Spring 2023

Gregory Bruich, Ph.D.
Dept. of Economics, Harvard University

Lab 5: Evaluating Education Policy using Regression Discontinuity Design

Methods/concepts: regression discontinuity design

LAB DESCRIPTION

In this lab, we will use *regression discontinuity design* to evaluate an academic probation policy at a Canadian University that places students on academic probation if their GPA falls below 1.60. The **probation.dta** data contains student-level data from this University. See Exhibit 1 for an excerpt from a letter that is sent by school administrators to students being placed on probation. For more details on the variables included in these data, see Table 1. A list and description of each of the Stata and R commands needed for this lab are contained in Table 2 and Table 3. For more background on the data and institutional setting, see Lindo, Sanders, and Oreopoulos (2010).

QUESTIONS

1. In this lab, we will seek to estimate the causal effect of the probation policy on educational outcomes (e.g., on time graduation). Why *don't* we want to compare *all students* who are on probation with *all students* who are not on probation to evaluate this University's program?

2. What is the "running variable" in this research design?

3. Now use the variables in the **probation.dta** file to validate the research design. Visual evidence is always the best evidence, so draw the following graphs and include them in your solutions:

    a. Binned scatter plots to check for smoothness of 2-3 predetermined characteristics

    b. Histograms to check for a spike in the density just above or just below the threshold. Use at least 200 bins.

4. Now study the effect of the academic probation policy on an outcome of your choice, such as on-time graduation. Start with a graphical analysis replicating the example in Figure 1 below, focusing on students within +/- 1.2 grade points of the 1.60 GPA threshold. That is, use a *bandwidth* of 1.2 grade points. Include your graph in your solutions.

5. Next we want to quantify any discontinuities that we saw in our binned scatter plots at the threshold. For the outcome variable that you examined in the previous question, focusing still on students within +/- 1.2 grade points of the 1.60 GPA threshold:

    a. Estimate a linear regression of the outcome variable on the running variable, but restrict the data to only observations that are to left of threshold with `GPA < 1.6` and `GPA >= 0.4`

    $$Y_i = \alpha_0 + \alpha_1 GPA_i + u_i \text{ if } GPA_i < 1.60 \text{ and } GPA_i \geq 0.4$$

    Calculate the predicted value from this regression at `GPA = 1.6`. That is, calculate: $\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 \times 1.6$. Include your calculation in your solutions.

b. Run a separate regression of the outcome variable on the running variable, but restrict the data to only observations that are to the right of the threshold with `GPA >= 1.6` and `GPA <= 2.8`:

$$Y_i = \gamma_0 + \gamma_1 GPA_i + v_i \text{ if } GPA_i \geq 1.60 \text{ and } GPA_i \leq 2.8$$

Calculate the predicted value from this regression at `GPA = 1.6`. That is, calculate: $\hat{Y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \times 1.6$ using this new regression. Include your calculation in your solutions.

c. Calculate the difference in predicted values in a. and b. Recall from Lab 2 that this is called a *predicted effect*. Include your calculation in your solutions.

6. Next we will show that the difference you calculated above exactly equals the regression coefficient from a *multivariable regression*. The code is provided for you in Table 2 and the starter script. The regression is as follows:

$$Y_i = \beta_0 + \beta_{RD} above_i + \beta_2 dist\_from\_cut_i + \beta_3 interaction_i + v_i$$

where Y_i is an outcome, $above_i$ is an indicator for being *above* the probation GPA threshold:

$$above_i = \begin{cases} 1 \text{ if } GPA_i - 1.60 \geq 0 \\ 0 \text{ if } GPA_i - 1.60 < 0 \end{cases}$$

The variable $dist\_from\_cut_i = GPA_i - 1.6$ is the difference between the GPA and the probation GPA threshold of 1.60. The variable $interaction_i = above_i \times dist\_from\_cut_i$ equals the product between the indicator $above_i$ and distance from the threshold $dist\_from\_cut_i$.

