Inferencia, Causalidad y Políticas Públicas
ECO-60116

Week 04: Regresión Discontinua

Eduard F. Martinez Gonzalez, Ph.D.

Departamento de Economía, Universidad Icesi

September 19, 2025

Roadmap

1. Regresión Discontinua (RD)
   - Motivación
   - Definición Formal
   - Estimación
   - Supuestos
2. Impact Assessment: TOP - El Campo Emprende
   - Contexto
   - Datos y Estrategía de Identificación
   - Resultados
3. Aplicación en R

RECAP: IV

- **Objetivo:** identificar el efecto causal de X sobre Y cuando hay *endogeneidad* (omitidas, reversa, medición).

RECAP: IV

- **Objetivo:** identificar el efecto causal de X sobre Y cuando hay *endogeneidad* (omitidas, reversa, medición).

- **Instrumento válido (Z):**
  - *Relevancia:* $\mathrm{Cov}(Z, X) \neq 0$
  - *Exclusión:* $\mathrm{Cov}(Z, \varepsilon) = 0$.

RECAP: IV

- **Objetivo:** identificar el efecto causal de X sobre Y cuando hay *endogeneidad* (omitidas, reversa, medición).

- **Instrumento válido (Z):**
  - *Relevancia:* $\mathrm{Cov}(Z, X) \neq 0$
  - *Exclusión:* $\mathrm{Cov}(Z, \varepsilon) = 0$.

- **Estimación:**
  - 1ª etapa $X_i = \phi + \gamma Z_i + \nu_i$
  - 2ª etapa $Y_i = \alpha + \beta \hat{X}_i + \epsilon_i$
  - **Atajo (Wald, Z binaria):** $\hat{\beta} = \dfrac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]}$

RECAP: IV

- **Objetivo:** identificar el efecto causal de X sobre Y cuando hay *endogeneidad* (omitidas, reversa, medición).

- **Instrumento válido (Z):**
  - *Relevancia:* $\mathrm{Cov}(Z, X) \neq 0$
  - *Exclusión:* $\mathrm{Cov}(Z, \varepsilon) = 0$.

- **Estimación:**
  - 1ª etapa $X_i = \phi + \gamma Z_i + \nu_i$
  - 2ª etapa $Y_i = \alpha + \beta \hat{X}_i + \epsilon_i$
  - **Atajo (Wald, Z binaria):** $\hat{\beta} = \dfrac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]}$

- **Diagnósticos clave:** F de instrumentos excluidos > 10; signos consistentes entre 1ª etapa, forma reducida y 2SLS.

RECAP: IV

- **Objetivo:** identificar el efecto causal de X sobre Y cuando hay *endogeneidad* (omitidas, reversa, medición).

- **Instrumento válido (Z):**
  - *Relevancia:* $\mathrm{Cov}(Z, X) \neq 0$
  - *Exclusión:* $\mathrm{Cov}(Z, \varepsilon) = 0$.

- **Estimación:**
  - 1ª etapa $X_i = \phi + \gamma Z_i + \nu_i$
  - 2ª etapa $Y_i = \alpha + \beta \hat{X}_i + \epsilon_i$
  - **Atajo (Wald, Z binaria):** $\hat{\beta} = \dfrac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]}$

- **Diagnósticos clave:** F de instrumentos excluidos > 10; signos consistentes entre 1ª etapa, forma reducida y 2SLS.

- **Reporte:** mostrar 1ª etapa, forma reducida y 2SLS; discutir plausibilidad de exclusión; pruebas de sobreidentificación (si $L > K$) informan consistencia conjunta (no prueban exclusión).

Roadmap

1. Regresión Discontinua (RD)
   - Motivación
   - Definición Formal
   - Estimación
   - Supuestos
2. Impact Assessment: TOP - El Campo Emprende
   - Contexto
   - Datos y Estrategía de Identificación
   - Resultados
3. Aplicación en R

Arbitrariedad en el método de asignación

¿Qué es la Regresión Discontinua (RD)?

Es una **estrategia cuasi-experimental** que imita un experimento aleatorio cuando la **asignación al tratamiento depende de superar un umbral específico** en una variable continua. El valor específico del umbral que determina la asignación al tratamiento suele ser **arbitrario**, definido por decisiones administrativas.

