Assignment 4

Jorge Daniel Guevara Acevedo

Sunday, June 14^{th}, 2020

`https://github.com/jorge9711/RDD`

1 Github repo and summary (worth 2 points)

1. Download Hansen$_dwi.dtafromgithubatthefollowingaddress.$

use https://github.com/scunning1975/causal-inference-class/raw/master/hansen$_dwi, clear$

Create a new github repo named "RDD". Inside the RDD directory, put all the subdirectories we've discussed in class. Post the link to the repo so I can see it's done as discussed in your assignment.

2. In the writing subdirectory, place your assignment. For the first part of this assignment, read Hansen's paper in the /articles directory of the main class github entitled "Hansen AER". Briefly summarize this paper. What is his research question? What data does he use? What is his research design, or "identification strategy"? What are his conclusions?

Hansen's paper explores the effects of harsh punishments and sanctions over the recidivism of offenders. He uses the data available in administrative records form the Washington State Impaired Driver Testing Program, 1999-2007. Using the cutoffs established by the law, he uses a Regression Discontinuity design. Also, present a reason for choosing the felony, in order to avoid problems of non-independence and manipulation of the data.

After the analysis using data from demographic characteristics and police ex-ante information, he concludes the effect of the harsh and the penalty itself decreases the probability of drunk driving in the short and the long run.

2 Replication (worth 6 points)

3. In the United States, an officer can arrest a driver if after giving them a blood alcohol content (BAC) test they learn the driver had a BAC of 0.08 or higher. We will only focus on the 0.08 BAC cutoff. We will be ignoring the 0.15 cutoff for all this analysis. Create a dummy equaling 1 if bac1>= 0.08 and 0 otherwise in your do file or R file.

Code used:
$genDUI = (bac1 >= 0.08)$

4. The first thing to do in any RDD is look at the raw data and see if there's any evidence for manipulation ("sorting on the running variable"). If people were capable of manipulating their blood alcohol content (bac1), describe the test we would use to check for this. Now evaluate whether you see this in these data? Either recreate Figure 1 using the bac1 variable as your measure of blood alcohol content or use your own density test from software. Do you find evidence for sorting on the running variable?

