chinese dataset food consumption declines, but not other categories of consumption. The sample includes all urban male household heads who are labor force participants, so, for example, homemakers are not included. Some workers may retire earlier than the mandatory retirement age, and some may be re-employed after the official retirement. Also, the mandatory retirement policy may not be strictly enforced in the private sector compared to the state sector, including state-owned enterprises (SOEs) and government units. As a result, the change in the retirement rate is less than one at 60, which entails fuzzy design RD models.

Data in this analysis are from the China Urban Household Survey (UHS), and are collected by the National Bureau of Statistics (NBS) every year to monitor consumption in China and to construct the consumer price index (CPI). Complete data from five provinces and one municipality from 1997 to 2006 are used.[7] The UHS questionnaires changed a few times over the years, but the survey questions are relatively consistent for the period 1997–2006. In addition, the pension system in China changed in 1997. In particular, the Chinese government adopted a system that combines individual

[6] Those who have jobs that are risky, harmful to their health or extremely physically demanding can retire 5 years before the official retirement ages, i.e. 45 for blue-collar female workers and 55 for male workers. Male workers who become disabled and hence are unable to do their work can apply to retire at 50, while disabled female workers can retire at 45. Civil servants also qualify for early retirement if they have worked for 30 years and are within 5 years of their retirement age.

[7] The five provinces are Liaoning, Zhejiang, Guangdong Shanxi and Sichuan, and the one city is Beijing.

DOI: 10.1002/jae

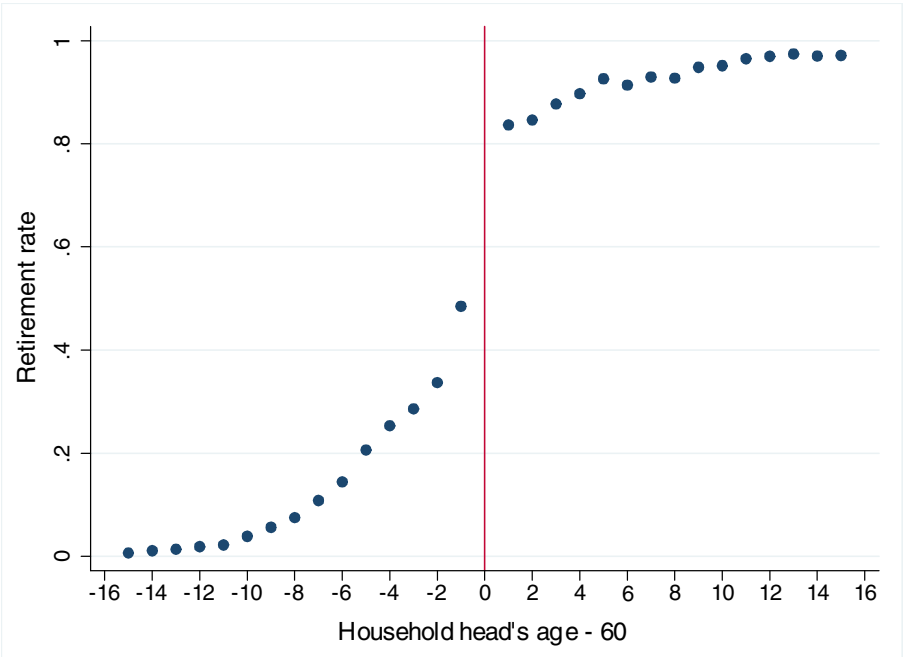

Figure 3. The age (in years) profile of retirement rates for male household heads, UHS 1997–2006

accounts and social pooling to provide retirement funds. Before that, pensions were provided entirely by employers.[8]

In China, eligible male workers can start their retirement paperwork at the beginning of the month they turn 60. Typically the paperwork is processed right away and eligible workers start to receive their pension the following month after they turn 60.

For the sample period this paper looks at, individual employment status reflects that in the last month.[9] In theory, a household head can retire any time during the year, so household consumption at 60 is generally a mixture of pre- and post-retirement consumption. I therefore exclude observations at the cutoff age 60 in the estimation, assuming that the retirement rate change induced by the mandatory retirement policy is fully realized at 61. This ensures that all individuals who are observed below the cutoff age of 60 are drawn from the pre mandatory retirement age profile $h_0(X)$, and all the individuals who are observed above the cutoff age are drawn from the post mandatory retirement age profile $h_1(X)$. I then estimate the polynomial models using data from ages 59 and below and ages 61 and above, and evaluate changes at 60 by extrapolating these regression curves to the cutoff age of 60.

