

Hansen replication

Weidi Hou

24 Feb 2021

Github repo and summary

1. 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. Save the Hansen_dwi.dta file into your new /data subdirectory. Note: The outcome variable is “recidivism” or “recid” which is measuring whether the person showed back up in the data within 4 months.

<https://github.com/wdhou/RDD>

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?

1. What is his research question?

Additional sanctions experienced by drunk drivers at BAC thresholds are effective in reducing repeat drunk driving.

2. What data does he use?

Administrative records on 512,964 DUI BAC tests in the state of Washington from 1995 to 2011.

3. What is his research design, or “identification strategy”?

- Defining the parameter: $E[BAC_1 - BAC_0 \mid X=0.08]$, $E[BAC_1 - BAC_0 \mid X=0.15]$
- Identifying the parameter(assumption): Sufficient conditions include the continuity of the underlying conditional regression and distribution functions. In short, these assumptions imply that both the unobservables and observables are expected to remain unchanged across the threshold with only treatment status (or the probability of treatment) changing. Another assumption is that some drunk drivers are randomly lucky, having a BAC barely below the threshold, while other drunk drivers are randomly unlucky and have a BAC barely above the threshold.
- Estimation: The author used local linear regression discontinuity design to estimate the effect of having a BAC above the DUI or aggravated DUI threshold on recidivism, with the slopes allowed to change at the discontinuities, as shown in equation (1):

$$y_i = X'_i\gamma + \alpha_1DUI_i + \alpha_2BAC_i + \alpha_3BAC_i \times DUI_i + u_i$$

The main results are based on a local-linear regression discontinuity design with a rectangular kernel. In the regression models the BAC variable is rescaled around the relevant threshold, either 0.08 or 0.15, X_i is a vector of controls, and y_i is a measure of recidivism.

- Density test: Firstly, The frequency figure (Figure 1) was showed to illustrate that there is no obvious jump at the cutoff points 0.08 and 0.15. Then the McCrary test implies p-values of 0.59 and 0.38 respectively at the 0.08 and 0.15 thresholds. Frandsen (2013) recently offered an alternative density test based on a local approximation to a binomial distribution that may offer improvements both in consistency and finite sample performance when the data are discrete or rounded. That test estimates a p-value of 0.795 at 0.08, and 0.886 at 0.15, again revealing no evidence of manipulation. Likewise, the histogram also shows little evidence of non-random heaping, which can also create bias in regression discontinuity designs (Barreca, Lindo, and Waddell 2011).
- Balance test: Table 2 of the paper estimates of the effect of having a BAC over the DUI and aggravated DUI thresholds on predetermined characteristics which should be unaffected by BAC thresholds, employing the same regression model as equation (1), with the control variables utilized as the dependent variable. The regression models estimated are local linear models as described above, and employ a bandwidth of 0.05 and a rectangular kernel for weighting. The author fail to reject the null that the predetermined characteristics are unrelated to the BAC cutoffs for DUI and aggravated DUI.
- bandwidth: Table 3 shows the outcome of treatment effect. panel A utilizes a bandwidth of 0.05 while panel B uses a bandwidth of 0.025, and all regressions use a rectangular kernel for weighting. From Appendix Figure 3, For every possible bandwidth from 0.005 to 0.068, the estimated effect of having a BAC above the DUI or aggravated DUI threshold is presented along with the 95 percent confidence interval. The point estimates are relatively stable across nearly all bandwidths. Except for particularly small bandwidths, generally those less 0.02, the estimates are also statistically significant (at least at the 95 percent level). The stability of the estimates across various bandwidths suggests that the linear specification is a reasonable choice for modeling the effect of BAC on recidivism.

4. What are his conclusions?

First, he verifies the effectiveness of punishments and sanctions in reducing recidivism among drunk drivers, finding evidence that having a BAC above either the 0.08 DUI threshold or the 0.15 aggravated DUI is associated with reduced repeat drunk driving both in the short and long term.

Second, several mechanisms which could explain the reductions in repeat driving associated with a BAC over the legal thresholds such as incapacitation, rehabilitation, and deterrence. Lastly, although public debate has often focused on the BAC limit, including

recent proposal by the NTSB to lower the BAC limit to 0.05, there are also other viable policies in addition to lowering the BAC limit.

Reproducing somewhat Hansen's results

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 $\text{bac1} \geq 0.08$ and 0 otherwise in your do file or R file.

Using data set Hansen.dta:

```
. use "/Users/macbookpro/Desktop/spring/causal/rep_hw1/hansen_dwi.dta"
```

Create a dummy treatment variable:

```
. gen dui=0

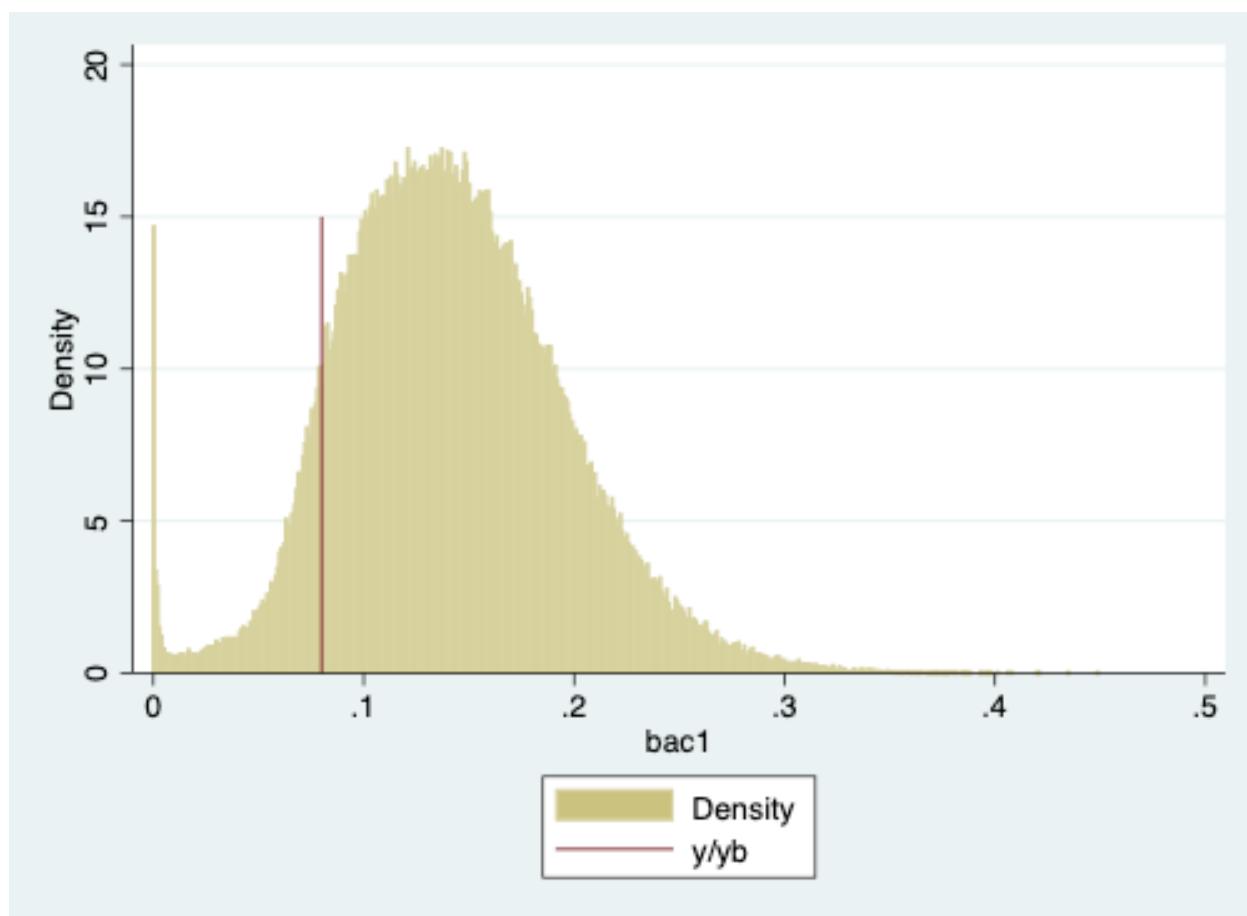
. replace dui=1 if bac1>=0.08
(191,548 real changes made)
```