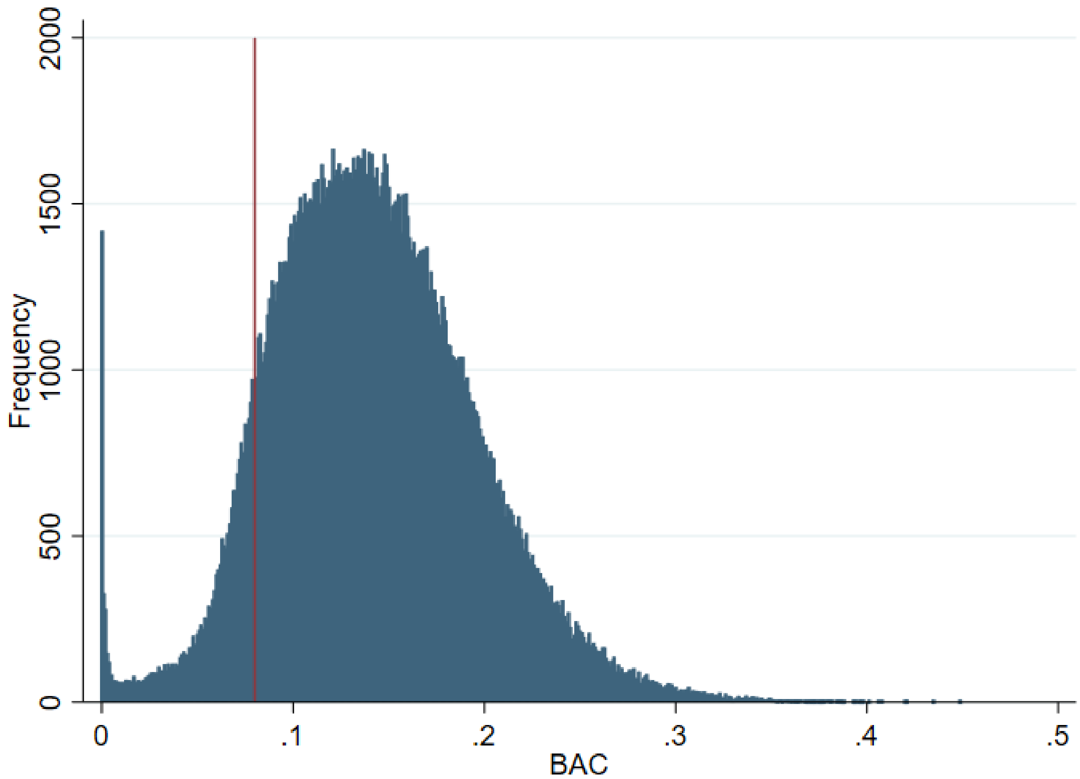


Figure 1. Punishment and Deterrence: Evidence from drunk driving. Benjamin Hansen.

The methods to evaluate this kind of data are the discontinuity density of observations for the assignment variable. The most popular is the McCray test (2008), that proofs the continuity of the distribution of the data. By the eyeball, there is no evidence of manipulation at the cutoff.

If we reproduce McCray test, the results vary between the conventional and robust tests. For the conventional one we find the same value Hansen found on his paper, but with the robust approach it seems evidence of manipulation (discontinuity) at the cutoff, mostly at the right (Graphic in the do file)

```
. do "C:\Users\jorge\AppData\Local\Temp\STD34b8_000000.tmp"

. rddensity bac1, c(0.08) plot all
Computing data-driven bandwidth selectors.

RD Manipulation Test using local polynomial density estimation.

     Cutoff c = .08 |Left of c   Right of c          Number of obs =       214558
   -----------------+----------------------          Model         = unrestricted
      Number of obs |     23010       191548          BW method     =         comb
 Eff. Number of obs |      8895        13730          Kernel        =   triangular
     Order est. (p) |         2            2          VCE method    =    jackknife
    Order bias (q)  |         3            3
        BW est. (h) |     0.011        0.012

Running variable: bac1.
   ------------------------------------------
             Method |          T        P>|T|
   -----------------+------------------------
       Conventional |     0.5337       0.5936
             Robust |     2.2032       0.0276
   ------------------------------------------
```

5. The second thing we need to do is check for covariate balance. Recreate Table 2 Panel A but only white male, age and accident (acc) as dependent variables. Use your equation 1) for this. Are the covariates balanced at the cutoff? It's okay if they are not exactly the same as Hansen's.

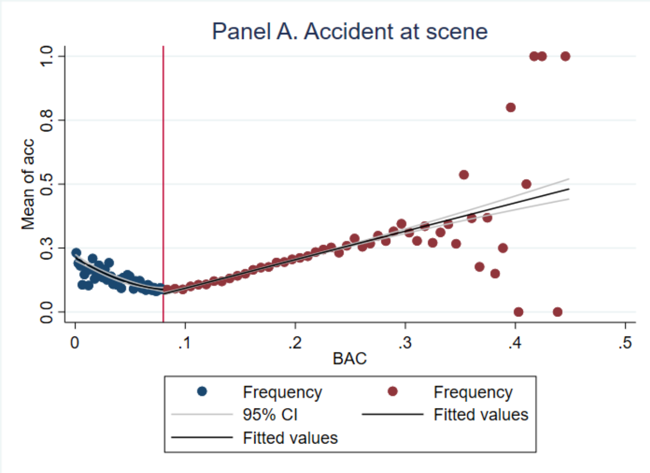
Regression Discontinuity Estimates for the Effect of Exceeding BAC Thresholds

