

# Diseño de Regresión Discontinua (RDD)

## Econometría Avanzada

---

Rony Rodriguez-Ramirez

DIME | The World Bank

05 September 2021



# Table of contents

1. Intro
2. Randomización local
3. Detalles del RDD
4. Implementación en la práctica
5. Ejemplos generales
6. Fuzzy RDD
7. Ejemplos de Fuzzy RDD
8. Conclusiones

# RDD Introducción

---

# RDD Introducción

- Introducido por primera vez para estudiar el impacto del reconocimiento de méritos por Thistlethwaite y Campbell (1960)
- Sin embargo, solo comenzó a llamar la atención en la economía desde fines de la década de 1990.
- Al igual que los ECA, el RDD consiste en comparar 2 variables de resultado: la variable de resultado de los individuos que fueron tratados y la variable de resultado de los individuos que no fueron tratados.
- El RDD se considera una estrategia de evaluación que proporciona resultados tan convincentes como las estimaciones de experimentos aleatorios, sabiendo que los experimentos aleatorios se consideran ampliamente como el estándar de oro de la evaluación de impacto.

# Regla de tratamiento

- El tratamiento de la población depende de si una variable observada  $X$ , supera un valor crítico  $c$ .
- La variable  $X$  es llamada variable de asignación (assignment variable or forcing variable).
- La regla de tratamiento viene dada por:

$$\begin{cases} D = 0, & \text{if } X < c \\ D = 1, & \text{if } X \geq c \end{cases}$$

donde  $D = 0$  significa que la población no es tratada y  $D = 1$  significa que la población es tratada.

- A esto se le llama como sharp RDD.

# RDD Introducción

- Pero, ¿qué es una discontinuidad?
  - Una ruptura brusca en los valores de una función.
  - Matemáticamente, estamos hablando de una ecuación por partes.

$$f(x) = \begin{cases} \frac{1}{2}x, & x < 5 \\ 2 + \frac{1}{2}x, & x \geq 5 \end{cases}$$

# Uso del RDD

- No tenemos un RCT y nos preocupan las variables endógenas.
- RDD utiliza la asignación de discontinuidad exógena para estimar los efectos causales, es decir, las observaciones de un lado de la ruptura terminan siendo muy similares a las del otro lado.
- Básicamente, terminamos con grupos equilibrados, pero no tenemos una asignación de tratamiento aleatoria.

# Randomización local

---



# RD como randomización local

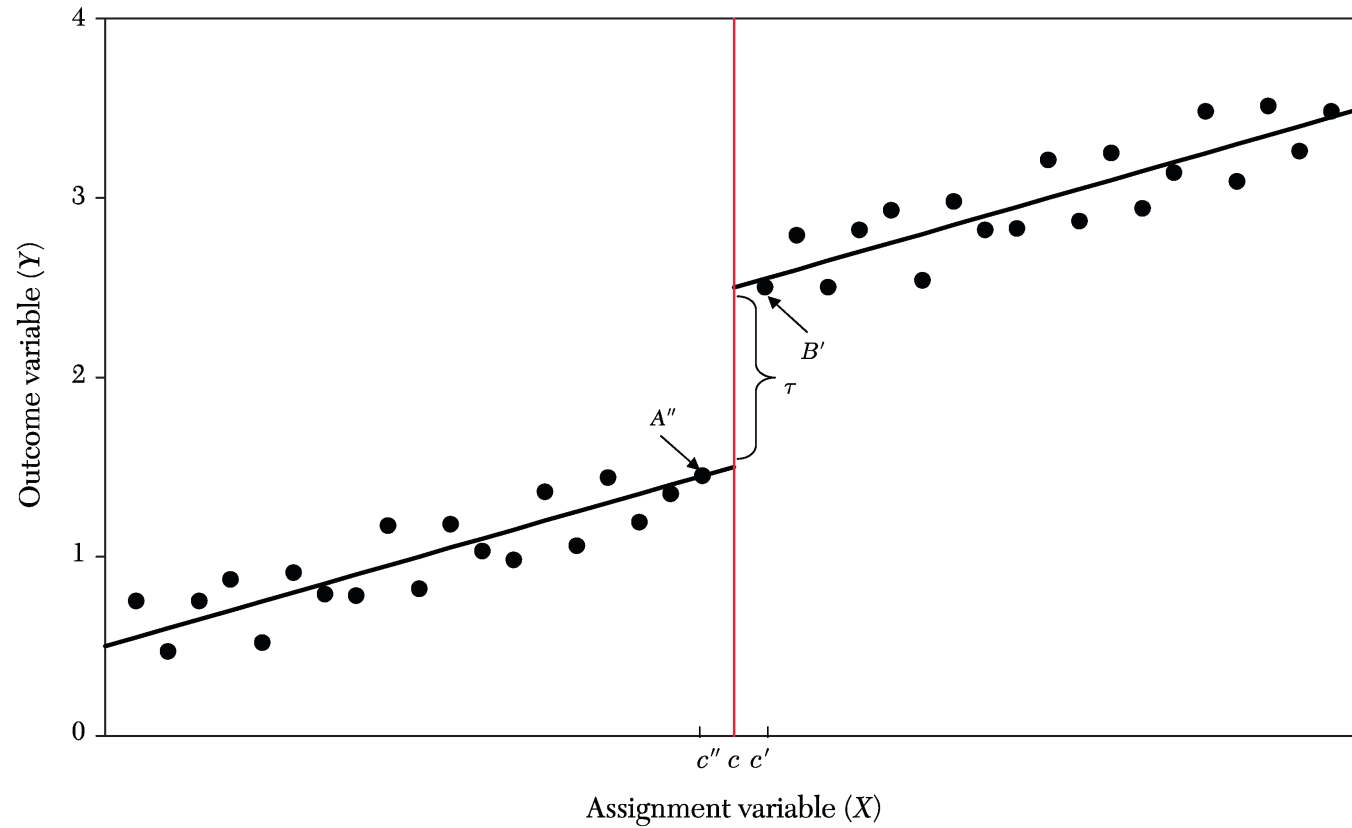
- Aleatorización local (de Angrist y Pischke)
- El estado del tratamiento es una función determinista de una variable  $\mathbf{a}$ , de modo que una vez que conocemos  $\mathbf{X}$ , sabemos  $\mathbf{D}_x$ .
- El estado del tratamiento es una función discontinua de  $\mathbf{a}$ , porque no importa qué tan cerca esté  $\mathbf{X}$  del límite,  $\mathbf{D}_x$  permanece sin cambios hasta que se alcanza el límite.

$$D_x = \begin{cases} 1 & \text{if } X \geq \text{cutoff} \\ 0 & \text{if } X < \text{cutoff} \end{cases}$$

# RD como randomización local

- A la variable de asignación  $X$  la llamamos *variable de forzamiento* o *variable de ejecución*.
- La variable de asignación (o forzamiento) no es ortogonal a:
  - ya sea a las características observadas de los individuos (socioeconómico, estado de sus padres, etc.).
  - ni a las características no observadas de los individuos (coeficiente intelectual, preferencia, etc.).

# RD como randomización local



# RD detalles

---

# Un setup lineal

Supongamos la siguiente ecuación:

- La ecuación (1) incluye la variable ficticia de tratamiento, ya que lo utilizaremos como un control lineal para nuestra variable de ejecución.

