PROBLEM SET 10: REGRESIÓN DISCONTINUA

GARCÍA OJEDA - HAUSVIRTH - HAYDUK - SALVATIERRA

a. En este inciso corrimos una regresión OLS simple que relaciona el tamaño de la clase, con las calificaciones tanto en Matemática como en Lengua. Los resultados que observamos muestran que ambos coeficientes son positivos, por lo que un aumento en el tamaño de clase está asociado con un incremento en las notas promedios de matematica y lengua. En particular, un aumento de una unidad en el tamaño de la clase está asociado con un aumento promedio de 0.159 puntos en las notas promedio de matemática, mientras que un aumento de una unidad en el tamaño de la clase está asociado con un aumento promedio de 0.080 puntos en las notas promedio de lengua.

OLS - Tamaño de clase y calificaciones

	(1)	(2)
	Nota promedio en matemática	Nota promedio en lengua
	b/se	b/se
Tamaño de la clase	0.159**	0.080
	(0.052)	(0.043)

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

En cuanto a los supuestos de OLS para identificar relaciones causales, no resulta creíble el supuesto de identificación $E[X^{'}\epsilon]$=0, lo cual significa que el tamaño de la clase no está correlacionado con otros factores que afectan a las notas. Esto no resulta creíble dado que hay variables omitidas como calidad de los profesores, nivel socioeconomico de la escuela y alumnos, habilidades de los alumnos, etc que afectan tanto al tamaño de clases como a las notas.

b. Ahora corremos nuevamente la regresión anterior pero controlando por *Enrollment* y por *Percent Disadvantaged*, y notamos que los coeficientes arrojan resultados contrarios a la subsección anterior, pero ninguno es estadísticamente significativo.

OLS - Tamaño de clase y calificaciones con controles

	(1)	(2)
	Nota promedio en matemática	Nota promedio en lengua
	b/se	b/se
Tamaño de la clase	-0.047	-0.057
	(0.066)	(0.050)

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

c. Las escuelas en Israel siguen la **Maimonides rule**, que establece un número máximo de 40 alumnos por aula; y una vez que este umbral es superado, se debe crear una nueva clase para dividir el curso, y que la cantidad de alumnos en cada clase quede por debajo del umbral establecido.

Regression discontinuity (RDD) explota la discontinuidad en la asignación al tratamiento para identificar efectos causales. En este caso concreto, utilizando RDD podemos explotar la discontinuidad que hay en el tamaño de las aulas cuando la cantidad de alumnos inscriptos es mayor a 40 y encontrar el efecto del tamaño de clases en las notas promedio de matemática y lengua.
De acuerdo con los datos observados, realizamos un scatter plott que relaciona el tamaño de las aulas con la cantidad de inscriptos por año, y notamos que existe una relación lineal positiva entre ambas variables, y que es continua hasta llegar a los 40 alumnos. Allí, donde se encuentra el *cutoff*, se dividen las aulas proporcionalmente para no superar el umbral. Podemos presuponer por los datos, que se trataría de un diseño **sharp**, ya que no se observa divisiones en las clases cuando hay menos inscripciones, y tampoco se observan aulas con mas de 40 alumnos.

d. Para utilizar RDD, es necesario que se cumplan dos supuestos:

(1) Los individuos no pueden manipular la variable de discontinuidad
(2) Las características de los individuos son funciones suaves de la asignación al tratamiento.

Para corroborar que estos supuestos se cumplan y para validar el enfoque hay varias pruebas o test de validez que se pueden realizar.
En primer lugar, realizamos un **Density Discontinuity Test** para verificar que la denisidad de nuestra *running variable*, en este caso el enrollment, es continua en el punto de corte. Es relevante realizar dicha prueba para testear si los individuos maniputaron su posición o decisión en el punto de corte; ya que de ser así, se viola el supuesto de identificación número 1 mencionado previamente. Por los resultados obtenidos, vemos que en todas las ventanas, el p-valor es muy bajo, por lo que concluimos que la densidad de x no es continua en el punto de corte.

