Diseños de regresión discontinua

Diseño e implementación de experimentos en ciencias sociales

Departamento de Economía (UdelaR)

Regresión Discontinua: motivación

Motivación

Objetivo: efecto causal de un programa de becas sobre rendimiento académico de los estudiantes

Mecanismo de asignación:

- Postulantes toman un examen, reciben un puntaje X
- Las becas son asignadas a los que obtienen $X >= c$

Diferencias en rendimiento académico entre estudiantes por encima y por debajo de c

1. Efecto causal de la beca
2. *Confounders*: IQ, horas de estudio, motivación, características del hogar, etc.,...

Motivación

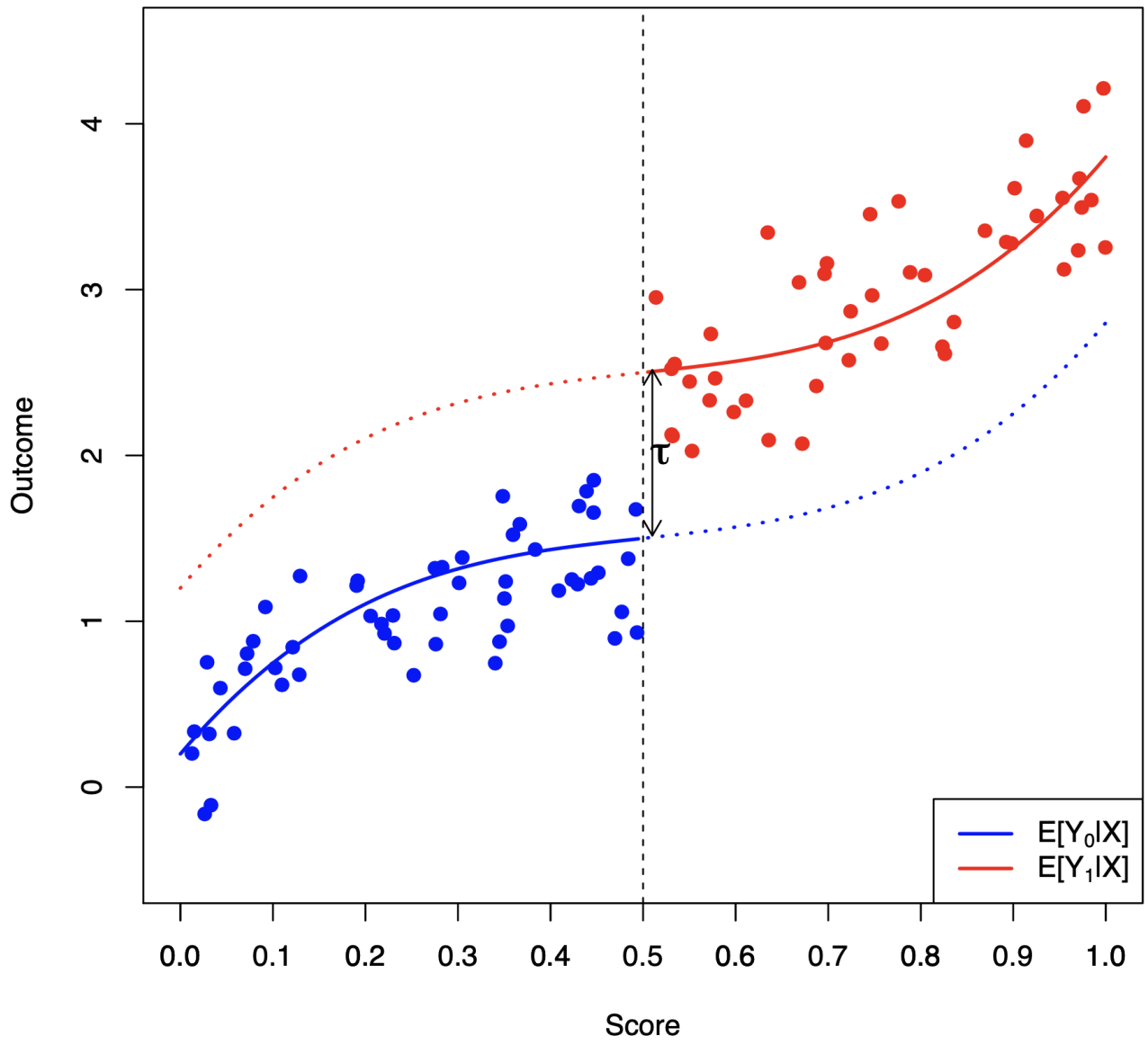
- Muchos de factores de auto-selección (*counfounding*) son no-observables
- Problema fundamental de superposición entre tratados y controles
- DRD explotan el mecanismo discontinuo de asignación
- Suponemos:
  1. Estudiantes no tienen control exacto sobre su puntaje
  2. Los inobservables no "saltan" en el punto de corte (*cutoff*)

Notación y estructura

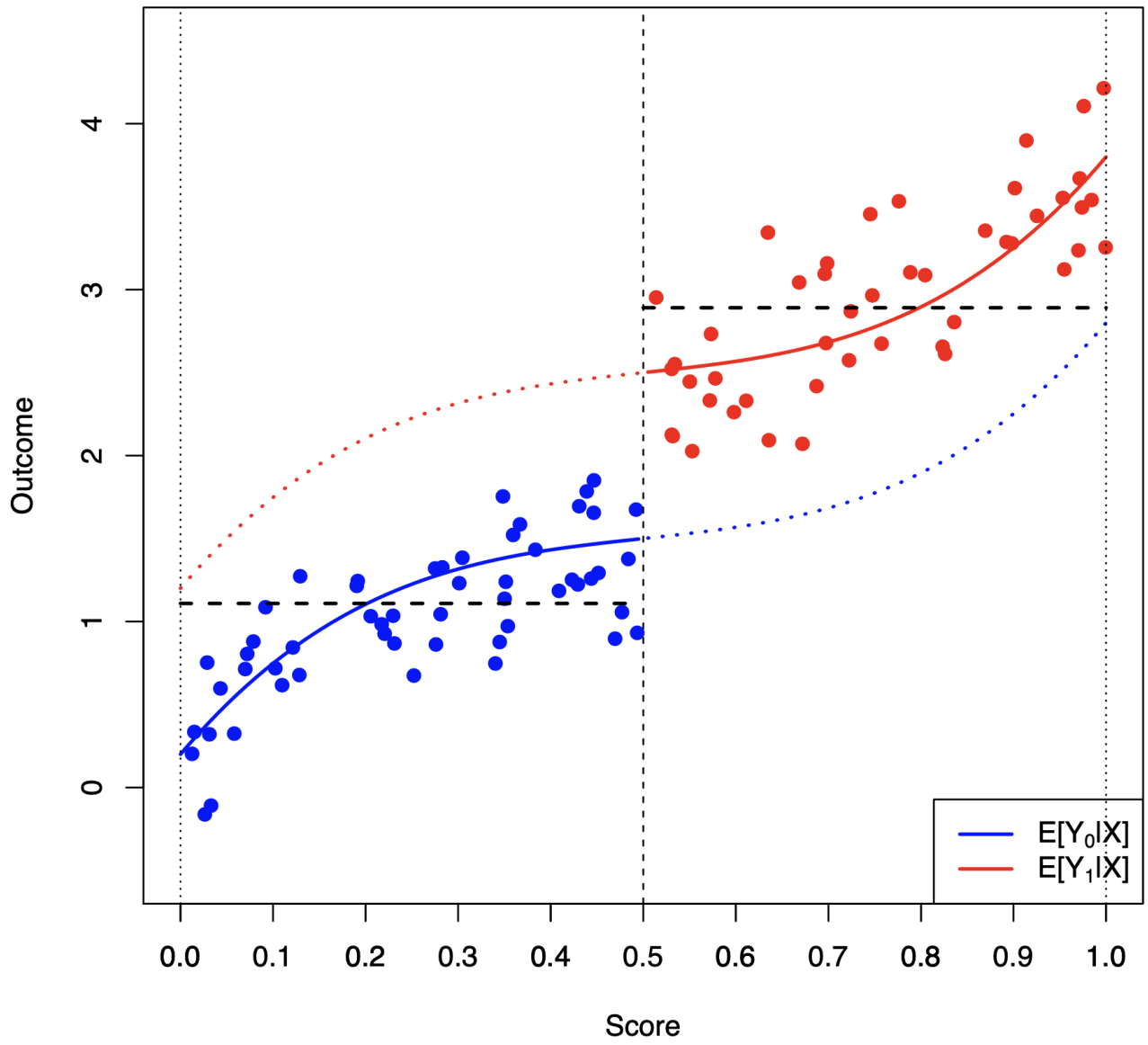
DRD notación y estructura

- Resultados potenciales: (Y_{i1}, Y_{i0}), con $\tau_i = Y_{i1} - Y_{i0}$
- Variable continua X (puntaje)
- Indicador de tratamiento: $T_i = T_i(X_i) = 1$ si es tratado, 0 de lo contrario
- Resultado observado: $Y_i = Y_{i1}T_i + Y_{i0}(1 - T_i)$
- RD explota una discontinuidad en $P[T_i = 1|X_i]$ en algún punto de corte c
- Diseño nítido: $P[T_i = 1|X_i] = 1(X_i >= c)$

