**Github repo and summary (worth 2 points)**

1. Download Hansen\_dwi.dta from github at the following address.

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. 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.

1. 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?

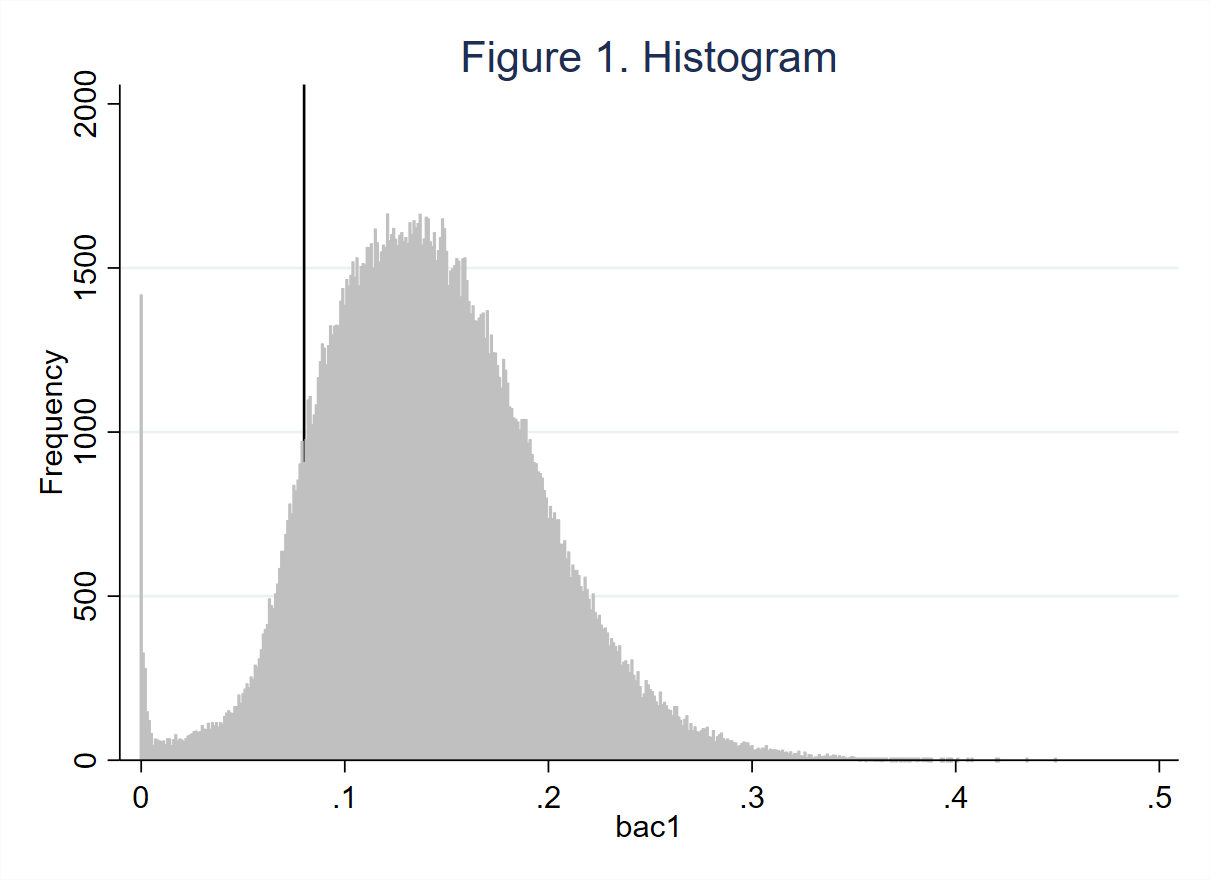
The authors explore the effects of punishment severity over the commission of future crimes. Specifically, they try to respond whether the blood alcohol content (BAC) limits and the increasingly harsh sanctions that surpassing such limits implies have any effect on reducing future drunk driving. To do this, the employ administrative records on 512,964 driving under the influence (DUI) BAC tests in the state of Washington from 1995 to 2011, checking if individuals incurred on DIU within a four-year window after their original BAC test.