Estimate this multivariable regression (using the code provided in Table 2, Table 3, and the starter script) and confirm that $\hat{\beta}_{RD}$ exactly equals your answer in 5c.

7. Using the *standard error* on the coefficient $\hat{\beta}_{RD}$ from the regression in the previous question, what do you conclude about the statistical significance of the discontinuity at the threshold?

8. Putting together all the analyses you did above, what do you conclude about the effectiveness of this college's academic probation program? What caveats would you put on your conclusions?

9. The files to submit for this lab are:

a. Your well annotated do-file/.R/.rmd file replicating all your analyses above (with enough comments that a principal investigator on a research project would be able to follow and understand what each step of the code is doing). Please submit this to Gradescope.

b. For Stata users, please submit a log-file with the log showing the output generated by your final do-file. Please submit this file to the same gradescope assignment as the do-file. (Please do not submit a .smcl file: we can only read .log files in gradescope).

c. A PDF version of the solutions to the above questions. For graphs, save them as .png files and insert them into the document. Please submit this file to the same gradescope assignment as the do-file/.R/.rmd and log-file. (Please do not submit a word document: we can only read PDFs in gradescope).

EXHIBIT 1
Excerpt of Letter Set by University to Students being Placed on Probation

*Dear < **first name** >:*

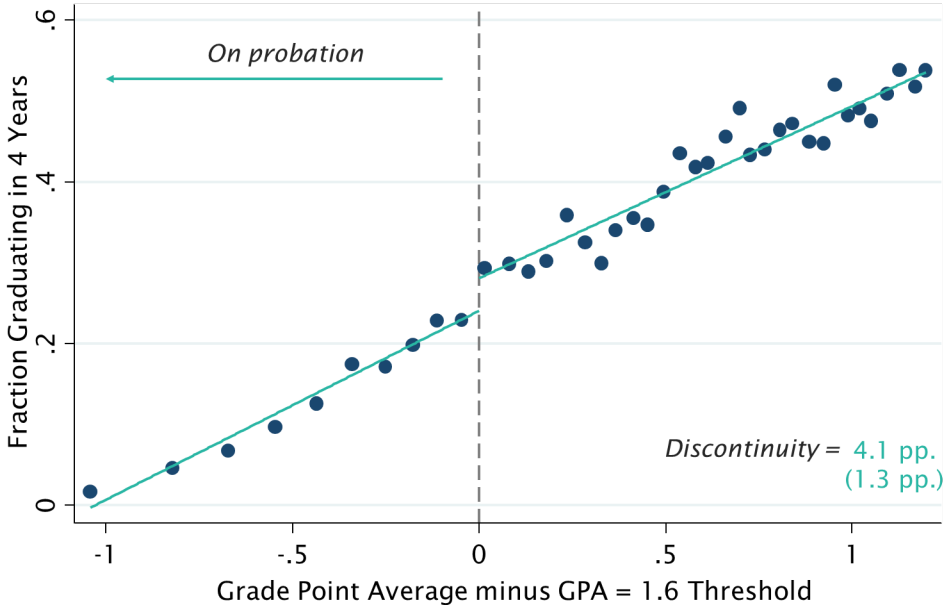
Your academic record indicates that you are experiencing challenges with your studies at xxxxxxxxxxx. As a result, you have been placed "On Probation" at the end of the xxxxxxx session. "On Probation" is an academic status applied to a student if he or she:

1. *Is having difficulty achieving a term average of at least **1.60** GPA*
2. *Is having difficulty meeting performance expectations and/or deadlines as outlined by the course instructor.*
3. *Is having difficulty achieving the minimum grades required for graduation.*

A student who at the end of any session during which they are on probation has a sessional GPA of less than 1.6 shall be suspended. Therefore, it is imperative that you seek assistance to improve your academic standing to avoid further sanction.

We know that you are capable of academic success, based on your academic record at admission. Let us review your goals and help you develop a plan to achieve them. You have the opportunity and available support to be successful. Please utilize our services to ensure your future success.