Arbitrariedad en el método de asignación

¿Qué es la Regresión Discontinua (RD)?

Es una **estrategia cuasi-experimental** que imita un experimento aleatorio cuando la **asignación al tratamiento depende de superar un umbral específico** en una variable continua. El valor específico del umbral que determina la asignación al tratamiento suele ser **arbitrario**, definido por decisiones administrativas.

- Focalización de programas sociales (ej. *Familias en Acción*).
- Becas según promedio acumulado.
- Puntajes de pruebas estandarizadas para acceso a educación superior.
- Beneficios tributarios para firmas con ciertas características (ej. menos de N empleados).
- Tamaño de la población de los municipios.

El Método de Asignación en RD Nítida

Elementos básicos del diseño:

- z_i: La **variable de asignación** (running variable). Es continua y observable.
  Ej: Puntaje en un examen.

El Método de Asignación en RD Nítida

Elementos básicos del diseño:

- z_i: La **variable de asignación** (running variable). Es continua y observable.
  Ej: Puntaje en un examen.

- z_0: El **umbral** o punto de corte. Es un valor conocido.
  Ej: El puntaje mínimo para aprobar (e.g., 60).

El Método de Asignación en RD Nítida

Elementos básicos del diseño:

- z_i: La **variable de asignación** (running variable). Es continua y observable.
  Ej: Puntaje en un examen.

- z_0: El **umbral** o punto de corte. Es un valor conocido.
  Ej: El puntaje mínimo para aprobar (e.g., 60).

- D_i: La variable de tratamiento (indicador binario de elegibilidad).

El Método de Asignación en RD Nítida

Elementos básicos del diseño:

- z_i: La **variable de asignación** (running variable). Es continua y observable.
  Ej: Puntaje en un examen.

- z_0: El **umbral** o punto de corte. Es un valor conocido.
  Ej: El puntaje mínimo para aprobar (e.g., 60).

- D_i: La variable de tratamiento (indicador binario de elegibilidad).

Regla de asignación determinística y discontinua en el umbral:

$$D_i = \begin{cases} 1 & \text{si } z_i \geq z_0 \quad \text{(Tratado)} \\ 0 & \text{si } z_i < z_0 \quad \text{(Control)} \end{cases}$$

Importante: En RD nítida la asignación es **determinística y discontinua en** z_0, ¡pero **NO es aleatoria**!

Contexto Apropiado para Aplicar RDN

Para que este método sea válido, deben cumplirse tres condiciones esenciales:

1. **Asignación Determinística:** La elegibilidad depende **únicamente** de la variable z_i y el umbral z_0. *Ej: puntaje mínimo para recibir una beca.*

Contexto Apropiado para Aplicar RDN

Para que este método sea válido, deben cumplirse tres condiciones esenciales:

1. **Asignación Determinística:** La elegibilidad depende **únicamente** de la variable z_i y el umbral z_0. *Ej: puntaje mínimo para recibir una beca.*

2. **Cumplimiento Perfecto:** Todos los elegibles reciben el tratamiento y los no elegibles no lo reciben (*perfect compliance*). Esto distingue la **RD nítida** de la difusa.

Contexto Apropiado para Aplicar RDN

Para que este método sea válido, deben cumplirse tres condiciones esenciales:

1. **Asignación Determinística:** La elegibilidad depende **únicamente** de la variable z_i y el umbral z_0. *Ej: puntaje mínimo para recibir una beca.*