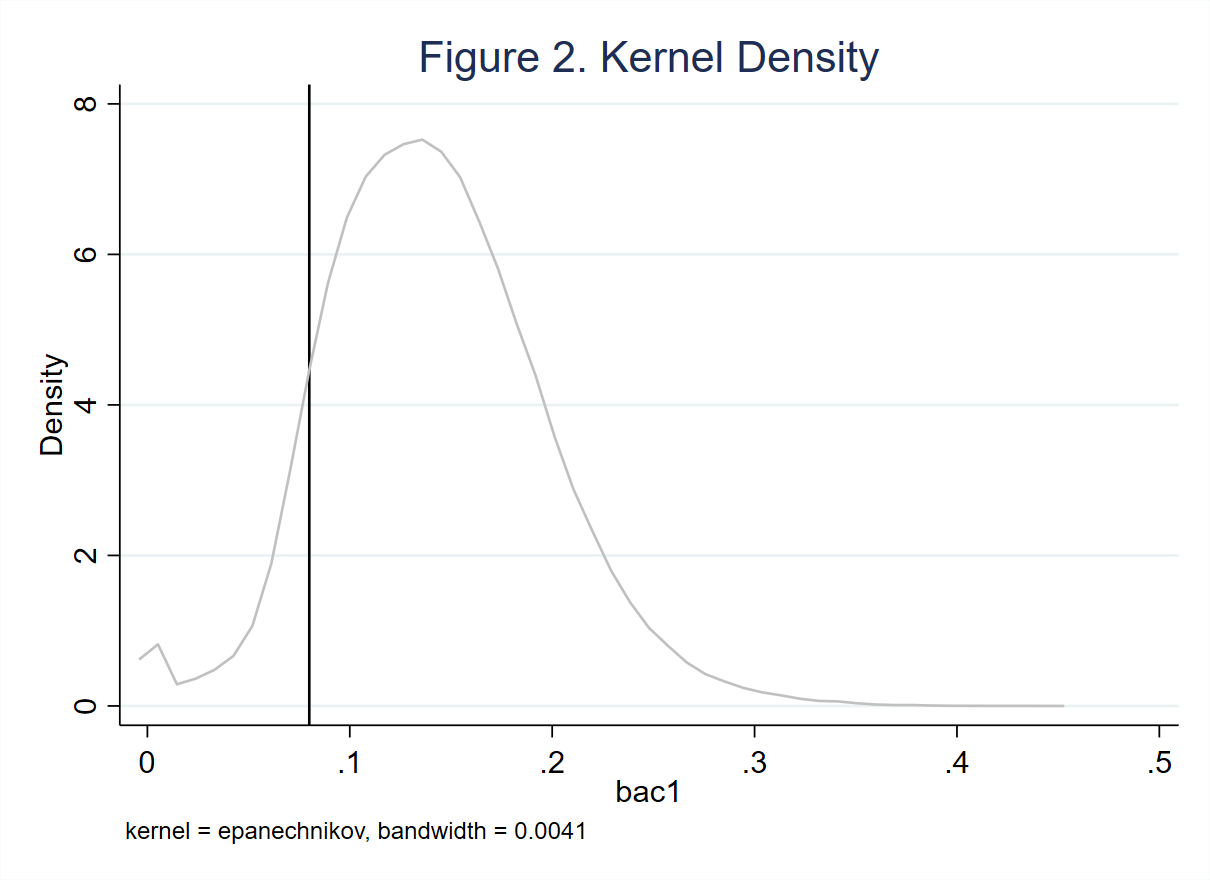
Given that WA applies both a 0.08 and a 0.15 threshold to determine the severity of the punishments for DUI, the authors draw upon a regression discontinuity design using these values as cutoffs. They take advantage of both the exogeneity of these thresholds and the fact that people can not accurately distinguish between two sufficiently close BACs. Therefore, having a BAC slightly over or under the limit is as good as random. The study finds that having a BAC above the 0.08 threshold decreases recidivism by 2 percentage points (p.p) and a BAC over 0.15 does so in an extra 1p.p. Hence, the authors conclude that the severity of the sanctions has a causal negative effect over recidivism.

