Assignment 4 : Causal Inference and Research design

Juan Sebastian Benavides - 1026304205

June 13, 2020

1 GitHub Repository

The data was succesfully downloaded from the provided Github account, and properly organized in the following GitHub account: https://github.com/Sebastian-Benavides1999/RDD.git

2 Summary

This paper uses a regresion discontinuity design to test the effect of harsher punishments and sanctions on driving under the influence (DUI). The author's dataset contains the administrative records on 512,964 DUI stops from the state of Washington, recorded between 1995 and 2011.

The RDD model proposed leverages the fact that US Law establishes discrete thresholds that determine both the current and the potential future punishments for drunk drivers (Strict RDD). Specifically, in Washington State, a blood alcohol level above 0.08 is considered a DUI while a figure above 0.15 is considered an aggravated DUI which results in higher fines, increased jail time, and a longer license suspension period.

The first step the author makes is to test the compliance of the dataset with the regression disconinuity general assumptionswhich include the continuity of the underlying conditional regression and distribution functions. In order to test these, the author graphically analyzes the dataset, performs McCrary test over the blood alcohol level variable and checks for covariate balance. He concludes that the assumptions are reasonably met, and thus procedes to establish his research design.

The RD model proposed is the following:

$$y_i = X_i'\gamma + \alpha_1 * DUI_i + \alpha_2 * BAC_i + \alpha_3 * BAC_i * DUI_i + u_i \tag{1}$$

Where X denotes a vector of control variables such as white (dummy:being of white ethnicity), age, man (dummy:being a man), accident (dummy:the fact that the alcohol test was performed at an accident scene). The variable y is defined as recidivism and is measured as a dummy variable which is 1 if the suspect was pulled over for an alcohol test again within 4 years of the initial check up due to suspicious driving style or random police checkpoint examination.

The results obtained suggest that having a blood alcohol level above the DUI threshold (and thus being punished) reduces recidivism by up to 2 percentage points (17 percent). Likewise the author finds that having a BAC over the aggravated DUI threshold (and thus having a tougher punishment) reduces recidivism by an additional percentage point (9 percent). The results lead the author to conclude that the additional sanctions experienced by drunk drivers at BAC thresholds are effective in reducing drunk driving.

3 Data preparation

The dummy variable creation was performed with the folllowing line of code: Datos$D = as.numeric(Datos$bac1 >= 0.08). It can be seen in the R file attached to this article.

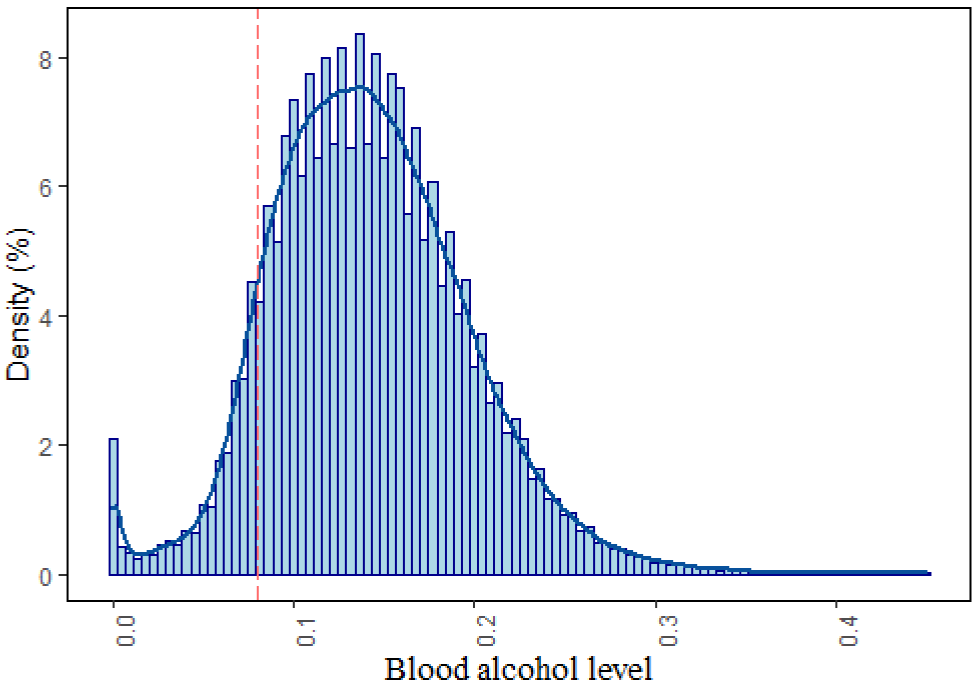
4 Manipulation test

In this section I perform two differnt tests to check for sorting in the running variable. The fisrt test is simply graphically analyzing the blood alcohol level (bac1) distribution. The Figure 1 shows such graph.