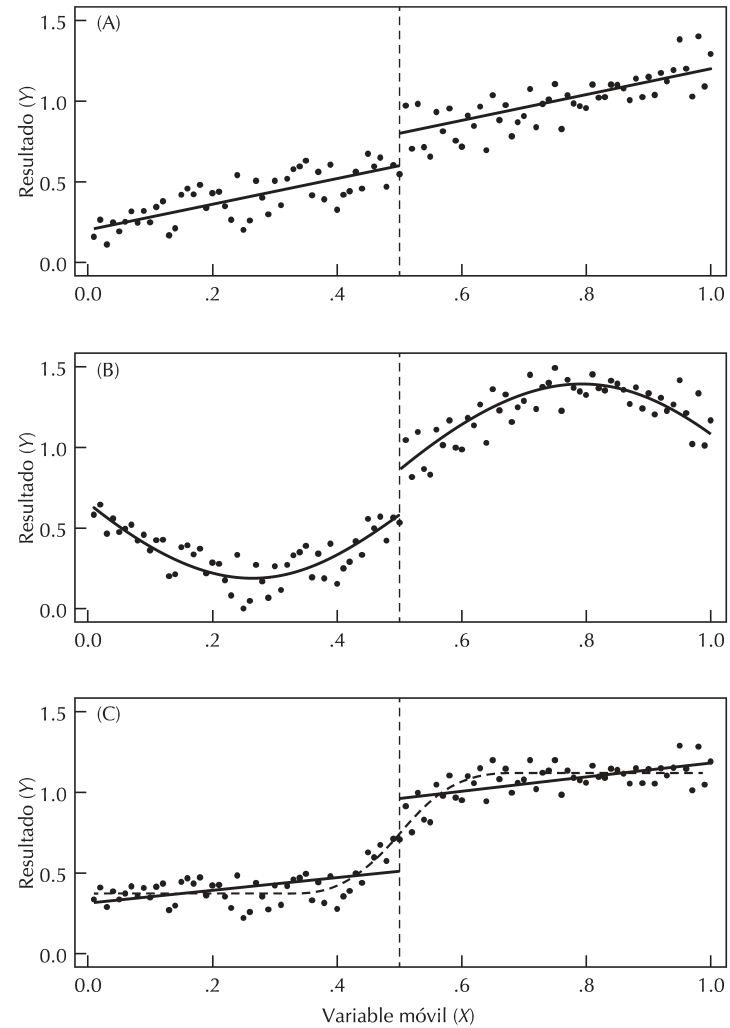
$$y = \alpha + D\tau + x\beta + \varepsilon.$$

- ¿Controlaremos otras variables?
- Aunque el tratamiento no se asigna al azar, sabemos de dónde proviene.

# Un setup lineal

- Para medir el impacto causal del premio al mérito, el RDD consiste en calcular, para el mismo individuo, el  $y$  que obtendría con y sin el tratamiento.
- Para hacerlo, el RDD se centra en las personas que obtienen una puntuación  $c$  y las siguientes razones:
  - $B'$  (que está relacionado con una puntuación  $c'$  ubicada justo encima de  $c$ ) sería una suposición razonable del valor de  $y$  del individuo que puntúa  $c$  en caso de que reciba el tratamiento;
  - $A''$  (que está relacionado con una puntuación  $c''$  ubicada justo debajo de  $c$ ) sería una estimación razonable del valor de  $y$  del individuo que puntúa  $c$  en el caso contrafactual en el que no recibe el tratamiento.
- Como consecuencia, el RDD considera  $\tau = B' - A''$  como el impacto causal de los premios al mérito en los logros académicos.

# RD detalles



# RD detalles

Se deben cumplir dos condiciones para considerar  $\tau$  como el impacto causal:

- **Primero:** El tratamiento de la población debe depender de si una variable observada excede un valor crítico denotado  $c$ .
- **Segundo:** Los individuos no tienen un control preciso sobre la variable de asignación (o forzamiento).



# RD detalles: Explicación intuitiva

- "Sin control preciso" significa que entre los que puntúan cerca del umbral, es una cuestión de "suerte" en qué lado del umbral aterrizan.
- Los que fracasan marginalmente y los que pasan marginalmente son idénticos.
- Suponemos, sin embargo, que todos los individuos son incapaces de controlar con precisión el puntaje justo alrededor del umbral.

# Formulación de un diseño RD

- Consideremos la formulación sencilla de un diseño RD:

$$\begin{aligned}y &= D\tau + W\delta_1 + U \\ D &= 1[X \geq c] \\ X &= W\delta_2 + V.\end{aligned}$$

Donde  $W$  es el vector de todas las características predeterminadas y observables del individuo que podrían afectar el resultado y/o la variable de asignación  $X$ .

- La heterogeneidad individual en el resultado se describe completamente por el par de variables aleatorias no observables  $(W, U)$ .

# Formulación de un diseño RD

- Este modelo parece una configuración de variable ficticia endógena estándar, excepto que observamos la variable de asignación,  $X$ .
- Esto nos permite relajar la mayoría de los demás supuestos que se suelen hacer en este tipo de modelo:
  - Primero, permitimos que  $W$  se determine de forma endógena siempre que se determine antes que  $V$ .
  - En segundo lugar, no tomamos ninguna postura sobre si algunos elementos de  $\delta_1$  o  $\delta_2$  son cero (restricciones de exclusión para IV).
  - En tercer lugar, no hacemos suposiciones sobre las correlaciones entre  $W, U$  y  $V$ .

# Sin control preciso

- Si los individuos tienen un control impreciso sobre  $X$ , tenemos la sorprendente implicación de que la variación en el estado del tratamiento será aleatoria en una vecindad del umbral. Para ver esto, tenga en cuenta que según la regla de Bayes, tenemos:

$$\begin{aligned} P_r[W = w, U = u | X = x] \\ = f(x | W = w, U = u) \frac{Pr[W = W, U = u]}{f(x)} \end{aligned}$$

donde  $f(\cdot)$  y  $f(\cdot | \cdot)$  son densidades marginal y conditional de  $X$ .

- Por lo tanto, cuando  $f(\cdot | \cdot)$  es continuo en  $x$ , la distribución  $Q$  y  $U$  condicional en  $X$  será continua en  $x$ .

# Sin control preciso

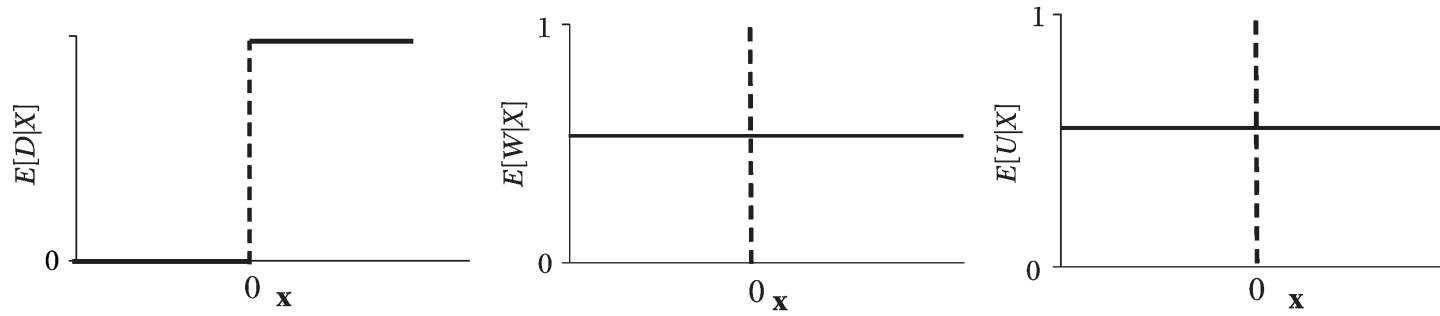
- Es importante enfatizar que, incluso en el caso de un "control impreciso", el individuo aún puede tener cierto control sobre  $\mathbf{X}$ : a través de sus esfuerzos, puede optar por cambiar la distribución hacia la derecha.
- La densidad para alguien con  $\mathbf{W} = \mathbf{w}, \mathbf{U} = \mathbf{u}$ , pero bien puede ser diferente, con una media, varianza o forma de densidad diferente, para otros individuos, con diferentes niveles de capacidad, que toman diferentes decisiones.
- Suponemos, sin embargo, que todos los individuos son incapaces de controlar con precisión la puntuación justo alrededor del umbral.

