Regression discontinuity

Paul Schrimpf

UBC
Economics 326

March 29, 2018

References

- Angrist and Pischke (2014) chapter 4
- Lee and Lemieux (2010)

Section 1

Regression discontinuity

Birthdays and funerals

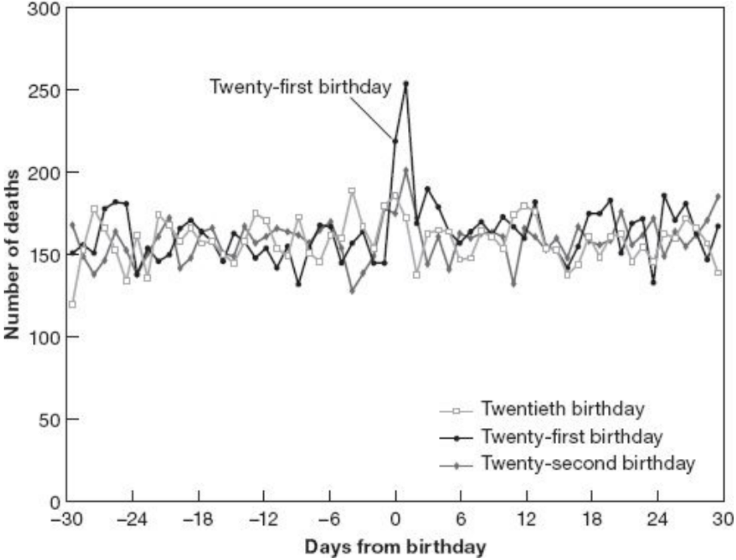
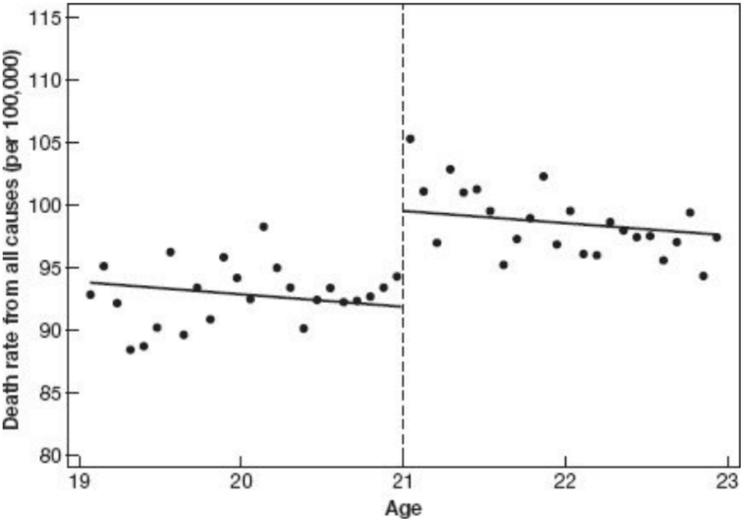


Figure 4.2
A sharp RD estimate of MLDA mortality effects



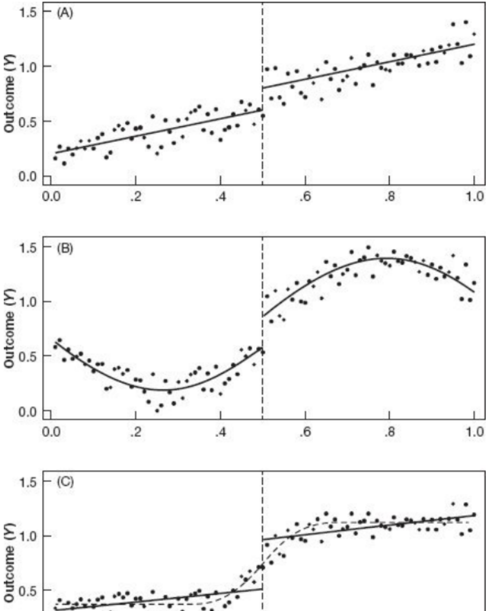
Regression discontinuity

- Regression

$$death_i = \beta_0 + \beta_1 LegalToDrink_i + \beta_2 age_i + \epsilon_i$$

- $LegalToDrink = \begin{cases} 1 & \text{if } age \geq 21 \\ 0 & \text{if } age < 21 \end{cases}$ is a known, discontinuous function of age
- Assume $\mathrm{E}[death|LegalToDrink = 1, age]$ and $\mathrm{E}[death|LegalToDrink = 0, age]$ are continuous functions of age
- Then β_1 is an estimate of $\mathrm{E}[death|LegalToDrink = 1, age = 21] - \mathrm{E}[death|LegalToDrink = 0, age = 21]$, the causal effect of drinking becoming legal at age 21 on death rates