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? Explain your results. Compare what you found to what Hansen found.

I use the frequency figure to see whether there is a big jump on the cutoff

```
. hist bac1, bin(1000) addplot(pci 0 0.08 15 0.08)
(bin=1000, start=0, width=.000449)

. graph export frequency.png, width(500) replace
(file /Users/macbookpro/Desktop/frequency.png written in PNG format)
```



BAC distribution

Bac1 Frequency Figure contains a histogram displaying the number of observations in each measured bac1 level from 1999–2007. From the figure, I didn't find obvious jump at the threshold 0.08. Therefore, the distribution of bac1 shows little evidence of sorting to one side of the thresholds at 0.08.

In addition, I use the local polynomial density estimators to implement manipulation test:

```
. rddensity bac1, c(0.08) p(1)
Computing data-driven bandwidth selectors.
```

Point estimates and standard errors have been adjusted for repeated observations.
> ns.

(Use option nomasspoints to suppress this adjustment.)

RD Manipulation test using local polynomial density estimation.

```

      c =      0.080 | Left of c Right of c      Number of obs =      2145
> 58
-----|-----
> ed      Number of obs |      23010      191548      Model = unrestrict
> mb      Eff. Number of obs |      14132      27467      BW method = co
> ar      Order est. (p) |          1          1      Kernel = triangul
> fe      Order bias (q) |          2          2      VCE method = jackkni
      BW est. (h) |      0.022      0.022

```

Running **variable**: bac1.

Method	T	P> T
Robust	1.3243	0.1854

P-values of binomial tests. (H0: **prob** = .5)

Window Length / 2	<c	>=c	P> T
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000
0.000	909	0	0.0000

From the stata output, the p-value is 0.1854. So we can't reject the null: There is no sorting on the running variable. This gets the same result as the previous Frequency Figure.

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.

```

. quietly rdrobust white bac1, c(0.08) h(0.03 0.13) kernel(uniform)

. mat whiteoc=e(b)

. quietly rdrobust male bac1, c(0.08) h(0.03 0.13) kernel(uniform)

. mat maleoc=e(b)

```

```
. quietly rdrobust age bac1, c(0.08) h(0.03 0.13) kernel(uniform)

. mat ageoc=e(b)

. quietly rdrobust acc bac1, c(0.08) h(0.03 0.13) kernel(uniform)

. mat accoc=e(b)

. local f %6.3f
```

Variables	White	Male	age	acc
DUI	0.002 (0.005)	0.008 (0.006)	-0.610 (0.167)	-0.013 (0.004)
Controls	No	No	No	No
Observations	214,558	214,558	214,558	214,558

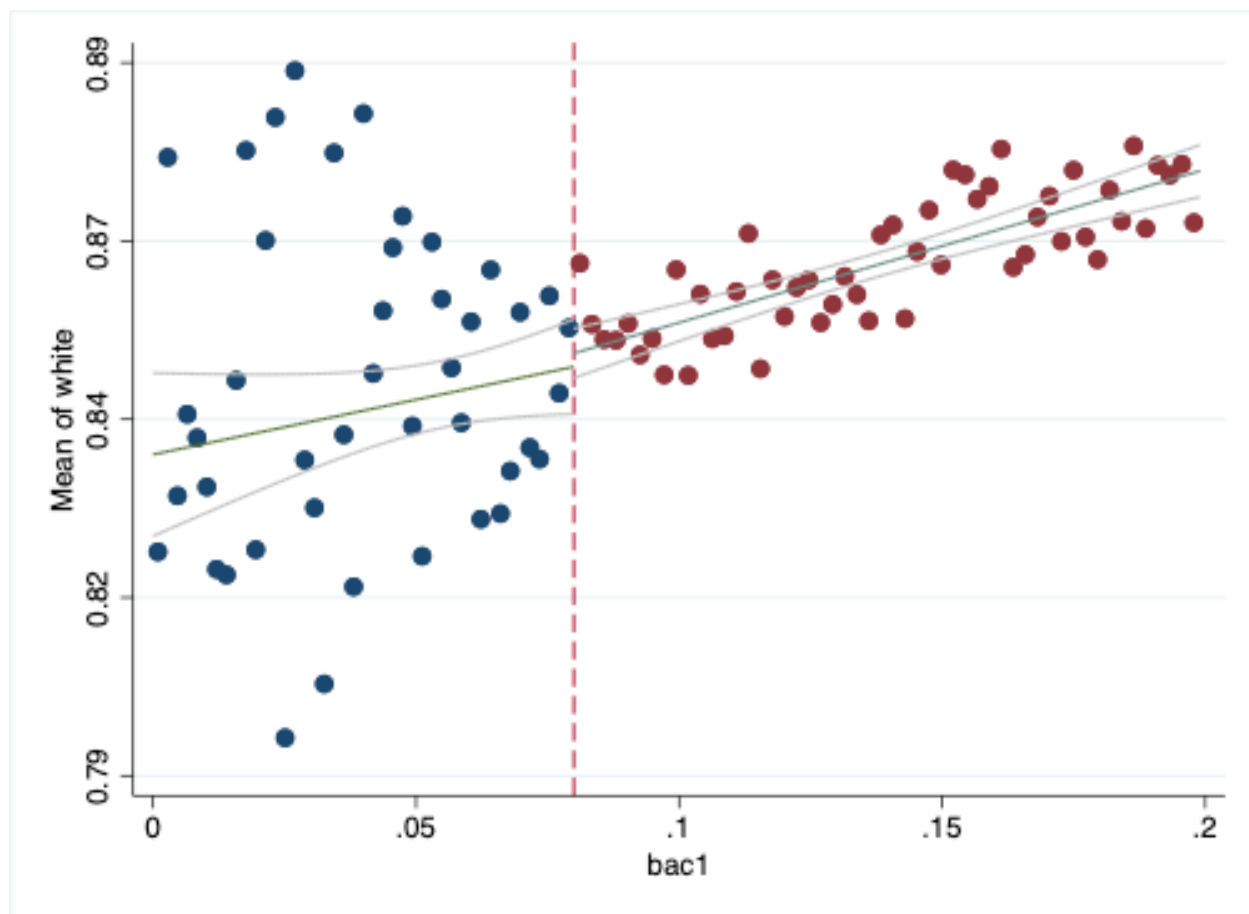
From the outcomes, I found that the covariates white and male are covariates balanced at the cutoff 0.08, but covariates age and acc are not. This result is a little different from that of Hansen's. I think the main reason is that I didn't get the full dataset as Hansen. Afterall, in order to get convincing result in RDD, we need to make sure there is no endogenous cutoff.