Figures 3 and 4 show the age profiles (age cell means) of the retirement rate and the logarithm of household food expenditure. Food expenditure is in 1996 constant Chinese Yuan. The retirement profile shows an obvious jump and a mild slope change crossing the retirement cutoff age 60 (normalized to 0 in the figures). The jump represents an exogenous change in the retirement rate induced by the retirement policy and provides identification of the retirement impact on food

[8] Including or excluding the implementation year 1997 does not make much difference in the estimation results. This could be because individuals' retirement status is recorded at the end of the year, which should be after the reform. Also, all specifications control for year fixed effects, which should pick up mean differences across years.

[9] In particular, information on employment status is collected and updated every month, starting on the 21st of the previous month to the 20th of the current month.

DOI: 10.1002/jae

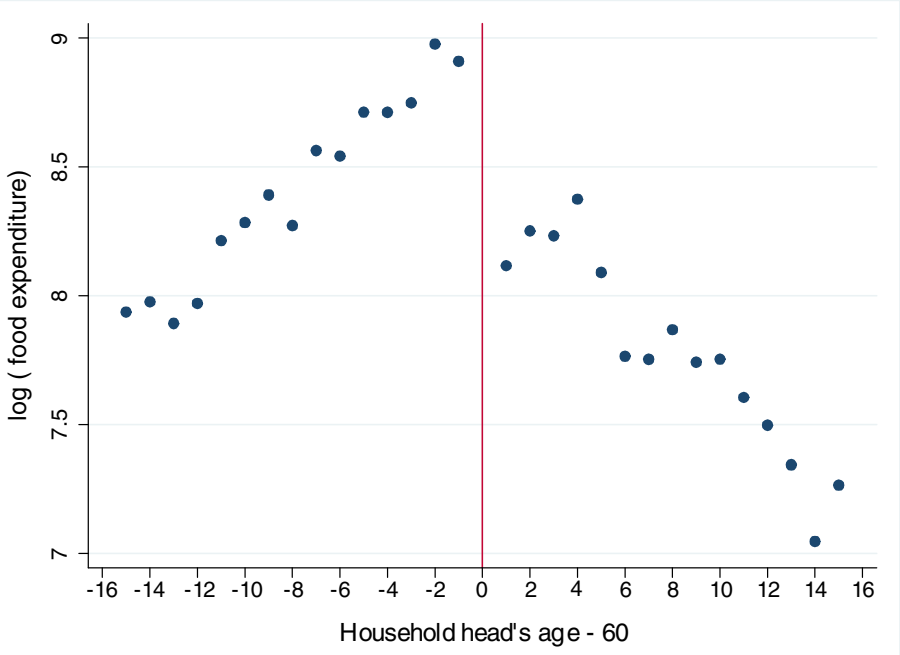

Figure 4. The age (in years) profile of log food expenditure, UHS 1997–2006

consumption. The modest slope change implies that rounding bias in the retirement rate jump at the threshold may not be zero, but is possibly rather small.

In contrast to the retirement profile, the food consumption profile shows an obvious drop crossing the threshold age, along with a substantial change in slope. Before 60, food consumption increases steadily with age, while after 60 food consumption declines rapidly. Given this large change in slope, one should expect substantial rounding bias in the estimated change in food consumption at the retirement age of 60.

Define the outcome Y as the logarithm of household food expenditure. Let T be a dummy indicating whether a household head has retired or not. Let c be the cutoff age 60, and X be the recorded age in years minus 60. Log food consumption Y and retirement T are specified as polynomial models as in equations (2) and (7). These models are estimated using three different window widths, i.e. 6, 10 and 15 years above and below the cutoff. The sample sizes corresponding to the three window widths are 12,866, 22,296 and 33,754, respectively. In particular, linear models (setting d_p and c_p for $p = 2, 3, 4$ in equation (2) to zero) are adopted for log food consumption, while third-order polynomials (setting r_4 and s_4 in equation (7) to zero) are used for the retirement rate, except that when using the short 6 years window a quadratic model is adopted in that case.[10] These polynomial orders are chosen based on goodness-of-fit measures and significance of the coefficients on higher-order terms.

The estimation results are reported in Table II. The top panel in Table II presents the uncorrected (or naive) estimates, while the bottom panel presents the bias-corrected estimates. The uncorrected and bias-corrected retirement effects in this case are given by $\tau'_f = c_0/s_0$ and $\tau_f = \left(c_0 - \frac{1}{2}c_1\right) / \left(s_0 - \frac{1}{2}s_1 + \frac{1}{6}s_2\right)$, respectively. I also try controlling for different covariates. The estimates on the right side of Table III (noted as (1) in the table) control for year fixed effects, family size, family size squared and education levels, including college or above, high school and less than

[10] Adopting a quadratic model for the retirement equation in this case means also setting s_3 to zero.