Nonlinearity vs discontinuity



Nonlinearity vs discontinuity

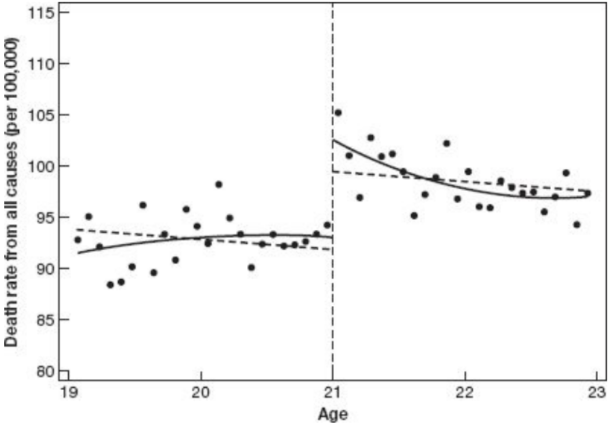
- Can be hard to tell nonlinearity from discontinuity
- Solution: flexibly model $\mathrm{E}[death|LegalToDrink, age]$ as function of age
    - Different slopes on each side of discontinuity:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 a_i + +\beta_3 D_i(a_i - 21) + \epsilon_i$$

    - Include powers of age :

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 a_i + \beta_3 a_i^2 + +\beta_4 D_i(a_i - 21) + \beta_5 D_i(a_i - 21)^2 + \epsilon_i$$

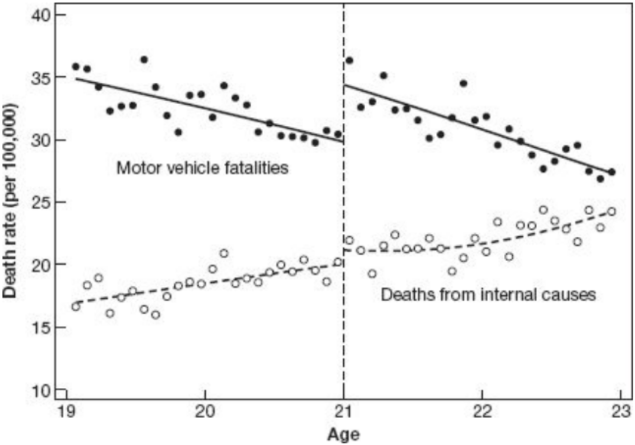
    - Limit sample to small range around discontinuity "bandwidth"



Notes: This figure plots death rates from all causes against age in months. Dashed lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months. The solid lines plot fitted values from a regression of mortality on an over-21 dummy and a quadratic in age, interacted with the over-21 dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

Dependent variable	Ages 19–22 (1)	Ages 19–22 (2)	Ages 20–21 (3)	Ages 20–21 (4)
All deaths	7.66 (1.51)	9.55 (1.83)	9.75 (2.06)	9.61 (2.29)
Motor vehicle accidents	4.53 (.72)	4.66 (1.09)	4.76 (1.08)	5.89 (1.33)
Suicide	1.79 (.50)	1.81 (.78)	1.72 (.73)	1.30 (1.14)
Homicide	.10 (.45)	.20 (.50)	.16 (.59)	−.45 (.93)
Other external causes	.84 (.42)	1.80 (.56)	1.41 (.59)	1.63 (.75)
All internal causes	.39 (.54)	1.07 (.80)	1.69 (.74)	1.25 (1.01)
Alcohol-related causes	.44 (.21)	.80 (.32)	.74 (.33)	1.03 (.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.