P-values of Binomial Tests (H0: prob = 0.5)

Window Length / 2	$< c$	$>= c$	$P > \|T\|$
5.000	26	103	0.0000
7.000	38	148	0.0000
9.000	51	190	0.0000
11.000	69	238	0.0000
13.000	92	284	0.0000
15.000	104	323	0.0000
17.000	134	363	0.0000
19.000	160	399	0.0000
21.000	190	440	0.0000
23.000	204	478	0.0000

En segundo lugar, se puede realizar la prueba de validez de **Covariate Balance**. Esta prueba lo que hace es testar que las covariables no estén relacionadas con el tratamiento, y por lo tanto, sean funciones suaves a la discontinuidad. Esto se relaciona directamente con el supuesto dos mencionado previamente. Es relevante testearlo ya que si hay un salto en covariables en el punto de corte, no se puede atribuir el cambio observado únicamente al tratamiento.
En nuestro caso, el test no aplica, ya que no tenemos covariables en la base de datos previamente determinadas.

Finalmente, se puede también realizar un test de **Different bandwidths or windows**, en donde se prueba si los resultados del RDD son consistentes al cambiar el ancho de banda o la ventana de datos utilizados en el análisis. En el punto f de este trabajo, se realiza dicha comparativa y se analizan los resultados esperados.

e. Realizamos el análisis de RDD utilizando un enfoque de continuity-based approach, para las calificaciones de lengua y de matemática respectivamente. Para analizar cual es el mejor ajuste de datos, testeamos las regresiones con polinomios del 1 al 4.

Estimaciones RD para las notas promedio de matemática

Método	**Polinomio 1**	**Polinomio 2**	**Polinomio 3**	**Polinomio 4**
Tamaño de Clase	-2.3488	-7.3669	-8.587	-8.2437
	(3.6134)	(4.6743)	(6.0708)	(6.6029)

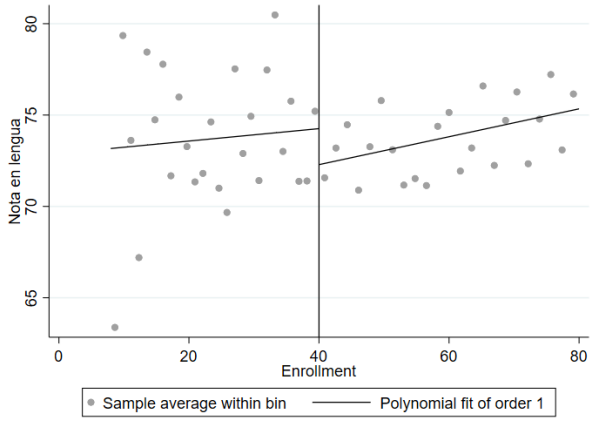
Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimaciones RD para las notas promedio de Lengua

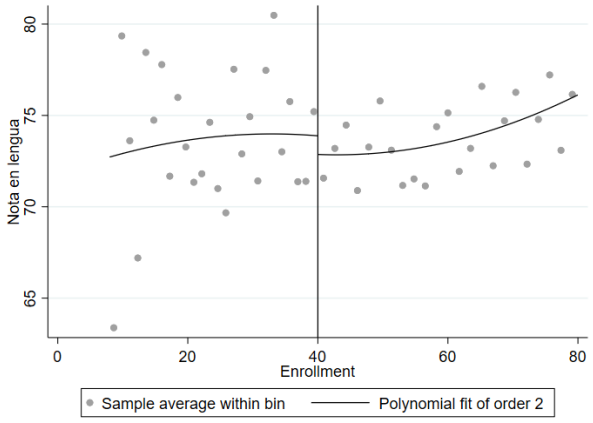
Método	**Polinomio 2**	**Polinomio 3**	**Polinomio 4**	**Polinomio 5**
Convencional	-2.5733	-9.0088	-14.182	-15.308
	(2.773)	(3.7657)	(5.2539)	(6.0685)