# Énfasis: randomización local

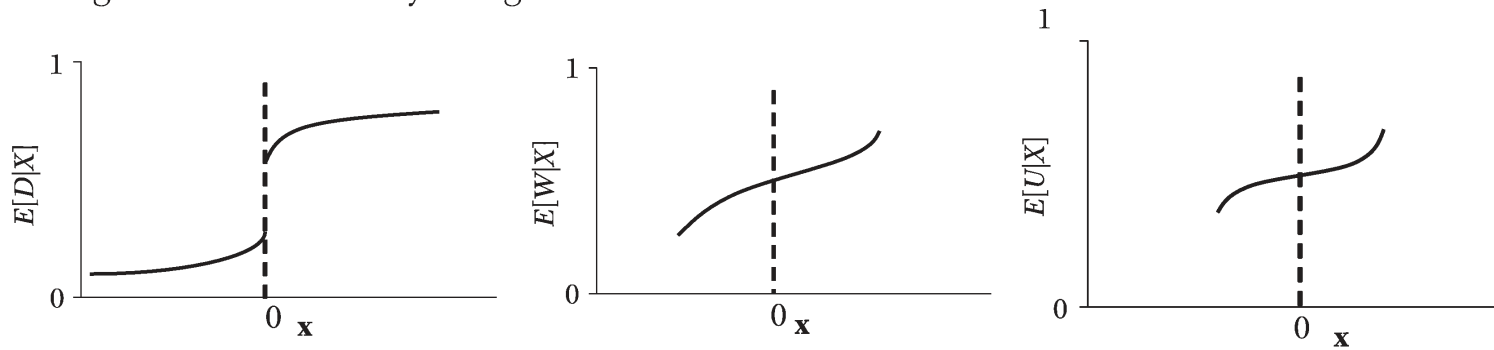
- Si los individuos tienen un control impreciso sobre  $X$  como se definió anteriormente, el tratamiento es "tan bueno como" asignado al azar alrededor del punto de corte, en otras palabras, tenemos "Aleatorización local".
- En otras palabras, la suposición conductual de que los individuos no manipulan  $X$  con precisión alrededor del umbral tiene la predicción de que el tratamiento es aleatorio localmente.
- La brecha de discontinuidad en el punto de corte identifica por tanto el efecto de tratamiento de interés.
- El diseño de RD es un primo mucho más cercano de los experimentos aleatorios que otros métodos de la competencia (como el emparejamiento en métodos observables y IV).

# RCTs y RDDs

## A. Randomized Experiment



## B. Regression Discontinuity Design



# RD detalles

- Podemos tener la siguiente especificación de nuestro RDD:

$$y = \alpha + D\tau + x\beta_1 + x^2\beta_2 + \varepsilon,$$

- donde  $x\beta_1 + x^2\beta_2$  es una función cuadrática de nuestra variable de ejecución.
- ¿Por qué cree que podríamos querer estimar esta ecuación?



# RD detalles

- También podríamos tener la siguiente especificación:

$$y = \alpha + D\tau + (x - x_0)\beta_1 + [(x - x_0)D]\beta_2 + \varepsilon,$$

- Hemos reemplazado  $x$  con  $x - x_0$  de la ecuación (1) y agregamos un término de interacción  $(x - x_0)D$ . Es decir, centrando la variable de ejecución restando el punto de corte.

# Implementación en la práctica

---

# Implementación en la práctica

Tres puntos importantes:

1. Representación gráfica
2. Estimación de la regresión a partir de modelos lineales y no lineales.
3. Una "lista de verificación" recomendada para la implementación.

# Representación gráfica

- Una gran ventaja del diseño de RD sobre los métodos de la competencia es su transparencia, que se puede ilustrar mediante métodos gráficos.
- Una forma estándar de graficar los datos es dividir la variable de asignación en varios contenedores, asegurándose de que haya dos contenedores separados a cada lado del punto de corte.
- Luego, el valor promedio de la variable de resultado se puede calcular para cada contenedor y graficar contra los puntos medios de los contenedores.

# Representación gráfica

- Para algún ancho de banda  $h$ , y para cierto número de contenedores  $K_0$  y  $K_1$  a la izquierda y derecha del valor de corte, respectivamente, la idea es construir contenedores  $b_k, b_{k+1}$ , para  $k = 1, 2, \dots, K_0 + K + 1$  dónde:

$$b_k = c - (k_0 - k + 1)h$$

- El valor promedio de la variable a estimar en el contenedor (bin) es:

$$\bar{y} = \frac{1}{N_k} \sum_{i=1}^N y_i 1[b_k < X_i \leq b_{k+1}]$$

- El número de observaciones en cada contenedor es:

$$N_k = \sum_{i=1}^N 1[b_k < X_i \leq b_{k+1}]$$

# Representación gráfica

- Hay varias ventajas importantes al graficar los datos de esta manera antes de comenzar a ejecutar regresiones:
  1. El gráfico proporciona una forma sencilla de visualizar cuál es la forma funcional de la función de regresión.
  2. La comparación de los resultados medios a la izquierda y a la derecha del punto de corte proporciona una indicación de la magnitud del efecto del tratamiento.
  3. El gráfico también muestra si hay saltos comparables inesperados en otros puntos.
- Tenga en cuenta que el impacto visual del gráfico generalmente se mejora al trazar también un modelo de regresión relativamente flexible, como un modelo polinomial.

# Estimando la regresión

- Intuitivamente, habrá dos pasos para estimar el efecto del tratamiento.
- Primero, estimar dos regresiones separadas a cada lado del punto de corte.
- En segundo lugar, calcular la diferencia entre el valor predicho de  $y$  en  $X = c$  de la regresión a la derecha del punto de corte, y la  $y$  en  $X = c$  de la regresión a la izquierda del punto de corte.

# Estimando la regreión

- Los modelos de regresión a la izquierda y a la derecha del punto de corte son:

$$y = \alpha_l + f_l(X - c) + \varepsilon,$$

$$y = \alpha_r + f_r(X - c) + \varepsilon$$

donde  $f_l(\cdot)$  y  $f_r(\cdot)$  son forma funcionales.

- Por tanto, el efecto del tratamiento es  $\tau = \alpha_r - \alpha_l$ .



# Estimando la regresión

- Una forma más directa de estimar el efecto del tratamiento es ejecutar una regresión agrupada en ambos lados del punto de corte:

$$y = \alpha_l + \tau D + f(X - c) + \varepsilon,$$

donde  $f(X - c) = f_l(X - c) + D[f_r(X - c) - f_l(X - c)]$  y  $\tau = \alpha_r - \alpha_l$ .

- Una ventaja del enfoque combinado es que produce directamente estimaciones y errores estándar del efecto del tratamiento  $\tau$ .

# El caso lineal

- El caso lineal donde  $f_l(X - c) = \beta_l(X - c)$  y  $f_r(X - c) = \beta_r(X - c)$ , la regresión agrupada es:

$$y = \alpha_l + \tau D + \beta_l(X - c) + (\beta_r - \beta_l)D(X - c) + \varepsilon.$$

- Se recomienda dejar que la función de regresión difiera en ambos lados del punto de corte al incluir términos de interacción entre  $D$  y  $X - c$ .