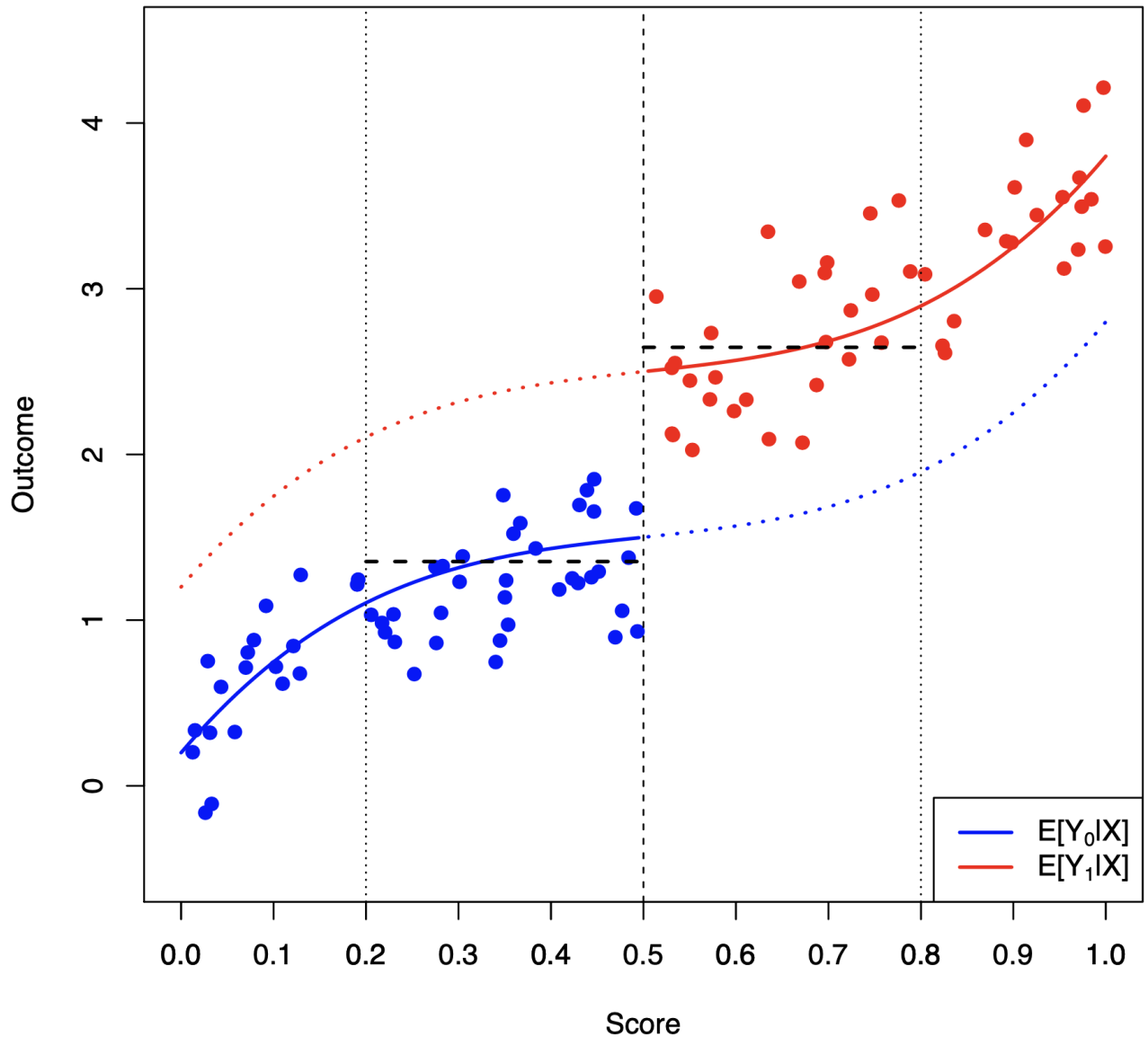
DRD intuición



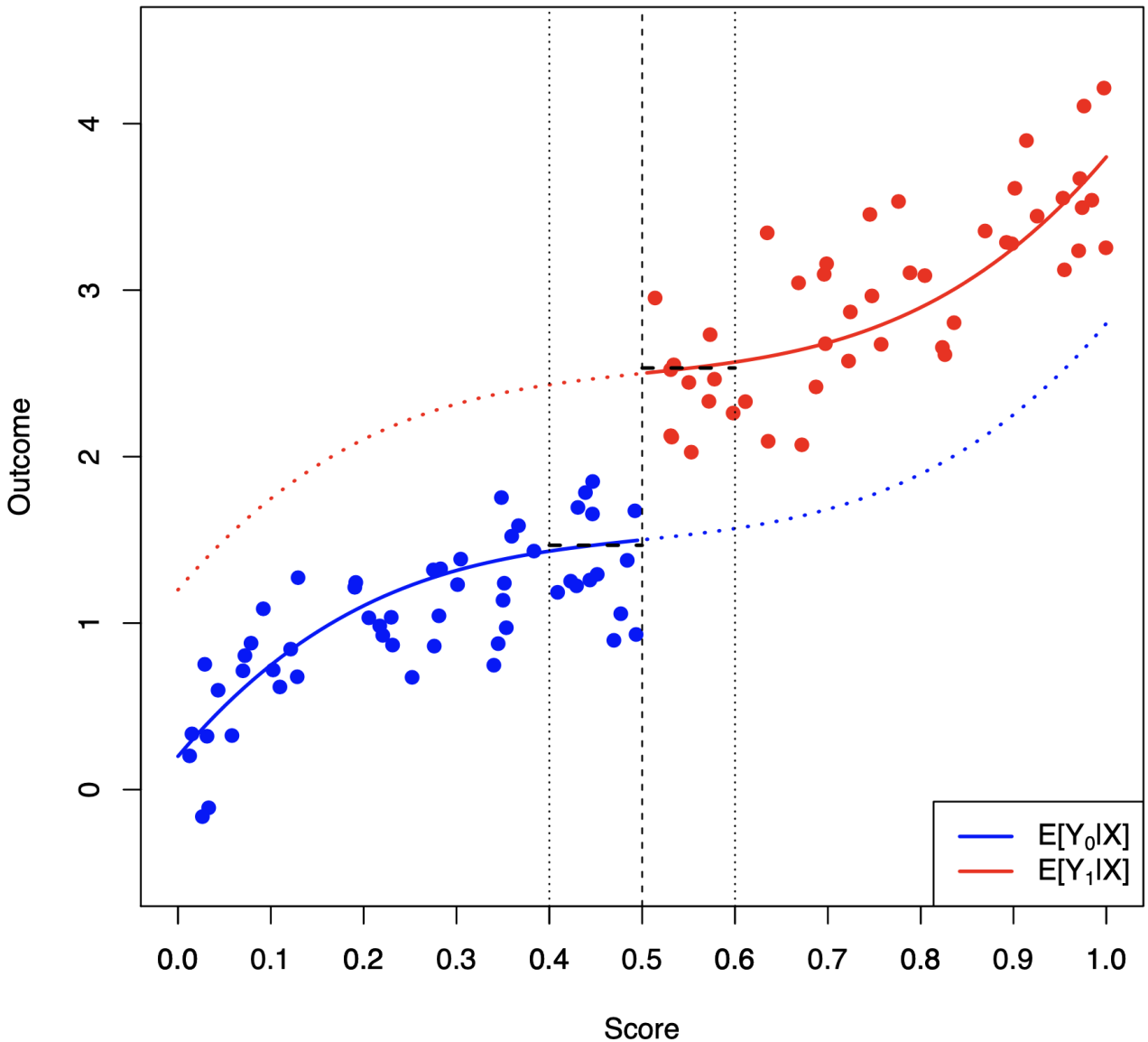
DRD intuición



DRD intuición



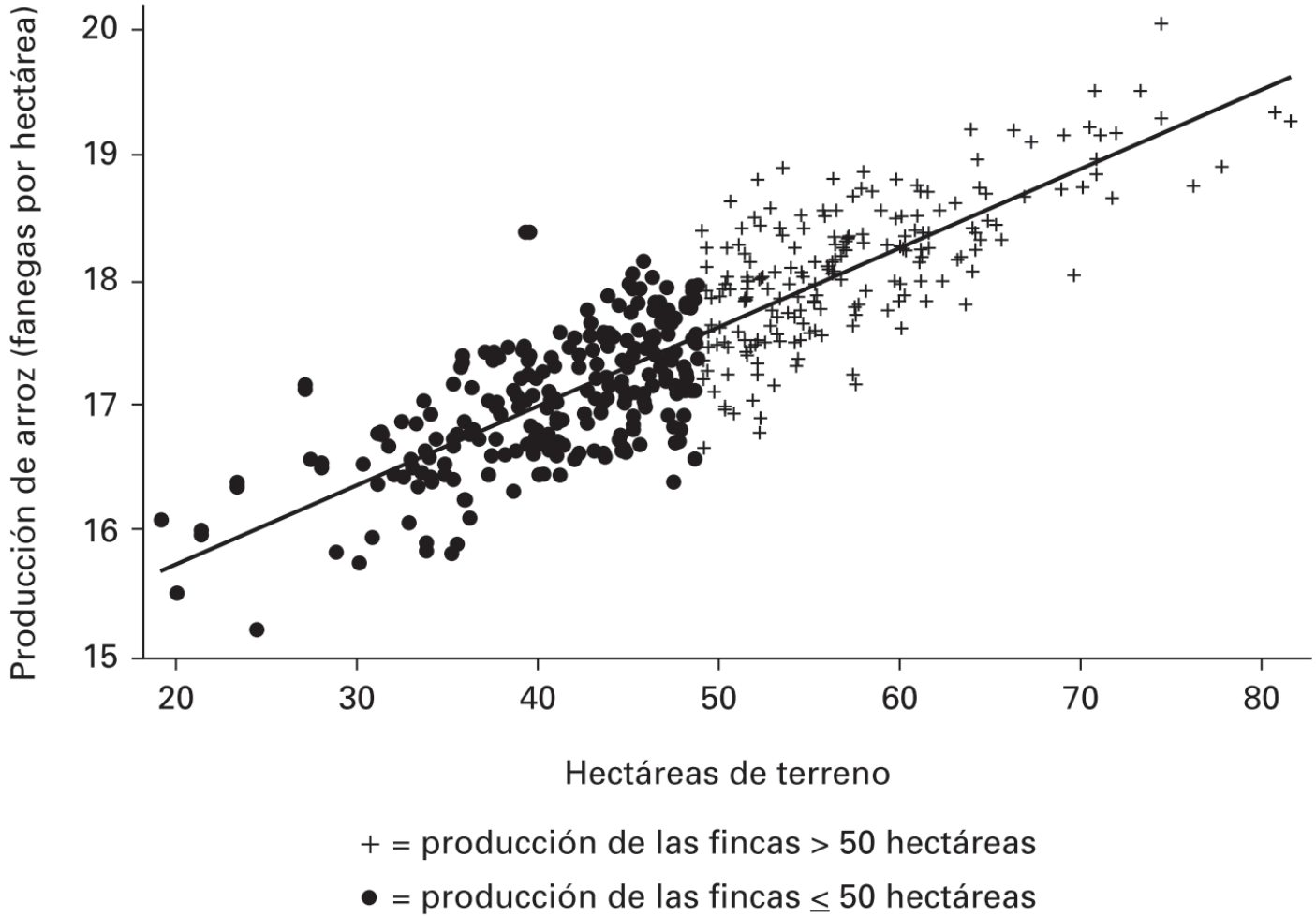
DRD intuición



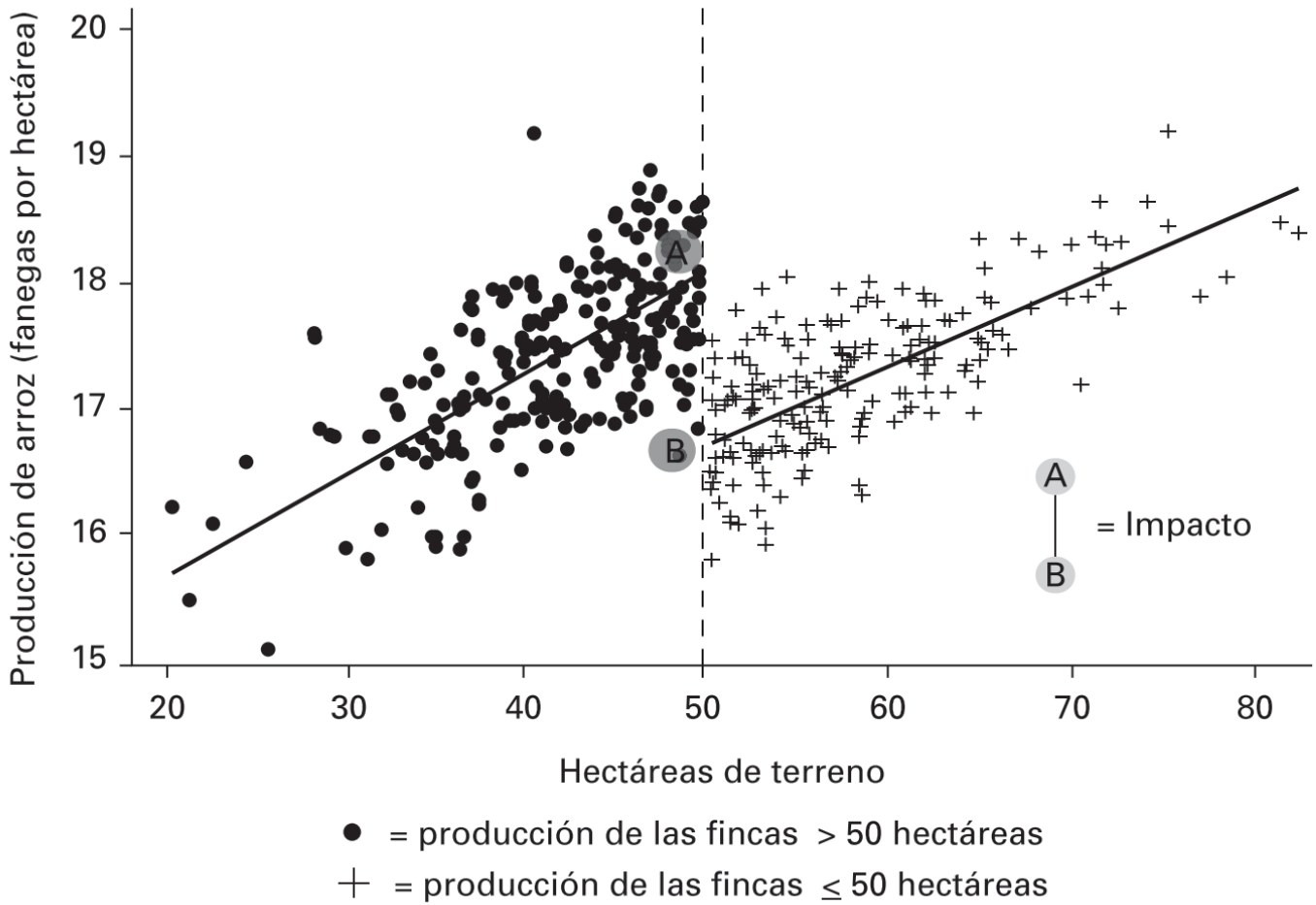
DRD intuición

- Efecto del tratamiento solo se identifica en el punto (*no paramétrico*)
  - Único punto de superposición (en el límite)
  - Interpretación local de la RD
- Estimación en puntos lmítrofes
  - En realidad tenemos cero observaciones en $X_i = c$
  - Tenemos que basarnos en extrapolación
  - El modelo generalmente está mal especificado

DRD ejemplo


Producción de arroz (fanegas por hectárea)
20
19
18
17
16
15
20
30
40
50
60
70
80
Hectáreas de terreno
+ = producción de las fincas > 50 hectáreas
• = producción de las fincas ≤ 50 hectáreas

DRD ejemplo


Producción de arroz (fanegas por hectárea)
20
19
18
17
16
15
20
30
40
50
60
70
80
Hectáreas de terreno
A
B
A
B
= Impacto
● = producción de las fincas > 50 hectáreas
+ = producción de las fincas ≤ 50 hectáreas

Identificación

Identificación

Identificación no paramétrica en RD:

Suponemos:

1. diseño "nítido" (*sharp*): $T_i = 1(X_i >= c)$
2. (suavidad): $E[Y_{i0}|X_i = x], E[Y_{i1}|X_i = x]$ es continuo en $x = c$

Por tanto,

$$E[\tau_i|X_i = c] = \lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]$$

Identificación

Diferencia de medias (*naive*) tiene el mismo sesgo de selección en la estimación del efecto que vimos en clases anteriores cuando hay *confounders*.

$$\begin{aligned}\Delta(h) &= E\{Y_i|X_i \in [c, c+h]\} - E\{Y_i|X_i \in [c-h, h]\} \\ &= E\{Y_{(1)}|X_i \in [c, c+h]\} - E\{Y_{(0)}|X_i \in [c-h, h]\} \\ &= E\{\tau|X_i \in [c, c+h]\} + Bias(h)\end{aligned}$$

donde

$$Bias(h) = E\{Y_{(0)}|X_i \in [c, c+h]\} - E\{Y_{(0)}|X_i \in [c-h, h]\}$$

- Si ocurre que $E[Y_{i(t)}|X_i = x]$ es continuo en x=c para $t = 0, 1$,

$$\lim_{h \downarrow 0} \delta(h) = E[\ \tau_i|X_i = c]$$

Estimación e inferencia

Estimación

- Estimar funciones de regresión a la izquierda y a la derecha del punto de corte

Estimación

- Estimar funciones de regresión a la izquierda y a la derecha del punto de corte
- Global:
  - Estimar un polinomio de orden p en toda la muestra
  - Sensible a errores de especificación
  - Comportamiento errático en los puntos de corte

Estimación

- Estimar funciones de regresión a la izquierda y a la derecha del punto de corte
- Global:
  - Estimar un polinomio de orden p en toda la muestra
  - Sensible a errores de especificación
  - Comportamiento errático en los puntos de corte
- "Paramétrico flexible":
  - Estimar un polinomio dentro de un ancho de banda *ad-hoc*

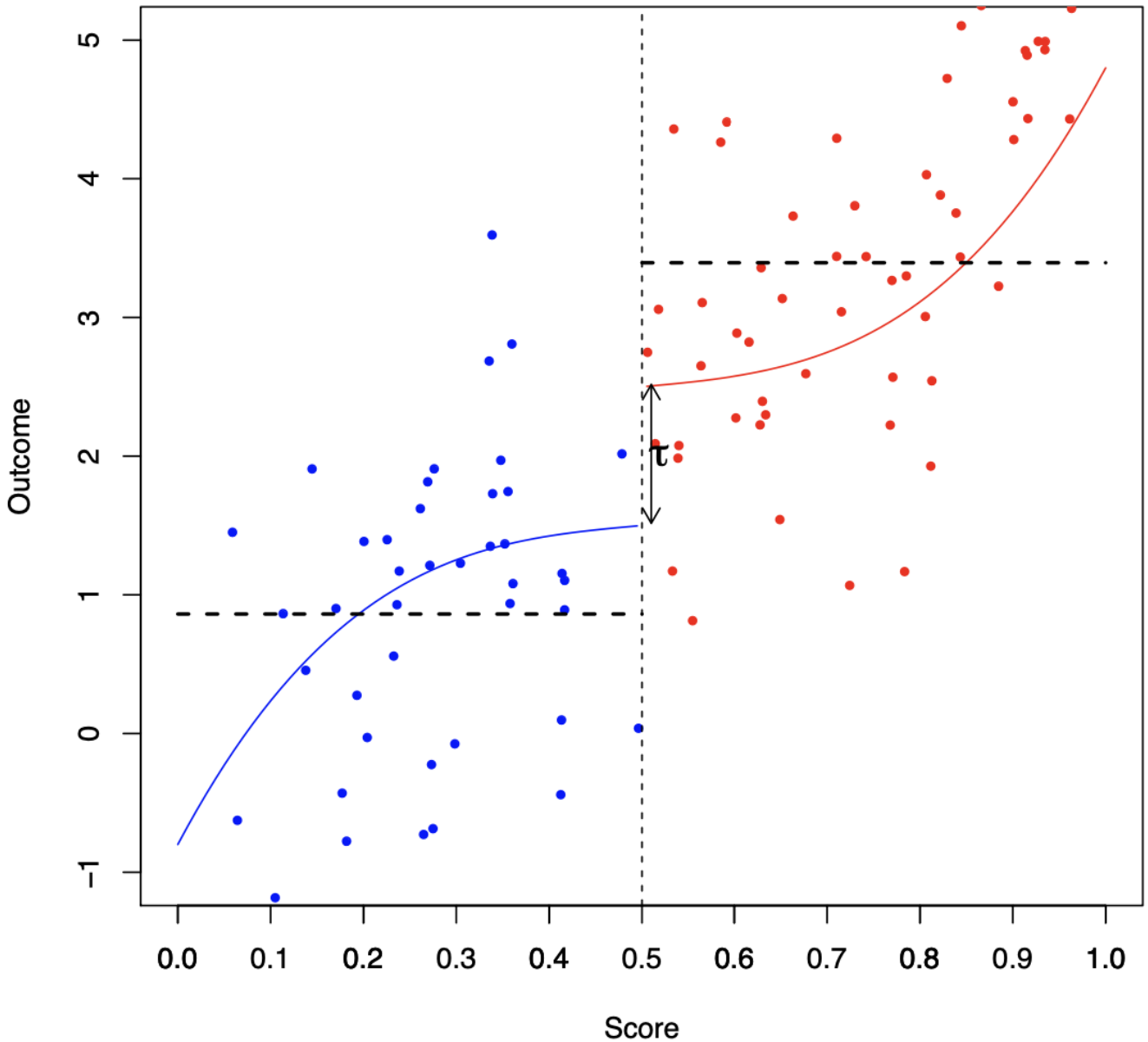
Estimación

- Estimar funciones de regresión a la izquierda y a la derecha del punto de corte
- Global:
  - Estimar un polinomio de orden p en toda la muestra
  - Sensible a errores de especificación
  - Comportamiento errático en los puntos de corte
- "Paramétrico flexible":
  - Estimar un polinomio dentro de un ancho de banda *ad-hoc*
- Polinomio local:
  - Selección de ancho de banda basada en datos, no paramétrica
  - Tiene en cuenta problemas de especificación al realizar inferencias

