Regression Discontinuity Applications with Rounding Errors in the Running Variable

Yingying Dong

University of California Irvine

© 2013 SOLE Meetings
$Y = \text{outcome}$

$X^* = \text{a continuous running variable}; \ c = \text{a known threshold value}$

$T = \text{a binary treatment indicator}$

Sharp design: $T = T^* = 1(X^* - c \geq 0)$. 
E.g., $Y = \text{educational outcome}, \ T = \text{attending summer school}, \ X^* = -\text{test score}$.

Assuming everything else is continuous at $X^* = c$,

$$\tau = \lim_{x^* \downarrow c} E(Y | X^* = x^*) - \lim_{x^* \uparrow c} E(Y | X^* = x^*) = \text{LATE at } X^* = c.$$
Motivations

Many RD applications use rounded integer-valued running variables $X$.

- E.g., *age in years, birth weight in ounces* (Barreca et al, JPE forthcoming), *an integer valued test score, calendar year or quarter* etc.
- Other: interval censoring, e.g., using the midpoint of reported income brackets

Types of rounding:

- Rounding down: Age in years (65 means true age is $[65, 66]$). Similarly for calendar year or quarter.
- Ordinary rounding: Birth weight in ounces, an integer test score, or midpoint of income brackets.
- Rounding up: ...
Motivations

Discuss rounding down first, using age in years as a leading example.

- Standard practice: to estimate parametric regressions (polynomials) of $Y$ on reported discrete age $X$ above and below the cutoff $c$.

- This leads to biased estimates of the true RD treatment effect, even if the functional form is assumed to be known and is correctly specified.
  - The rounded running variable is like a mismeasured true running variable.
  - Unlike the classical measurement error, rounding errors can lead to either overestimates or underestimates.
  - i.e., discrete data may reveal a larger or smaller discontinuity than what it really is.
Propose simple bias corrections:

- do not require instrumental variables (IVs).
- utilize the first few low order moments of the rounding error within each discretization cell.
  - E.g., for age in years, can calculate the moments assuming a uniform distribution of birthdates within a year, or using census data.
- Bounds can be easily constructed without assuming any particular distribution for the rounding error.
- Can easily test existence of the rounding bias.
Identification

- $Y(t)$ for $t = 0, 1$, potential outcomes when treated or not treated (Rubin, 1974).

- $g_t(X^*) = E(Y(t) \mid X^*)$ for $t = 0, 1$, conditional means of potential outcomes *conditioning on the true continuous age*. Sharp design:
  - $E(Y \mid X^*, T^* = 1) = g_1(X^*)$
  - $E(Y \mid X^*, T^* = 0) = g_0(X^*)$
  - $\tau = g_1(0) - g_0(0)$ is the true RD treatment effect. WLG, let $c = 0$.

- Analogously, $h_t(X)$ for $t = 0, 1$, conditional means of potential outcomes *conditioning on the rounded age*.
  - $\tau' = h_1(0) - h_0(0)$ is the (biased) discrete data RD treatment effect.

- $e = X^* - X$ is the rounding error. $0 \leq e < 1$. Define its $k$th moment $\mu_k = E(e^k)$. 

Assumptions:

A1: \( T = I(X^* \geq 0) \).
A2: \( g_0(X^*) \) and \( g_1(X^*) \) are continuous at \( X^* = 0 \).
A3: \( g_0(X^*) \) for \( X^* < 0 \) and \( g_1(X^*) \) for \( X^* \geq 0 \) are polynomials of possibly unknown degree \( J \).
A4: \( h_0(X) \) is identified for all \(- (J + 1) \leq X < 0\), and \( h_1(X) \) is identified for all \( 0 \leq X \leq J \).
A5: \( I(X \geq 0) = I(X^* \geq 0) \).
A6: For all integers \( k \leq J \), \( E(e^k | X) = E(e^k) \), and these \( J \) moments are identified.