**Replication (worth 6 points)**.[[1]](#footnote-1)

1. 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.
2. 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?

To check for possible sorting on the running variable it is standard to use a McCrary density test. If the results of the assignment variable are not being manipulated by the individuals, then its density function should be continuous. On the contrary, if individuals are willingly choosing to alter their results in order to receive or avoid treatment, then the density function would present a heap at the cutoff. Therefore, the null hypothesis for the McCrary test states that the density function is continuous at the cutoff, while the alternative claims it is not. To carry on this test, the assignment variable must be partitioned into bins and the frequency of each of these beans must be calculated. Then, those frequencies are used as dependent variables in local linear regressions.

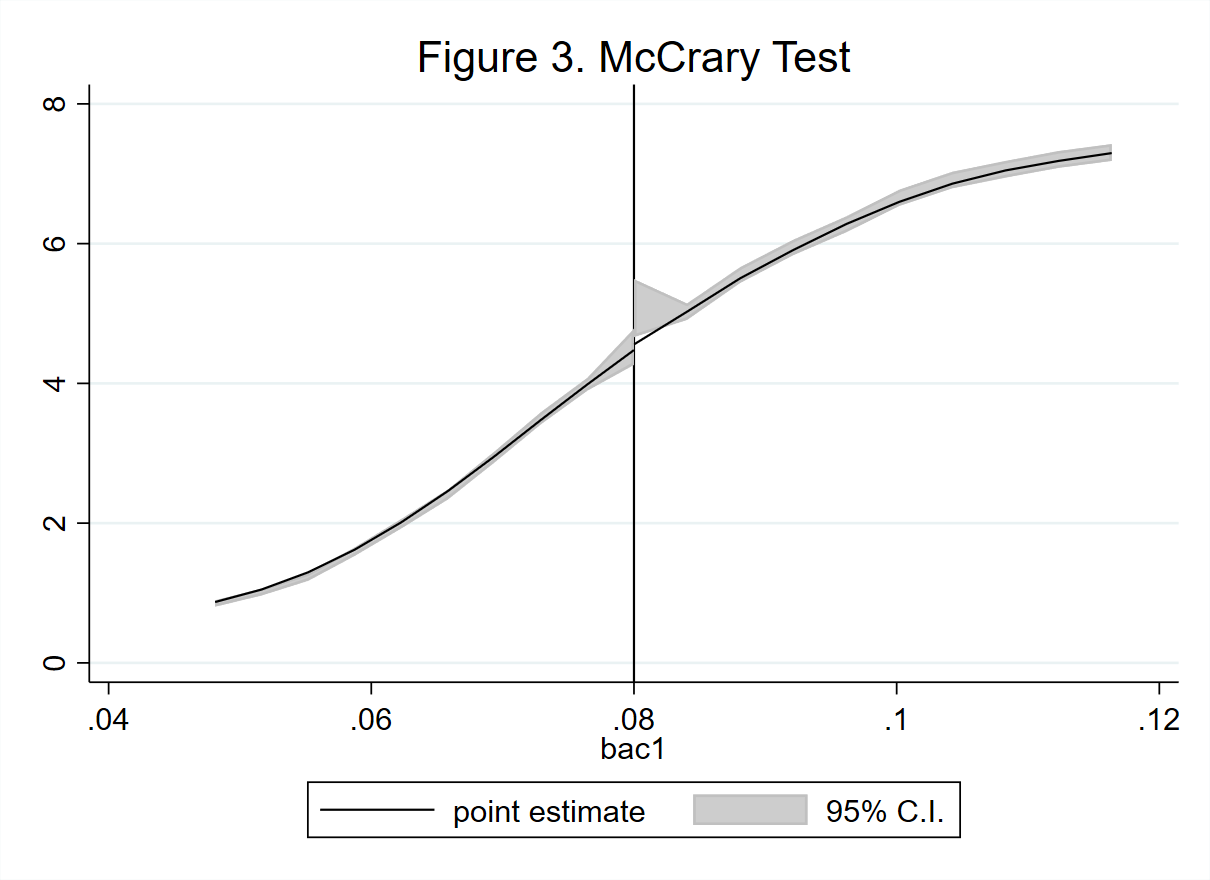




Neither the histogram in figure 1 nor the kernel density estimate in figure 2 show clear signs of discontinuity of the density function of bac1 at the cutoff. Nonetheless, contrary to the conventional McCrary test -and to the results presented