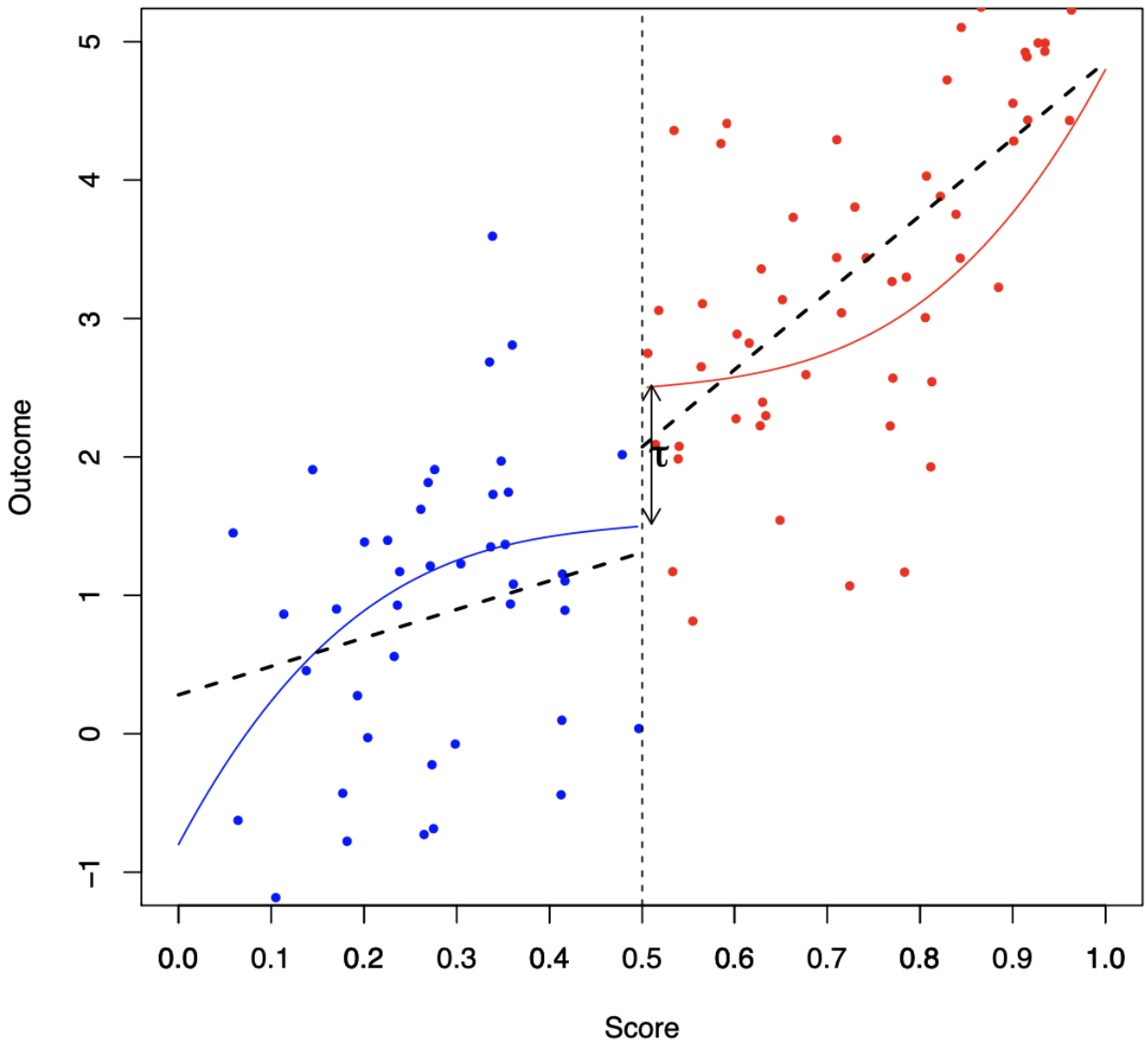
Estimación global

$$E[Y_{i(t)}|X_i] = \alpha_t + \beta_t(X_i - c)$$

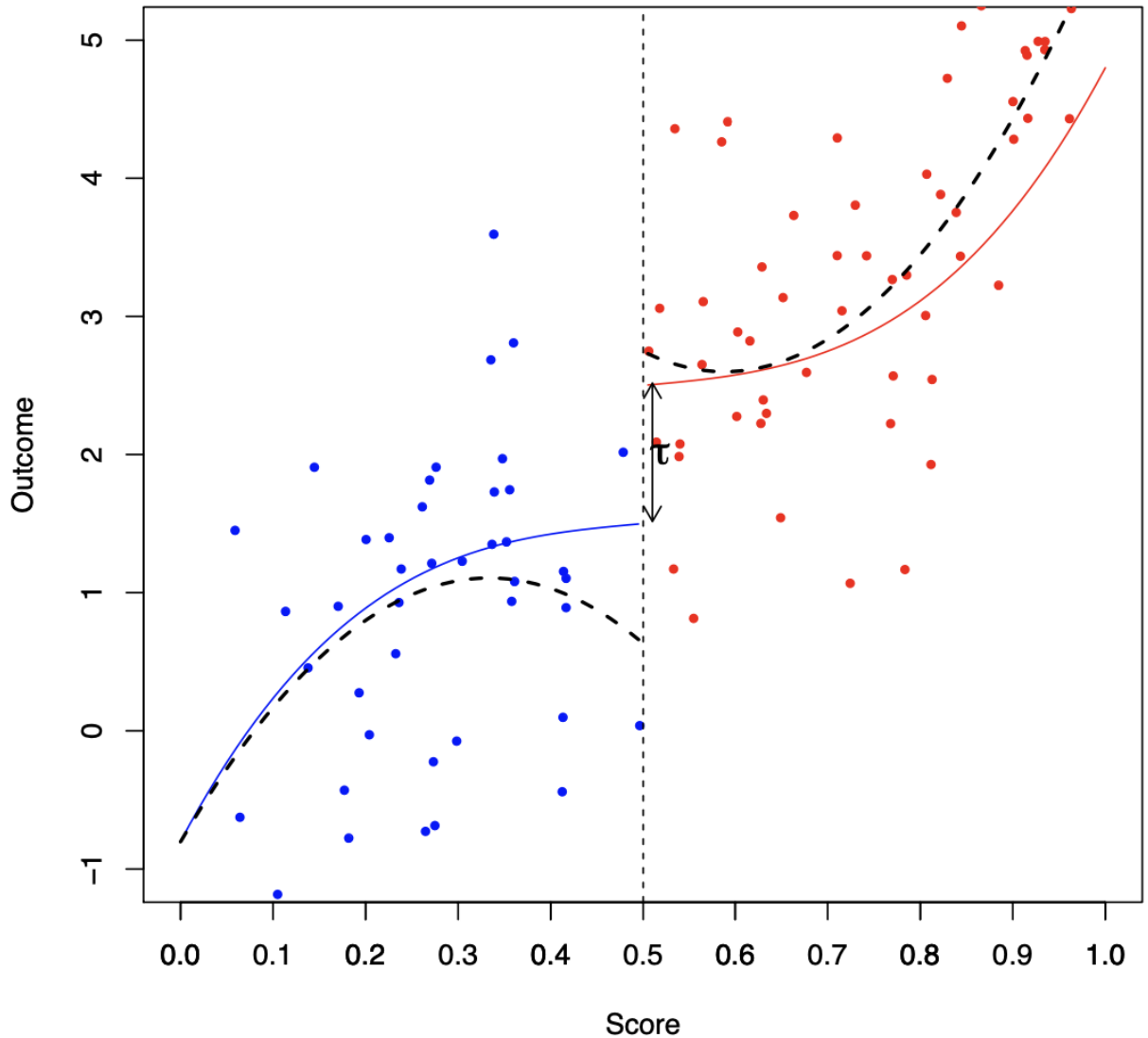
Estimación global (orden 0)



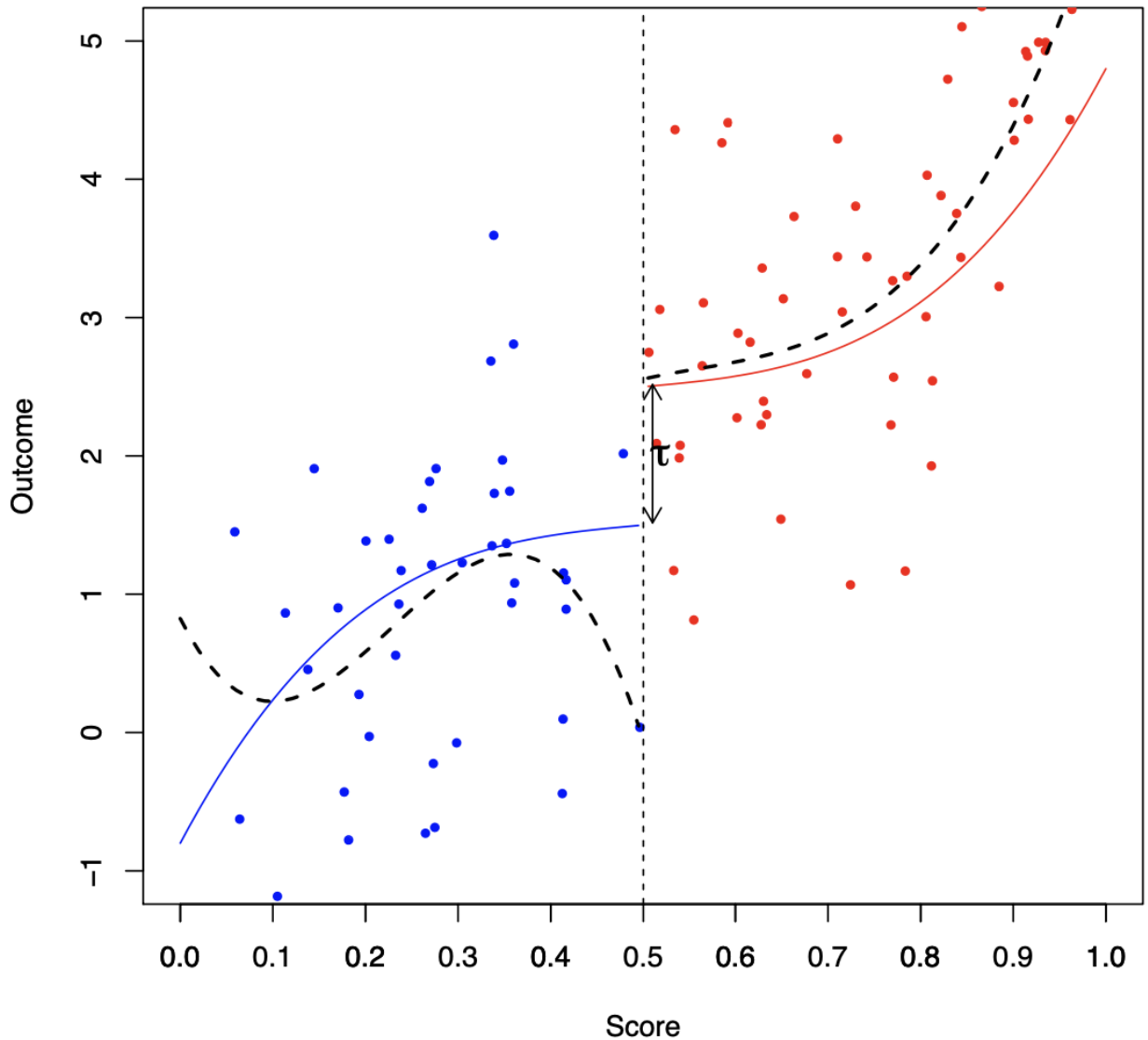
Estimación global (orden 1)