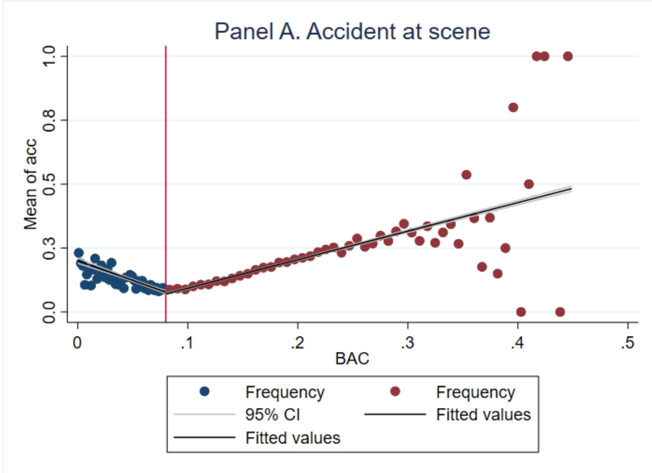
	Driver			
Characteristics	Male	White	Age	Accident
Panel A. DUI threshold	-0.0183815	0.0044543	-6.224398	-0.1544195
	(0.0197867)	(0.0175152)	(0.5861375)	(0.0152688)
Controls	No	No	No	No
Observations	89,967	89,967	89,967	89,967

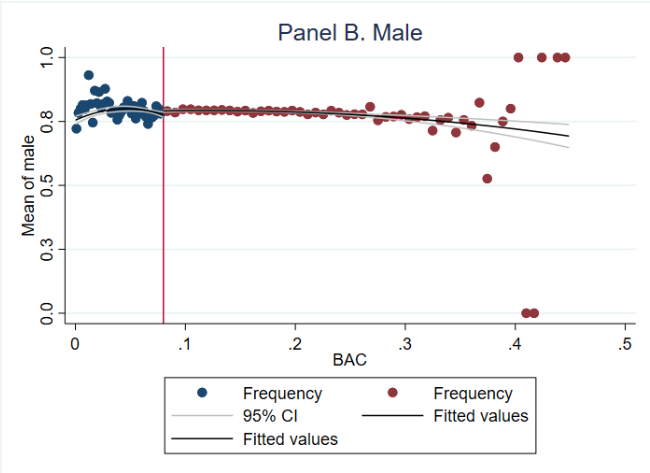
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

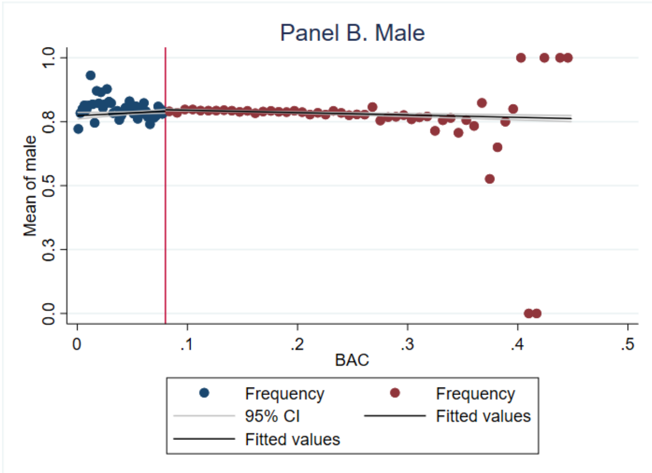
It is important to point out that the results obtained in this replication, differs from Hansen's estimators.
Two of the estimators are statistically significant. The problem is that, there is no complete validity in his regression discontinuity design because the significance make us think that covariates are not balanced at the cutoff (With this sample and variables. Also the method of rectangular kernell is ignored in our results, what could make a bigger differences between results).