DOI: 10.1002/jae

Table III. Effects of retirement on food consumption at the mandatory retirement age 60

	(1)			(2)		
	(a)	(b)	(a)/(b)	(a)	(b)	(a)/(b)
Naive estimates						
$[-6, +6]$	**−0.046**	0.193	**−0.237**	**−0.041**	0.191	**−0.213**
	(0.017)***	(0.024)***	(0.085)***	(0.016)**	(0.024)***	(0.085)**
$[-10, +10]$	**−0.054**	0.187	**−0.289**	**−0.054**	0.188	**−0.286**
	(0.013)***	(0.022)***	(0.078)***	(0.013)***	(0.022)***	(0.078)***
$[-15, +15]$	**−0.054**	0.211	**−0.257**	**−0.055**	0.209	**−0.261**
	(0.011)***	(0.014)***	(0.049)***	(0.010)***	(0.014)***	(0.048)***
Bias-corrected estimates						
$[-6, +6]$	**−0.034**	0.215	**−0.157**	**−0.029**	0.214	**−0.134**
	(0.017)**	(0.022)***	(0.075)**	(0.016)**	(0.022)***	(0.074)*
$[-10, +10]$	**−0.044**	0.207	**−0.213**	**−0.045**	0.207	**−0.219**
	(0.013)***	(0.020)***	(0.061)***	(0.013)***	(0.020)***	(0.061)***
$[-15, +15]$	**−0.044**	0.233	**−0.188**	**−0.046**	0.232	**−0.198**
	(0.011)***	(0.014)***	(0.042)***	(0.011)***	(0.014)***	(0.042)***

Note: Estimates are based on male household heads, UHS 1997–2006; (a) represents change in the log food consumption at 65; (b) represents change in the retirement rate at 65; (a)/(b) represents the effect of retirement on food consumption; (1) controls for year dummies, family size, family size squared, and education levels; (2) only controls for year dummies. Bootstrapped standard errors are in the parentheses; *significant at the 10% level; **significant at the 5% level; ***significant at the 1% level.

high school (the default). As a comparison, the estimates on the left half of the table (noted as (2) in Table III) control only for year fixed effects.

The preferred specification is the one that uses data 10 years above and below the cutoff age of 60 and controls for the full set of covariates as discussed above. Household food consumption crucially depends on family size and permanent income (proxied by education levels here), so including these covariates can help reduce the sample variation in log food consumption and hence provides more precise estimates.

The uncorrected estimates of the retirement effect range from −0.213 to −0.289, representing a drop of 21.3% to 28.9% in food consumption at retirement among those male household heads who retire due to the mandatory retirement policy. In contrast, the bias-corrected estimates indicate a smaller 13.4% to 21.9% drop in food consumption at retirement, so accounting for rounding of age leads to a decrease of over one fourth in the estimated retirement effects on food consumption. Note that in this case the denominator is biased down and the numerator biased up, so the impact of rounding error on the ratio is magnified.

Table A2 in the online Appendix shows the estimated biases and their bootstrapped standard errors. The estimated biases range from 0.064 to 0.080. All are statistically significant at the 1% level. These estimates indicate that the naive uncorrected estimates overestimate the retirement effect on food consumption by 6.4–8.0 percentage points.

Table IV gives the bounds on the bias-corrected estimates. The naive estimates are the lower bounds, so ignoring the rounding error overestimates the negative effect of retirement on consumption.

Overall, the estimated retirement effects are consistent with the existing evidence documented for many developed Western countries, i.e. food consumption drops significantly when male household heads retire at the mandatory retirement age, and that households do not seem to smooth food expenditures at retirement even though the age of retirement is fully anticipated. RD models using age in years as the running variable overestimate the food consumption drop at retirement. Because of the substantial change in the slope of the food consumption profile around the

DOI: 10.1002/jae

Table IV. Bounds for the bias-corrected estimates of the retirement effects on consumption

	(1)			(2)	
$[-6,+6)$	-0.237	$[-0.237,$	$-0.095)$	-0.213	$[-0.213,$	$-0.073)$
$[-10,+10)$	-0.289	$[-0.289,$	$-0.160)$	-0.286	$[-0.286,$	$-0.172)$
$[-15,+15)$	-0.257	$[0.257,$	$-0.136)$	-0.261	$[-0.261,$	$-0.150)$

Note: Estimates are based on male household heads, UHS 1997–2006; bounds are provided in parentheses next to the naive estimates; (1) controls for year dummies, family size, family size squared, and education levels; (2) only controls for year dummies.

mandatory retirement age, correcting for the rounding bias has sizable effects on the estimated retirement effects in this case. Applying the proposed correction appears to be both statistically and economically important.