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.

Linear fit of BAC AND CHARACTERISTICS

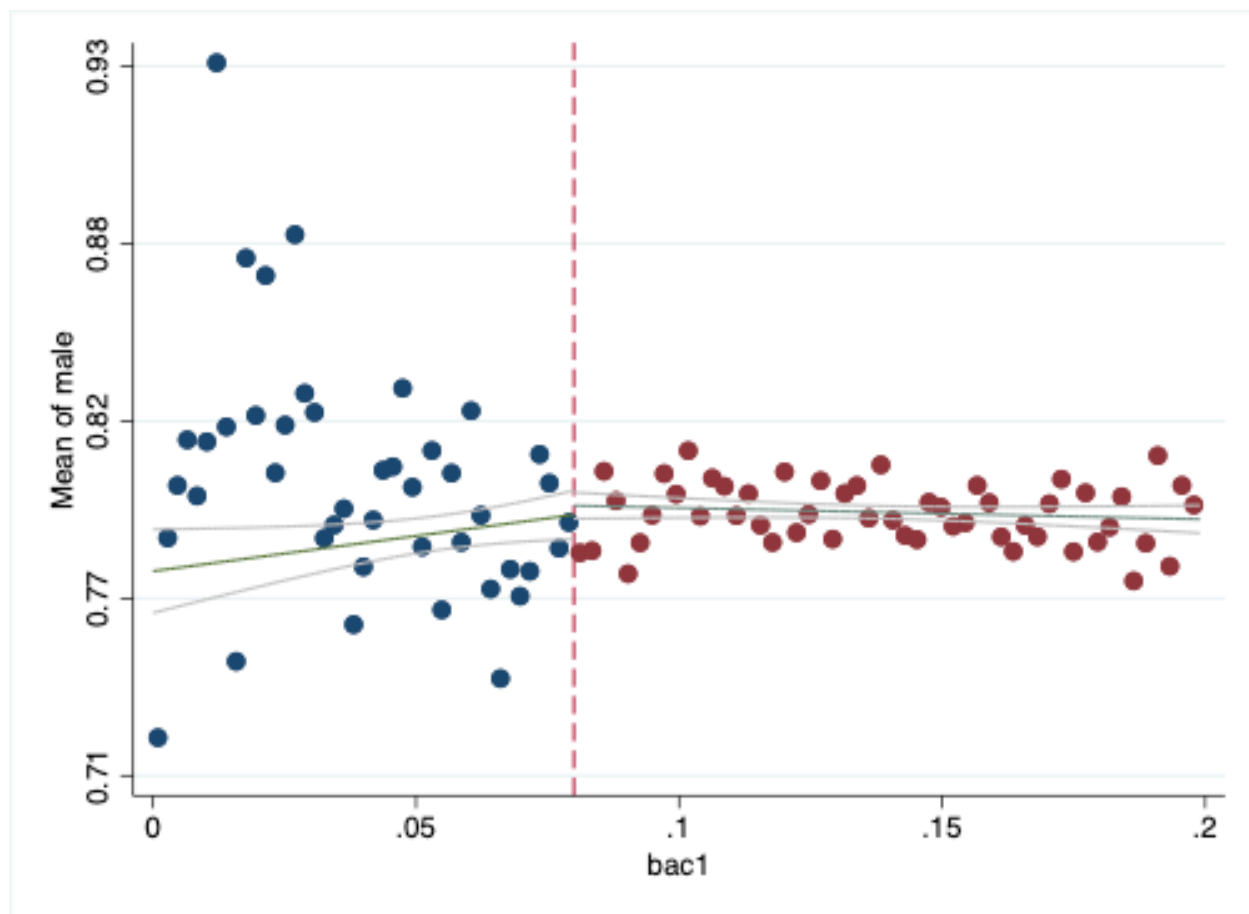
```
. quietly cmogram white bac1 if bac1<=0.2, cut(0.08) scatter lineat(0.08) lfitc
> i

. graph export whitet21.png, width(500) replace
(file /Users/macbookpro/Desktop/whitet21.png written in PNG format)
```