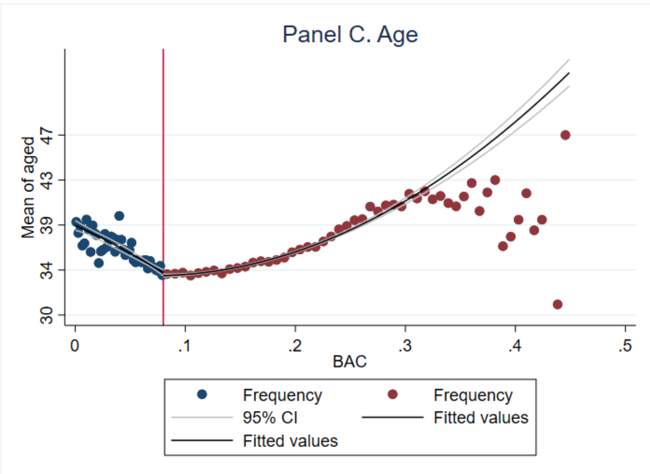
6. Recreate Figure 2 panel A-D. You can use the -cmogram- command in Stata to do this. Fit both linear and quadratic with confidence intervals. Discuss what you find and compare it with Hansen's paper.

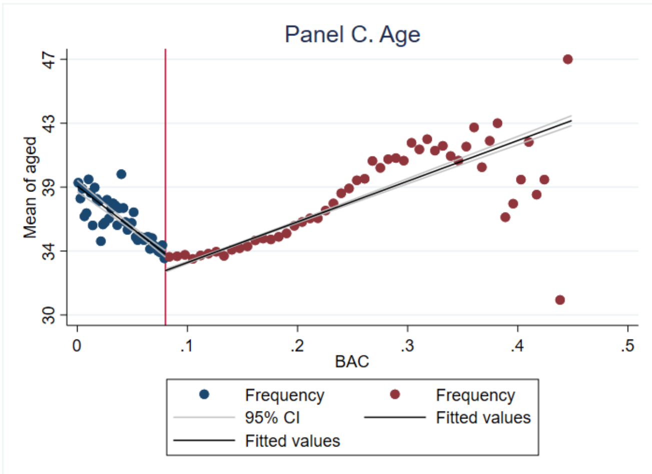

Panel A. Accident at scene
Mean of acc
0.0
0.3
0.5
0.8
1.0
0
.1
.2
.3
.4
.5
BAC
Frequency
Frequency
95% CI
Fitted values
Fitted values

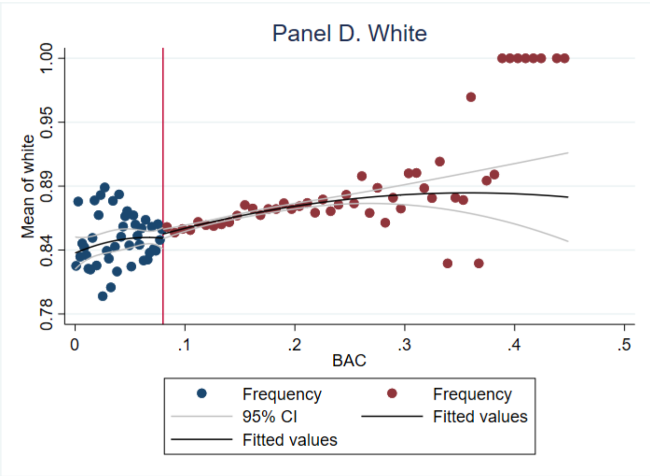

Panel A. Accident at scene
Mean of acc
0.0
0.3
0.5
0.8
1.0
0
.1
.2
.3
.4
.5
BAC
Frequency
Frequency
95% CI
Fitted values
Fitted values