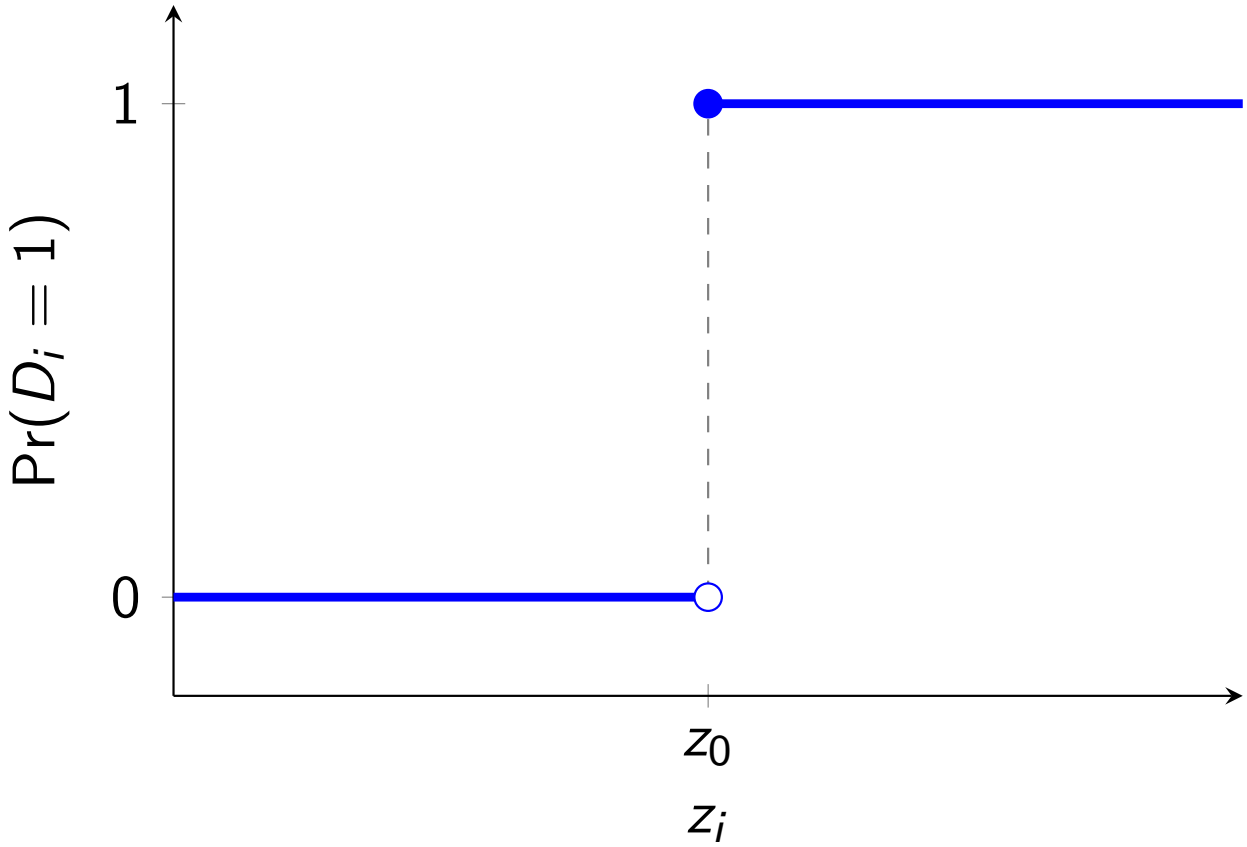
2. **Cumplimiento Perfecto:** Todos los elegibles reciben el tratamiento y los no elegibles no lo reciben (*perfect compliance*). Esto distingue la **RD nítida** de la difusa.

3. **No Manipulación:** Los individuos no pueden alterar estratégicamente z_i para quedar justo del lado tratado. *Ej: no se puede "inflar" el puntaje para pasar el umbral.*

En resumen: la RD nítida funciona cuando el **umbral genera un salto real y exógeno** en la probabilidad de recibir el tratamiento.

La Discontinuidad es la Clave

La característica principal de la RD Nítida es un **salto** en la probabilidad de tratamiento de 0 a 1 en el umbral z_0.