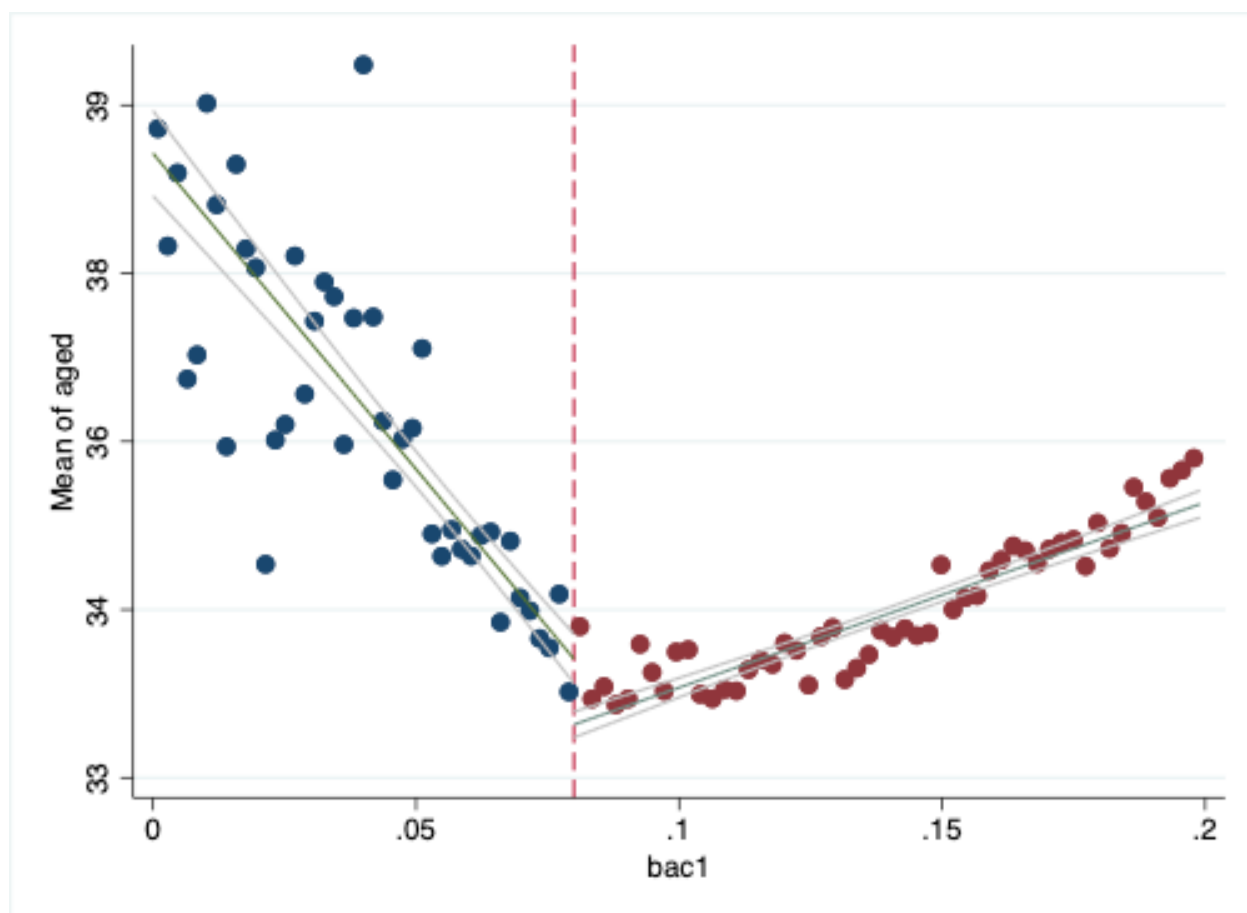
Panel A. White

```
. quietly cmogram male bac1 if bac1<=0.2,cut(0.08) scatter lineat(0.08) lfitci
. graph export malet21.png, width(500) replace
(file /Users/macbookpro/Desktop/malet21.png written in PNG format)
```



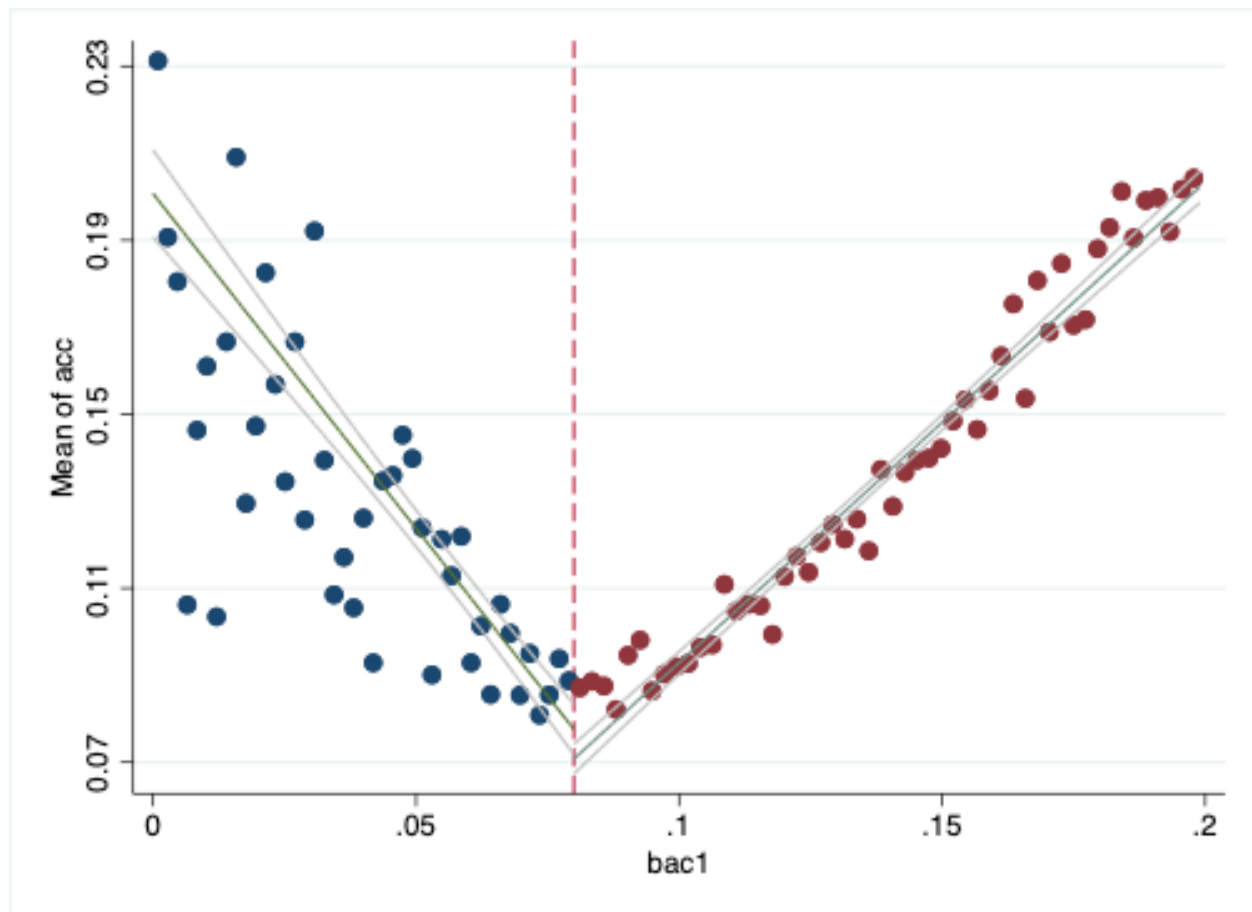
Panel B. Male

```
. quietly cmogram age bac1 if bac1<=0.2,cut(0.08) scatter lineat(0.08) lfitci
. graph export aget21.png, width(500) replace
(file /Users/macbookpro/Desktop/aget21.png written in PNG format)
```

Panel C. Age

```
. quietly cmogram acc bac1 if bac1<=0.2, cut(0.08) scatter lineat(0.08) lfitci
. graph export acct21.png, width(500) replace
(file /Users/macbookpro/Desktop/acct21.png written in PNG format)
```