FIGURE 1
Fraction Graduating in Four Years versus GPA Threshold for Academic Probation



NOTE—The figure plots the fraction of students graduating in four years along the y-axis versus the student's GPA relative to the 1.60 threshold for academic probation along the x-axis. This figure is a binned scatter plot: the blue dots show the average for all students falling into bins defined by the variable along the x-axis. Students to the left of the dashed line are below the threshold and therefore would be placed on probation.

DATA DESCRIPTION, FILE: probation.dta

The data consist of $N = 44{,}362$ college students at a Canadian University. For more information, see Lindo, Jason M., Nicholas J. Sanders, Philip Oreopoulos (2010) "Ability, Gender, and Performance Standards: Evidence from Academic Probation," *American Economic Journal: Applied Economics* 2(2): 95-117, April 2010.

TABLE 1
Variable Definitions

	Variable (1)	Description (2)	Obs. (3)	Mean (4)	St. Dev. (5)	Min (6)	Max (7)
1	*GPA*	Grade point average in 1st year (running variable)	44,362	2.511	0.895	0	4
2	*hsgrade_pct*	High school grade percentile (Baseline Characteristic)	44,362	50.17	28.86	1	100
3	*totcredits_year1*	Credits attempted in first year (Baseline Characteristic)	44,362	4.573	0.511	3	6.5
4	*age_at_entry*	Age at entry (Baseline Characteristic)	44,362	18.67	0.743	17	21
5	*male*	Male (Baseline Characteristic)	44,362	0.383	0.486	0	1
6	*bpl_north_america*	Born in North America (Baseline Characteristic)	44,362	0.871	0.335	0	1
7	*english*	English is first language (Baseline Characteristic)	44,362	0.714	0.452	0	1
8	*loc_campus1*	At Campus 1 (Baseline Characteristic)	44,362	0.584	0.493	0	1
9	*loc_campus2*	At Campus 2 (Baseline Characteristic)	44,362	0.173	0.379	0	1
10	*loc_campus3*	At Campus 3 (Baseline Characteristic)	44,362	0.242	0.429	0	1
11	*gradin4*	Graduated by year 4 (Outcome)	30,017	0.447	0.497	0	1
12	*gradin5*	Graduated by year 5 (Outcome)	24,581	0.675	0.468	0	1
13	*gradin6*	Graduated by year 6 (Outcome)	19,757	0.753	0.431	0	1
14	*left_school*	Left university after 1st evaluation (Outcome)	44,362	0.0490	0.216	0	1
15	*probation_year1*	On probation after 1st year (Outcome)	44,362	0.160	0.367	0	1
16	*suspended_ever*	Ever suspended (Outcome)	44,362	0.0803	0.272	0	1
17	*nextGPA*	GPA in next enrolled term (Outcome)	44,362	2.758	0.949	0	4
18	*probation_ever*	Ever on academic probation (Outcome)	44,362	0.196	0.397	0	1