Notes: This figure plots death rates from motor vehicle accidents and internal causes against age in months. Lines in the figure plot fitted values from regressions of mortality by cause on an over-21 dummy and a quadratic function of age in months, interacted with the dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

Section 2

Using discontinuity as an instrument

Discontinuity as an instrument

- Earlier regressions give good estimates of the effect of reaching the legal drinking age on mortality
- What is the causal effect of drinking (or binge drinking) on mortality?
- Idea : use discontinuity as instrument

Carpenter and Dobkin (2009)

"The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age"

- Data on drinking and age from National Health Interview Survey (NHIS), 1997-2005
- Data on mortality and age from National Center for Health Statistics, 1997-2004

Model

- Equation of interest:

$$mortality = \beta_0 + \beta_1 drinking + controls + \epsilon$$

- First stage :

$$drinking = \alpha_0 + \alpha_1 Legal + \alpha_2 age + controls + u$$

- Reduced form :

$$mortality = \pi_0 + \pi_1 Legal + \pi_2 age + controls + v$$

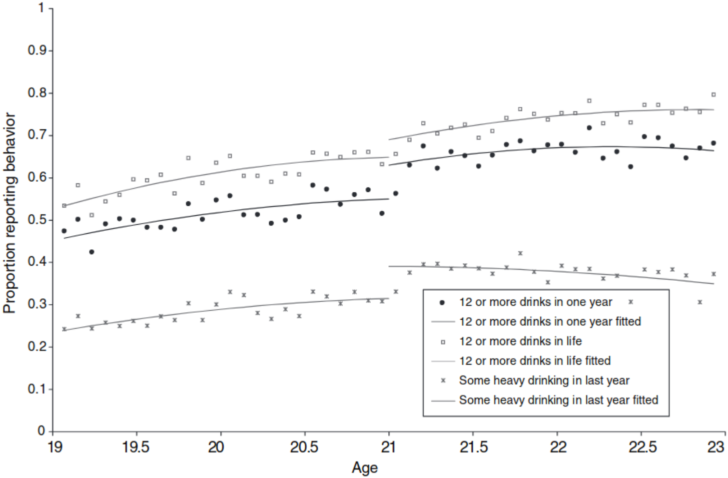


FIGURE 1. AGE PROFILE OF DRINKING PARTICIPATION

Notes: NHIS Sample Adult 1997–2005. Cells are the proportion of people in a 30-day block that report the behavior. The regression line is a second-order polynomial fitted on unweighted individual observations on either side of the age 21 cutoff.

TABLE 1—ALCOHOL CONSUMPTION: PARTICIPATION

	(1)	(2)	(3)	(4)	(5)
12 or more drinks in lifetime					
Over 21	0.0418	0.0316	0.0268	0.0198	0.0199
	(0.0242)	(0.0301)	(0.0292)	(0.0423)	(0.0179)
Observations	16,107	16,107	16,107	16,107	
R^2	0.02	0.03	0.10	0.10	
Prob > Chi-Squared			0.00	0.61	
12 or more drinks in one year					
Over 21	0.0796	0.0657	0.0611	0.0603	0.0461
	(0.0254)	(0.0313)	(0.0301)	(0.0438)	(0.0218)
Observations	16,107	16,107	16,107	16,107	
R^2	0.02	0.03	0.11	0.11	
Prob > Chi-Squared			0.00	0.56	
Any heavy drinking in last year					
Over 21	0.0761	0.0527	0.0492	0.0262	0.0398
	(0.0248)	(0.0304)	(0.0291)	(0.0430)	(0.0201)
Observations	16,107	16,107	16,107	16,107	
R^2	0.01	0.01	0.10	0.10	
Prob > Chi-Squared			0.00	0.67	
Covariates	N	N	Y	Y	N
Weights	N	Y	Y	Y	N
Quadratic terms	Y	Y	Y	Y	N
Cubic terms	N	N	N	Y	N
LLR	N	N	N	N	Y

Notes: The first column of each panel contains the regression from the corresponding figure. Robust standard errors are in parentheses. Covariates include dummies for census region, race, gender, health insurance, employment status, twenty-first birthday, twenty-first birthday + 1 day, and looking for work. Weights are the NHIS adult sample weights and reduce the precision of the regressions significantly as the weights vary substantially across observations. People reporting five or more drinks on one day (Nt necessarily in one sitting) are coded as heavy drinkers. The first four columns give the estimates from polyNmial regressions on age interacted with a dummy for being over 21. The age variable is centered on 21, so the Over 21 variable gives us an estimate of the discontinuous increase at age 21. In the fifth column, we present the results of a local linear regression procedure with a rule-of-thumb bandwidth. For this procedure, we follow Fan and Gijbels (1996) and fit a fourth order polyNmial separately on each side of the age-21 cutoff. We use the fit of this regression to estimate the average second derivative of the expectation function (D) and the mean squared error of this function (σ^2). The rule-of-thumb bandwidth is $h = c(\sigma^2 R/D)$, where c is a constant that depends on the kernel ($c = 3.44$ for a triangular kernel), and R is the range of the running variable (i.e., the range of ages used to estimate the polyNmial on each side). We then use this bandwidth, and a triangular kernel, to fit local linear regressions on each side of age 21, and estimate the limit of the expectation function from the left and the right of age 21. The local linear regressions have two fewer observations because the twenty-first birthday and the day after the twenty-first birthday have been dropped.