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

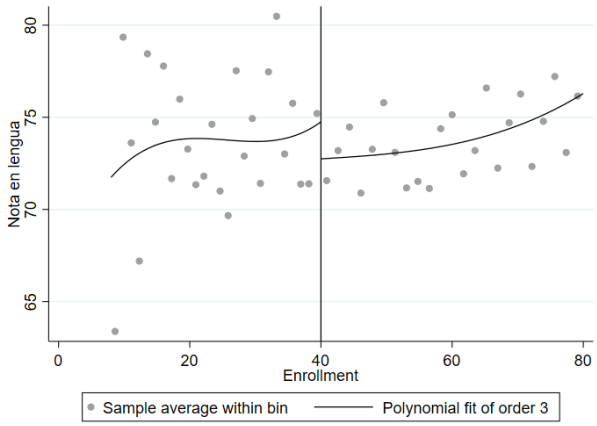
Luego, realizamos un gráfico para cada ajuste de polinomio



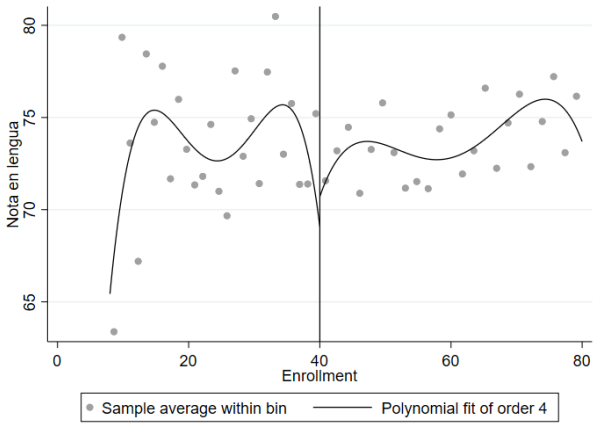
(A) Polinomio 1



(B) Polinomio 2

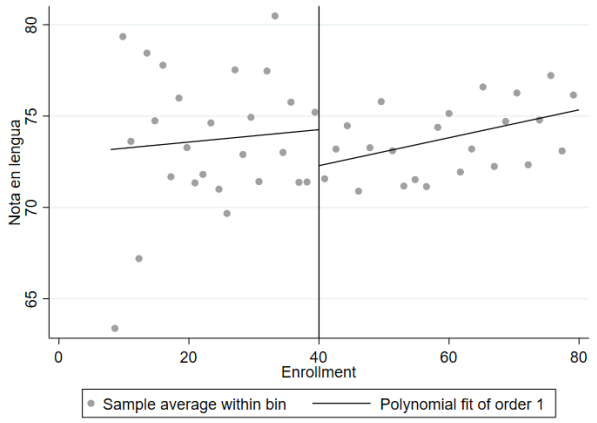


(C) Polinomio 3

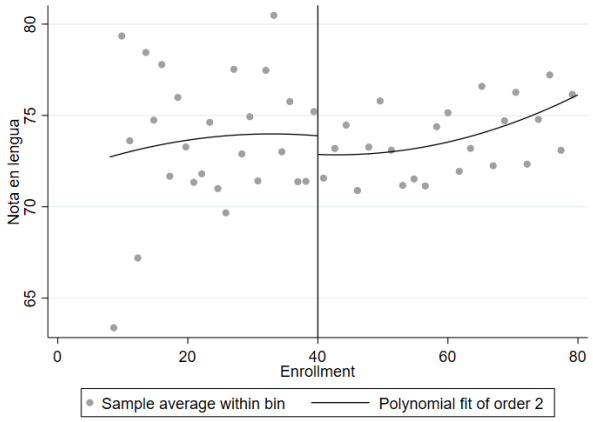


(D) Polinomio 4

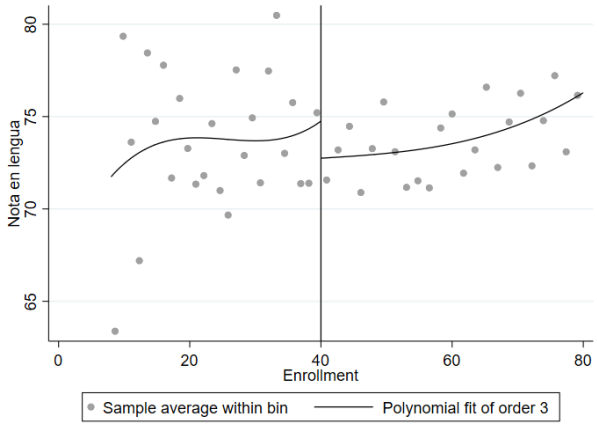
Figura 1: RDD para Lengua



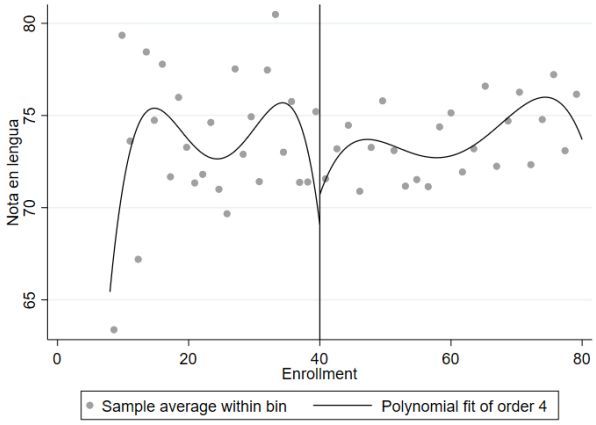
(A) Polinomio 1



(B) Polinomio 2



(C) Polinomio 3



(D) Polinomio 4

Figura 12: RDD para Matemática

Para determinar el polinomio mas adecuado, hay que considerar tanto la flexibilidad como la simplicidad. Hay ventajas y desventajas de elegir un grado de polinomio u otro, ya que existe un trade off entre una aproximación mejor de los datos, y el riesgo de caer en overfitting. El polinomio de grado 1 (lineal) aunque es el mas utilizado en la literatura, no logra captar de manera correcta las relaciones no lineales existentes en los datos, lo que lleva a una mala especificación del *data generating process*; mientras que en el otro extremo, un polinomio de grado 4 está sobreajustando los datos de la muestra, lo que genera mucho ruido y lo hace muy complejo. En ninguno de los dos casos se logran captar las tendencias reales, y ambos dos son difíciles de interpretar.
Por otro lado, polinomios intermedios como lo son el 2 y el 3, permiten capturar patrones no lineales moderados, y concretamente el polinomio de grado 2, lo hace sin añadir complejidad excesiva, y con menos riesgo de introducir sobreajuste. Es por esta razón que consideramos que el polinomio de grado 2 es el mas adecuado en un análisis RDD.