# Un principio

- Sin embargo, en la práctica, las estimaciones en las que la pendiente de regresión o, más generalmente, la función de regresión  $f(X - c)$  a menudo se limitan a ser iguales en ambos lados del punto de corte, a menudo se informan.
- Una posible justificación para hacerlo es que si la forma funcional es de hecho la misma en ambos lados del límite, entonces se obtienen estimaciones más eficientes del efecto del tratamiento al imponer esa restricción.
- Sin embargo, una especificación tan limitada no debe constituir la base central del enfoque empírico, sino que solo debe considerarse como una estimación adicional que debe informarse en aras de la exhaustividad.

# Regresiones lineales locales y ancho de banda

- ¿Cómo podemos saber qué forma funcional usar para  $f_r(X - c)$  y  $f_l(X - c)$ ?
  - Primero, si estimamos las regresiones lineales dentro de una ventana dada de ancho  $h$  alrededor del punto de corte, se puede suponer que  $f_r(X - c)$  y  $f_l(X - c)$  son lineales:

$$y = \alpha_l + \tau D + \beta_l(X - c) + (\beta_r - \beta_l)D(X - c) + \varepsilon,$$

donde  $c - h \leq X \leq c + h$ .

- Esta regresión lineal local es una regresión del núcleo con un núcleo rectangular. Es un tipo de muchas estrategias no paramétricas (locales).

# Notas sobre kernel

- Tenga en cuenta que estimar una regresión lineal estándar (paramétrica) utilizando puntos de datos que están dentro de  $\pm h$  del límite es equivalente a estimar una regresión lineal local con ancho de banda  $h$  y un núcleo rectangular.
- Un kernel es una función de ponderación que se utiliza en algunas técnicas de estimación no paramétricas y semiparamétricas.
- Estos pesos son distintos de cero dentro de un intervalo dado y cero fuera de él, con un patrón dentro de los intervalos que depende del tipo de kernel utilizado.
- Un núcleo rectangular pondera todas las observaciones en un intervalo igual.

# Elección de un ancho de banda

- Tenga en cuenta que elegir un ancho de banda  $h$  implica encontrar un óptimo equilibrio entre precisión y sesgo.
- Por un lado, el uso de un ancho de banda mayor produce estimaciones más precisas a medida que se dispone de más observaciones para estimar la regresión.
- Por otro lado, es menos probable que la especificación lineal sea precisa si la expectativa condicional subyacente no es lineal, ya que la especificación lineal solo puede proporcionar una aproximación cercana para la expectativa condicional no lineal en un ancho de banda pequeño.

# Elección de un ancho de banda

- Tenga en cuenta que elegir un ancho de banda  $h$  implica encontrar un equilibrio óptimo entre precisión y sesgo.
- Por un lado, el uso de un ancho de banda mayor produce estimaciones más precisas a medida que se dispone de más observaciones para estimar la regresión.
- Por otro lado, es menos probable que la especificación lineal sea precisa si la expectativa condicional subyacente no es lineal, ya que la especificación lineal solo puede proporcionar una aproximación cercana para la expectativa condicional no lineal en un ancho de banda pequeño.

# Enfoque paramétrico vs no paramétrico

Una forma de pensar en estos dos enfoques es la siguiente:

1. El enfoque paramétrico intenta elegir el modelo correcto para ajustarse a un conjunto de datos dado, mientras que el enfoque no paramétrico intenta elegir el conjunto de datos correcto para ajustarse a un modelo dado.
2. La sugerencia aquí es no depender de un método o especificación en particular.



# Prueba de la validez del diseño RD

- Primero, inspeccionando el histograma de la variable de asignación.
  - Tenga en cuenta que no podemos probar el supuesto de "control impreciso" directamente, ya que solo observaremos una observación sobre la variable de asignación por individuo en un momento dado.
  - Pero una prueba intuitiva de este supuesto es si la distribución agregada de la variable de asignación es discontinua, ya que una mezcla de densidades continuas a nivel individual es en sí misma una densidad continua.

# Prueba de la validez del diseño RD

- En segundo lugar, examinar si las covariables de referencia observadas están equilibradas "localmente" a ambos lados del umbral.
  - Es natural realizar un análisis DR gráfico y una estimación formal, reemplazando la variable dependiente con cada una de las covariables de referencia observadas en  $W$ .
  - Una discontinuidad indicaría una violación del supuesto subyacente que predice la asignación aleatoria local.

# Prueba de validez del diseño RD

- En tercer lugar, incorporar covariables en la estimación: si el diseño de RD es válido, el otro uso de las covariables de línea de base es solo para reducir la variabilidad muestral en las estimaciones de RD, pero no tiene ningún impacto en las estimaciones de RD.
  - Un enfoque es simplemente agregar  $\mathbf{W}$  a la regresión y demostrar que la inclusión de estos regresores no afectará la consistencia del estimador de RD.
  - Otro enfoque es, podemos "residualizar" la variable dependiente, es decir, restar de  $\mathbf{Y}$  una predicción de  $\mathbf{Y}$  basada en las covariables de línea de base  $\mathbf{W}$ , y luego realizar un análisis de RD en los residuos. Si el diseño de RD es válido, este procedimiento proporciona una estimación consistente del mismo parámetro de RD de interés.

# El Checklist recomendado

1. Evaluar la posibilidad de manipulación de la variable de asignación mostrando su distribución.
2. Presente el gráfico RD principal utilizando promedios locales agrupados.
3. Grafique una especificación polinomial de referencia.
4. Explore la sensibilidad de los resultados a un rango de anchos de banda y un rango de órdenes al polinomio.
5. Lleve a cabo un análisis de RD paralelo en las covariables de la línea de base.
6. Explore la sensibilidad de los resultados a la inclusión de covariables de referencia

**Recordemos:** La inclusión de covariables de línea de base, sin importar cuán altamente correlacionadas estén con el resultado, no debería afectar la discontinuidad estimada, si se cumple el supuesto de no manipulación.