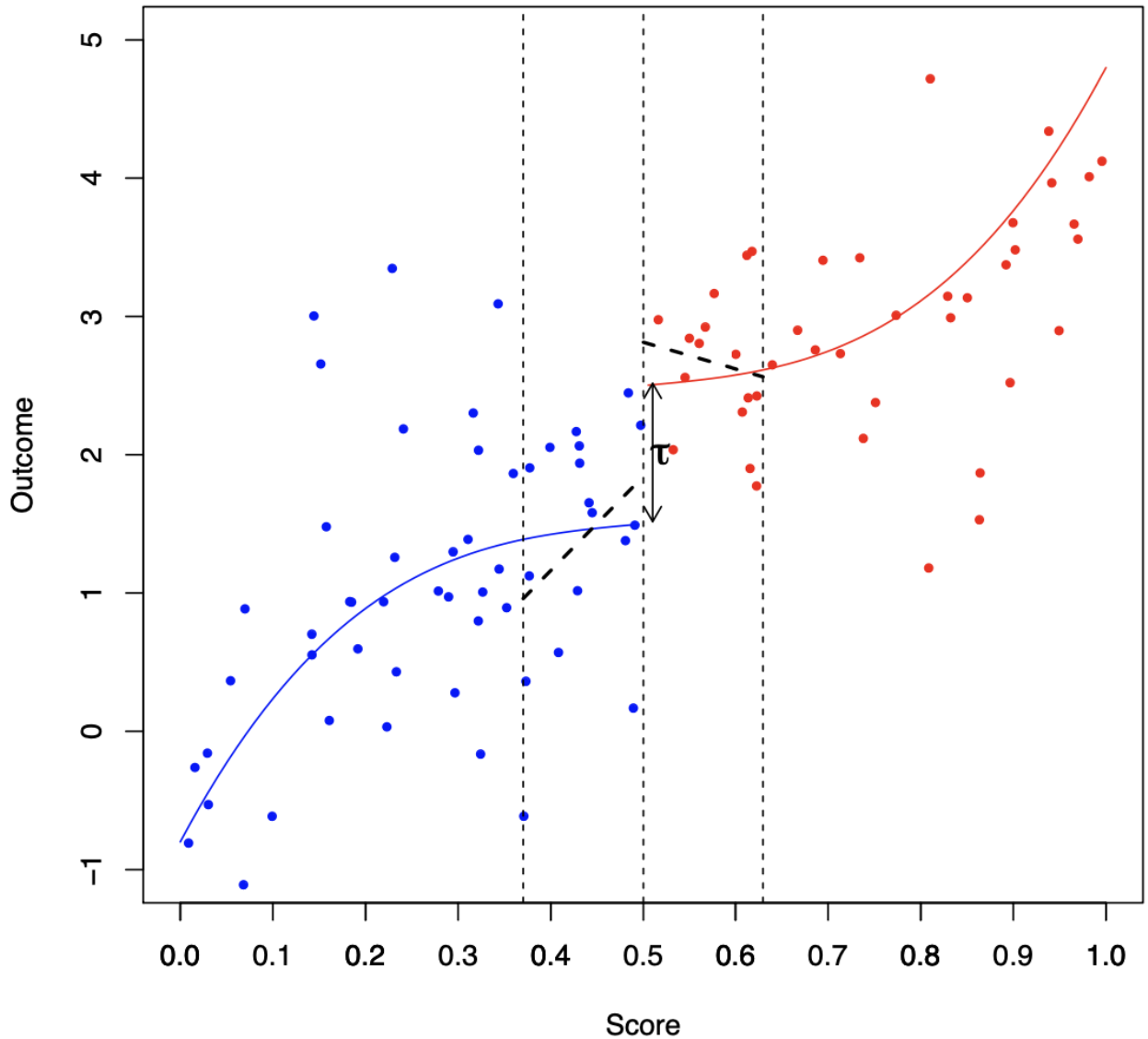
Estimación global (orden 2)



Estimación global (orden 3)



Estimación local



Estimación local

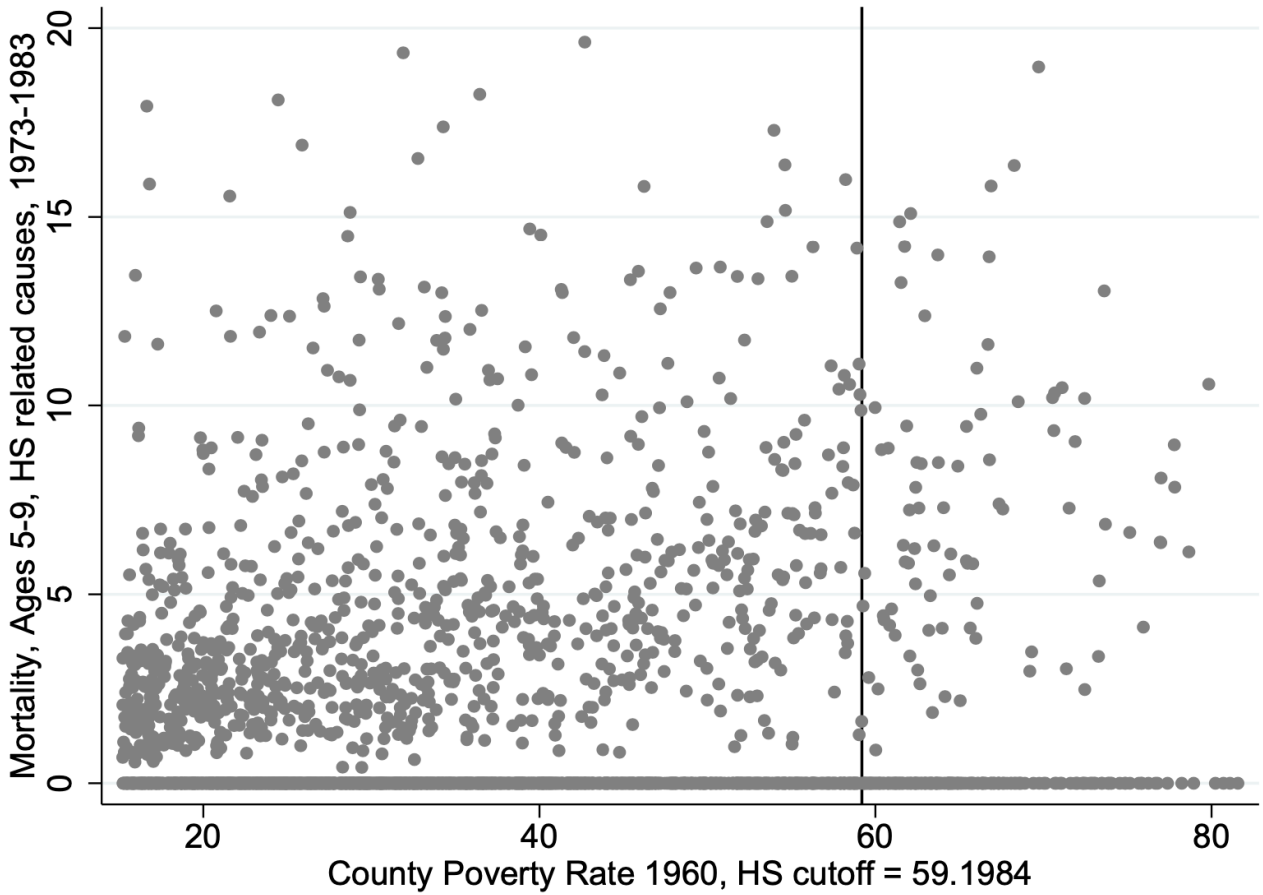
- Estimación no paramétrica debe dar cuenta del sesgo
- Si $h = 0$, el estimador es insesgado
- Pero $h = 0$ implica que no hay observaciones dentro de la ventana
- Menor h implica un pequeño sesgo, pero menos observaciones (mayor varianza)
- *Optimal bandwith*