by the authors-, the McCrary robust test rejects the null, hence suggesting that the density function of the running variable is not continuous at the 0.08 threshold, which indicates that individuals might be manipulating their BACs.

Table 1. McCracry Test

|  |  |  |
| --- | --- | --- |
| Method | T | P>T |
| Conventional | 0.534 | 0.594 |
| Robust | 2.203 | 0.028 |
|  | | |



1. 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.

Table 2. Regression Discontinuity Estimates for the Effect of Exceeding BAC Thresholds on Predetermined Characteristics

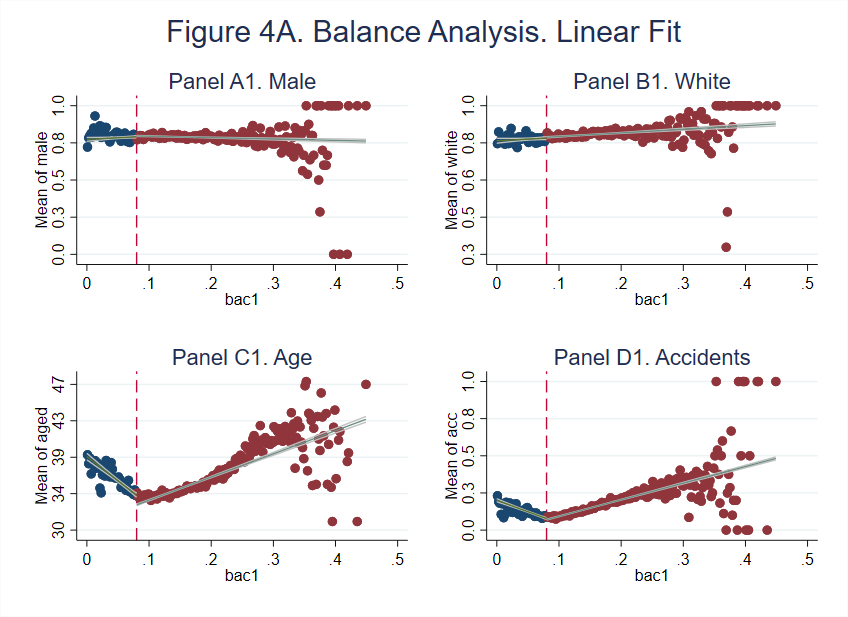
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) |
| VARIABLES | male | white | aged | acc |
|  |  |  |  |  |
| D | 0.0307\*\*\* | 0.00271 | -7.787\*\*\* | -0.219\*\*\* |
|  | (0.00750) | (0.00653) | (0.215) | (0.00682) |
| bac1 | 0.218\* | 0.154 | -56.36\*\*\* | -1.540\*\*\* |
|  | (0.112) | (0.0980) | (3.223) | (0.0977) |
| Dbac1 | -0.311\*\*\* | 0.0170 | 83.40\*\*\* | 2.656\*\*\* |
|  | (0.114) | (0.0994) | (3.268) | (0.0995) |
|  |  |  |  |  |
| Observations | 214,558 | 214,558 | 214,558 | 214,558 |
| Mean | 0.7895 | 0.8615 | 34.9573 | 0.1472 |
| Controls | No | No | No | No |
| R-squared | 0.000 | 0.001 | 0.013 | 0.021 |