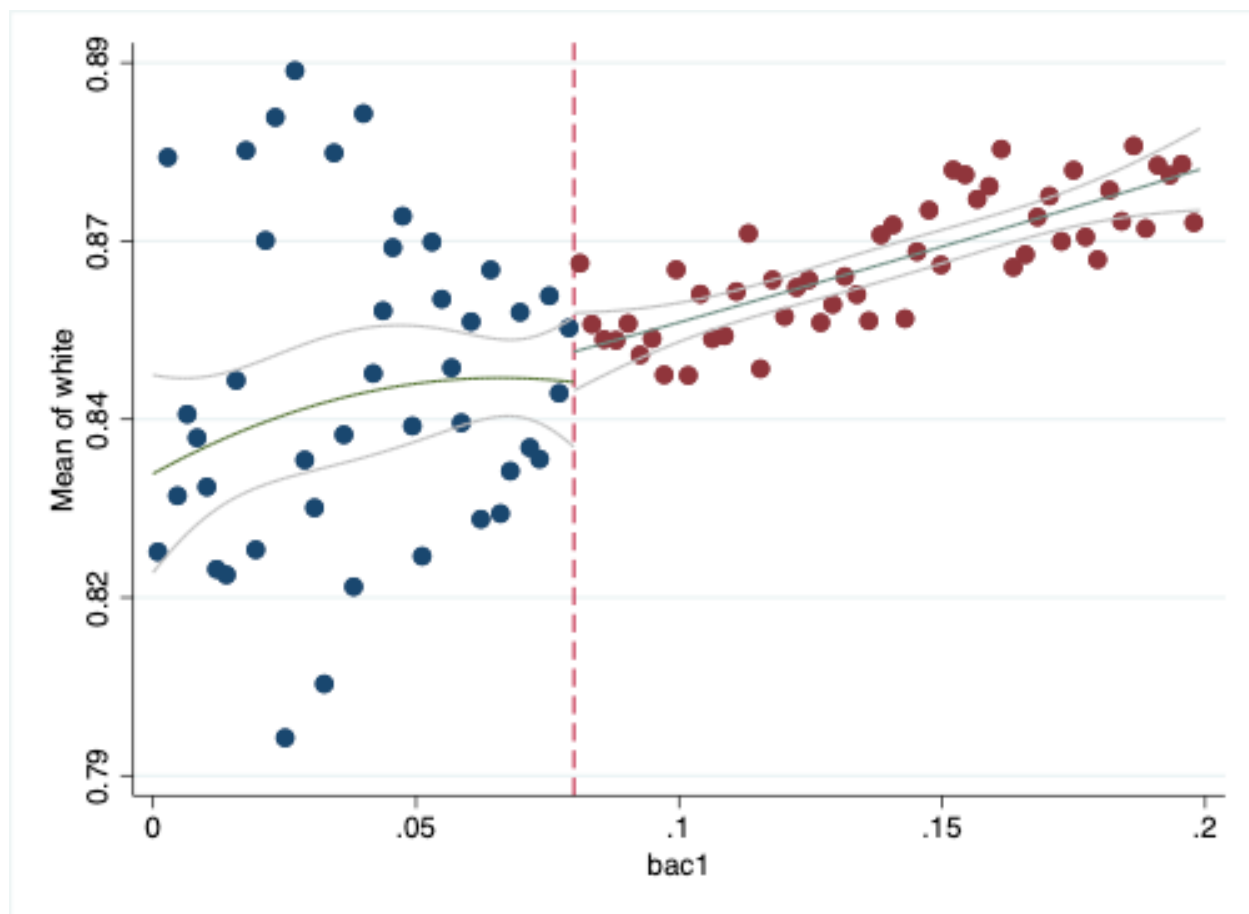
Panel D. Acc

From above figures, I found that the CI of white and male variables are the same on either side of the cutoff, which means the smoothness holds. However, as showed in the balance test, the variables age and acc are not the same on each side of the cutoff, and their confidence intervals didn't overlap at the cutoff line. This means that when we use the linear fit, we got a slightly different results as Hansen. I think the reason leading to it is that we didn't get the whole data set.

Quadratic fit of BAC AND CHARACTERISTICS

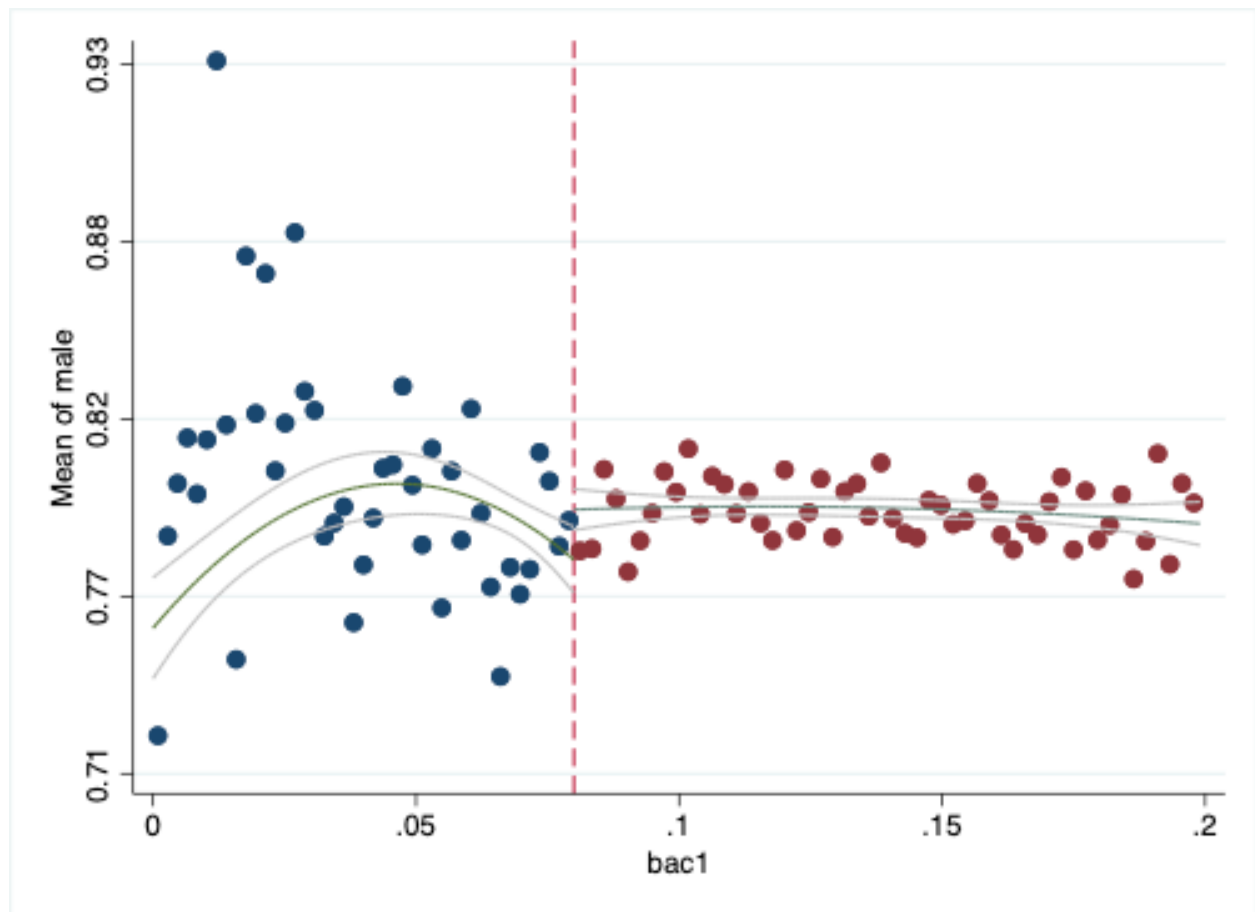
```
. quietly cmogram white bac1 if bac1<=0.2, cut(0.08) scatter lineat(0.08) qfitc
> i
```

```
. graph export whitet2q.png, width(500) replace
(file /Users/macbookpro/Desktop/whitet2q.png written in PNG format)
```