Theorem

Let assumptions A1 to A6 hold. Then \( \tau \) is identified even if \( X^* \) is not observed.
Given A1 - A3, the true model for $Y$ given the continuous $X^*$ can be written as

$$Y = \sum_{j=0}^{J} a_j X^* j + \sum_{j=0}^{J} b_j X^* j T^* + \epsilon^*$$ (1)

$$= \sum_{j=0}^{J} a_j (X + e)^j + \sum_{j=0}^{J} b_j (X + e)^j T^* + \epsilon^*$$ (2)

The true treatment effect is $\tau = b_0$. The true model for $Y$ given the discrete rounded $X$ has the form:

$$Y = \sum_{j=0}^{J} d_j X^j + \sum_{j=0}^{J} c_j X^j T^* + \epsilon$$ (3)

The (biased) discrete data treatment effect $\tau' = c_0$.

Define coefficient vectors $A = (a_0, a_1, ..., a_J)'$, $B = (b_0, b_1, ..., b_J)'$, $D = (d_0, d_1, ..., d_J)'$ and $C = (c_0, c_1, ..., c_J)'$. 

Yingying Dong (UCI)  
RD with Rounded Discrete Running Variable
Corollary

Let assumptions A1 to A6 hold. Then (i) the coefficients in the true underlying model A and B are identified by $A = M^{-1}D$ and $B = M^{-1}C$, where $M$ is a $J + 1$ by $J + 1$ matrix and has the element $(j)_k \mu_{j-k}$ in row $k + 1$ and column $j + 1$ for all $j, k$ satisfying $J \geq j \geq k \geq 0$. (ii) the bias is $\tau' - \tau = \sum_{j=1}^{J} b_j \mu_j$.

$$b_j(X + e)^j = \ldots + b_j(j)_k e^{i-k} X^k + \ldots$$ for any $J \geq j \geq k$;

For the intercept, i.e., $k = 0$, $\tau' = \sum_{j=0}^{J} b_j \mu_j$.

If $e$ is uniform, so $\mu_k = 1/(k + 1)$, then the bias is

$$\tau' - \tau = (1/2)b_1 + (1/3)b_2 + \ldots + 1/(J + 1)b_J.$$
Example: when $J = 4$, then first estimate

$$Y = d_0 + d_1 X + d_2 X^2 + d_3 X^3 + d_4 X^4 + (c_0 + c_1 X + c_2 X^2 + c_3 X^3 + c_4 X^4) T^*$$

- Can add other covariates to the regression if desired.

The (biased) discrete data treatment effect $\tau' = c_0$.

Assuming age distribution is uniform within a year, then the true treatment effect

$$\tau = c_0 - (1/2)c_1 + (1/6)c_2 - (1/30)c_4.$$  

For any distribution of $e$, the true treatment effect is

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

A general formula that works for any $J$ is provided in the paper.
Can test the rounding bias:

- Use an ordinary $t$ test for $H_0 : \tau - \tau' = 0$.
  - E.g., for $J \leq 4$ with a uniform $e$, test $H_0 : -c_1/2 + c_2/6 - c_4/30 = 0$.
- If don’t know moments of the $e$ distribution, could still do a standard $F$ test for $H_0 : c_1 = c_2 = c_3 = c_4 = 0$. 
Bounds on the treatment effect can be easily obtained without assuming any particular distribution of rounding errors.

- $0 \leq e < 1$, so $\mu_j = E(e^j)$ satisfy $1 > \mu_1 \geq \mu_2 \geq ... \geq \mu_j \geq 0$.

- With estimates of $c_j$ for any $j \leq J$, one can search the min and max of the estimated $\tau$ over the set of values of $\mu_j$ to obtain bounds.

Example 1: $J = 1$, $\tau = c_0 - \mu_1 c_1$, bounds are given by $c_0$ and $c_0 - c_1$.

Example 2: $J = 2$, $\tau = c_0 - \mu_1 c_1 + (2\mu_1^2 - \mu_2) c_2$ bounds are given by the min and max of $\tau$ over the set $1 > \mu_1 \geq \mu_2 \geq 0$. 
The fuzzy design RD local treatment effect $\tau_f$ is

$$
\tau_f = \frac{\tau_Y}{\tau_T},
$$

$\tau_Y$ = the discontinuity in the conditional mean outcome at the cutoff.