# Ejemplos

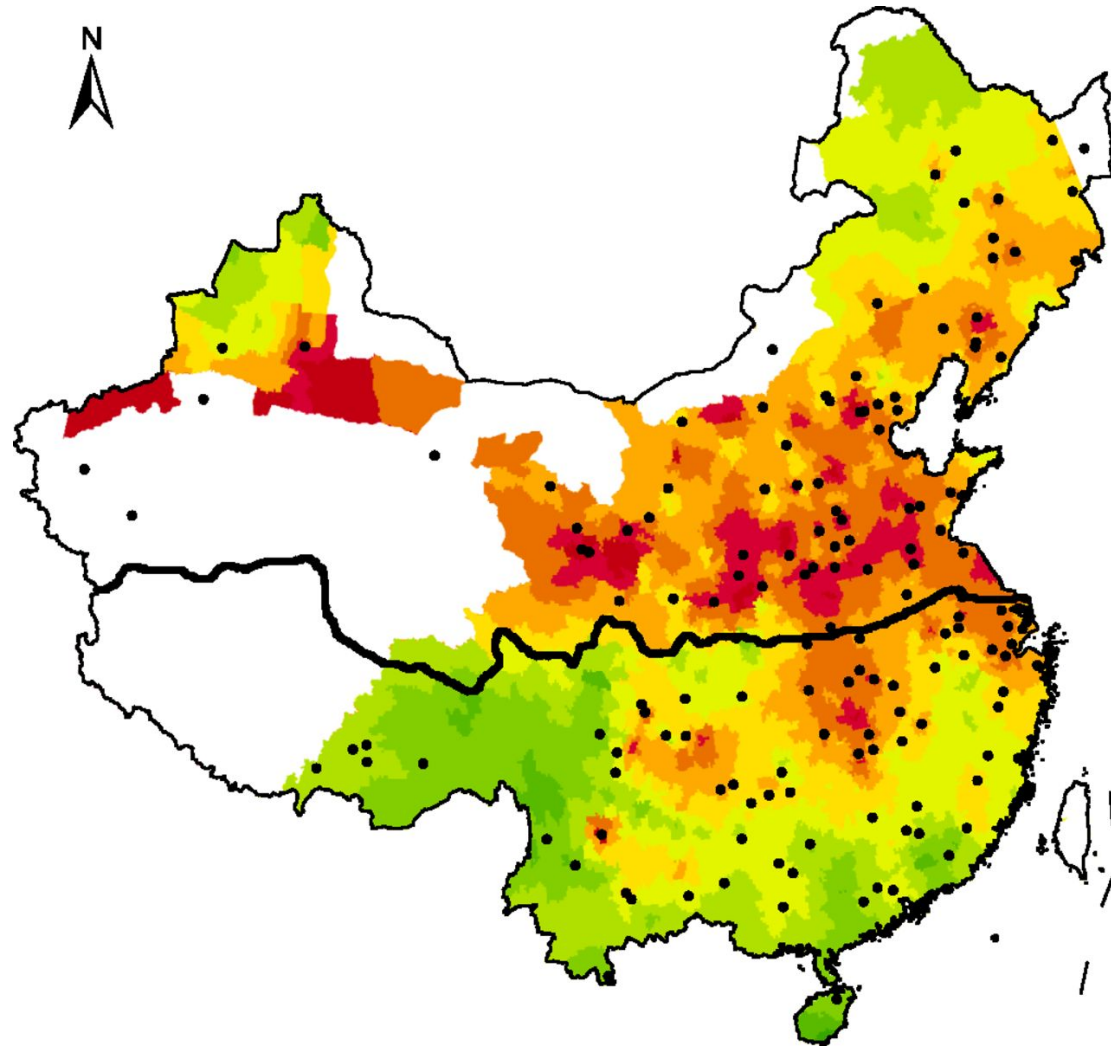
---

# Exposición sostenida a la contaminación del aire

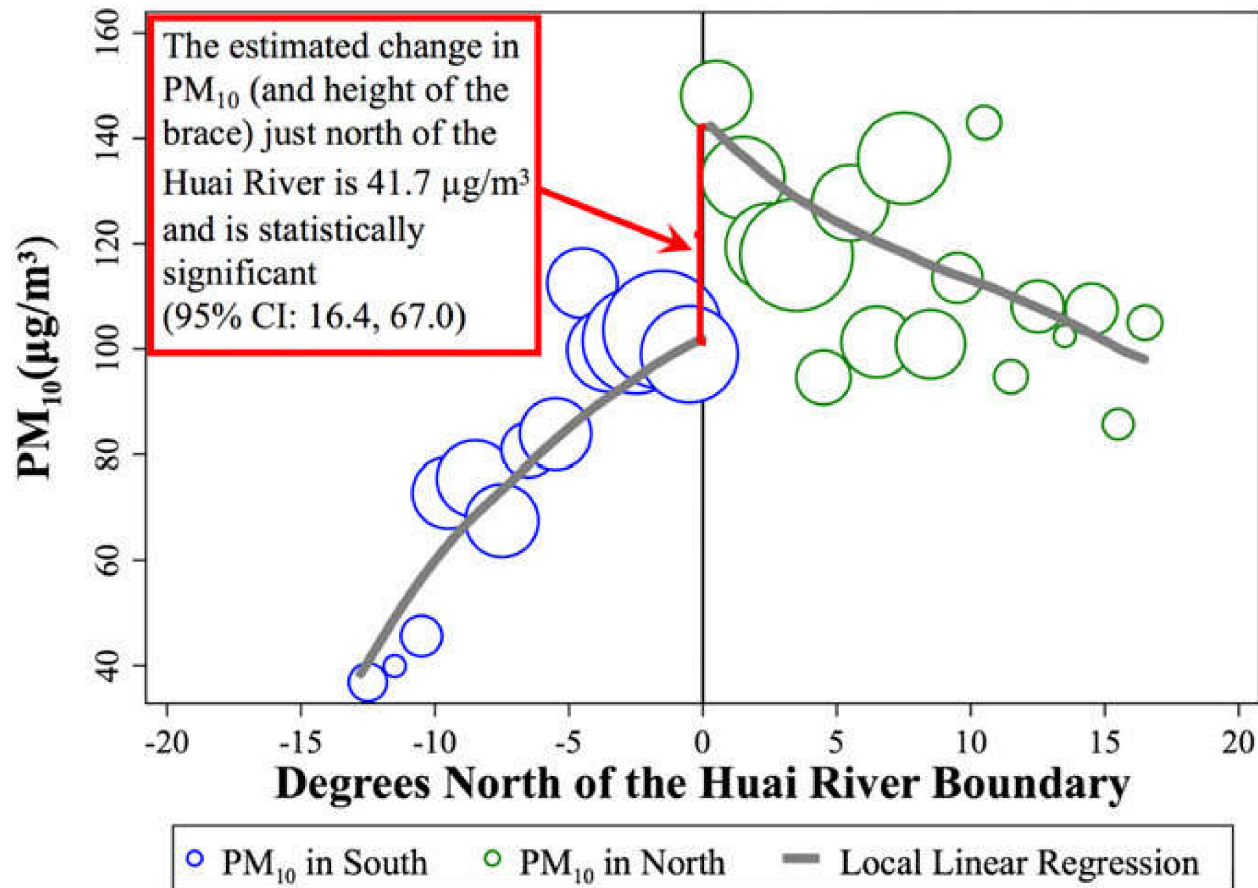
Ebenstein, A. et al. (2016):

- Estima el efecto de la exposición sostenida a partículas menores de  $10\text{ }\mu\text{m}$  (PM10) en la esperanza de vida.
- El diseño de la investigación cuasiexperimental se basa en la política del río Huai de China
- La política se instituyó durante la década de 1950 cuando los recursos económicos se asignaron a través de la planificación central y dictaba que las áreas al norte del río Huai recibían carbón gratuito o altamente subsidiado para calefacción interior.
  - Construcción de una infraestructura de calefacción centralizada a carbón solo en ciudades al norte del río Huai.
  - No hay un sistema equivalente en las ciudades del sur.

# Exposición sostenida a la contaminación del aire



# Exposición sostenida a la contaminación del aire



**Fig. 2.** Fitted values from a local linear regression of PM<sub>10</sub> exposure on distance from the Huai River estimated separately on each side of the river.



# Orígenes legales y VIH femenino

Anderson (2017):

- Ella demuestra que el origen legal de estos países anteriormente colonizados determina significativamente las tasas actuales de VIH femenino.
- Centrándose en las diferencias en los derechos de propiedad de las mujeres bajo dos códigos legales.
  - Derecho consuetudinario vs Derecho civil.
- Ella estima la siguiente ecuación:

$$Y_{rcepi} = \alpha_0 + \alpha_1 L_{rc} + \alpha_2 X_{rc} + \alpha_3 X_{rcep} + \alpha_4 X_{rcepi} + f(BD_{rcep}) + \delta_e + \gamma_r + \lambda_t + \epsilon_{rcepi},$$

- Nuestras principales variables son:  $L_{rc}$  and  $f(BD_{rcep})$ ; indicador de derecho consuetudinario y un polinomio de segundo orden de la distancia desde el centroide de cada píxel hasta la frontera nacional más cercana con diferentes orígenes legales.