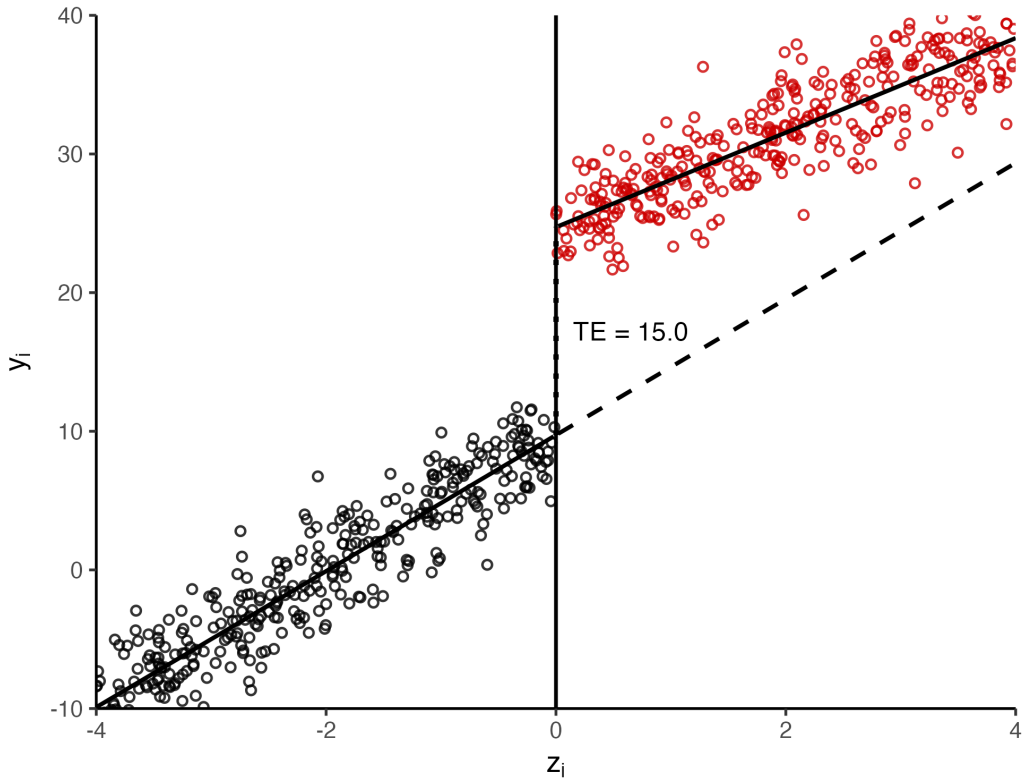


- Hay una discontinuidad en la probabilidad de recibir el tratamiento en z_0

La Discontinuidad es la Clave

Pregunta principal: Si la probabilidad del tratamiento es discontinua, ¿qué pasa con la variable de resultado y_i alrededor de z_0?

Figure: Ejemplo Gráfico: Salto en la Variable de Resultado



Un "salto" en y_i en la "**vecindad**" del umbral es indicativo de que el tratamiento tiene un efecto causal (*Natura non facit saltum*, Darwin,C.).

Roadmap

1. Regresión Discontinua (RD)
   - Motivación
   - Definición Formal
   - Estimación
   - Supuestos
2. Impact Assessment: TOP - El Campo Emprende
   - Contexto
   - Datos y Estrategía de Identificación
   - Resultados
3. Aplicación en R

Resultados potenciales en RDN

Cada unidad i tiene dos posibles resultados:

$$y_i = \begin{cases} y_i^1 & \text{si recibe el tratamiento} \\ y_i^0 & \text{si no recibe el tratamiento} \end{cases}$$

Método de asignación:

$$D_i = \begin{cases} 1 & \text{si } z_i \geq z_0 \quad \text{(Tratado)} \\ 0 & \text{si } z_i < z_0 \quad \text{(Control)} \end{cases}$$

Implicaciones:

1. $y_i^0, y_i^1 \not\perp D_i$: la asignación al tratamiento **no es aleatoria**, depende del valor de z_i.

2. $y_i^0, y_i^1 \perp D_i \mid z_i$: **condicional en** z_i, la asignación es determinística. *Esto es lo que permite identificar un salto en el umbral.*

Importante: En RDN la independencia surge **sólo en el entorno del umbral**.

Resultados potenciales y regresión lineal

1. **Resultado sin tratamiento:**

$$E[y_i^0 \mid z_i] = \beta_0 + f(z_i)$$

donde $f(z_i)$ captura la relación sistemática entre z_i y el resultado en ausencia de tratamiento.

2. **Efecto del tratamiento:**

$$\tau = y_i^1 - y_i^0$$

Si asumimos que τ es constante, hablamos de un efecto promedio global. Más generalmente:

$$\tau_z = E[y_i^1 - y_i^0 \mid z_i = z]$$

es un efecto causal que puede depender de z.

3. **Modelo empírico:** Definimos un componente idiosincrático:

$$\epsilon_i = y_i^0 - E[y_i^0 \mid z_i]$$

y escribimos la regresión:

$$y_i = \beta_0 + f(z_i) + \tau D_i + \epsilon_i.$$

Interpretación: $f(z_i)$ controla la tendencia en z_i; el coeficiente τ captura el **salto causal** en el umbral.