Ejemplos

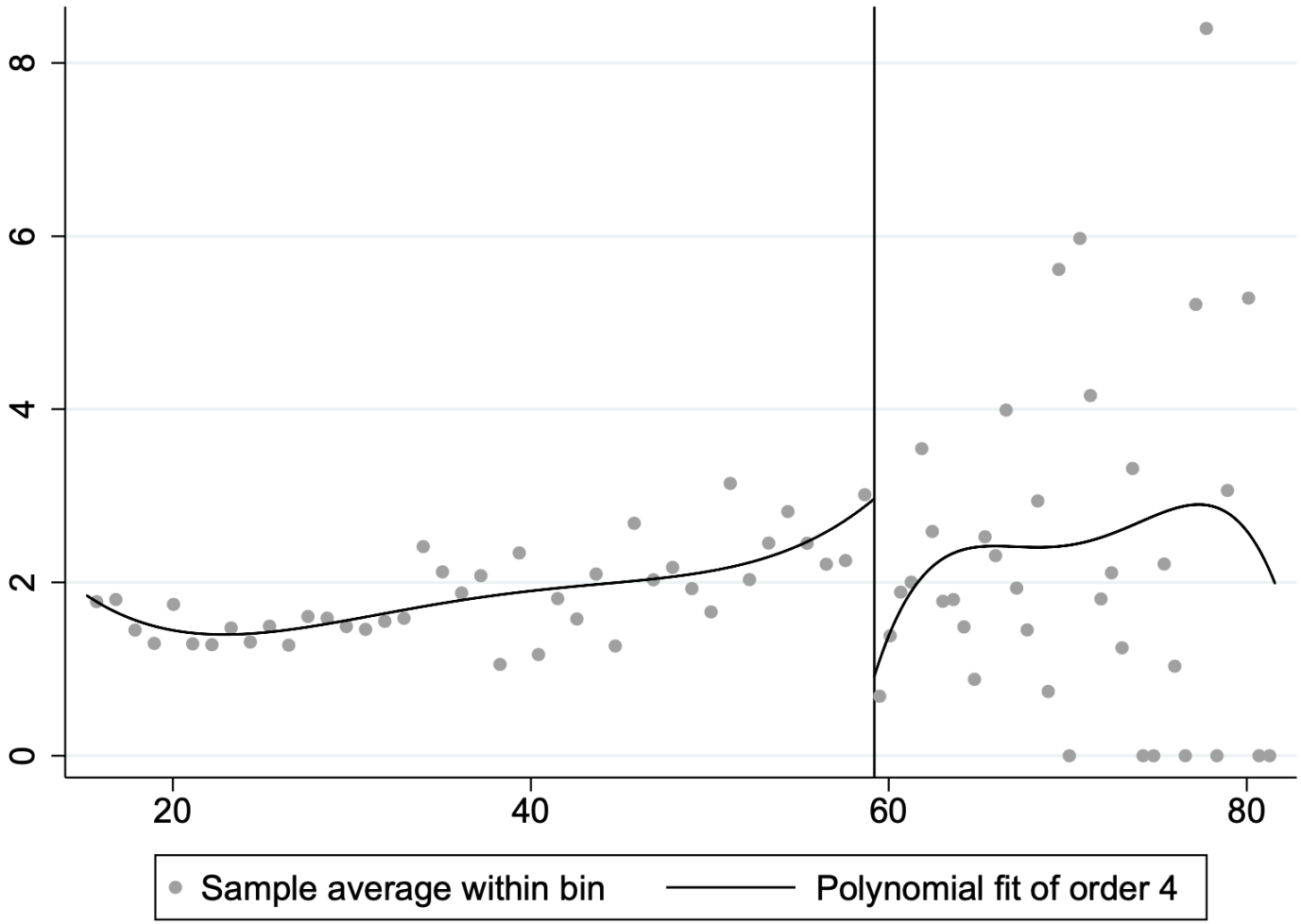
Ejemplo gráfico

Programa nutricional para niños de 5-9 años.