# Orígenes legales y VIH femenino

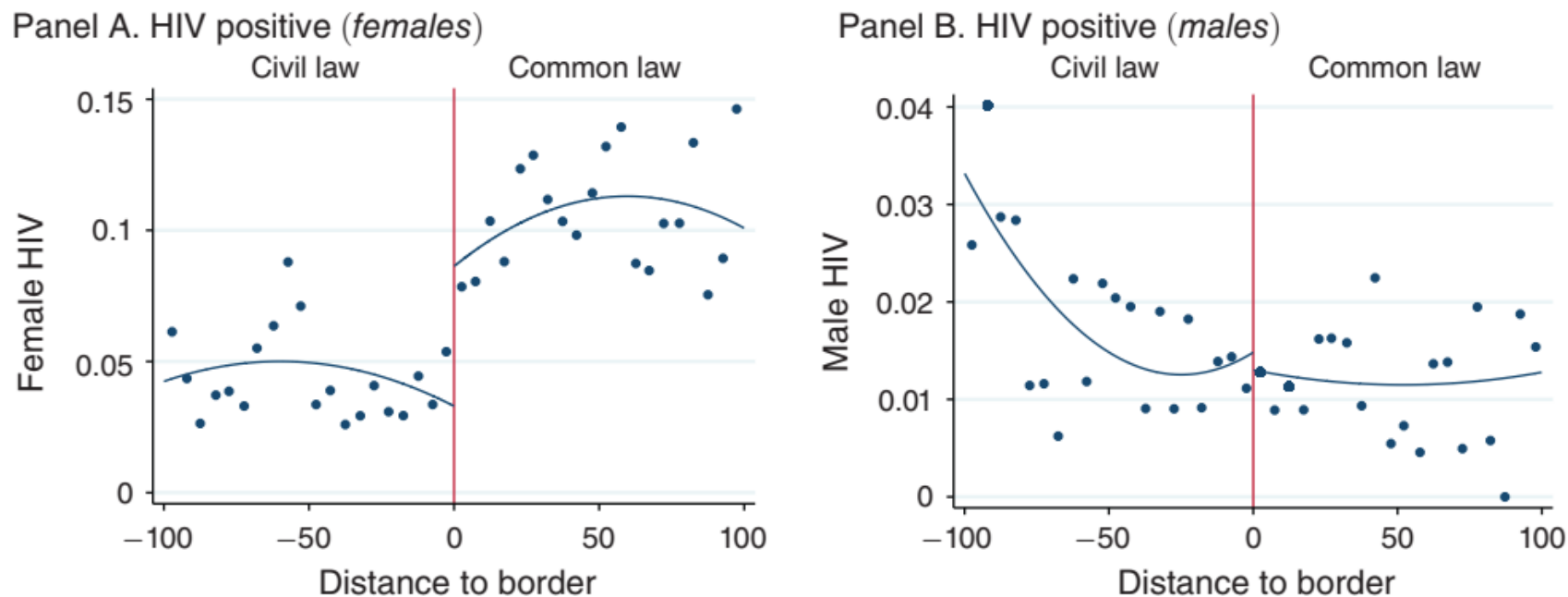
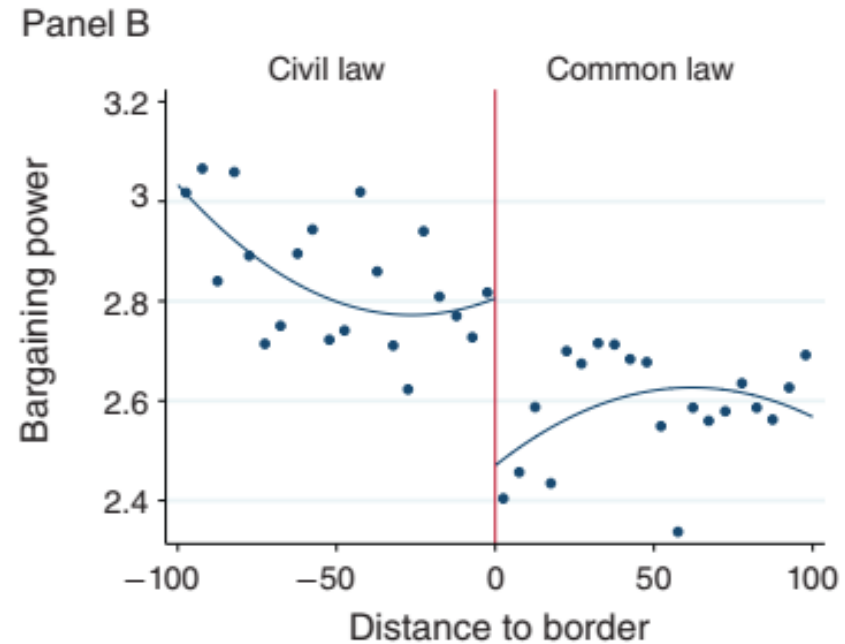
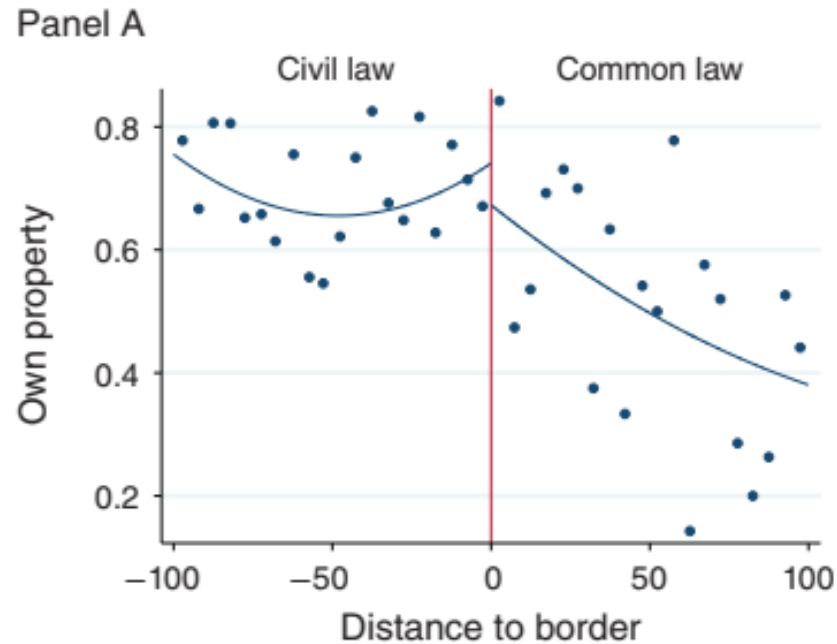


FIGURE 2. HIV POSITIVE

# Orígenes legales y VIH femenino



# Orígenes legales y VIH femenino

*Table 1 - HIV Positive (Females aged 15-49)*

Variable	Whole Sample	$\leq 150$ km	$\leq 100$ km	Non-Muslim/ Non-Polygynous	Muslim/ Polygynous
	$\leq 200$ km			$\leq 100$ km	$\leq 100$ km
Common Law	0.016*** (0.006)	0.018*** (0.006)	0.019*** (0.007)	0.016*** (0.006)	0.007 (0.013)
Observations	118903	99511	77336	55507	21829

Notes: \*\*\* indicates statistical significance at the 1% level. Standard errors are clustered at the ethnic and country level using the approach of Cameron, Gelbach, and Miller (2011). All estimations include: country, individual, and pixel controls; region fixed effects; ethnic fixed effects; 2nd-order RD polynomial of distance to national border; and the year of the survey. Refer to the Appendix for details on the data.

# Fuzzy RDD

---

# Fuzzy RDD

- Hemos discutido varios ejemplos donde el umbral determina (perfectamente) el estado del tratamiento (es decir, RD aguda).
  - Ahora, analicemos Fuzzy RD donde el umbral cambia la probabilidad de tratamiento, pero menos de 100 pp
  - Recuerde el concepto de cumplimiento. Al igual que en RCT, la asignación de tratamiento no incluye el cumplimiento perfecto.
- Por lo tanto, podemos usarlo como un IV (estimaciones de ITT).