8. EXTENSIONS: OTHER FORMS OF ROUNDING OR NON-INTEGER THRESHOLD

So far, the analysis has focused on rounding down to the nearest integer, as in the case of how age is typically reported. However, Theorem 1 and Corollary 1 do not actually specify or require X to be X^* rounded down to the nearest integer. In particular, the assumptions that involve X are Assumptions 4, 5 and 6. While these assumptions are plausible for discretization based on rounding down, they do not require this type of rounding, and they may be applied to other types of rounding.

Still, Assumption 5 requires that there should be no mismeasurement in the crossing threshold dummy $I(X \geq 0)$. This may not hold in other common types of rounding, such as rounding up or ordinary rounding, i.e. rounding either up or down, whichever is closer. In the following I discuss these alternative forms of rounding and provide simple extensions of the previous approach to handle these cases. In particular, I show that one can simply discard observations at the cutoff, because the crossing threshold dummy is mismeasured only at that point, i.e. the true crossing threshold dummy $I\,(X^* \geq 0 \mid X = 0)$ could be 0, while the observed crossing threshold dummy $I(X \geq 0 \mid X = 0)$ is always 1. In these cases, observations at the cutoff contains both above- and below-threshold outcomes, i.e. they contains data generated by both the pre- and post-cutoff regression functions, $h_0(X)$ and $h_1(X)$.

To illustrate, suppose that age is now recorded by ordinary rounding and that the threshold c is age 65. Then individuals who are over 64.5 and under 65 will have their true crossing threshold status $I(X^* \geq 0) = 0$, while at the same time they will have their recorded age be 65 (based on ordinary rounding) and hence their observed crossing threshold status $T^* = I(X \geq 0) = 1$. These individuals are misclassified regarding their crossing threshold status. This will tend to bias downward the treatment probability change at the cutoff.

Discretization by ordinary rounding or rounding up with an integer cutoff can only cause $I\,(X^* \geq 0) \neq I\,(X \geq 0)$ at $X = 0$. In the above example, by ordinary rounding, everyone over age 65.5 will have both their true age and their rounded age be above the cutoff, and hence both X^* and X positive. Similarly, everyone strictly under age 64.5 will have both their true age and their rounded age be below the cutoff and hence both X^* and X negative. Discarding observations at the cutoff can ensure that one only uses observations truly above the cutoff to estimate $h_0(X)$ and those truly below the estimate $h_1(X)$.

Another way in which rounding can cause the crossing threshold dummy to be mismeasured is when the running variable is rounded to integer values while the threshold c is not an integer. For example, the age at which people born in the years 1938–1942 qualify for full social security benefits in the USA (called the full retirement age by the social security administration) ranges from 65 years and 2

DOI: 10.1002/jae

months to 65 years and 10 months. In particular, for those who were born in 1939, the full retirement age is 65 years and 4 months, i.e. 65.33 years. Individuals who are 65.33–66 years old will have passed the full retirement age, given their recorded age of 65 (assuming rounding down), yet they will be mistaken as still being below the cutoff.

Note that X is normalized by subtracting the cutoff c, and so will be non-integer valued if the cutoff is a non-integer. Assuming rounding down, in this case $I(X \geq 0) \neq I(X^* \geq 0)$ only for the one observable value of X right under the cutoff, i.e. the one value that lies in the interval $-1 < X < 0$, because the observations that have X^* right above the non-integer cutoff will be rounded down to below the cutoff. Similarly, if X is discretized by always rounding up, then $I(X \geq 0)$ can fail to equal $I(X^* \geq 0)$ only for the one observable value of X right above the cutoff, i.e. the one value that lies in the interval $0 < X < 1$. Further, if X is discretized by ordinary rounding, then $I(X \geq 0)$ can fail to equal $I(X^* \geq 0)$ for the one observable value of X either right under or right above the cutoff, depending on whether the cutoff is positive or negative.

In all these cases, the observed outcomes for X at or right next to the cutoff contain a mix of observations whose true X^* is right above and right under the cutoff, so treating them as if they are all at or above the cutoff or all below the cutoff leads to a biased estimate of the true treatment effect, in addition to the rounding bias that involves all points away from the cutoff.