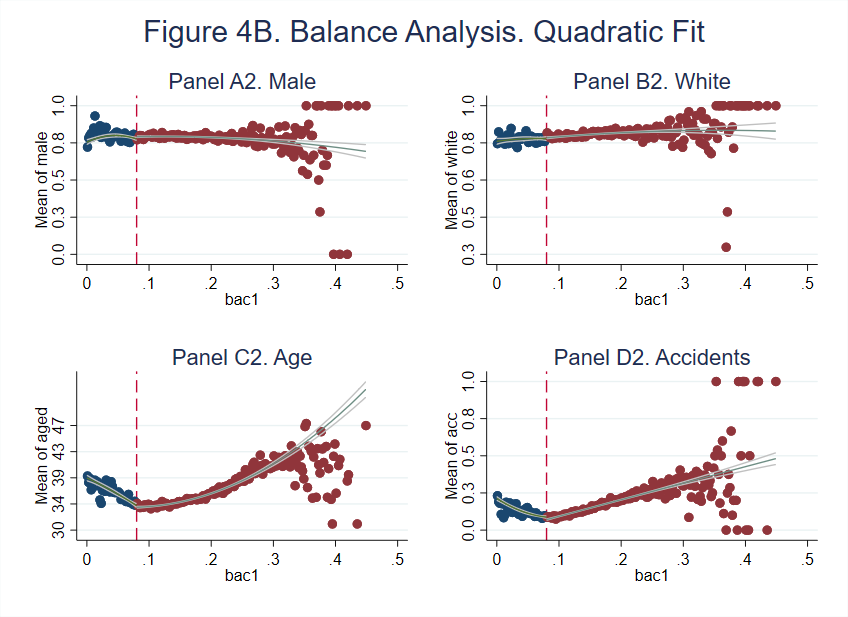
Robust standard errors in parentheses

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

The results of the balance test are rather different from those reported by Hansen. Table 2 shows the regression estimates for the covariates as dependent variables and the treatment, the running variable, and their interaction as explanatory variables. The estimation suggests that only for the variable “white” there is no statistically significant discontinuity at the 0.08 cutoff. On the other hand, the estimation for male presents a significant yet not large discontinuity at the cutoff and a different slope after this, as the interaction coefficient shows. However, for aged and acc the discontinuity at the cutoff is not only statistically different from zero, but also of considerable magnitude. The change in slope after the cutoff for these variables is also large. Hence, the balance check does not provide too strong a proof in favor of the continuity of the potential outcomes function, which is consistent with the findings mentioned in point 4.

1. 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.





Figures 4A and 4B present the graphic results of the balance analysis. The graphs for all four variables are visibly similar to the ones presented by the authors. For example, the graphs for white exhibit a slightly less pronounced yet upward trend than that showed in the paper. Age and accident both describe an initially negative slope that turns positive nearby the cutoff and stays like that thereafter. Finally, male follows a rather steadier trend here than in the paper, but this is more likely to be due to the greater number of observations in our analysis and the high level of dispersion male variable. To finish, a comparison between the linear and the quadratic fits suggest that while male and white are better adjusted by the linear regression, age and acc are best fitted by the quadratic one.

Nevertheless, the graphs here obtained display more pronounced “jumps” for all covariates than those in the study, excluding white. Even though the quadratic fit seems to smoothen these jumps, they are still quite noticeable. This indicates that there might be manipulation of the running variable, as has been warned in the previous points.

1. 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, use heteroskedastic robust standard errors.