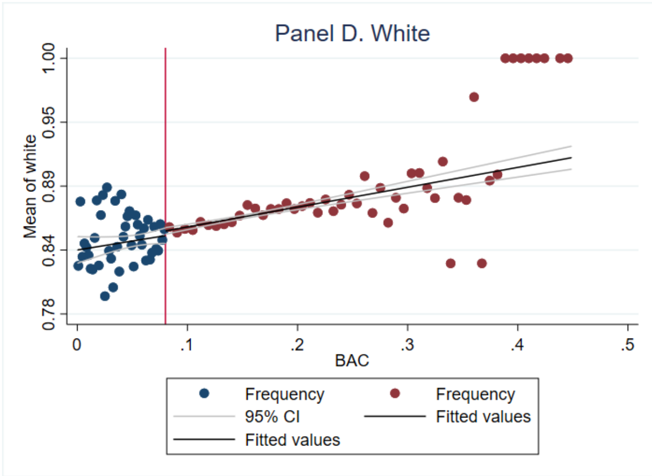

Panel B. Male
Mean of male
0.0
0.3
0.5
0.8
1.0
0
.1
.2
.3
.4
.5
BAC
Frequency
Frequency
95% CI
Fitted values
Fitted values


Panel B. Male
Mean of male
0.0
0.3
0.5
0.8
1.0
0
.1
.2
.3
.4
.5
BAC
Frequency
Frequency
95% CI
Fitted values
Fitted values


Panel C. Age
Mean of aged
30
34
39
43
47
0
.1
.2
.3
.4
.5
BAC
Frequency
Frequency
95% CI
Fitted values
Fitted values


Panel C. Age
Mean of aged
30
34
39
43
47
0
.1
.2
.3
.4
.5
BAC
Frequency
Frequency
95% CI
Fitted values
Fitted values





We may see tiny differences between Hansen's and replicated graphs, basically for the size of the sample used to replicate.
Also is important to point out about the dispersion of the data for BACs superior to 0.2. That dispersion could lead to suspicious about data that may be bad registered or be part of extreme unusual cases.

7. Estimate equation (1) with recidivism (recid) as the outcome. This corresponds to Table 3 column 1, but since I am missing some of his variables, your sample size will be the entire data set of 214,558. Nevertheless, replicate Table 3, column 1, Panels A and B. Note that these are local linear regressions and Panel A uses as its bandwidth 0.03 to 0.13. But Panel B has a narrower bandwidth of 0.055 to 0.105. Your table should have three columns and two A and B panels associated with the different bandwidths.:

    (a) Column 1: control for the bac1 linearly
    (b) Column 2: interact bac1 with cutoff linearly
    (c) Column 3: interact bac1 with cutoff linearly and as a quadratic
    (d) For all analysis, use heteroskedastic robust standard errors.

Estimates for the Effect of Exceeding the 0.08 BAC Threshold on Recidivism

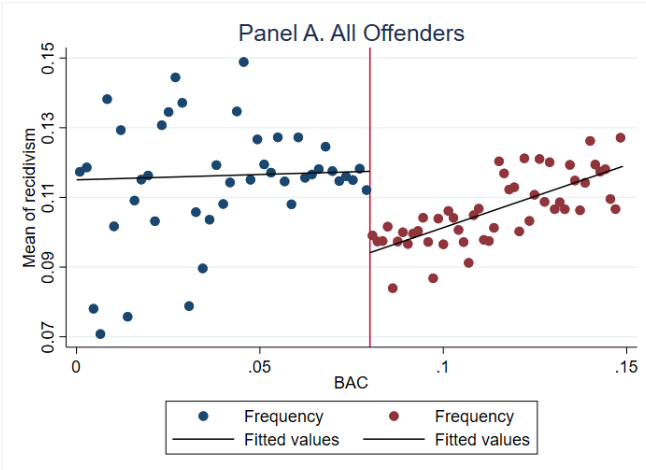
	recidivism BAC linearly	recidivism BAC*cutoff	recidivism BAC^2*cutoff
Panel A. BAC[0.03, 0.13]			
DUI	-0.0273***	-0.0591***	0.113
	(0.00403)	(0.0152)	(0.0843)
Controls	Yes	Yes	Yes
Observations	89,967	89,967	89,967
Panel B. BAC[0.055, 0.105]			
DUI	-0.0219***	-0.0643*	0.371
	(0.00558)	(0.0350)	(0.422)
Controls	Yes	Yes	Yes
Observations	46,957	46,957	46,957