Another way to understand the problem of mismeasuring the crossing threshold dummy is to think of its role as an instrumental variable (IV) in RD models. It is well known that the standard fuzzy design RD estimator can be interpreted as a local IV estimator, using the crossing threshold dummy $I(X^* \geq 0)$ as an instrument for the treatment T. With these alternative types of rounding, the observed crossing threshold dummy $I(X \geq 0)$ is mismeasured relative to the true instrument $I(X^* \geq 0)$, and this mismeasurement will introduce bias in the estimated treatment effect, in addition to the bias caused by rounding as in Theorem 1.

To consistently handle all the above cases, consider the following simple extension to Theorem 1 and Corollary 1.

Corollary 2. Let Assumptions 1–4 and 6 hold. Assume that if $X \geq 1$, then $X^* > 0$, and that if $X \leq -1$, then $X^* < 0$, then the conclusions of Theorem 1 and Corollary 1 hold, replacing equation (2) with

$$Y = \sum_{j=0}^{J} d_j X^j + \sum_{j=0}^{J} c_j X^j T^* + \varepsilon \text{ for all } X \geq 1 \text{ or } X \leq -1 \tag{9}$$

Since these alternative forms of rounding cause trouble only for observations at one value of X such that $-1 < X < 1$, Corollary 2 states that one can fix the problem by just discarding those observations from the estimation. A similar approach, dropping observations for which the running variable is mismeasured in an RD model, has been proposed by Barreca *et al.* (2010). In particular, Barreca *et al.* (2010) find that birth weights are disproportionately represented at multiples of round numbers (i.e. 100 g and ounce multiples), which caused biased RD treatment effect estimates when using birth weight as a running variable. To deal with the problem, those authors suggest discarding observations corresponding to rounded birth weight.

Ordinary rounding may not be common for reporting age, but may be more likely in other applications, such as when the running variable is a test score or when one uses the midpoints of reported income or wealth intervals. If one has the ordinary rounding problem and apply Corollary 2, then e will range from -0.5 to $+0.5$, instead of ranging from 0 to 1. If e is uniformly distributed in this interval then $\mu_k = E\left(e^k\right) = \int_{-0.5}^{0.5} e^k de = \left[(0.5)^{k+1} - (-0.5)^{k+1}\right] / (k+1)$, which is zero for all odd

DOI: 10.1002/jae

values of k, so many more elements of the matrix M will be zero than before, and hence the bias from rounding in this case is likely to be smaller.[11]

Corollary 2 can be extended to fuzzy designs in the same way as Corollary 1, by being applied in both the numerator and denominator of the fuzzy design treatment effect $\tau = \tau_Y / \tau_T$.

9. CONCLUSIONS

Using a rounded and hence discrete running variable has been common in applications of RD models. This is frequently due to data availability. This paper contributes to the growing RD literature by addressing issues associated with this common practice. In particular, when the running variable is rounded and hence is discrete, the standard RD estimation yields biased estimates of the RD treatment effects, even if the functional form of the model is correctly specified. In practice, this rounding or discretization bias can be very easily corrected. This paper presents simple formulas to fix this bias and hence provides consistent estimates of RD treatment effects given only rounded data of the running variable. The proposed approach does not require instrumental variables, but instead uses information regarding the distribution of rounding errors within the discretization cell, e.g. the distribution of ages within a year in the case of using age in years as a running variable. This can be easily obtained from census data, and often close to uniform.

In one empirical application, I investigate the effect of Medicare eligibility at 65 on insurance coverage in the USA. Higher-frequency age data (age in months) are available, and so provide a benchmark. I show that the proposed method to correct the rounding bias works well and produces estimates that are consistent with having and using data where age is more accurately measured.

In another empirical application, I provide an RD analysis that exploits the mandatory retirement policy in China to test for the presence of, and estimate the magnitude of, a retirement-consumption puzzle in China. In this case, the food consumption profile around the mandatory retirement age of 60 for male workers has relatively large slope changes, so the rounding bias is sizable and the bias correction is empirically important.

The proposed methodology is extended to cases involving non-integer cutoffs or other common forms of rounding, such as ordinary rounding or rounding up to the nearest integer. This paper's empirical applications focus on age in years, but the proposed approach can be used in other RD applications where the running variable is similarly rounded. Examples include using calendar years or integer-valued test scores as a running variable or dealing with the heaping problem (at ounce multiples) of birth weight when it is used as a running variable. Further, the proposed approach may also be similarly applied in regression models where one only has interval data on a regressor.

Instead of estimating the mean treatment effect at the cutoff, Frandsen *et al.* (2012) estimate quantile RD treatment effects for standard fuzzy design RD models. One interesting topic for future research is then to extend the current results to the case of estimating quantile RD treatment effects.