TABLE 2
Stata Commands

STATA command	Description
*clear the workspace clear all version 17 *change working directory and open data cd "C:\Users\gbruich\Ec 50\Lab 5\" use probation.dta, clear *Display all variables in the data describe *Report detailed information on all variables codebook	This code shows how to clear the workspace, change the working directory, and open a Stata data file. To change directories on either a mac or windows PC, you can use the drop down menu in Stata. Go to file -> change working directory -> navigate to the folder where your data is located. The command to change directories will appear; it can then be copied and pasted into your .do file. The describe and codebook commands will report information on what is included in the data set loaded into memory.
*Create running variable, centered at GPA = 1.60 gen dist_from_cut = GPA - 1.6	This code shows how to create a new variable *dist_from_cut* the equals GPA minus the threshold 1.60.
*Install binscatter ssc install binscatter, replace *Draw graph (command all goes on one line) binscatter yvar dist_from_cut if inrange(dist_from_cut, -1.2, 1.2), rd(0) line(lfit) *Save graph graph export figure1_linear.png, replace *Draw with quadratic best fit line binscatter yvar dist_from_cut if inrange(dist_from_cut, -1.21.2), rd(0) line(qfit) *Save graph graph export figure1_quadratic.png, replace	The first command installs binscatter, which only has to be done once. The second command produces a binned scatter plot of *yvar* against *dist_from_cut* with a linear best fit line, restricting the graph to observations with *dist_from_cut* in [-1.2, 1.2]. Some options are: 1. `if inrange(dist_from_cut, -1.2, 1.2)` restricts graph to observations +/- 1.2 from threshold 2. `rd(0)` allows break at x-axis value of 0 3. `linetype(lfit)` shows a linear best fit line. The third line saves the graph. The fourth line shows how to change the best fit line to be quadratic by changing *line(lfit)* to *line(qfit)*. To find all the options for binscatter, type "help binscatter" For example, the option `nq(100)` would divide data into 100 equal size groups for purposes of binning (the default is 20 bins)
*Histogram histogram yvar graph export histogram_yvar.png, replace *Histogram, changing number of bins to 200 histogram yvar, bin(200) graph export histogram_yvar.png, replace	These commands create and save histograms of a variable "yvar" which is a placeholder for the name of a variable in your data set. The first line creates a histogram (letting Stata decide how many bins to use). The second line saves the graph as a .png file. Alternatively, we could instead use the `histogram` type in twoway graphs. The second block of code changes the options by adding ", bin(200)" which will override the default binning and group the data into 200 buckets.

*Create running variable, centered at GPA = 1.60 gen dist_from_cut = GPA - 1.6 * Create indicator for being above probation threshold gen T = 0 replace T = 1 if dist_from_cut >= 0 * Interact dist_from_cut with non-probation gen interaction = dist_from_cut * T *Estimate regression (all goes on one line) regress yvar T dist_from_cut interaction if inrange(dist_from_cut, -1.2, 1.2), r	These commands show how to run a regression to quantify the discontinuity in *yvar* at the 1.60 GPA threshold. We first create a new variable *dist_from_cut* the equals GPA minus the threshold 1.60. We then generate an indicator variable *T* for *dist_from_cut* being positive. We next generate *interaction* that is the product between *dist_from_cut* and the indicator. Finally, we run a regression of *yvar* on these three variables, restricting the regression to observations with *dist_from_cut* between -1.2 and 1.2. The coefficient of interest is coefficient on *T*, the indicator for being above probation threshold. The `, robust` option computes standard errors that allow for unequal variances.
*close any possibly open log-files cap log close *start a log file log using lab5.log, replace *commands go here *close and save log file log close	These commands show how to start and close a log file, which will save a text file of all the commands and output that appears on the command window in stata. The first line is short for "capture log close" which will close any open log files, and otherwise just proceed to the next step. Then the "log using lab5.log, replace" starts the log file and changes the default in two ways. First, it changes the file type to have a .log file extension, which creates a plain text log file (which is readable in Gradesope so is important!). Second, it also adds the ", replace" option which will save over any other log file that has the same name. This is usually what you want. The rest of your lab code can go below the "log using lab5.log, replace" line. At the end of your do-file you can include the last line which is "log close" which will close and save the log-file.

TABLE 3
R Commands

R command	Description

#clear the workspace
rm(list=ls()) # removes all objects from the environment
cat('\014') # clears the console
 	
#Install and load haven package
if (!require(haven)) install.packages("haven"); library(haven)
 	
#Change working directory and load stata data set
setwd("C:/Users/gbruich/Ec 50/Lab 5")
dat <- read_dta("probation.dta")
 	
#Report detailed information on all variables
summary(dat)
	This sequence of commands shows how to open Stata datasets in R. The first block of code clears the work space. The second block of code installs and loads the "haven" package. The third block of code changes the working directory to the location of the data and loads in probation.dta. To change the working directory in R Studio, you can also use the drop down menu. Go to session -> set working directory -> choose working directory.
	
	The easiest way to open a Stata data set in R Studio is to use the drop down menu. Go to file, then import data set, and finally browse to locate the file you want to open. This option will be available after you install the haven package.
	
	The summary command will report information on what is included in the data set loaded into memory, including information on the number of missing observations NAs for each variable.