Roadmap

1. Regresión Discontinua (RD)
   - Motivación
   - Definición Formal
   - Estimación
   - Supuestos
2. Impact Assessment: TOP - El Campo Emprende
   - Contexto
   - Datos y Estrategía de Identificación
   - Resultados
3. Aplicación en R

Problema de identificación (comparación ingenua)

Supongamos el modelo:

$$y_i = \beta_0 + f(z_i) + \tau D_i + \epsilon_i, \qquad D_i = \mathbf{1}\{z_i \geq z_0\}$$

Por tanto, para un $\eta > 0$ pequeño:

1. $E[y_i \mid z_i = z_0 - \eta] = E[y_i^0 \mid z_i = z_0 - \eta] = \beta_0 + f(z_0 - \eta)$.
2. $E[y_i \mid z_i = z_0 + \eta] = E[y_i^1 \mid z_i = z_0 + \eta] = \beta_0 + f(z_0 + \eta) + \tau$.

Problema de identificación (comparación ingenua)

Supongamos el modelo:

$$y_i = \beta_0 + f(z_i) + \tau D_i + \epsilon_i, \qquad D_i = \mathbf{1}\{z_i \geq z_0\}$$

Por tanto, para un $\eta > 0$ pequeño:

1. $E[y_i \mid z_i = z_0 - \eta] = E[y_i^0 \mid z_i = z_0 - \eta] = \beta_0 + f(z_0 - \eta)$.
2. $E[y_i \mid z_i = z_0 + \eta] = E[y_i^1 \mid z_i = z_0 + \eta] = \beta_0 + f(z_0 + \eta) + \tau$.

Diferencia ingenua:

$$E[y_i \mid z_0{+}\eta] - E[y_i \mid z_0{-}\eta] = \tau + \underbrace{f(z_0{+}\eta) - f(z_0{-}\eta)}_{\textbf{Sesgo de selección}}$$

Problema de identificación (comparación ingenua)

Supongamos el modelo:

$$y_i = \beta_0 + f(z_i) + \tau D_i + \epsilon_i, \qquad D_i = \mathbf{1}\{z_i \geq z_0\}$$

Por tanto, para un $\eta > 0$ pequeño:

1. $E[y_i \mid z_i = z_0 - \eta] = E[y_i^0 \mid z_i = z_0 - \eta] = \beta_0 + f(z_0 - \eta)$.
2. $E[y_i \mid z_i = z_0 + \eta] = E[y_i^1 \mid z_i = z_0 + \eta] = \beta_0 + f(z_0 + \eta) + \tau$.

Diferencia ingenua:

$$E[y_i \mid z_0{+}\eta] - E[y_i \mid z_0{-}\eta] = \tau + \underbrace{f(z_0{+}\eta) - f(z_0{-}\eta)}_{\textbf{Sesgo de selección}}$$

$\Rightarrow$ Sin controlar $f(\cdot)$, la diferencia no identifica τ.

Robustez y Verificación de los Supuestos

1. **Ausencia de programas simultáneos:** En el entorno del umbral z_0, no deben existir otros programas P_j tales que

$$D_i^{(j)} = \mathbf{1}\{z_i \geq z_0\}, \quad j \neq 1,$$

2. **No manipulación de la variable de asignación:** La densidad de z_i debe ser continua en el umbral:

$$\lim_{z \uparrow z_0} f_Z(z) = \lim_{z \downarrow z_0} f_Z(z),$$

donde $f_Z(\cdot)$ es la función de densidad de z_i. Un agrupamiento anómalo en torno a z_0 implica manipulación.