The red dashed line represents the relevant 0.08 blood alcohol level threshold. From a visual point of view, there seems to be no evidence of sorting around the threshold. However, it is important to recognize that since the sheer data amount is so big (512964 observations), there could exist some discontinuities that cannot be spotted visually. That is why we perform a data driven test to test for sorting.

I performed two similar tests whose purpose is to compary the density of observations just below and just above the threshold. The general idea is that if the data is not manipulated, then the density of observations should not be significantly different between just below the threshold and just above the threshold. In this particular case, it means to test whether suspects were able to adjust ther BAC as to barely be below the threshold, and whether the state troopers systematically manipulated the observation so as to get less (approximate to 0.79) or more (approximate to 0.8) infractions.

Figure 1: Running variable distribution



The first test is the McCrary test, and the second one uses the local polynomial density estimator proposed in Cattaneo, Jansson and Ma (2019). The later test can be seen as superior to the Mccrary test, since it uses a robust bias-corrected statistic as opposed to McCrary's conventional test statistic. The results are reported in Table 1.

Number of obs	214558
Model	Unrestricted
Kernel	Triangular
Bandwidth method	Cattaneo et al (2019)
VCE method	jacknife
Cutoff	0.08

Table 1: Sorting tests

Method	P> ‖T‖	T
Conventional (McCrary)	0.5936	0.5337
Robust (Cattaneo et al.)	0.0276	2.2032

The results obtained cast doubt on Hansen's original results. Under the tradtional McCrary test there is no evidence of manipulation. However, under the more updated Cattaneo et al. methodology, there seems to be evidence of sorting towards the higher limit of the threshold.

5 Covariate balance

In the following section I intend to check for covariate balance. For this I recreate Table 2 Panel A with white male, age and accident (acc) as dependent variables. In order to do this, I use the first equation provided by Hansen to predict the control variables instead of recidvidism in order to find if a regression discontinuity set up would yield psitive results. If the RDD were to give significant results, this would cast doubt on the smoothness assumption. The following table contains the results.

Table 2: Covariate balance

	Dependent variable:			
	Male	White	Age	Accident
	(1)	(2)	(3)	(4)
DUI	-0.007	0.007	0.106	0.004
	(0.011)	(0.009)	(0.315)	(0.008)
Observations	89967	89967	89967	89967
Mean at bac1=0.079	0.7885	0.8528	33.77	0.08874
Controls	No	No	No	No

Note: *p<0.1; **p<0.05; ***p<0.01

It is important to note that the replication used 0.05 bandwidhts and rectangular kernels just as Hansen's paper. Nevertheless, using the same bandwidhts yields a different amount of observations from that reported by Hansen, hence the difference in the point estimators (but not on the significance). These results suggest that the covariates are balanced at the cutoff and consequently provide no evidence of a violation in the smoothness assumption.

6 Covariate balance: Graphic

This section will replicate Figure 2 panel A-D from the original paper. Both linear and quadratic fits are provided, with confidence intervals. Figure 2 shows the covariate balance test performed with linear fit, while Figure 3 contains the test with quadratic fit. Both groups of graphs contain the confidence intervals.

From a graphic standpoint, there is little to no evidence of a violation in the smoothness assumption. That statement holds for both linear and quadratic fits. These results are consitent to those found by the data-driven balance tests, and are very similar to Hansen's results. It is important to note that the replication used exactly the same binwidths for data agrupations, but that it was imposible to replicate