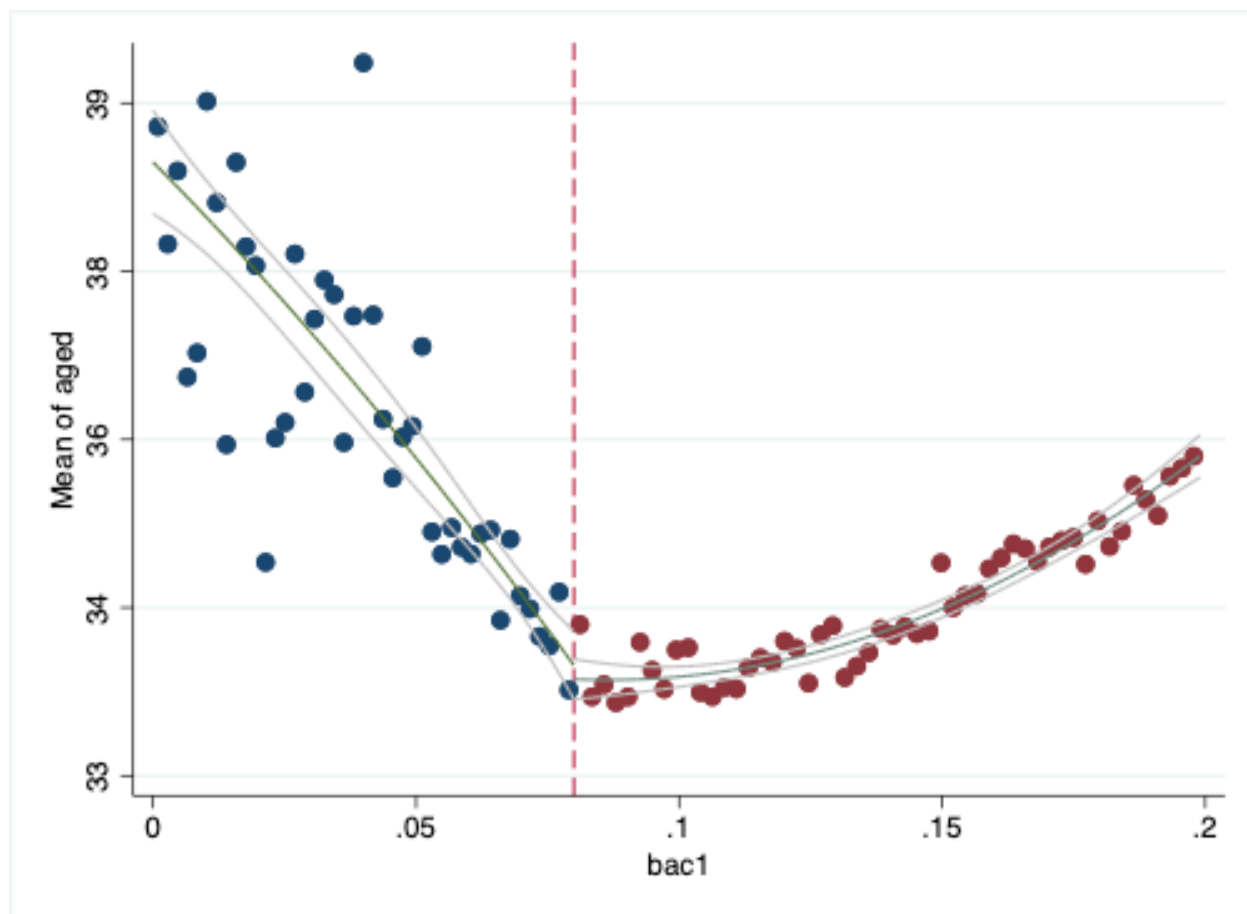
Panel A. White

```
. quietly cmogram male bac1 if bac1<=0.2,cut(0.08) scatter lineat(0.08) qfitci
. graph export malet2q.png, width(500) replace
(file /Users/macbookpro/Desktop/malet2q.png written in PNG format)
```



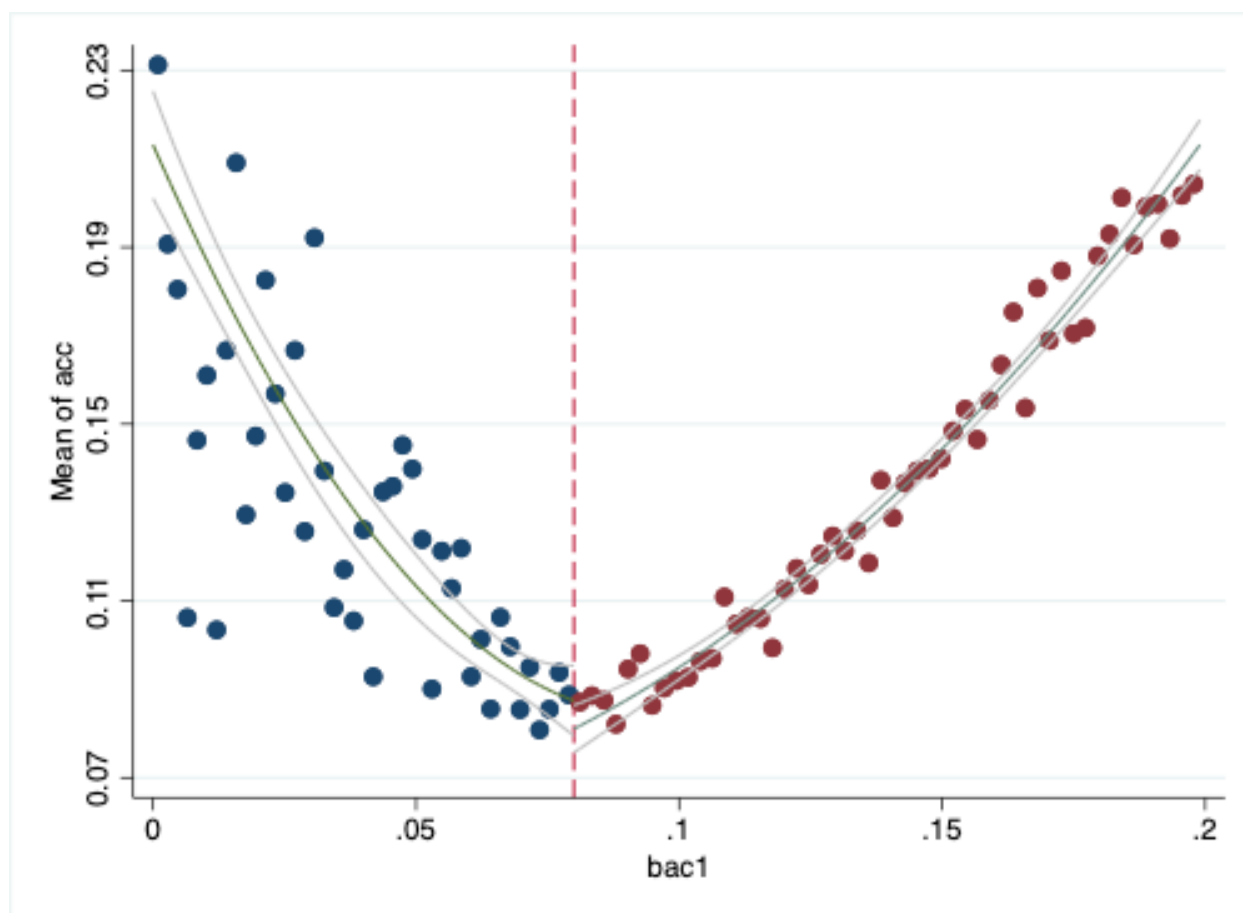
Panel B. Male

```
. quietly cmogram age bac1 if bac1<=0.2, cut(0.08) scatter lineat(0.08) qfitci
. graph export aget2q.png, width(500) replace
(file /Users/macbookpro/Desktop/aget2q.png written in PNG format)
```