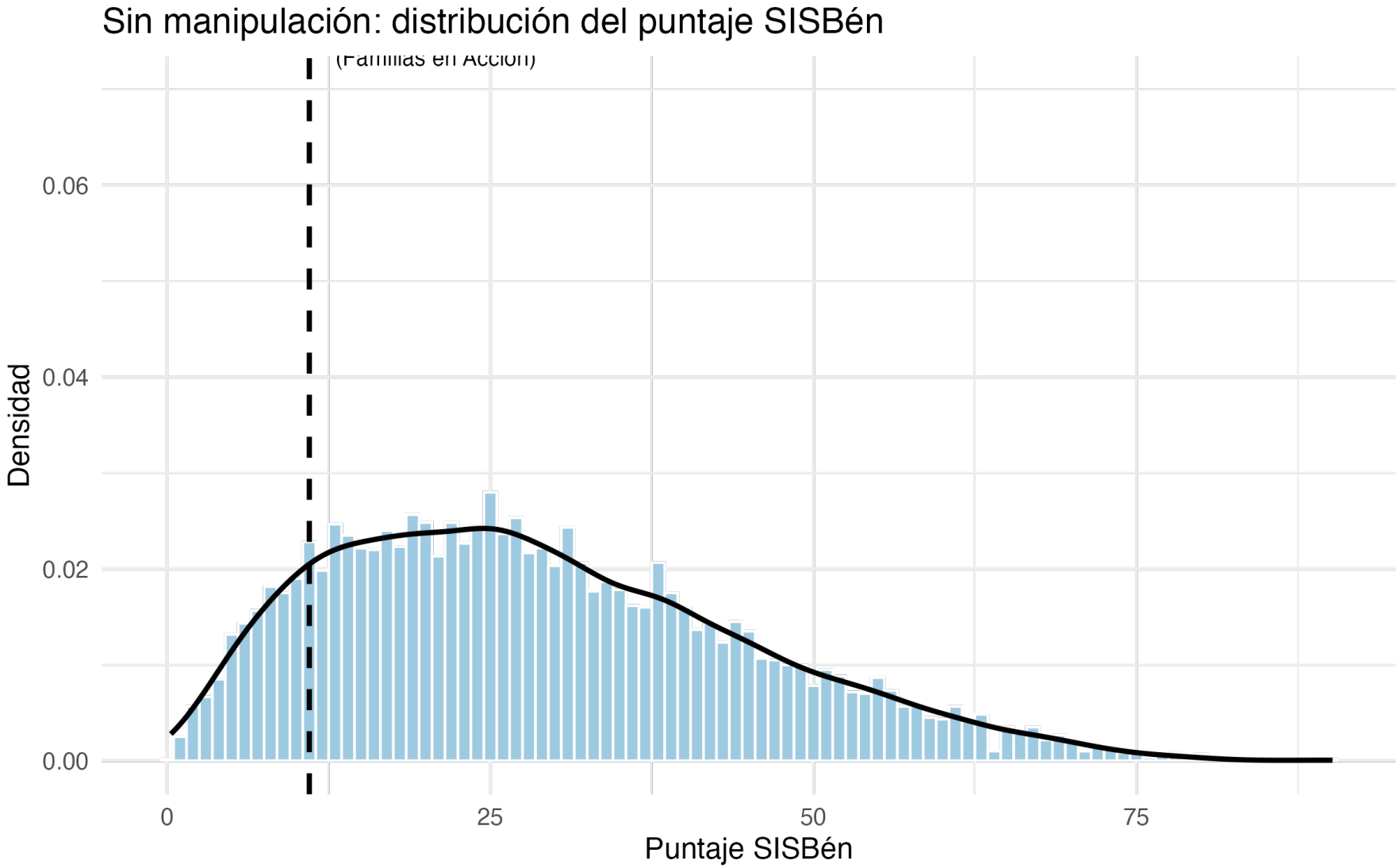
3. **Continuidad en covariables predeterminadas:** Para cualquier variable W_i no afectada por el tratamiento:

$$\lim_{z \uparrow z_0} E[W_i \mid z] = \lim_{z \downarrow z_0} E[W_i \mid z],$$

Pruebas típicas de robustez:

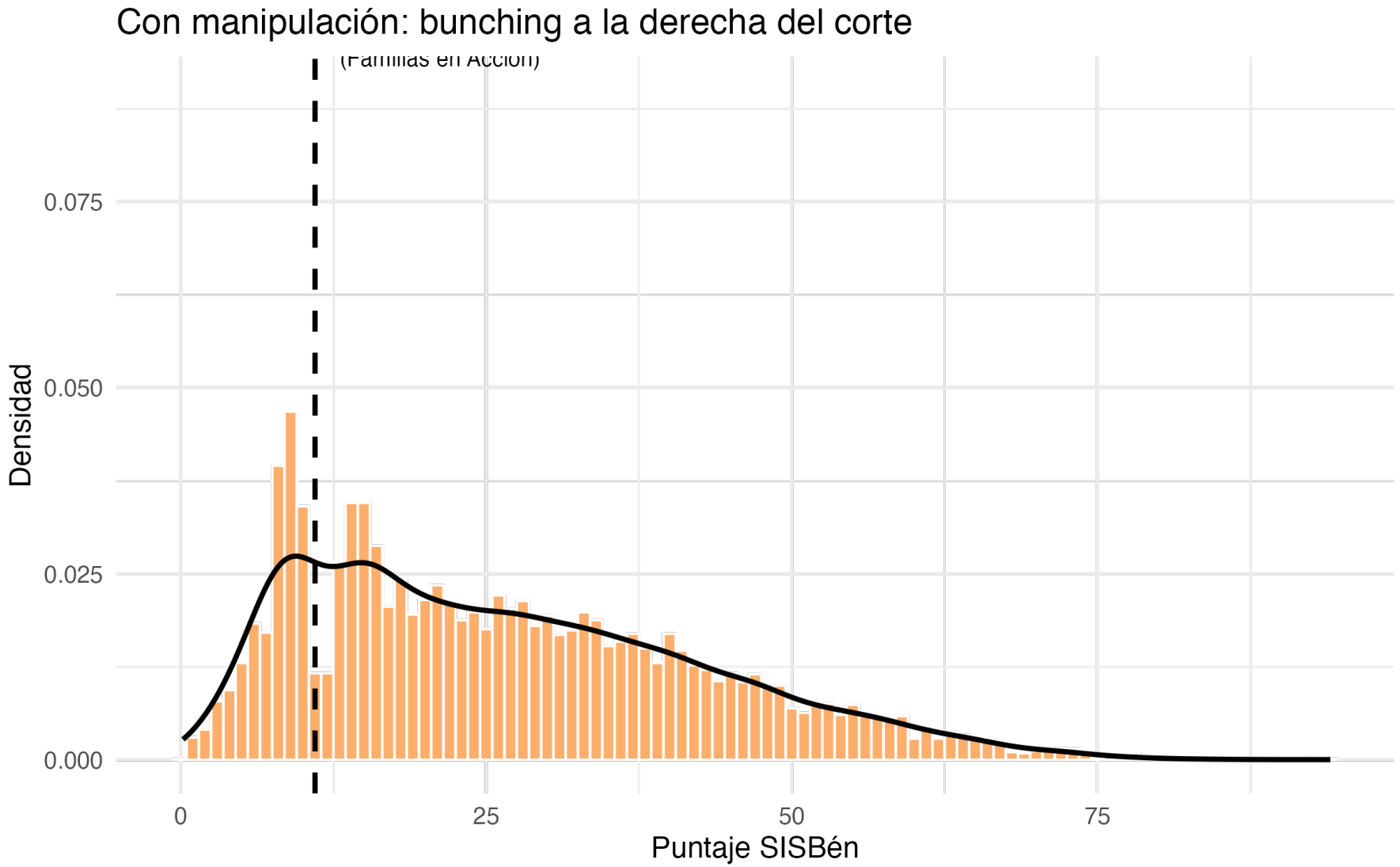
- **Robustez a cambios en el ancho de banda:** el estimador debe ser estable al variar el rango $[z_0 - h,\ z_0 + h]$.
- **Robustez a cambios en la forma funcional:** los resultados deben sostenerse frente a distintas especificaciones de $f(z)$ (lineal, polinómica, local).

Distribución del puntaje (sin manipulación)



Nota: Distribución continua en torno al corte del SISBén. No hay evidencia de manipulación (densidad suave).

Distribución del puntaje (con manipulación)



Nota: Se observa un **bunching** a la derecha del corte del SISBén, lo que sugiere manipulación de puntajes para acceder a Familias en Acción.

Continuidad en covariables predeterminadas

Un requisito clave de validez en RD es que las **covariables predeterminadas** (no afectadas por el tratamiento) sean continuas en torno al umbral z_0.

