**2023 Causal Inference Workshop: Stata, R, and Python Sessions**

**Day 4. Regression discontinuity**

Dataset: Causal Inference Workshops (All Years)\Stata and R Materials\2015-Hansen-rounded-DUI-RDD-2023-07-26-lf.dta

Underlying article: Benjamin Hansen, Punishment and Deterrence: Evidence from Drunk Driving, *American Economic Review* 105(4), 1581-1617 (2015).

**Answer Sheet**

**Question 1**. Prepare the dataset.

**STATA Code for Step 1**

**\*Generate DUI threshold**. By Washington State law, DUI guilt is determined by minimum score of two BAC tests.

replace bac1=bac2 if bac2<bac1

rename bac1 BAC

drop bac2

label var BAC "Blood Alcohol Level"

**\*Generate DUI flag**

gen dui = 0

replace dui = 1 if BAC>0.080 // DUI threshold

**\*Generate centered BAC**

gen BAC\_c = BAC - 0.080 // centered blood alcohol content variable

**Question 2.** Look for evidence of manipulation of the running variable by plotting the density of *BAC* with a histogram near the cutoff using Hansen’s bandwidth of 0.05 on each side. Consider *BAC* as a discrete random variable, and drop observations outside the range [0.03, 0.13]. Is there evidence that drivers are managing their blood alcohol level, close to the threshold, to avoid being caught for drunk driving? If they were, what would it look like in the data?

**STATA Code for Step 2**

**\*Drop observations outside the bandwidth**

drop if BAC>0.13 | BAC<0.03

**\*Plot the density of BAC near the cutoff using a histogram.**

histogram BAC, discrete width(0.001) color(gray)/\*

\*/ addplot(pci 0 0.08 20 0.08 , lcol(black)) /\*

\*/ xlabel(0.04 "0.04" 0.06 "0.06" 0.08 "0.08" 0.10 "0.10" 0.12 "0.12", labsize(medium)) /\*

\*/ xtitle("BAC", size(large) height(5)) /\*

\*/ ytitle("Density", size(large) height(5)) /\*

\*/ title("BAC Histogram", size(large)) /\*

\*/ legend(off)

graph export "$desktop\histogram\_density\_discrete\_BAC.png", as(png) width(4840) height(3160) replace



Answer: From the histogram of the density of BAC near the cutoff there is no evidence of manipulation of blood alcohol content test values. If there was manipulation, there would be a jump in the distribution of the running variable below the cutoff, and a drop above the cutoff.

**Warning**: In developing this problem, we considered asking you to use the rddensity package. It will not work! It assumes that the running variables continuous. Close (as in this example, where BAC is measured in increments of 0.001) is not good enough.

**Question 3.** Another way to test for whether treatment is randomly assigned near the threshold is to assess whether the treatment variable predicts drivers’ predetermined characteristics or other outcomes (in the paper, prior car accidents). This is done by estimating the following equation (equation (1) on page 1588), for each characteristic or outcome *Y*i:

**Step 3a.** Regress male, white, age, and car\_accident on the saturated model above interacting the running variable with the treatment variable, within a bandwidth around the threshold (use Hansen’s bandwidth of 0.05 on either side of the threshold*;* see footnote to Table 2). Use robust standard errors. Are the covariates balanced near the 0.08 threshold?

**STATA Code for Step 3a**

local outcomes = “male white age car\_accident”

foreach y of local outcomes {

**\*OLS**

reg `y' i.dui BAC\_c i.dui#c.BAC\_c if BAC>=0.03 & BAC<=0.13, robust

outreg2 using "$desktop\OLS\_covariate\_balancing\_for\_RD.xls", bdec(6) sdec(6) ctitle("`y', [0.03, 0.13]”) append

}

Answer: Covariates are balanced near the threshold. DUI does not predict neither individuals predetermined characteristics nor other outcomes that might drive police to test drivers.