Panel C. Age

```
. quietly cmogram acc bac1 if bac1<=0.2,cut(0.08) scatter lineat(0.08) qfitci
. graph export acct2q.png, width(500) replace
(file /Users/macbookpro/Desktop/acct2q.png written in PNG format)
```



Panel D. Acc

For the above figures, we used quadratic fit. At this time, all demographic factors such as age, race and gender are stable across the DUI punishment thresholds. Likewise key sources of information that could drive the police to administer a breath test including the BAC measured at a portable breath test (if taken), the presence of an accident at the scene is also unchanged, which is the same as Hansen's.

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 dataset 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.:

1. Column 1: control for the bac1 linearly
2. Column 2: interact bac1 with cutoff linearly
3. Column 3: interact bac1 with cutoff linearly and as a quadratic
4. For all analysis, estimate uncertainty using heteroskedastic robust standard errors. [ed: But if you want to show off, use Kolesár and Rothe's 2018 "honest" confidence intervals (only available in R).

Panel A:

```
. quietly reg recidivism bac1 male white acc age if bac1>=0.03 & bac1<=0.13

. mat linear1=e(b)

. quietly rdrobust recidivism bac1, c(0.08) h(0.03 0.13) kernel(uniform) covs(
> male white acc age)

. mat rd1=e(b)

. quietly rdrobust recidivism bac1, c(0.08) h(0.03 0.13) kernel(uniform) p(2)
> covs(male white acc age)

. mat rdquadratic1=e(b)

. local f %6.3f
```

Panel B:

```
. quietly reg recidivism bac1 male white acc age if bac1>=0.055 & bac1<=0.105

. mat linear2=e(b)

. quietly rdrobust recidivism bac1, c(0.08) h(0.055 0.105) kernel(uniform) cov
> s(male white acc age)

. mat rd2=e(b)

. quietly rdrobust recidivism bac1, c(0.08) h(0.055 0.105) kernel(uniform) p(2)
> ) covs(male white acc age)

. mat rdquadratic2=e(b)

. local f %6.3f
```

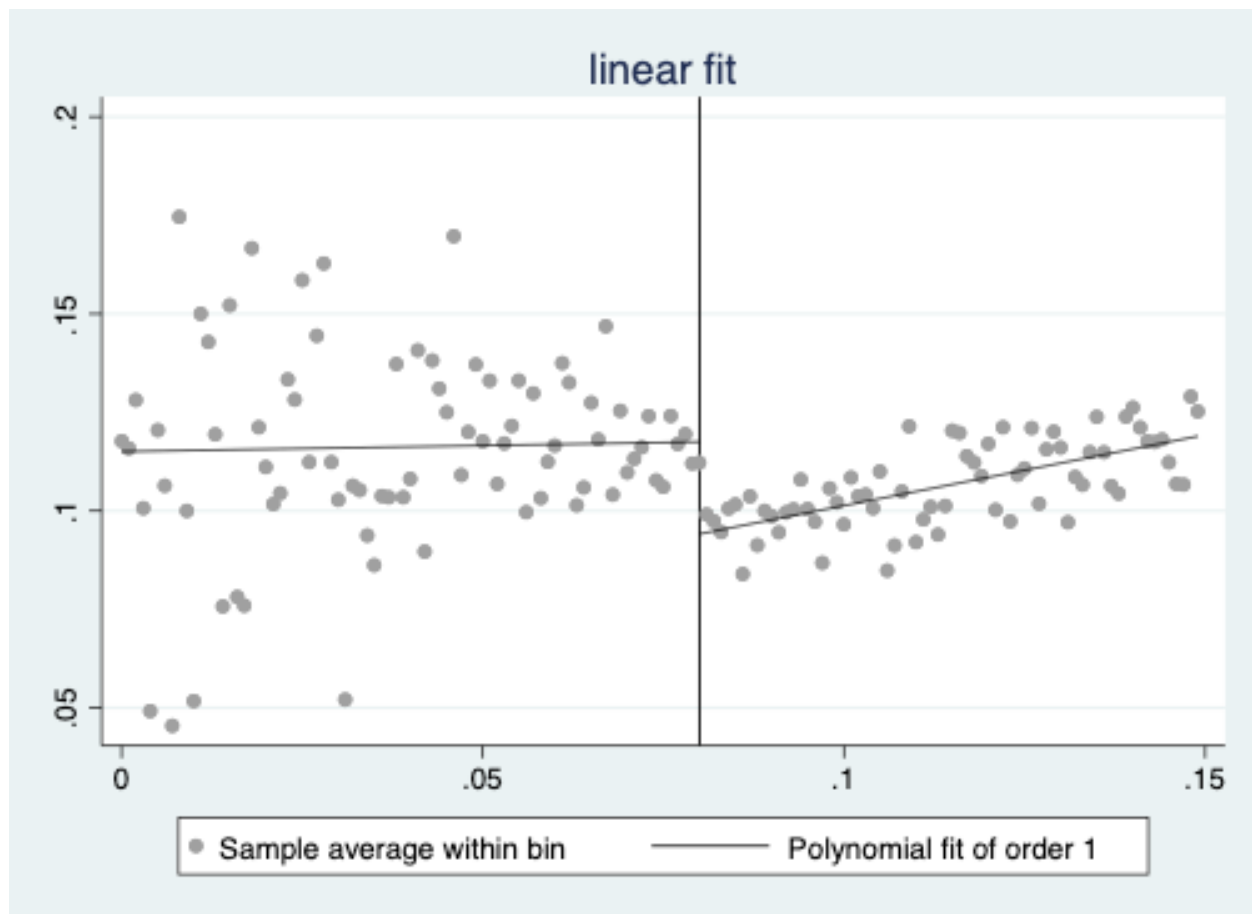
	bac1 Linearly	interact liearly	interact liearly and as quadratic
panel A.	BAC \in [0.03,0.13]		
DUI	-0.075 (0.047)	-0.020*** (0.004)	-0.018*** (0.006)
controls	Yes	Yes	Yes
observations	214,558	214,558	214,558
panel B.	BAC \in [0.0055,0.105]		
DUI	-0.476***	-0.020***	-0.020***