# Fuzzy RDD

- Supongamos el siguiente modelo donde  $D_i$  denota el tratamiento como antes:

$$P[D_i = 1|x_i] = \begin{cases} g_0(x_1), & \text{if } x_i \geq x_0 \\ g_1(x_1), & \text{if } x_i < x_0 \end{cases}, \text{ where } g_1(x_0) \neq g_0(x_0).$$

- ¿Cuál sería nuestra principal suposición con respecto a  $g_1(x_0)$  and  $g_0(x_0)$  ?
- Pueden ser cualquier cosa siempre que difieran (y cuanto más mejor) en  $x_0$ .

# Fuzzy RDD

Fuzzy RD conduce naturalmente a una estrategia de estimación 2SLS simple:

- Supongamos que  $g_1(x_0)$  y  $g_0(x_0)$  pueden describirse mediante polinomios de orden  $p$ th:

$$\begin{aligned} E[D_i|x_i] &= \gamma_{00} + \gamma_{01}x_1 + \gamma_{02}x_i^2 + \cdots + \gamma_{0p}x_i^p \\ &\quad + [\gamma_0^* + \gamma_1^*x_i + \gamma_2^*x_i^2 + \cdots + \gamma_p^*x_i^p]T_i \\ &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + \cdots + \gamma_{0p}x_i^p \\ &\quad + \gamma_0^*T_i + \gamma_1^*x_iT_i + \gamma_2^*x_i^2T_i + \cdots + \gamma_p^*x_i^pT_i \end{aligned}$$

- La variable dummy  $T_i$  indica el punto de discontinuidad en  $E[D_i|x_i]$  donde  $T_i = 1(x_i \geq x_0)$ .



# Fuzzy RDD

El estimador de RD difuso más simple usa solo  $T_i$  como un instrumento, sin los términos de interacción (con los términos de interacción en la lista de instrumentos, también nos gustaría dejar las interacciones en la segunda etapa).

- La primera etapa en este caso sería:

$$D_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \cdots + \gamma_p x_i^p + \pi T_i + \xi_{1i},$$

- donde  $T_i$ , es el instrumento excluido que proporciona poder identificativo con efectos de primera etapa dados por  $\pi$ .

# Fuzzy RDD

La primera etapa que utiliza los términos de interacciones se puede escribir como:

$$\begin{aligned} D_i = & \gamma_{00} + \gamma_{01}\tilde{x}_1 + \gamma_{02}\tilde{x}_i^2 + \cdots + \gamma_{0p}\tilde{x}_i^p \\ & + \gamma_0^*T_i + \gamma_1^*\tilde{x}_i T_i + \gamma_2^*\tilde{x}_i^2 T_i + \cdots + \gamma_p^*\tilde{x}_i^p T_i \end{aligned}$$

- ¿Cuáles son los instrumentos excluidos?
- $T_i, \tilde{x}_i T_i, \tilde{x}_i^2 T_i, \dots, \tilde{x}_i^p T_i$ .
- ¿Cuál sería nuestra segunda etapa?

# Fuzzy RDD

- Una forma más fácil de ver el 2SLS:

$$\begin{aligned}y &= \alpha + \tau D + f(X - c) + \varepsilon, \\D &= \alpha + \delta T + g(X - c) + \epsilon.\end{aligned}$$

# Fuzzy RD: Takeaways

- Piense en la RD difusa como la versión IV de la RD aguda.
- Se aplican todas las lecciones IV, incluidos los requisitos de la restricción de exclusión.
- Los efectos del tratamiento en ese contexto son tardíos para las personas cercanas al umbral para quienes el umbral induce un cambio en el estado del tratamiento.
- Verifique los umbrales de placebo.

# Ejemplos: Fuzzy RDD

---

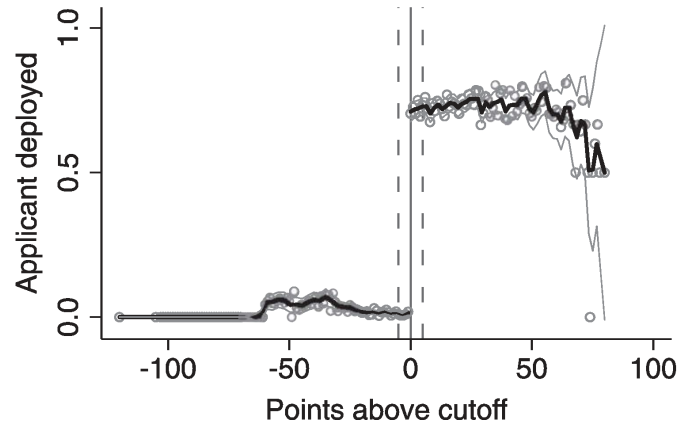
# Finanzas familiares

Clemens y Tiongson (2017) utilizan un RDD para estudiar una discontinuidad de políticas en Filipinas que asignó de manera cuasialeatoria la migración familiar parcial y temporal para trabajos con salarios altos en Corea:

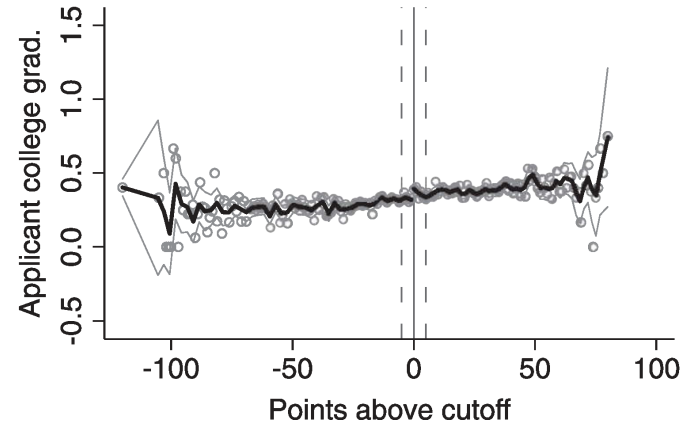
- Migración y cambios en el gasto de los hogares.
- Acuerdo bilateral entre Filipinas y Corea del Sur que permite la participación de trabajadores filipinos en el sistema de permisos de empleo de Corea.
- Diseño:
  - Se requirió la solicitud de trabajos de EPS para aprobar una prueba de idioma coreano (KLT).
  - Se requirió una puntuación de 120 puntos o más para obtener un permiso de trabajo.
- Resultados: gastar más en salud y educación, reducir los préstamos y aumentar los ahorros.

# Finanzas familiares

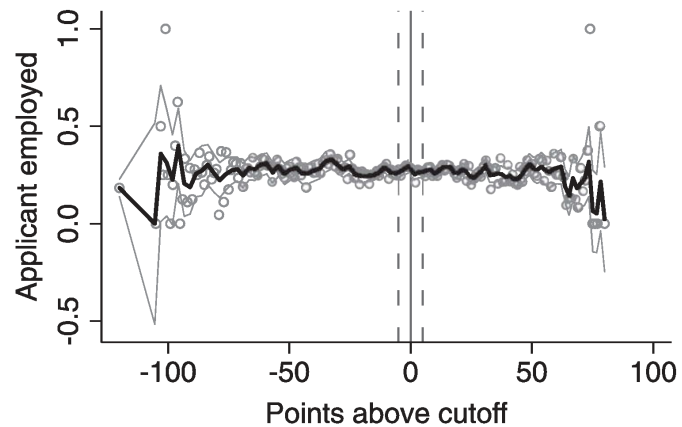
FIGURE 1.—DISCONTINUITIES IN SAMPLING UNIVERSE