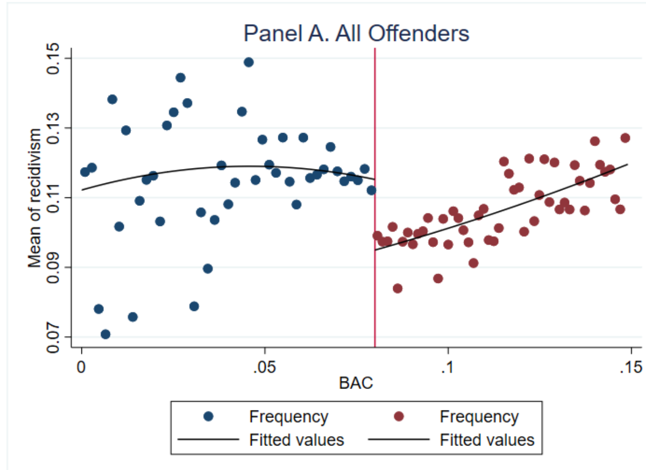
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

8. Recreate the top panel of Figure 3 according to the following rule:

(a) Fit linear fit using only observations with less than 0.15 bac on the bac1.

(b) Fit quadratic fit using only observations with less than 0.15 bac on the bac1





3 Appendix

Replication Table 3 (bandwidth 0.03 to 0.13)

VARIABLES	(1) recidivism	(2) recidivism	(3) recidivism
white	0.0162***	0.0162***	0.0162***
	(0.00280)	(0.00280)	(0.00280)
aged	-0.000847***	-0.000854***	-0.000854***
	(8.49e-05)	(8.50e-05)	(8.50e-05)
acc	0.00444	0.00421	0.00418
	(0.00345)	(0.00345)	(0.00345)
male	0.0332***	0.0332***	0.0332***
	(0.00233)	(0.00233)	(0.00233)
DUI	-0.0273***	-0.0591***	0.113
	(0.00403)	(0.0152)	(0.0843)
bac1	0.321***	-0.0429	2.902*
	(0.0748)	(0.187)	(1.637)
0b.DUI#co.bac1		0	0
		(0)	(0)
1.DUI#c.bac1		0.438**	-4.210**
		(0.204)	(2.111)
0b.DUI#c.sq_bac			-24.72*
			(13.74)
1.DUI#c.sq_bac			8.014
			(6.276)
Constant	0.0853***	0.109***	0.0262
	(0.00672)	(0.0131)	(0.0473)
Observations	89,967	89,967	89,967
R-squared	0.004	0.004	0.004

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Replication Table 3 (bandwidth 0.055 to 0.105)

VARIABLES	(1) recidivism	(2) recidivism	(3) recidivism
white	0.0176***	0.0176***	0.0176***
	(0.00381)	(0.00381)	(0.00381)
aged	-0.000756***	-0.000758***	-0.000758***
	(0.000115)	(0.000115)	(0.000115)
acc	0.00431	0.00422	0.00423
	(0.00497)	(0.00497)	(0.00497)
male	0.0357***	0.0357***	0.0358***
	(0.00317)	(0.00317)	(0.00317)
DUI	-0.0219***	-0.0643*	0.371
	(0.00558)	(0.0350)	(0.422)
bac1	0.188	-0.196	6.167
	(0.201)	(0.383)	(8.120)
0b.DUI#co.bac1		0	0
		(0)	(0)
1.DUI#c.bac1		0.547	-10.52
		(0.449)	(10.61)
0b.DUI#c.sq_bac			-46.06
			(58.75)
1.DUI#c.sq_bac			25.20
			(36.58)
Constant	0.0862***	0.113***	-0.104
	(0.0154)	(0.0278)	(0.278)
Observations	46,957	46,957	46,957
R-squared	0.004	0.004	0.004

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1