$\tau_T$ = the discontinuity in the treatment probability at the cutoff.

- Apply the bias correction to both the numerator and the denominator.

For 4th (or lower) order polynomials of $Y$ and $T$, estimate

$$
Y = \sum_{j=0}^{4} d_j X^j + \sum_{j=0}^{4} c_j X^j T^* + \varepsilon \text{ and } T = \sum_{j=0}^{4} r_j X^j + \sum_{j=0}^{4} s_j X^j T^* + \tilde{\varepsilon}.
$$

Then discrete data RD treatment effect is $\tau'_f = \frac{c_0}{s_0}$. For a uniform $\varepsilon$, the correct treatment effect is

$$
\tau_f = \frac{c_0 - (1/2)c_1 + (1/6)c_2 - (1/30)c_4}{s_0 - (1/2)s_1 + (1/6)s_2 - (1/30)s_4}.
$$
Application 1: Medicare and Insurance Coverage

How much does the health insurance coverage rate increase at the Medicare eligibility age 65 in the US?

- Use data from the US Health and Retirement Study (HRS).

- Both age in months and age in years are available; can verify the accuracy of the proposed bias correction.

Use data from 1992 to 2007. The samples have 60,290 to 135,582 observations, depending on the window width $[-6, 6]$, $[-9, 9]$, $[-12, 12]$ or $[-15, 15]$. 
$Y =$ whether one has any health insurance  
$X =$ reported age in years minus 65  
$T = T^* =$ whether one is eligible for Medicare

**Figure:** The age (in years) profile of health insurance coverage rates, HRS 1992 - 2008

**Figure:** The age (in months) profile of health insurance coverage rates, HRS 1992 - 2008
### Table 1 Estimated increases in the health insurance coverage rate at age 65

<table>
<thead>
<tr>
<th></th>
<th>3rd order polynomial</th>
<th>4th order polynomial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)  (2)  (3)</td>
<td>(1)  (2)  (3)</td>
</tr>
<tr>
<td>Naive estimates using age in years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[-6, +6)</td>
<td>0.128 0.125 0.124</td>
<td>0.107 0.106 0.107</td>
</tr>
<tr>
<td></td>
<td>(0.015)*** (0.015)*** (0.015)***</td>
<td>(0.035)*** (0.034)*** (0.035)***</td>
</tr>
<tr>
<td>[-15, +15)</td>
<td>0.124 0.126 0.126</td>
<td>0.129 0.128 0.129</td>
</tr>
<tr>
<td></td>
<td>(0.006)*** (0.006)*** (0.006)***</td>
<td>(0.009)*** (0.009)*** (0.009)***</td>
</tr>
<tr>
<td>Naive estimates using age in months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[-6, +6)</td>
<td>0.119 0.118 0.119</td>
<td>0.112 0.113 0.113</td>
</tr>
<tr>
<td></td>
<td>(0.008)*** (0.008)*** (0.008)***</td>
<td>(0.011)*** (0.011)*** (0.011)***</td>
</tr>
<tr>
<td>[-15, +15)</td>
<td>0.119 0.121 0.120</td>
<td>0.119 0.120 0.120</td>
</tr>
<tr>
<td></td>
<td>(0.005)*** (0.005)*** (0.005)***</td>
<td>(0.006)*** (0.006)*** (0.006)***</td>
</tr>
<tr>
<td>Bias-corrected estimates using age in years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[-6, +6)</td>
<td>0.118 0.117 0.117</td>
<td>0.112 0.113 0.113</td>
</tr>
<tr>
<td></td>
<td>(0.009)*** (0.009)*** (0.009)***</td>
<td>(0.014)*** (0.014)*** (0.014)***</td>
</tr>
<tr>
<td>[-15, +15)</td>
<td>0.117 0.118 0.119</td>
<td>0.119 0.120 0.120</td>
</tr>
<tr>
<td></td>
<td>(0.005)*** (0.005)*** (0.005)***</td>
<td>(0.006)*** (0.006)*** (0.006)***</td>
</tr>
</tbody>
</table>