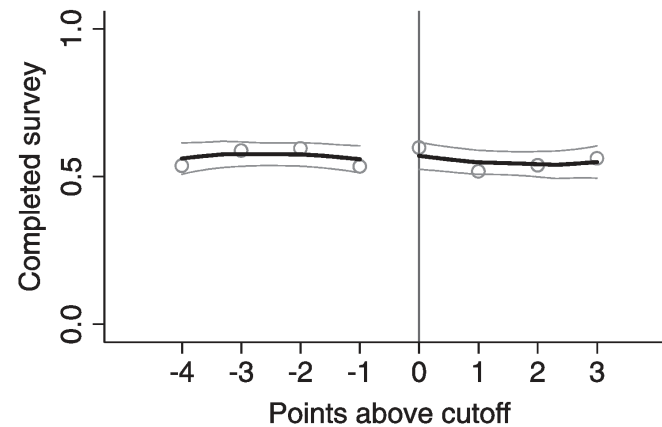
N = 23448, bandwidth = 2.113



N = 23448, bandwidth = 2.149



N = 23448, bandwidth = 2.211



Good addresses only. N = 1532, bandwidth = 2.360

# Finanzas familiares

TABLE 2.—IMPACTS ON HOUSEHOLDS: SPENDING, SAVING, AND BORROWING

Outcome	Parametric				Nonparametric			
	$\hat{\beta}_{ITT}$	SE	$\hat{\beta}_{TOT}$	SE	$\hat{\gamma}_{ITT}$	SE	$\hat{\gamma}_{TOT}$	SE
All households								
$asinh$ expenditures: Total	0.179**	(0.073)	0.364**	(0.147)	0.129	(0.124)	0.280	(0.271)
Food	0.026	(0.090)	0.053	(0.181)	0.004	(0.079)	0.009	(0.165)
Quality of life	0.181**	(0.078)	0.368**	(0.158)	0.146*	(0.078)	0.304*	(0.162)
Education and health	1.182***	(0.317)	2.399***	(0.661)	0.977***	(0.322)	2.035***	(0.679)
Durables	0.382	(0.445)	0.776	(0.893)	0.571	(0.794)	1.250	(1.738)
$asinh$ savings	0.887*	(0.471)	1.800*	(0.963)	0.801	(0.490)	1.668	(1.030)
Borrowed: family, for business?	−0.053	(0.045)	−0.107	(0.091)	−0.081*	(0.044)	−0.169*	(0.094)
Borrowed: nonfamily, for business?	−0.027	(0.038)	−0.055	(0.076)	0.086	(0.071)	0.188	(0.159)
Borrowed: family, for nonbusiness?	−0.034**	(0.015)	−0.068**	(0.032)	−0.034**	(0.017)	−0.072**	(0.036)
Borrowed: nonfamily, for nonbusiness?	−0.050	(0.041)	−0.101	(0.081)	−0.047	(0.042)	−0.097	(0.087)
Households with applicant initially married, only								
$asinh$ Expenditures: Total	0.117	(0.106)	0.247	(0.221)	0.161	(0.105)	0.321	(0.209)
Food	0.008	(0.154)	0.018	(0.321)	−0.044	(0.206)	−0.097	(0.459)
Quality of life	0.088	(0.105)	0.186	(0.218)	0.144	(0.108)	0.287	(0.214)
Education and health	0.912**	(0.430)	1.935**	(0.940)	0.711*	(0.423)	1.418	(0.870)
Durables	−0.300	(0.663)	−0.637	(1.393)	0.815	(1.173)	1.810	(2.602)
$asinh$ savings	−0.234	(0.728)	−0.496	(1.516)	0.544	(1.349)	1.205	(3.042)
Borrowed: family, for business?	−0.108	(0.073)	−0.229	(0.154)	−0.136**	(0.068)	−0.270*	(0.140)
Borrowed: nonfamily, for business?	−0.073	(0.062)	−0.154	(0.133)	0.107	(0.116)	0.237	(0.256)
Borrowed: family, for nonbusiness?	−0.055**	(0.023)	−0.116**	(0.053)	−0.028	(0.059)	−0.061	(0.132)
Borrowed: nonfamily, for nonbusiness?	−0.107*	(0.063)	−0.226*	(0.132)	−0.042	(0.121)	−0.092	(0.267)
Households with applicant initially unmarried, only								
$asinh$ Expenditures: Total	0.231**	(0.100)	0.457**	(0.198)	0.086	(0.169)	0.186	(0.359)
Food	0.045	(0.111)	0.088	(0.216)	−0.057	(0.109)	−0.125	(0.237)
Quality of life	0.247**	(0.111)	0.487**	(0.222)	0.169	(0.194)	0.365	(0.417)
Education and health	1.425***	(0.461)	2.813***	(0.921)	1.192**	(0.472)	2.603**	(1.029)
Durables	0.951	(0.608)	1.879	(1.194)	0.406	(1.076)	0.878	(2.320)
$asinh$ savings	1.738***	(0.599)	3.432***	(1.245)	1.247**	(0.616)	2.722**	(1.387)
Borrowed: family, for business?	−0.003	(0.059)	−0.006	(0.114)	−0.022	(0.105)	−0.047	(0.226)
Borrowed: nonfamily, for business?	−0.003	(0.049)	−0.006	(0.094)	0.060	(0.086)	0.130	(0.192)
Borrowed: family, for nonbusiness?	−0.015	(0.021)	−0.030	(0.041)	−0.049	(0.047)	−0.109	(0.104)
Borrowed: nonfamily, for nonbusiness?	−0.009	(0.053)	−0.018	(0.103)	−0.003	(0.101)	−0.007	(0.221)

$asinh$  is inverse hyperbolic sine.  $N_{fail} = 460$ ;  $N_{pass} = 439$ . Parametric regressions include predetermined controls (see note 4). “Treatment” = household ever had a member in Korea. ITT = intent-to-treat effect. TOT = treatment-on-treated effect. \*\*\* $p < 0.01$ , \*\* $p < 0.05$ , \* $p < 0.10$ . “Food” = food, beverages, and tobacco. “Health and education” = school, medicine, and medical care. “Quality of life” = fuel, transportation, household and personal care, clothing, recreation, family occasions, gifts. “Durables” = durable goods, taxes, home improvement. “Savings” includes deposits in banks, paying off loans, making loans to others.



# Conclusiones

---

# Conclusiones

- Los diseños de RD pueden ser válidos en el entorno muy general de "selección en inobservables", lo que permite una correlación arbitraria entre  $U$ ,  $V$  y  $W$ , pero al mismo tiempo no requiere un IV.
- Todo lo que se necesita es que, condicional en  $W$ ,  $U$ , la densidad de  $V$  sea continua, y sigue el resultado de la aleatorización local.
- ¿Cómo un diseño de RD puede lograr esto, dados estos supuestos débiles? La respuesta está en lo absolutamente necesario en un diseño de RD: la observabilidad del índice latente  $X$ .
- Con el diseño de RD, el investigador puede evitar adoptar una postura firme sobre qué  $W$  incluir en el análisis, ya que el diseño predice que las  $W$  son irrelevantes e innecesarias para la identificación.

# Conclusiones

- Por lo tanto, puede ser más útil considerar los diseños de RD como una descripción de un *proceso de generación de datos* en particular, en lugar de un "método" o incluso un "enfoque".
- En prácticamente cualquier contexto con un resultado  $y$ , estado de tratamiento  $D$  y otras variables observables  $W$ , en principio, un investigador puede construir un estimador de control de regresión o IV y afirmar que se satisfacen los supuestos de identificación necesarios.
- Esto no es así con un diseño RD. Si se observa  $X$ , entonces no hay más remedio que intentar estimar la expectativa de  $y$  condicionada a  $X$  a cada lado del límite. Si no es así, entonces el diseño RD simplemente no se aplica. En este sentido, el diseño de RD obliga al investigador a analizarlo de una manera particular, y hay poco margen para la discreción del investigador, al menos desde el punto de vista de la identificación.

¡Muchas gracias!