**Table Step 3a**. Assessing manipulation with OLS for BAC values in the range [.03, .13]. \* = p<0.01

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dependent Variable | Male | White | Age | Car Accident |
| DUI | 0.0057 | 0.0032 | -0.2012 | -0.0013 |
| s.e. | (0.0055) | (0.0048) | (0.1582) | (0.0039) |
| BAC centered | -0.1315 | 0.1260 | -63.9859\* | -1.1609\* |
| s.e. | (0.2287) | (0.2023) | (6.8569) | (0.1767) |
| DUI\*BAC centered | 0.22717 | -0.0298 | 71.7623\* | 2.0344\* |
| s.e. | (0.2522) | (0.2225) | (7.4856) | (0.1939) |
| Observations | 93,899 | 93,899 | 93,899 | 93,899 |

**Step 3b.** Create a histogram of covariate means below and above the cutoff (see Hansen paper, Figure 2, panels A to D) using the Stata command *rdplot*. Plot with linear fit over covariates means. Does there appear to be any jump in the mean covariates at the cutoff? What should you conclude?

**STATA Code for Step 3b**

**\*Rdplot**

foreach y of local outcomes {

rdplot `y' BAC if BAC>=0.03 & BAC<=0.13, p(1) h(0.05 0.05) c(0.08) /\*

\*/ graph\_options(xtitle("BAC", height(5) size(large)) /\*

\*/ ytitle("Mean `y'", height(5) size(large)) /\*

\*/ title("Regression Discontinuity Estimator", size(large)) /\*

\*/ xlabel(0.03 "0.03" 0.05 "0.05" 0.07 "0.07" 0.09 "0.09" 0.11 "0.11" 0.13 "0.13")/\*

\*/ xmtick(0.02(0.01)0.14))

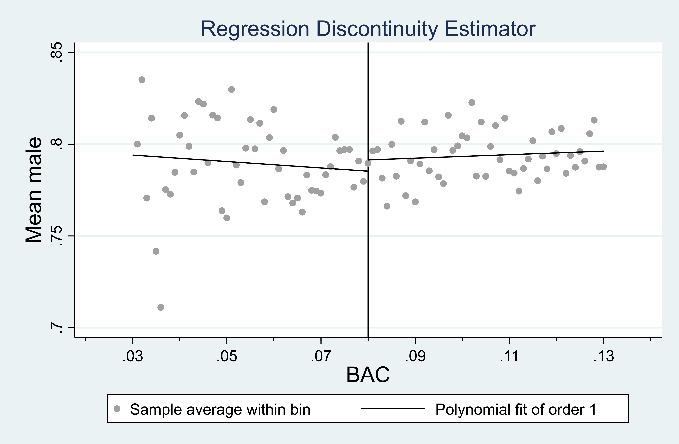
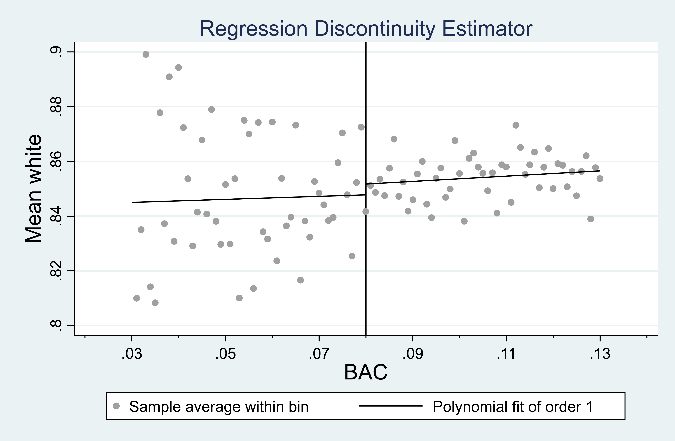
graph export "$desktop\graph\_means\_linear\_fit\_`y'.png", as(png) replace width(4840) height(3160)