Note: HRS 1992-2008; (1) does not control for covariates; (2) controls for year dummies; (3) controls for year dummies and additional demographic variables. Robust clustered standard errors are calculated by the delta method; * Significant at the 10% level; ** Significant at the 5% level; *** Significant at the 1% level.
Table 2 Bounds on the bias-corrected estimates of the health insurance rate increase at 65

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[-6,+6)</td>
<td>0.128</td>
<td>(0.109, 0.128)</td>
<td>0.125</td>
</tr>
<tr>
<td>[-15,+15]</td>
<td>0.124</td>
<td>(0.111, 0.124)</td>
<td>0.126</td>
</tr>
</tbody>
</table>

Note: HRS 1992-2008; All estimates are based on third order polynomials; Bounds are provided in the 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.
Application 2: The Retirement-Consumption Puzzle in China

Does food consumption have a similar significant decline at retirement in China as in many developed Western countries?

Exploiting the mandatory retirement in China for identification.

- The official retirement age for male workers is 60.
- May retire earlier before 60 or get re-employed after official retirement—a fuzzy design RD.
  - \( Y \) = the logarithm of household food expenditure
  - \( T \) = whether a male household head is retired or not
  - Sample sizes range from 12,866 to 33,754, for 6 to 15 years of windows.
Application 2: The Retirement Consumption Puzzle in China

Figure: The age (in years) profile of retirement rates for male household heads, UHS 1997 - 2006

Figure: The age (in years) profile of log food consumption, UHS 1997 - 2006
## Table 3 Effects of retirement on food consumption

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>[-6,+6]</td>
<td>-0.046</td>
<td>0.193</td>
</tr>
<tr>
<td></td>
<td>(0.017)<em><strong>(0.024)</strong></em>(0.085)***</td>
<td>(0.016)**</td>
</tr>
<tr>
<td>[-15,+15]</td>
<td>-0.054</td>
<td>0.211</td>
</tr>
<tr>
<td></td>
<td>(0.011)<em><strong>(0.014)</strong></em>(0.049)***</td>
<td>(0.010)<em><strong>(0.014)</strong></em>(0.048)***</td>
</tr>
<tr>
<td>[-6,+6]</td>
<td>-0.034</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>(0.017)**</td>
<td>(0.022)<em><strong>(0.075)</strong></em></td>
</tr>
<tr>
<td>[-15,+15]</td>
<td>-0.044</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>(0.011)<em><strong>(0.014)</strong></em>(0.042)***</td>
<td>(0.011)<em><strong>(0.014)</strong></em>(0.042)***</td>
</tr>
</tbody>
</table>

Note: 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.
### Application 2: The Retirement Consumption Puzzle in China

#### Table 4 Bounds for the bias-corrected estimates of the retirement effects on consumption

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[-6, +6)</td>
<td>[-0.237, -0.095)</td>
</tr>
<tr>
<td></td>
<td>[-15, +15)</td>
<td>[-0.257, -0.136)</td>
</tr>
</tbody>
</table>

Note: Male household heads, UHS 1997-2006; Bounds are provided in the 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.
Extensions: Other Forms of Rounding or Non-integer Threshold

Other forms of rounding: rounding up, ordinary rounding, or non-integer cutoff

- Assumption A5 $I(X \geq 0) = I(X^* \geq 0)$ is violated; there is misclassification of crossing threshold status.

- $T^* = I(X^* \geq 0)$ is only mis-measured near the cutoff $X = 0$; can drop observations near the cutoff.

**Corollary**

Let assumptions A1 to A4 and A6 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 for $Y = \sum_{j=0}^{J} d_j X^j + \sum_{j=0}^{J} c_j X^j T^* + \epsilon$ for all $X \geq 1$ or $X \leq -1$. 
Conclusions

- Standard RD estimation based on a rounded discrete running variable yields biased estimates of the true RD treatment effect, even if the true functional form is correctly specified.
- Provides a simple formula to correct this bias.
  - does not require IVs; uses low order moments of the distribution of the rounding error within discretization cells.
- Can test the existence of rounding bias by a simple $t$ or $F$ test.
- Can easily construct bounds without assuming any distribution for rounding errors.