Table 3—Regression Discontinuity Estimates for the Effect of Exceeding the 0.08 BAC Threshold on Recidivism

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1) | (2) | (3) |
| VARIABLES | No Inter | Inter | Inter sq |
| **Panel A: bac1 in [0.03,0.13]** |  |  |  |
| D | -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) |
| Dbac1 |  | 0.438\*\* | -4.210\*\* |
|  |  | (0.204) | (2.111) |
| bac1sq |  |  | -24.72\* |
|  |  |  | (13.74) |
| Dbac1sq |  |  | 32.73\*\* |
|  |  |  | (15.10) |
| Constant | 0.0853\*\*\* | 0.109\*\*\* | 0.0262 |
|  | (0.00672) | (0.0131) | (0.0473) |
|  |  |  |  |
| Observations | 89,967 | 89,967 | 89,967 |
| Mean | 0.1069 | 0.1069 | 0.1069 |
| Controls | Yes | Yes | Yes |
| R-squared | 0.004 | 0.004 | 0.004 |
|  |  |  |  |
| **Panel B: bac1 in [0.055,0.105]** |  |  |  |
|  |  |  |  |
| D | -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) |
| Dbac1 |  | 0.547 | -10.52 |
|  |  | (0.449) | (10.61) |
| bac1sq |  |  | -46.06 |
|  |  |  | (58.75) |
| Dbac1sq |  |  | 71.27 |
|  |  |  | (69.21) |
| Constant | 0.0862\*\*\* | 0.113\*\*\* | -0.104 |
|  | (0.0154) | (0.0278) | (0.278) |
|  |  |  |  |
| Observations | 89,967 | 89,967 | 89,967 |
| Mean | 0.1052 | 0.1052 | 0.1052 |
| Controls | Yes | Yes | Yes |
| R-squared | 0.004 | 0.004 | 0.004 |
|  |  |  |  |

Robust standard errors in parentheses

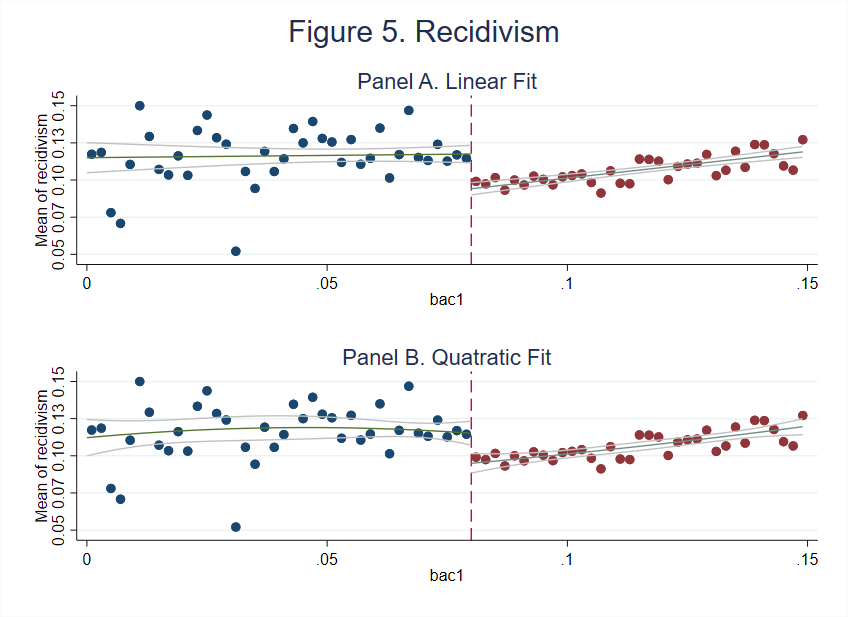
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 3 contains the results of the RD estimation for three different specifications using heteroskedasticity robust standard errors and controlling by age, sex, race and accidentality. Panel A employs a bandwidth from 0.03 to 0.13. The first estimation shows that both the running and the treatment variable are significant at the 1% level. For the marginal individual (that one within the neighborhood of the 0.08 cutoff), a harsher punishment, consequence of surpassing the BAC threshold, reduces the probability of recidivism within the next 4 months in 2.73 percentage points. When the interaction between bac1 and D is added to the specification, this effect increases in absolute value up to 5.91 p.p. Parallelly, the interaction term has an estimated coefficient of 0.438, which means that once an individual is above the cutoff, an increase in 1 unit on its BAC is associated with an increase of an extra 43.8 p.p relative to the case in which she is below the cutoff.