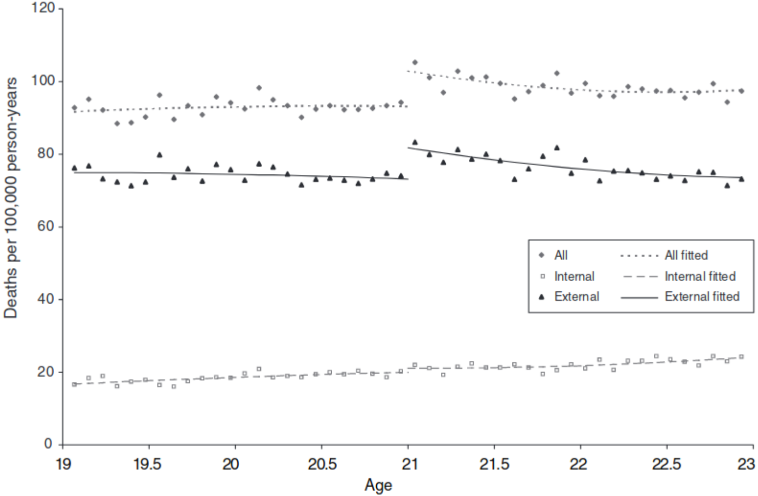


FIGURE 3. AGE PROFILE FOR DEATH RATES

Notes: Deaths from the National Vital Statistics Records. Includes all deaths that occurred in the United States between 1997–2003. The population denominators are derived from the census. See online Appendix C for a list of causes of death.

Table 4—Discontinuity in Log Deaths at Age 21

	(1)	(2)	(3)	(4)
Deaths due to all causes				
Over 21	0.096	0.087	0.091	0.074
	(0.018)	(0.017)	(0.023)	(0.016)
Observations	1,460	1,460	1,460	1,458
R^2	0.04	0.05	0.05	
Prob > Chi-Squared		0.000	0.735	
Deaths due to external causes				
Over 21	0.110	0.100	0.096	0.082
	(0.022)	(0.021)	(0.028)	(0.021)
Observations	1,460	1,460	1,460	1,458
R^2	0.06	0.08	0.08	
Prob > Chi-Squared		0.000	0.788	
Deaths due to internal causes				
Over 21	0.063	0.054	0.094	0.066
	(0.040)	(0.040)	(0.053)	(0.031)
Observations	1,460	1,460	1,460	1,458
R^2	0.10	0.10	0.10	
Prob > Chi-Squared		0.000	0.525	
Covariates	N	Y	Y	N
Quadratic terms	Y	Y	Y	N
Cubic terms	N	N	Y	N
LLR	N	N	N	Y

Notes: See Notes from Table 1. The dependent variable is the log of the number of deaths that occurred x days from the person's twenty-first birthday. External deaths include all deaths with mention of an injury, alcohol use, or drug use. The Internal Death category includes all deaths Nt coded as external. Please see Web Appendix C for the ICD codes for each of the categories above. The first three columns give the estimates from polyNmial regressions on age interacted with a dummy for being over 21.