Figure 2: Linear fit

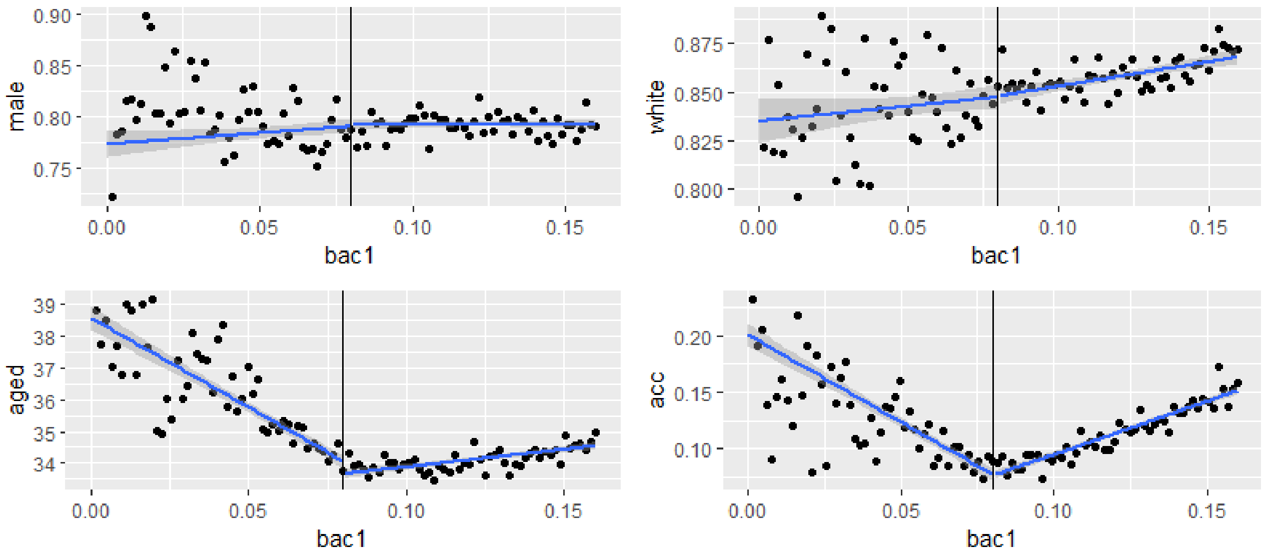
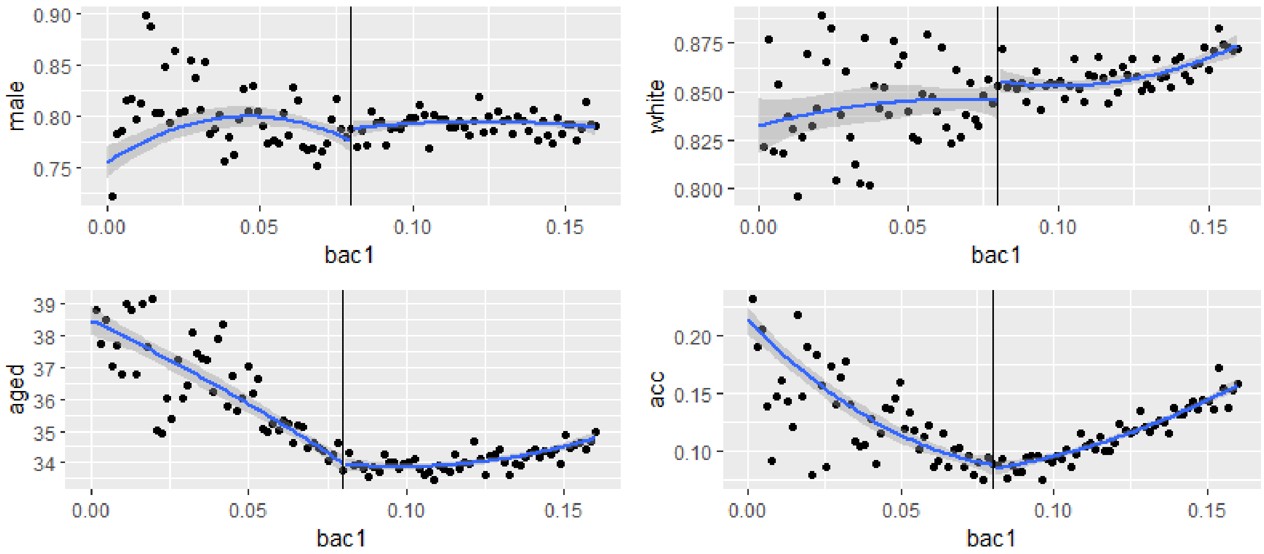


Figure 3: Quadratic fit



the bandwidths since Hansen only states that he used 'local' regressions.

Additionally, it is worth mentioning that even if some graphs like the quadratic fit for Male may suggest a slight degree of non smoothness, this could be due to using only a 2nd degree polynomial. The Cattaneo et al. test results suggest that if a higher degree polynomial were used, there would not be a discontinuity in the graphs.

7 RDD estimation

In this section I estimate the RDD model for three different specifications. The first specification controls only for the Blood Aclohol level (bac1) linearly. The second model interacts bac1 with the treatment variable(cutoff) and the third model interacts bac1 with the treatment both lienarly and quadratically. Since the original paper performed local lienar regressions, I estimate all 3 models for wide (0.03-0.13) and narrow (0.055 -0.105) bandwidths. Note that all specifications are controled for White, Male, Accident and Age.