}

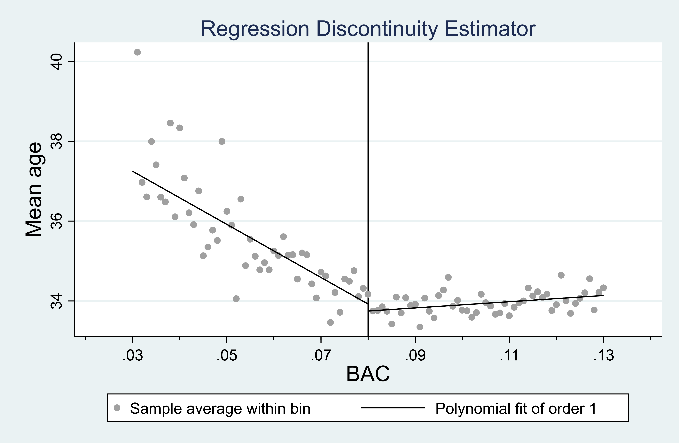
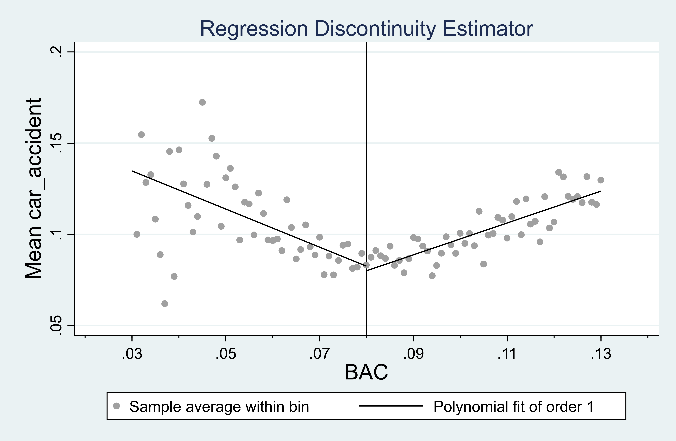
Answer: Results are consistent with what found using OLS. An eyeball analysis would suggest that there are no large jumps between mean values on the left and on the right of the cutoff. We can conclude that covariates are not being manipulated by drivers to avoid being caught while driving under influence.

**Figure Step 3b**. Visually assessing manipulation with *rdplot* for BAC values in the range [.03, .13].

**Panel A**. Male **Panel B**. White

**Panel C**. Age **Panel D**. Recent Car Accident

**Question 4**. Estimate the effect of getting arrested for a DUI on recidivism. Recidivism is an indicator which takes on the value of 0 if the drivers are not pulled over under suspicion of drunk driving within four years of the original offense, and takes on a value of 1 if they are subjected to a test or refuse a test by a police officer within this time period.

**Step 4a**. Estimate the equation below using OLS and a bandwidth of 0.05 near the cutoff. As covariates, use age, gender, race, and car\_accident. Use robust standard errors.

**Step 4b**. Re-estimate the effect of DUI including a quadratic in blood alcohol content.

**Step 4c**. Now repeat this analysis using a narrower bandwidth of 0 0.025.

**Step 4d**. Repeat the analysis in this question, but this time use the *rdrobust* command. Why is the treatment effect size calculated using OLS different from the one obtained from *rdrobust*? Interpret the meaning of your results.

**STATA Code for Step 4a-d**

**\*Replace age with age-21 for regression intercept**

replace age=age-21

**\*Local for regressors**

local regressors = "age male white car\_accident year\_\*"

**\*Local linear regressions with linear BAC**

reg recidivism i.dui BAC\_c i.dui#c.BAC\_c `regressors' if BAC>=0.03 & BAC<=0.13, robust

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("OLS, Linear BAC, [0.03, 0.13]”) append

**\*Local linear regression with quadratic BAC**

reg recidivism i.dui BAC\_c c.BAC\_c#c.BAC\_c i.dui#c.BAC\_c i.dui#c.BAC\_c#c.BAC\_c `regressors' if BAC>=0.03 & BAC<=0.13, robust

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("OLS, Quadratic BAC, [0.03, 0.13]”) append

**\*Local linear regressions with linear BAC and narrower bandwidth**

reg recidivism i.dui BAC\_c i.dui#c.BAC\_c `regressors' if BAC>=0.055 & BAC<=0.105, robust

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("OLS, Linear BAC, [0.055, 0.105]”) append