Supuesto formal: Para cualquier covariable W_i predeterminada,

$$\lim_{z \uparrow z_0} E[W_i \mid z] = \lim_{z \downarrow z_0} E[W_i \mid z].$$

Esto asegura que las unidades justo por encima y por debajo de z_0 son comparables en características observables.

Procedimiento empírico:

- Estimar el modelo de RD reemplazando la variable dependiente Y_i por cada W_i predeterminada (ej. edad, género, educación de los padres).
- Cuando esté disponible, incluir el valor rezagado Y_{t-1} para comprobar que no haya saltos previos al tratamiento.

Resultado esperado: No debe detectarse un salto estadísticamente significativo en estas covariables, confirmando que la discontinuidad observada en Y_i proviene únicamente del tratamiento.

Roadmap

1. Regresión Discontinua (RD)
   - Motivación
   - Definición Formal
   - Estimación
   - Supuestos

2. Impact Assessment: TOP - El Campo Emprende
   - Contexto
   - Datos y Estrategía de Identificación
   - Resultados

3. Aplicación en R

Programa TOP – El Campo Emprende

- Iniciativa del Ministerio de Agricultura con apoyo de IFAD (2012–2023).
- Objetivo: mejorar ingresos y condiciones de vida en zonas rurales vulnerables y de posconflicto.
- Estrategia: apoyar a Organizaciones Productivas (POs) mediante Planes de Negocio.
- Criterios de elegibilidad: número de mujeres, jóvenes, condiciones de vulnerabilidad, exposición a violencia.
- Promedio del apoyo: COP 40 millones por grupo (USD 13,530 en 2018).

Diseño Muestral (Sección 3.1)

- Convocatorias: 2018 y 2019 usadas para el estudio (96 municipios).
- **Muestra:** 354 organizaciones en 42 municipios.
  - 177 tratadas y 177 de control.
  - 2,460 hogares encuestados + 354 cuestionarios a líderes de PO.
- Estrategia de identificación: **RDD en el umbral de elegibilidad**.
- Selección de hogares: aleatoria dentro de cada PO.
- Recolección: encuestas junio–julio 2023.

Metodología de Estimación

- Se aplicó un **Regression Discontinuity Design (RDD)**
- **Lógica:**
  - PO elegibles $\Rightarrow$ mayor probabilidad de ser tratadas.
- Estimaciones:

$$y_i = \beta_0 + \tau D_i + f(z_i) + \varepsilon_i$$

  donde D_i es la elegibilidad.

- Supuesto clave: continuidad local de potenciales resultados en torno al umbral.

Resultados: Movilidad Económica (Sección 5.1.1)

- Incremento de **34% en ingreso bruto per cápita** (C̃OP 1.6M; USD 387).
- Incremento de **48% en ingreso neto per cápita**.
- **46% más ingresos por actividades empresariales**.
- Efecto positivo en empleo y salarios (aumento de 35% en ingresos laborales).
- Activos del hogar: $+0.26$ desviaciones estándar en el índice de activos.
- Testimonios cualitativos: mayor resiliencia durante la pandemia, creación de nuevos mercados.

Síntesis de Evidencia

- TOP fortaleció la organización productiva y democratizó el acceso a activos.
- Los impactos son significativos en ingresos, empleo y acumulación de activos.
- El diseño de identificación RDD asegura alta validez interna, aunque se limita a efectos locales (LATE).
- Evidencia cualitativa confirma mejoras en capital humano y social (empoderamiento, cohesión).
- Implicación: El Campo Emprende constituye un **caso exitoso de política rural con base empírica rigurosa**.

Roadmap

1. Regresión Discontinua (RD)
   - Motivación
   - Definición Formal
   - Estimación
   - Supuestos
2. Impact Assessment: TOP - El Campo Emprende
   - Contexto
   - Datos y Estrategía de Identificación
   - Resultados
3. Aplicación en R

Replication Package

Repositorio en R (GitHub): Descarga directa del paquete de replicación (código y datos en R): `replication_package.zip`

Instrucciones: Antes de salir al break:

- Descarguen y descompriman el paquete.
- Ejecuten el script inicial para cargar librerías:

Script en R:

- *require(pacman)*
- *p_load(tidyverse, rio, fixest, broom)*

Cuando regresemos de la pausa, las librerías ya estarán cargadas y listas para usar.