f. Dado que elegimos el polinomio de grado 2, corremos nuevamente la regresión RDD, pero con *bandwith* $10, 15, 7$. Sabemos que esta elección tiene implícito un trade off entre ganar estadistical power, y que se mantenga creíble el supuesto de identificación. Un bandwidth pequeño, como $h = 7$, permite capturar con mayor precisión el efecto causal del tratamiento, ya que utiliza únicamente las

observaciones más cercanas al punto de corte, donde las condiciones de continuidad y aleatoriedad son más probables. Como contra cara, esta elección aumenta la varianza de la estimación, debido a que la muestra es muy pequeña (en nuestro caso, 33 observaciones por izquierda y 120 por derecha).

Por otro lado, un bandwidth grande, como $h = 15$, reduce la varianza al incluir más observaciones, pero introduce el riesgo de sesgo, ya que las observaciones más alejadas del punto de corte pueden no seguir la misma relación causal, comprometiendo la validez del estudio.

Finalmente podríamos argumentar que un bandwith de $h = 10$ es un punto intermedio y equilibra el trade off sesgo y varianza de la estimación.

g. En este punto se busca evaluar el efecto del tamaño de las clases sobre las calificaciones promedio en matemática y lengua, utilizando variables instrumentales. Este enfoque es necesario porque el tamaño de las clases podría ser endógeno, es decir, estar correlacionado con factores no observados que también afectan las calificaciones.

Regresión con Variables Instrumentales (2SLS) - Matemática y Lengua

	Nota promedio en Matemática	Nota promedio en Lengua
	b/se	b/se
Tamaño de la clase	0.159^{**}	0.072
	(0.053)	(0.045)

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

En concordancia con los resultados obtenidos en OLS, cuando utilizamos el método de variables instrumentales, el coeficiente en ambos coeficientes son positivos, lo que indica que el tamaño de la clase tiene un efecto sobre las notas, sin embargo para matemática esto resulta estadísticamente significativo, mientras que para lengua no.