Table 3: RDD estimation. Bandwidth = 0.05

	Dependent variable: Recidivism		
	Linear	Linear interaction	Quadratic interaction
	(1)	(2)	(3)
DUI	-0.0271875^{***} (0.00403924)	$\text{-}0.0595943^{***}$ (0.015230)	0.1053542 (0.084337)
bac1	0.3228383^{***} (0.07486272)	-0.0487633 (0.186923)	2.8333959^{*} (1.638995)
bac1^2			$\text{-}24.19013^{*}$ (13.755076)
DUI * bac1		0.4473878^{**} (0.204086)	-4.0354791^{*} (2.113454)
DUI * bac1^2			31.721^{**} (15.121)
Observations	89,967	89,967	89,967
Mean	0.103	0.103	0.103
Controls	Yes	Yes	Yes
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: RDD estimation. Bandwidth = 0.025

	Dependent variable: Recidivism		
	Linear	Linear interaction	Quadratic interaction
	(1)	(2)	(3)
DUI	-0.021643***	-0.0708943**	0.2395385
	(0.005582)	(0.034264)	(0.4088569)
bac1	0.1719897	-0.2607998	2.6699791
	(0.199134)	(0.369608)	(7.5575500)
bac1^2			-21.3341745
			(54.9371633)
DUI * bac1		0.6335865	-6.8423852
		(0.438524)	(10.1883637)
DUI * bac1^2			45.6808860
			(66.0165845)
Observations	89,967	89,967	89,967
Mean	0.103	0.103	0.103
Controls	Yes	Yes	Yes
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 4: Recidivism RDD

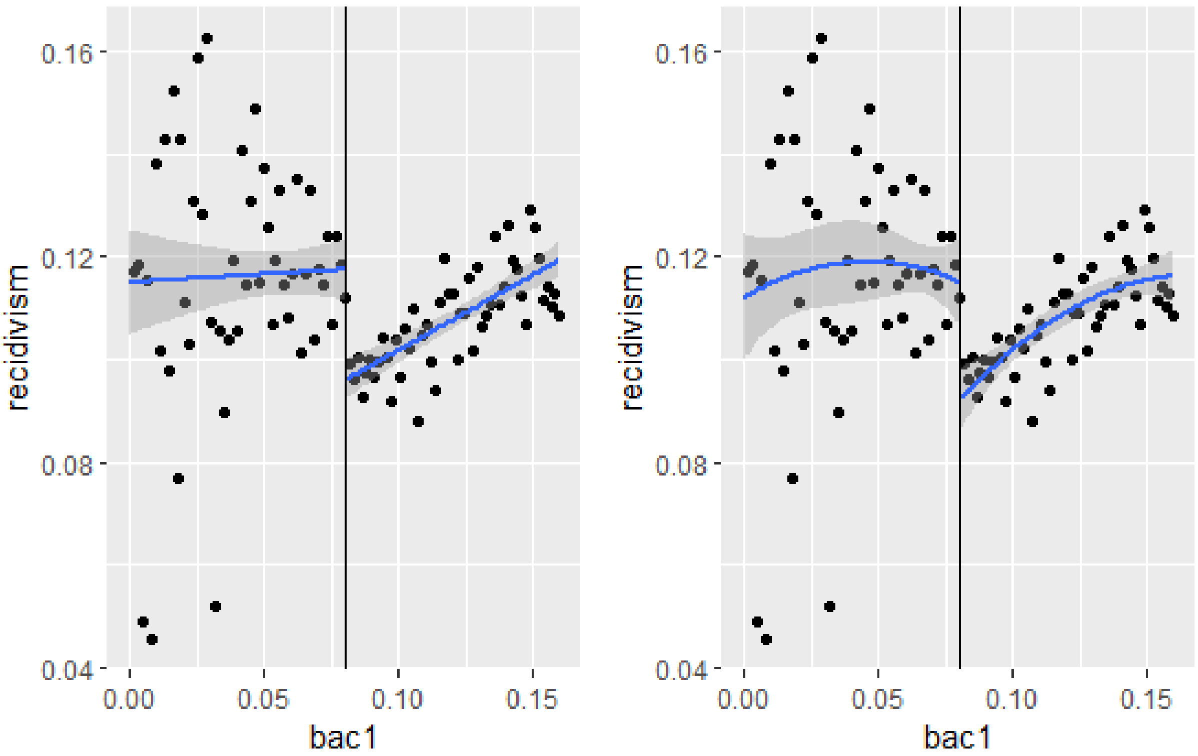


Table 3 and Table 4 show results consistent with those of Hansen. However, our increased specifications show how sensitive the significance of the discontinuity is: Its significance disappears with the quadratic fit of the model. All calculations were made with heteroskedastic robust errors.

8 RDD plotting

In this final section, the results of the third figure in Hansen's paper are replicated. This replication is plotted for both linear and quadratic fits, only for observations with bac1 below 0.15.

It is worth notting how sensitive the results are to the bandwidth selection: As Table 3 shows, the quadratic fit's coefficient for DUI is not significant for a bandwidth of 0.5, while Figure 5 shows the same coefficient to be significant for a 0.8 bandwidth.

Aside form the previous statement, these results are almost identical to those obtained by Hansen.