	(0.109)	(0.004)	(0.005)
controls	Yes	Yes	Yes
observations	214,558	214,558	214,558

8. Recreate the top panel of Figure 3 according to the following rule: Fit linear fit using only observations with less than 0.15 bac1. Fit quadratic fit using only observations with less than 0.15 bac1

```
. quietly rdplot recidivism bac1 if bac1<=0.15,c(0.08) p(1) graph_options(titl
> e(linear fit))
```

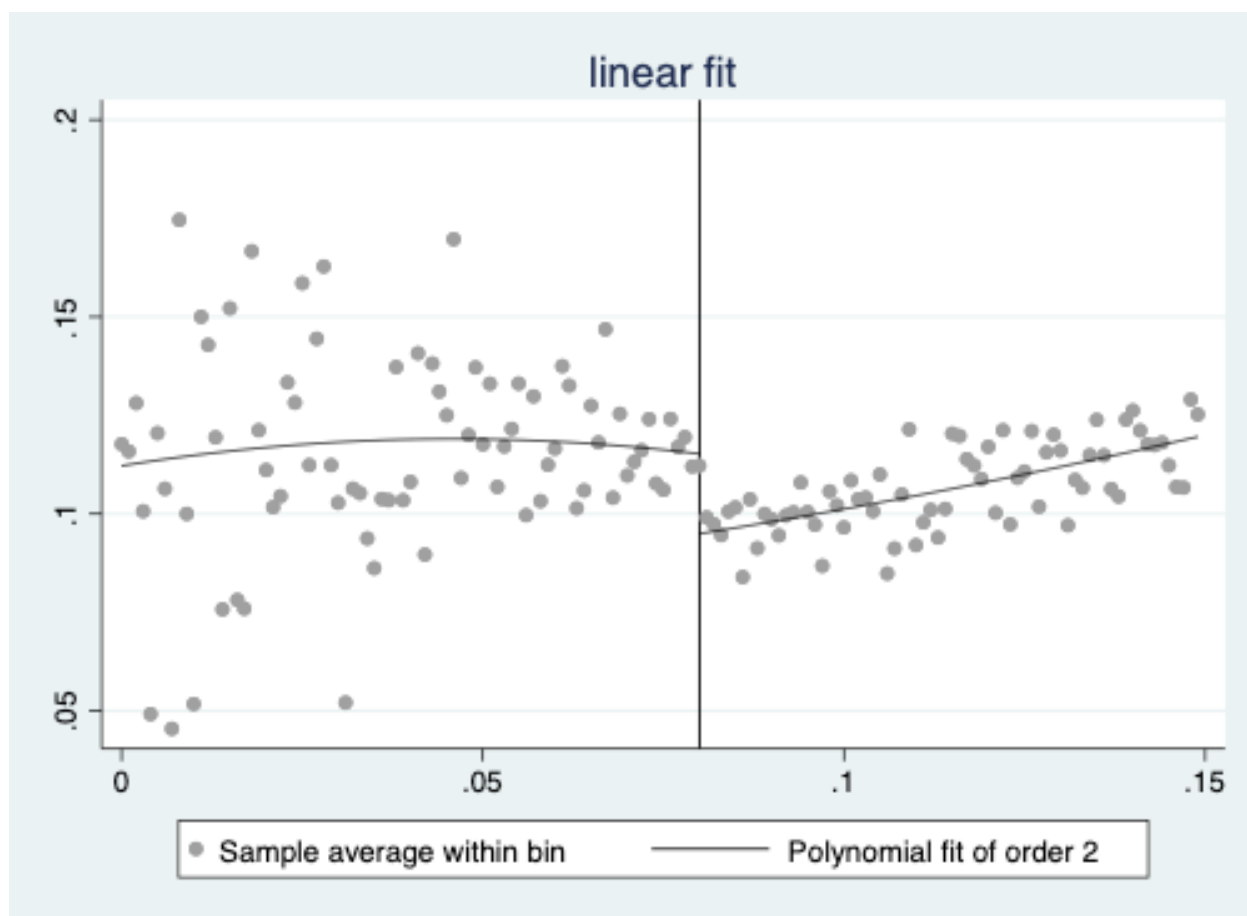
```
. graph export table3repln.png, width(500) replace
(file /Users/macbookpro/Desktop/table3repln.png written in PNG format)
```



Linear fit

```
. quietly rdplot recidivism bac1 if bac1<=0.15,c(0.08) p(2) graph_options(titl
> e(linear fit))
```

```
. graph export table3repqr.png, width(500) replace
(file /Users/macbookpro/Desktop/table3repqr.png written in PNG format)
```

Quadratic fit

9. Discuss what you learned from this exercise. What was the hypothesis you tested and what did you find? How confident are you in Hansen's original conclusion? Why/why not?

I learned the main steps of doing RDD. Specifically, I learned how to test the smoothness of potential outcomes. Since we couldn't get the counterfactual data, we needed to density test and balance test to check the smoothness. First, I recreated a frequency figure and then found no manipulation at the cutoff 0.08. In order to make sure, I used `rdrobust` command to test and I didn't find evidence of sorting on the running variable. Then, I used the balance test to check whether the predeterminants were independent from the cutoff. These above two tests provided evidence of smoothness. From my replication, my results were similar to those of Hansen's except for the balance test of variable "acc" and "age". I think the main reason leading to it was the missing data, I didn't get the full data set as Hansen. Second, I used local polynomial regression discontinuity estimation and got the similar outcomes as Hansen's original results. Therefore, I had confidence to believe Hansen's original conclusion.

From this exercise, I deeply studied the paper first, knew the main steps of RDD and learned how to use it in the real life. Furthermore, I learned how to use stata to make discontinuity regression and the way of using stata markdown to make my assignment neatly.