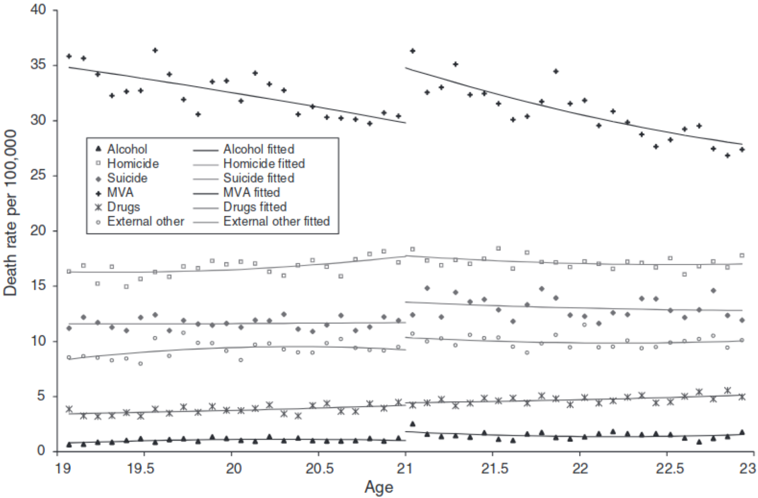


FIGURE 4. AGE PROFILES FOR DEATH RATES BY EXTERNAL CAUSE

Notes: See notes to Figure 3. The categories are mutally exclusive. The order of precedence is homicide, suicide, MVA, deaths with a mention of alcohol, and deaths with a mention of drugs. The ICD-9 and ICD-10 Codes are in Appendix C.

Table 5—Discontinuity in Log Deaths at Age 21 Due to External Causes

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	Alcohol				Homicide			
Over 21	0.388	0.346	0.406	0.441	0.009	0.002	−0.003	−0.014
	(0.119)	(0.116)	(0.156)	(0.117)	(0.045)	(0.045)	(0.061)	(0.041)
Observations	1,460	1,460	1,460	1,458	1,460	1,460	1,460	1,458
R^2	0.03	0.04	0.04			0.01	0.01	0.01
Prob > Chi-Squared		0.000	0.228			0.000	0.495	
	Suicide				Motor vehicle accidents			
Over 21	0.160	0.154	0.135	0.105	0.158	0.143	0.145	0.139
	(0.059)	(0.059)	(0.086)	(0.045)	(0.033)	(0.032)	(0.044)	(0.032)
Observations	1,460	1,460	1,460	1,458	1,460	1,460	1,460	1,458
R^2	0.02	0.02	0.02	0.15	0.16	0.16		
Prob > Chi-Squared		0.000	0.892	0.000	0.666			
	Drugs				Other external causes			
Over 21	0.070	0.067	0.004	−0.016	0.087	0.098	0.098	0.074
	(0.081)	(0.082)	(0.107)	(0.078)	(0.060)	(0.059)	(0.075)	(0.043)
Observations	1,460	1,460	1,460	1,458	1,460	1,460	1,460	1,458
R^2	0.04	0.04	0.04	0.01	0.01	0.01		
Prob > Chi-Squared		0.000	0.643			0.000	0.877	
Covariates	N	Y	Y	N	N	Y	Y	N
Quadratic terms	Y	Y	Y	N	Y	Y	Y	N
Cubic terms	N	N	Y	N	N	N	Y	N
LLR	N	N	N	Y	N	N	N	Y

Notes: See Notes from Table 4. There are 276 observations where there are N deaths coded as due to alcohol; for this variable 0.5 was added to the dependent variable before taking the log. There are 15 observations where there are N deaths coded as due to drug use; for this variable 0.5 was added to the count before taking the log.

IV

- Equation of interest:

$$mortality = \beta_0 + \beta_1 drinking + controls + \epsilon$$

- First stage :

$$drinking = \alpha_0 + \alpha_1 Legal + \alpha_2 age + controls + u$$

- Reduced form :

$$mortality = \pi_0 + \pi_1 Legal + \pi_2 age + controls + v$$

- First stage and reduced form estimated with different data, how to estimate β_1?

Section 3

General guidelines

- Regression discontinuity useful when policy of interest is discontinuous function of some "running variable"
  - Examples : age, exam scores, elections
  - Important that individuals cannot perfectly control being just above/below threshold (social assistance program with strict income limit would be problematic)
- Report results graphically and in tables
- Report a range of specifications, flexibly controlling for the running variable
- See Lee and Lemieux (2010)

References

Angrist, Joshua D and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Carpenter, Christopher and Carlos Dobkin. 2009. "The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age." *American Economic Journal: Applied Economics* 1 (1):164–82. URL http://www.aeaweb.org/articles?id=10.1257/app.1.1.164.

Lee, David S. and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2):281–355. URL http://www.aeaweb.org/articles?id=10.1257/jel.48.2.281.