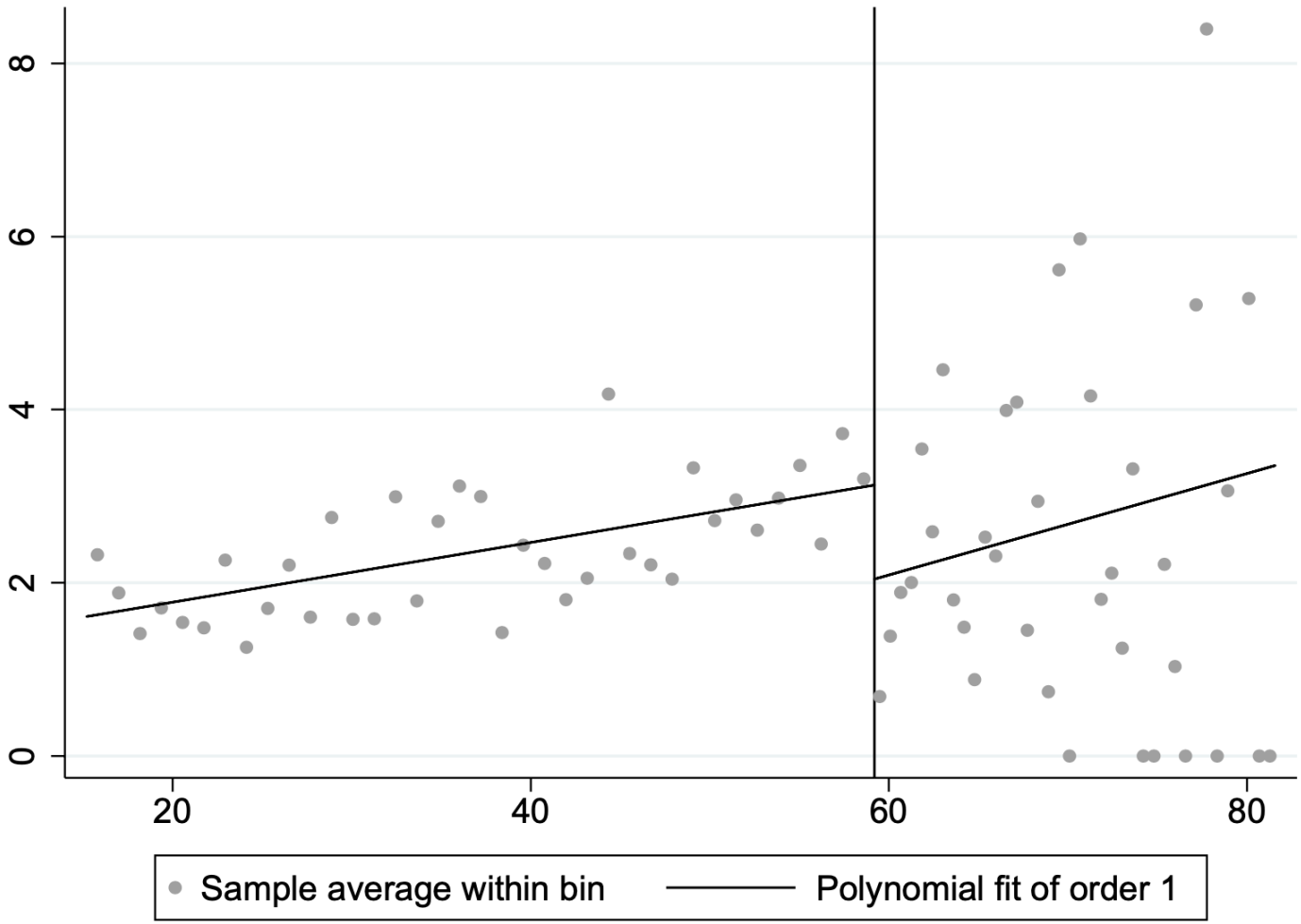
Ejemplo gráfico



Ejemplo gráfico

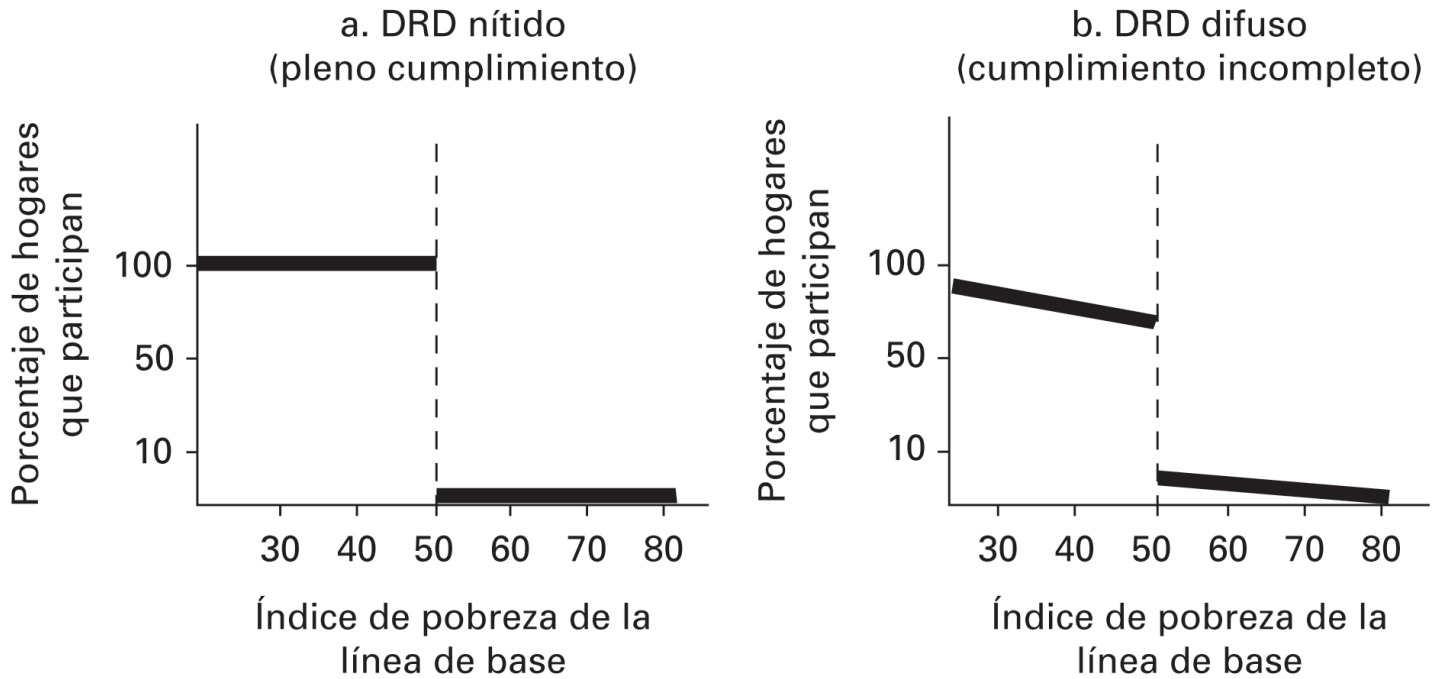


Ejemplo gráfico



Diseño difuso

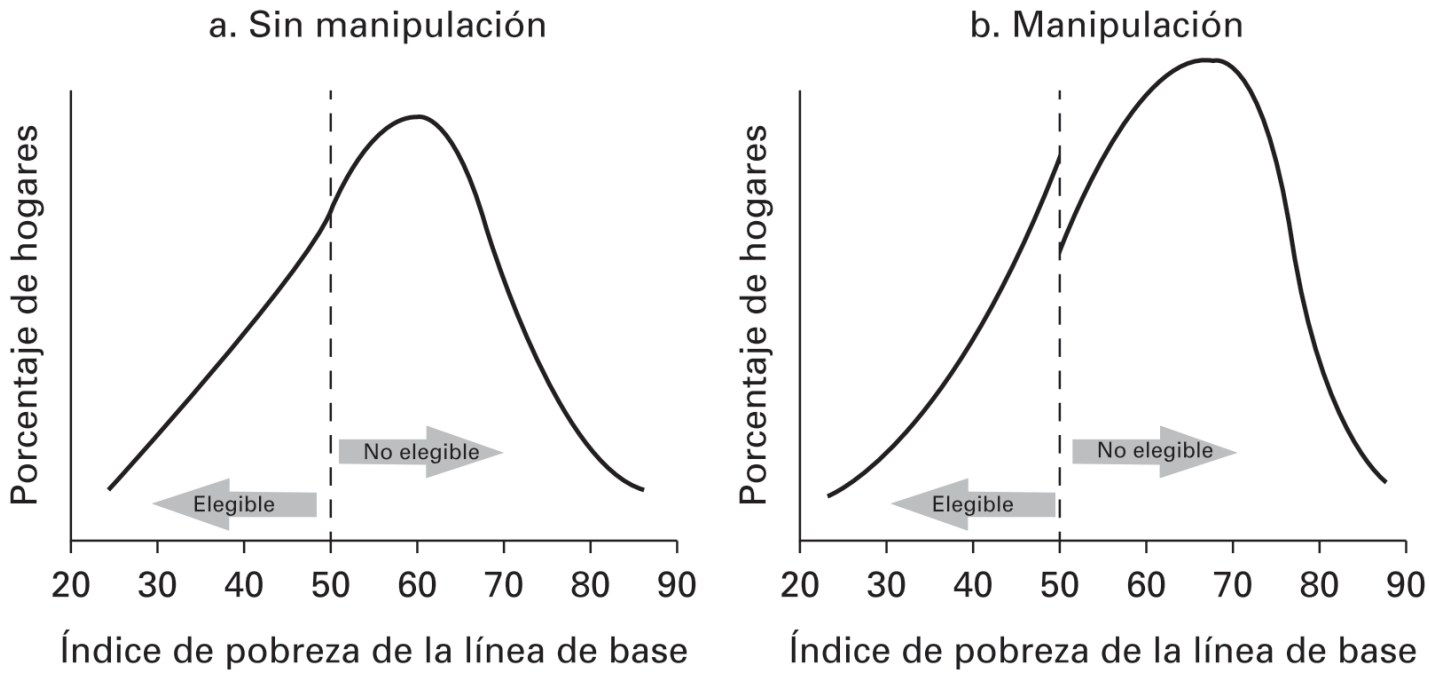
(Fuzzy) DRD


a. DRD nítido
(pleno cumplimiento)
Porcentaje de hogares que participan
100
50
10
30
40
50
60
70
80
Índice de pobreza de la línea de base
b. DRD difuso
(cumplimiento incompleto)
Porcentaje de hogares que participan
100
50
10
30
40
50
60
70
80
Índice de pobreza de la línea de base

Chequeos y validez

Chequeos y validez

Es importante que el índice de elegibilidad no sea manipulado en la cercanía de la puntuación límite



Lista de verificación

- ¿Es continuo el índice en torno la puntuación límite en el momento de la línea de base?

[1] Se utiliza la localización a la izquierda o la derecha del punto de corte como variable instrumental para la aceptación del programa en la primera etapa de una estimación de mínimos cuadrados en dos etapas.

Lista de verificación

- ¿Es continuo el índice en torno la puntuación límite en el momento de la línea de base?
- ¿Hay alguna evidencia de falta de cumplimiento de la regla que determine la elegibilidad para el tratamiento?
  - Compruébese que todas las unidades elegibles y ninguna unidad no elegible han recibido el tratamiento.
  - Si se encuentra falta de cumplimiento, habrá que combinar el DRD con un enfoque de variable instrumental para corregir esta *discontinuidad difusa*.[1]

[1] Se utiliza la localización a la izquierda o la derecha del punto de corte como variable instrumental para la aceptación del programa en la primera etapa de una estimación de mínimos cuadrados en dos etapas.

Lista de verificación

- ¿Hay alguna evidencia de que las puntuaciones del índice puedan haber sido manipuladas con el fin de influir en quien tenía derecho a beneficiarse del programa?
  - Compruébese si la distribución de la puntuación del índice es fluida en el punto límite.
  - Si se halla evidencia de una *concentración* de puntuaciones ya sea por encima o por debajo del punto límite, puede que esto sea una señal de manipulación.

Lista de verificación

- ¿Hay alguna evidencia de que las puntuaciones del índice puedan haber sido manipuladas con el fin de influir en quien tenía derecho a beneficiarse del programa?
  - Compruébese si la distribución de la puntuación del índice es fluida en el punto límite.
  - Si se halla evidencia de una *concentración* de puntuaciones ya sea por encima o por debajo del punto límite, puede que esto sea una señal de manipulación.
- ¿El umbral corresponde a un único programa que se está evaluando o está siendo usado por otros programas también?

Análisis estadístico de DRD

R

`rdrobust`: graphical presentation and local polynomial methods

`rddensity`: density discontinuity tests (manipulation testing)

`rdlocrand`: local randomization methods

`rdmulti`: analysis of RD with multiple cutoffs or scores

`rdpower`: power and sample size calculation