ACKNOWLEDGEMENTS

I would like to thank Emanuele Ciani for providing the Italian anagraphic records data, Baris Yoruk for providing the NLSY97 birth date data, and the three anonymous referees for helpful comments. All errors are my own.

[11] For example, in the first empirical application, when artificially rounding age to the nearest integer year based on age in months, it can be shown that the estimated insurance rate increase averages about 11.0% for the fourth-order polynomial using the 9 years window and averages about 11.6% using the 12 years window. These estimates are close to the estimates using age in months or the bias-corrected estimates using age in years. Further bias correction does not make much differences.

DOI: 10.1002/jae

REFERENCES

Aguila E, Attanasio O, Meghir C. 2011. Changes in consumption at retirement: evidence from panel data. *Review of Economics and Statistics* **93**(3): 1094–1099.

Ameriks J, Caplin A, Leahy J. 2007. Retirement consumption: insights from a survey. *Review of Economics and Statistics* **89**(2): 265–274.

Banks J, Blundell R, Tanner S. 1998. Is there a retirement-savings puzzle? *American Economic Review* **88**(4): 769–788.

Barreca A, Guldi M, Lindo JM, Waddell GR. 2010. Heaping-induced bias in regression-discontinuity designs. NBER Working Paper 17408.

Barreca A, Guldi M, Lindo JM, Waddell GR. 2011. Saving babies? Revisiting the effect of very low birth weight classification. *Quarterly Journal of Economics* **126**(4): 2117–2123.

Battistin E, Chesher A. 2011. Treatment effect estimation with covariate measurement error. CeMMAP Working Paper CWP25/09.

Battistin E, Brugiavini A, Rettore E, Weber G. 2009. The retirement consumption puzzle: evidence from a regression discontinuity approach. *American Economic Review* **99**(5): 2209–2226.

Behaghel L, Crepon B, Sedillot B. 2008. The perverse effects of partial employment protection reform: the case of French older workers. *Journal of Public Economics* **92**(3–4): 696–721.

Beresford GC. 1980. The uniformity assumption in the birthday problem. *Mathematics Magazine* **53**: 286–288.

Bernheim D, Skinner J, Weinberg S. 2001. What accounts for the variation in retirement wealth among U.S. households? *American Economic Review* **91**(4): 832–857.

Borella M, Moscarola FC, Rossi M. 2011. (Un)expected retirement and the consumption puzzle. Netspar Discussion Paper No. 10/2011-116.

Card D, Shore-Sheppard L. 2004. Using discontinuous eligibility rules to identify the effects of the federal Medicaid expansions on low income children. *Review of Economics and Statistics* **86**: 752–766.

Card D, Dobkin C, Maestas N. 2008. The impact of nearly universal insurance coverage on health care utilization: evidence from Medicare. *American Economic Review* **98**(5): 2242–2258.

Card D, Dobkin C, Maestas N. 2009. Does Medicare save lives? *Quarterly Journal of Economics* **124**(2): 597–636.

Carpenter C, Dobkin C. 2009. The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics* **1**(1): 164–182.

Chen S, van der Klaauw W. 2008. The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics* **142**(2): 757–784.

De Giorgi G. 2005. Long-term effects of a mandatory multistage program: the new deal for young people in the UK. Institute for Fiscal Studies Working Paper 05/08.

DiNardo J, Lee DS. 2004. Economic impacts of new unionization on private sector employers: 1984–2001. *Quarterly Journal of Economics* **119**: 1383–1442.

Edmonds EV. 2004. Does illiquidity alter child labor and schooling decisions? Evidence from household responses to anticipated cash transfers in South Africa. National Bureau of Economic Research Working Paper 10265.

Edmonds EV, Mammen K, Miller DL. 2005. Rearranging the family? Income support and elderly living arrangements in a low-income country. *Journal of Human Resources* **40**(1): 186–207.

Ferreira F. 2010. You can take it with you: Proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities. *Journal of Public Economics* **94**(9–10): 661–673.

Frandsen BR, Frölich M, Melly B. 2012. Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics* **168**(2): 382–395.

Haider SJ, Stephens M. 2007. Is there a retirement-consumption puzzle? Evidence using subjective retirement expectations. *Review of Economics and Statistics* **89**(2): 247–264.

Heitjan DF, Rubin DB. 1991. Ignorability and coarse data. *Annals of Statistics* **19**(4): 2244–2253.

Hsiao C. 1983. Regression analysis with a categorized explanatory variable. In *Studies in Econometrics, Time Series, and Multivariate Statistics*, Karlin S, Amemiya T, Goodman L (eds). Academic Press: New York.