However, once the interaction between the treatment variable and the square of bac1 is introduced to the model, the estimated coefficients for the treatment variable, the running variable and its square become statistically insignificant. Regardless, both the linear and the quadratic interaction terms are statistically nonzero, yet the linear is negative (-4.21) and the quadratic is negative (32.73). This indicates a change both in slope and concavity for observations over the 0.08 cutoff.

Panel B tells a similar story: when reducing even more the sample by imposing a bandwidth from 0.055 to 0.105, the first model -that that includes no interaction terms- shows that the treatment, interpreted as above, reduces the probability of recidivism by 2.19 p.p. Nonetheless, once the interaction between the treatment and the square variable (model 2) and its square (model 3) is incorporated to the specification, all coefficients become statistically 0, therefore suggesting that a harsher punishment has no significant causal effect over the probability of recidivism for the marginal individual at the cutoff.

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



When only restricting the sample to those observations with a BAC inferior to 0.15, the “jump” in the observed outcomes at the cutoff is more evident. There also seems to be a differentiated positive effect of a higher BAC for individuals with BACs over the cutoff. In general, individuals with BACs under the 0.08 threshold exhibit a systematically higher probability of recidivism than those with BACs over the threshold. This difference is more evident when comparing observations that are closer to the cutoff. Thus, it can be graphically seen that harsher punishments tend to lower the rates of recidivism, at least at the local (nearby the cutoff) level.

Additionally, both the linear and the quadratic fit adjust similarly to the data, reason why an obvious preference can not be stablished from the graphic analysis. However, as the quadratic adjustment does not clearly present a nonzero degree of concavity, it might be more appropriate to stay with the linear model. Table 4 presents the results of the same 3 models than the ones estimated in the previous point, this time using the sample restricted to BACs under 0.15. Once again, the RD estimates show a negative and statistically significant effect of the treatment over the probability or recidivism only when the interaction with the quadratic term is not included.

Table 4. RD Estimates

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1) | (2) | (3) |
| VARIABLES | recidivism | recidivism | recidivism |
|  |  |  |  |
| D | -0.0267\*\*\* | -0.0545\*\*\* | -0.0354 |
|  | (0.00351) | (0.00783) | (0.0374) |
| bac1 | 0.259\*\*\* | -0.0210 | 0.193 |
|  | (0.0432) | (0.0847) | (0.341) |
| Dbac1 |  | 0.385\*\*\* | -0.127 |
|  |  | (0.0986) | (0.734) |
| bac1sq |  |  | -2.605 |
|  |  |  | (4.012) |
| Dbac1sq |  |  | 3.888 |
|  |  |  | (4.895) |
| Constant | 0.0933\*\*\* | 0.109\*\*\* | 0.107\*\*\* |
|  | (0.00477) | (0.00637) | (0.00729) |
|  |  |  |  |
| Observations | 124,642 | 124,642 | 124,642 |
| Mean | 0.1091 | 0.1091 | 0.1091 |
| Controls | Yes | Yes | Yes |
| R-squared | 0.003 | 0.003 | 0.003 |

Robust standard errors in parentheses

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

1. Much of this advice applies to Stata commands, but you can check the R files for lmb.r to see ways of doing the same in R. [↑](#footnote-ref-1)