LISTING 1. Código Stata para el Problem Set 10

```
/*****************************************************************************

                          Problem Set 10: REGRESI N DISCONTINUA
                          Universidad de San Andr s
                             Econom a Aplicada
*****************************************************************************/

* Gaspar Hayduk; Juan Gabriel Garc a Ojeda; Elias Lucas Salvatierra;
   Martina Hausvirth

/*****************************************************************************

* 0) Set up environment
*============================================================================*

global main "\Users\marti\OneDrive\Documentos\Aplicada.stata\PS10"
global output "$main/output"
global input "$main/input"
cd "$output"

* Abrimos la base de datos:
use "$input/grades5.dta", clear

** TO INSTALL STATA PACKAGES:
net install rdrobust, from(https://raw.githubusercontent.com/rdpackages/
   rdrobust/master/stata) replace
net install rdlocrand, from(https://raw.githubusercontent.com/rdpackages/
   rdlocrand/master/stata) replace
net install rddensity, from(https://raw.githubusercontent.com/rdpackages/
   rddensity/3084126ee0e5401cef662e1b7b3f0d802c319e7a/stata) replace

* Inciso a)
*============================================================================*

label var classize "Tama o de la clase"
label var avgmath "Nota promedio en Matem tica"
label var avgverb "Nota promedio en Lengua"

eststo clear

*----CLASSIZE CONTRA NOTAS EN MATEMATICA
```

```
eststo: reg avgmath classize

*----CLASSIZE CONTRA NOTAS EN LENGUA
eststo: reg avgverb classize

*** Exportamos Resultados
esttab using "$output/table1.tex", se replace label noobs ///
keep(classize) ///
cells(b(star fmt(3)) se(par fmt(3)))

*Testeamos exogeneidad
estat ovtest

*Testeamos homocedasticidad
estat imtest, white

* Inciso b)
*================================================================================*

eststo clear
label var enroll "Enrollment"
label var tip_a "Percent disadvantaged"

*----CLASSIZE CONTRA NOTAS EN MATEMATICA
eststo: reg avgmath classize enroll tip_a

*----CLASSIZE CONTRA NOTAS EN LENGUA
eststo: reg avgverb classize enroll tip_a

*** Exportamos Resultados
esttab using "$output/table2.tex", se replace label noobs ///
keep(classize) ///
cells(b(star fmt(3)) se(par fmt(3)))

* Inciso c)
*================================================================================*

scatter classize enroll, title("Tama o de clase vs Inscripci n") ///
xline(40) msize(small)

* Inciso d)
```

```
*=============================================================================*

global x enroll
global y1 avgmath
global y2 avgverb
global y avgmath avgverb
global covs "tip_a"

* 1) Density discontinuity test
rddensity $x, c(40)

* Inciso e)
*=============================================================================*

* Matem tica: RDD con diferentes grados
rdrobust $y1 $x, c(40) p(1) masspoints(off) stdvars(on)
rdplot $y1 $x, c(40) p(1) title("RDD - Polinomio Grado 1") ///
graph_options(graphregion(color(white)) xtitle("Enrollment") ytitle("Nota
    en matem tica"))

* Lengua: RDD con diferentes grados
rdrobust $y2 $x, c(40) p(1) masspoints(off) stdvars(on)
rdplot $y2 $x, c(40) p(1) title("RDD - Polinomio Grado 1") ///
graph_options(graphregion(color(white)) xtitle("Enrollment") ytitle("Nota
    en lengua"))

* Inciso f)
*=============================================================================*

* Bandwidth 10
rdrobust $y1 $x, c(40) h(10) p(2) masspoints(off) stdvars(on)

* Inciso g)
*=============================================================================*

gen zs = enroll / (floor((enroll - 1) / 40) + 1)

* IV con Regla de Maim nides
ivregress 2sls avgmath (classize = zs)
ivregress 2sls avgverb (classize = zs)
```