Hullegie P, Klein TJ. 2010. The effect of private health insurance on medical care utilization and self-assessed health in Germany. *Health Economics* **19**(9): 1048–1062.

Hurd MD, Rohwedder S. 2008. The retirement consumption puzzle: actual spending change in panel data. RAND Working Paper WR-563.

Imbens GW, Lemieux T. 2008. Regression discontinuity designs: a guide to practice. *Journal of Econometrics* **142**: 615–635.

DOI: 10.1002/jae

Lalive R. 2007. Unemployment benefits, unemployment duration, and post-unemployment jobs: a regression discontinuity approach. *American Economic Review* **97**(2): 108–112.
Lalive R. 2008. How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics* **142**(2): 785–806.
Lalive R, van Ours JC, Zweimuller J. 2006. How changes in financial incentives affect the duration of unemployment. *Review of Economic Studies* **73**(4): 1009–1038.
Lee DS. 2008. Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics* **142**: 675–697.
Lee DS, Card D. 2008. Regression discontinuity inference with specification error. *Journal of Econometrics* **142**: 655–674.
Lee DS, Lemieux T. 2010. Regression discontinuity designs in economics. *Journal of Economic Literature* **48**: 281–355.
Lee DS, McCrary J. 2005. Crime, punishment, and myopia. National Bureau of Economic Research Working Paper 11491.
Lemieux T, Milligan K. 2008. Incentive effects of social assistance: a regression discontinuity approach. *Journal of Econometrics* **142**(2): 807–828.
Leuven E, Oosterbeek H. 2004. Evaluating the effect of tax deductions on training. *Journal of Labor Economics* **22**(2): 461–488.
Manski CF, Tamer E. 2002. Inference on regressions with interval data on a regressor or outcome. *Econometrica* **70**: 519–546.
Murphy R. 1996. An analysis of the distribution of birthdays in a calendar year. Available: http://www.panix.com/~murphy/bday.html [12 November 2013].
Oreopoulos P. 2006. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review* **96**(1): 152–175.
Pei Z. 2011. Regression discontinuity design with measurement error in the assignment variable. Working paper, Department of Economics, University of Princeton.
Robb AL, Burbridge JB. 1989. Consumption, income, and retirement. *Canadian Journal of Economics* **22**(3): 522–542.
Rubin D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**(5): 688–701.
Schwerdt G. 2005. Why does consumption fall at retirement? Evidence from Germany. *Economics Letters* **89**(3): 300–305.
Tsiatis A. 2006. *Semiparametric Theory and Missing Data*. Springer: Berlin.

APPENDIX

Proof of Theorem 1. X equals the integer part of X^* and hence it is a deterministic function of X^*; therefore, $E\left(Y \mid X^*, T^*\right) = E\left(Y \mid X^*, X, T^*\right)$. By this result and the law of iterated expectations, we have

$$E\left(Y \mid X, T^*\right) = E\left[E\left(Y \mid X^*, T^*\right) \mid X, T^*\right]$$

Therefore

$$E\left(Y \mid X, T^* = t\right) = E\left[E\left(Y \mid X^*, T^* = t\right) \mid X, T^* = t\right] = E\left[g_t\left(X^*\right) \mid X\right] = E[g_t\left(X + e\right) \mid X]$$

The second equality follows from $T^* = I\left(X \geq 0\right) = I\left(X^* \geq 0\right)$ being a deterministic function of X^* and $E(Y \mid X^*, I\left(X^* \geq 0\right) = t) = g_t\left(X^*\right)$ for $t = 0, 1$ given a sharp design.

Given Assumption 3, define the unknown polynomial coefficients b_{jt} by

$$g_t\left(X^*\right) = \sum_{j=0}^{J} b_{jt} X^{*j} \tag{A.1}$$

DOI: 10.1002/jae

and let B_t be the column vector of elements $b_{0t}, b_{1t}, \ldots, b_{Jt}$ for $t = 0, 1$. Then

$$g_t(X+e) = \sum_{j=0}^{J} b_{jt}(X+e)^j = \sum_{j=0}^{J}\sum_{k=0}^{j}\binom{j}{k} b_{jt} e^{j-k} X^k = \sum_{k=0}^{J}\sum_{j=k}^{J}\binom{j}{k} b_{jt} e^{j-k} X^k$$

where $\binom{j}{k}$ is the binomial coefficient $\frac{j!}{k!(j-k)!}$. Substituting this expression for g_t into the equation for $E(Y \mid X, T^* = t)$ gives

$$E\left(Y \mid X, T^* = t\right) = \sum_{k=0}^{J}\sum_{j=k}^{J}\binom{j}{k} b_{jt} E\left(e^{j-k} \mid X\right) X^k$$