**\*Local linear regression with quadratic BAC and narrower bandwidth**

reg recidivism i.dui BAC\_c c.BAC\_c#c.BAC\_c i.dui#c.BAC\_c i.dui#c.BAC\_c#c.BAC\_c `regressors' if BAC>=0.055 & BAC<=0.105, robust

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("OLS, Quadratic BAC, [0.055, 0.105]") append

**\*Rdrobust, Linear on BAC**

rdrobust recidivism BAC, c(0.08) p(1) h(0.05 0.05) covs(age white male car\_accident)

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("rdrobust, Linear BAC, [0.03, 0.13]”) append

**\*Rdrobust, Quadratic polynomial on BAC**

rdrobust recidivism BAC, p(2) c(0.08) h(0.05 0.05) covs(age white male car\_accident)

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("rdrobust, Quadratic BAC, [0.03, 0.13]”) append

**\*Rdrobust, Linear on BAC, narrower bandwidth**

rdrobust recidivism BAC, c(0.08) p(1) h(0.05 0.05) covs(age white male car\_accident)

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("rdrobust, Linear BAC, [0.055, 0.105]”) append

**\*Rdrobust, Quadratic polynomial on BAC, narrower bandwidth**

rdrobust recidivism BAC, p(2) c(0.08) h(0.05 0.05) covs(age white male car\_accident)

outreg2 using "$desktop\OLS\_DUI\_recidivism.xls", bdec(6) sdec(6) ctitle("rdrobust, Quadratic BAC, [0.055, 0.105]”) append

**Table Step 4a-d**. The tables below report the ATT coefficient calculated using both OLS and *rdrobust*, for bandwidth near the cutoff of 0.05 (Panel A) and 0.025 (Panel B). \* = p<0.01, *italics* = p<0.05.

**Panel A**. BAC bandwidth is 0.05.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear model in BAC | | Quadratic model in BAC | |
|  | OLS | *rdrobust* | OLS | *rdrobust* |
| DUI | -0.0198\* | -0.0186\* | -0.0103 | *-0.0161* |
| *s.e.* | (0.0042) | (0.0047) | (0.0059) | (0.0067) |
| Covariates | Yes | Yes | Yes | Yes |

**Panel B**. BAC bandwidth is 0.025.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear model in BAC | | Quadratic model in BAC | |
|  | OLS | *rdrobust* | OLS | *rdrobust* |
| DUI | -0.0162\* | -0.0171\* | -0.0101 | -0.0121 |
| *s.e.* | (0.0055) | (0.0062) | (0.0079) | (0.0092) |
| Covariates | Yes | Yes | Yes | Yes |

Answer: *Rdrobust* coefficients are estimated using a triangular kernel on BAC, so observations whose BAC value is away from the cutoff are given less weight than observations with BAC values near the cutoff. OLS applies the same weight (rectangular kernel) to every observation. The negative coefficient on DUI provides evidence that a prior drunk driving arrestfuture drunk driving likelihood during the next four years. However, the coefficient is insignificant with the narrower bandwidth and the quadratic model.

**Step 4e**. Plot the RD estimator using the *rdplot* command with a linear fit on both sides of the cutoff.

**STATA Code for Step 4e**

**\*Rdplot**

rdplot recidivism BAC if BAC>=0.03 & BAC<=0.13, c(0.080) p(1) h(0.05 0.05) covs(age white male car\_accident)/\*

\*/ graph\_options(xtitle("BAC", height(5) size(large)) /\*

\*/ ytitle("Mean Recidivism", height(5) size(large)) /\*

\*/ title("Regression Discontinuity Estimator", size(large)) /\*

\*/ ymtick(0(0.01)0.2) /\*

\*/ ylabel(0 "0.00" 0.04 "0.04" 0.08 "0.08" 0.12 "0.12" 0.16 "0.16" 0.20 "0.20", grid labsize(large)) /\*

\*/ xlabel(0.03 "0.03" 0.05 "0.05" 0.07 "0.07" 0.09 "0.09" 0.11 "0.11" 0.13 "0.13")/\*

\*/ xmtick(0.02(0.01)0.14))

graph export "$desktop\graph\_recidivism\_bac\_linear\_fit\_`y'.png", as(png) replace width(4840) height(3160)