#Create running variable, centered at GPA = 1.60 dat$dist_from_cut <- dat$GPA - 1.6	This code shows how to create a new variable *dist_from_cut* the equals GPA minus the threshold 1.60.

#Load packages
if (!require(tidyverse)) install.packages("tidyverse"); library(tidyverse)
if (!require(rdrobust)) install.packages("rdrobust"); library(rdrobust)
 	
#Subset data to observations in [-1.2, 1.2]
narrow <- subset(dat,dist_from_cut<=1.2 & dist_from_cut>=-1.2)
 	
#draw binned scatter plot with linear fit
rdplot(dat_narrow$yvar, #outcome variable
dat_narrow$dist_from_cut, #running variable
p = 1,
nbins = c(20, 20),
binselect = "es",
y.lim = c(0, 0.6),
x.label = "Grade Point Average minus 1.6",
y.label = "Outcome variable (yvar)"
)
 	
#Save graph
ggsave("figure1_linear.png")
	The first command installs rdrobust, which only has to be done once.
	
	The second command subsets the data to only observations with *dist_from_cut* between -1.2 and 1.2.
	
	The third block of code produces a binned scatter plot of *yvar* against *dist_from_cut* with a linear best fit line. The options shown are:
	p = 1, #p = 1 is linear best fit line. p = 2 is quadratic
	nbins = c(20, 20), #number of bins on each side of threshold
	binselect = "es", #option to use "equal spaced" binning
	y.lim = c(0, 0.6), #Set y-axis scale
	x.label = "Grade Point Average minus 1.6", #x axis label
	y.label = "Outcome variable (yvar)" #y axis label
	
	The fourth block of code saves the graph.

#Histogram using ggplot
if (!require(tidyverse)) install.packages("tidyverse"); library(tidyverse)
if (!require(ggplot2)) install.packages("ggplot2"); library(ggplot2)
 	
ggplot(dat) + geom_histogram(aes(x=yvar, y=..density..))
ggsave("histogram_yvar.png")
 	
#Use 200 bins, overriding default
ggplot(dat) + geom_histogram(aes(x=yvar, y=..density..), bins = 200)
	These commands create and save histograms of a variable "yvar" using ggplot. First start by installing the tidyverse library. Then use ggplot to draw the graph. The ggsave() line saves the graph as a .png file. The last line overrides the default to show 200 bins by adding the bins = 200 option.

```
#Load packages
if (!require(sandwich)) install.packages("sandwich"); library(sandwich)
if (!require(lmtest)) install.packages("lmtest"); library(lmtest)

#Create running variable, centered at GPA = 1.60
dat$dist_from_cut <- dat$GPA - 1.6

#Create indicator for being above probation threshold
dat$T <- 0
dat$T[which(dat$dist_from_cut >= 0)] <- 1

#Interact dist_from_cut with non-probation
dat$interaction <- dat$dist_from_cut*dat$T

##Subset data to [-1.2,1.2] with new variables added
dat_narrow <- subset(dat,dist_from_cut<=1.2 & dist_from_cut>=-1.2)


#Estimate regression
linear <- lm(yvar ~ T + dist_from_cut + interaction , data = dat_narrow)

#Report coefficients and standard errors
coeftest(linear, vcov = vcovHC(linear, type="HC1"))
```

These commands show how to run a regression to quantify the discontinuity in *yvar* at the 1.60 GPA threshold. We first create a new variable *dist_from_cut* the equals GPA minus the threshold 1.60.

We then generate an indicator variable *T* for *dist_from_cut* being positive. We next generate the a variable *interaction* that is the product between *dist_from_cut* and the indicator.

Then we subset the data to a new data frame with *dist_from_cut* between -1.2 and 1.2.

Finally, we run a regression of *yvar* on these three variables, restricting the regression to observations with *dist_from_cut* between -1.2 and 1.2. The coefficient of interest is coefficient on *T*, the indicator for being above probation threshold.

The `type=HC1` option computes standard errors that allow for unequal variances.