By Assumption 6 $E\left(e^k \mid X\right) = \mu_k = E\left(e^k\right)$ is known. By Assumptions 1 and 4 and the definition of $h_t(X)$, we have $h_0(X) = E(Y \mid X, T^* = 0)$ when $X < 0$ and $h_1(X) = E(Y \mid X, T^* = 1)$ when $X \geq 0$. Putting these equations together gives

$$h_t(X) = \sum_{k=0}^{J} c_{kt} X^k \text{ where } c_{kt} = \sum_{j=k}^{J}\binom{j}{k}\mu_{j-k} b_{jt} \tag{A.2}$$

By Assumption 6, the value of $h_t(X)$ is identified at J (or more) values of X for $t = 0, 1$. Since the above shows that $h_t(X)$ is a polynomial of order J, and any polynomial of order J is uniquely identified by its values at $J + 1$ points, it follows that the coefficients c_{kt} are identified. Note that one does not need to know the polynomial order J a priori, since given this paper's assumption the observed values of $h_t(X)$ will trace out the polynomial of proper degree, thereby identifying J.

Equation (A.2) shows the connection between the coefficients in the discrete data regression b_{jt} and the true continuous data regression c_{kt} for $j, k = 0, 1, \ldots, J$. The following express this relationship in matrix notation.

For $t = 0, 1$, define the upper triangular $J + 1$ by $J + 1$ matrix M as having the element $\binom{j}{k}\mu_{j-k}$ in row $k + 1$ and column $j + 1$ for all j, k satisfying $0 \leq k \leq j \leq J$. All elements of M below the diagonal are zero. Recall that C_t consists of elements c_{kt} for $k = 0, 1, \ldots, J$ and that B_t consists of elements b_{jt} for $j = 0, 1, \ldots, J$. Given the matrix M, $c_{kt} = \sum_{j=k}^{J}\binom{j}{k}\mu_{j-k} b_{jt}$, for $t = 0, 1$ can be rewritten as

$$C_t = MB_t$$

Note that the matrix M is non-singular, because it is triangular with all ones on its diagonal, and all finite values off its diagonal, as e is bounded between 0 and 1. Then M can be inverted to solve for B_t given the already identified constants C_t, i.e. $B_t = M^{-1}C_t$. For the special case that $e = 0$ with probability one, $B_t = C_t$. Then $\tau = g_1(0) - g_0(0) = b_{01} - b_{00}$ is identified. Also the conditional mean functions $g_t(X^*) = \sum_{j=0}^{J} b_{jt} X^{*j}$ for $t = 0, 1$ are identified.

Proof of Corollary 1. By construction $E(Y \mid X, T^*) = h_0(X) + [h_1(X) - h_0(X)]\, T^*$, and this equation along with equation (A.2) yields equation (2) with $d_j = c_{j0}$ and $c_j = c_{j1} - c_{j0}$ for $j = 0, \ldots, J$, and hence part (i) of the Corollary holds given the property that any polynomial of degree J is identified by $J + 1$ points.

By construction, $A = B_0 = [b_{00}, b_{10}, \ldots, b_{J0}]'$, and $B = [b_0, b_1, \ldots, b_J]' = B_1 - B_0 = [b_{01} - b_{00}, b_{11} - b_{10}, \ldots, b_{J1} - b_{J0}]'$, so $\tau = b_0$ is the first element of B. Similarly, by construction $D = C_0 = [c_{00}, c_{10}, \ldots, c_{J0}]'$ and $C = [c_0, c_1, \ldots, c_J]' = C_1 - C_0 =$

DOI: 10.1002/jae

$[c_{01} - c_{00}, c_{11} - c_{10}, \ldots, c_{J1} - c_{J0}]'$, so $\tau' = c_0$ is the first element of C. By Theorem1, $C_t = MB_t$, for $t = 0, 1$, which implies $D = MA$ and $C_1 - C_0 = M(B_1 - B_0)$, and hence $C = MB$, and M is invertible by Theorem 1, which gives part (ii). The first row of the matrix equation $C = MB$ is $c_0 = b_0 + \sum_{j=1}^{J} b_j \mu_j$ and so part (iii) holds.

Proof of Corollary 2. The assumptions of Corollary 2 imply that $T^* = I$ $(X \geq 0)$ for all values of $X \geq 1$ or $X \leq -1$, i.e. there is no mismeasurement error in T^*. Thus the steps of the proofs of Theorem 1 and Corollary 1 are repeated using all values of X except those having $-1 < X < 1$.

DOI: 10.1002/jae