Causal Inference with Regression Discontinuity Design

Lê Việt Phú
Fulbright School of Public Policy and Management

Ngày 23 tháng 7 năm 2020

Thử nghiệm tự nhiên/bán thử nghiệm (natural/quasi experiment)

Thử nghiệm tự nhiên là một tình huống khi một can thiệp chính sách hoặc một sự kiện xảy ra mà ở đó có sự phân định ngẫu nhiên nhóm đối chứng và nhóm hưởng lợi, mặc dù không đảm bảo tất cả các thuộc tính của hai nhóm hoàn toàn tương đồng.

- Với RCT thì nhóm nghiên cứu ngẫu nhiên hóa đối tượng hưởng lợi và đối chứng.
- Với NE thì nhóm nghiên cứu không kiểm soát được quá trình can thiệp.

Thử nghiệm ngẫu nhiên vs thử nghiệm tự nhiên

- Đánh giá thử nghiệm ngẫu nhiên có kiểm soát RCT đảm bảo hai điều kiện:
  1. Các nhóm tương đồng về các điều kiện quan sát được và không quan sát được trước khi can thiệp xảy ra (exchangeability condition).
  2. Quá trình can thiệp là ngẫu nhiên. Không có hiện tượng tự lựa chọn vào nhóm hưởng lợi hay đối chứng (random treatment assignment, no self selection into treatment).
- Thử nghiệm tự nhiên đảm bảo thỏa điều kiện 2, nhưng hầu như không bao giờ đảm bảo điều kiện 1.
  1. Khi xảy ra vấn đề không tuân thủ, chúng ta vẫn ước lượng được tác động can thiệp ITE và ITT với encouragement design.
  2. Với dữ liệu thử nghiệm tự nhiên, chúng ta ước lượng được tác động can thiệp trung bình nội tại (LATE).

Tại sao thử nghiệm tự nhiên cho phép thiết lập quan hệ nhân quả?

Cho phép xây dựng phản chứng để ước lượng tác động can thiệp.

Table 1 Similarities and differences between RCTs, NEs, and observational studies

Type of study	Is the intervention well defined?	How is the intervention assigned?	Does the design eliminate confounding?	Do all units have a nonzero chance of receiving the treatment?
RCTs	A well-designed trial should have a clearly defined intervention described in the study protocol.	Assignment is under the control of the research team; units are randomly allocated to intervention and control groups.	Randomization means that, in expectation, there is no confounding, but imbalances in covariates could arise by chance.	Randomization means that every unit has a known chance of receiving the treatment or control condition.
NEs	Natural experiments are defined by a clearly identified intervention, although details of compliance, dose received, etc., may be unclear.	Assignment is not under the control of the research team; knowledge of the assignment process enables confounding due to selective exposure to be addressed.	Confounding is likely due to selective exposure to the intervention and must be addressed by a combination of design and analysis.	Possibility of exposure may be unclear and should be checked. For example, RD designs rely on extrapolation but assume that at the discontinuity units could receive either treatment or no treatment.
Nonexperimental observational studies	There is usually no clearly defined intervention, but there may be a hypothetical intervention underlying the comparison of exposure levels.	There is usually no clearly defined intervention and there may be the potential for reverse causation (i.e., the health outcome may be a cause of the exposure being studied) as well as confounding.	Confounding is likely due to common causes of exposure and outcomes and can be addressed, in part, by statistical adjustment; residual confounding is likely, however.	Possibility of exposure is rarely considered in observational studies so there is a risk of extrapolation unless explicitly addressed.

Thiết kế hồi quy gián đoạn - Regression Discontinuity Design

Là một thiết kế nghiên cứu với giả định lựa chọn mẫu dựa trên đặc tính quan sát được (selection on observables). RDD dùng các tiêu chí can thiệp được thiết lập không phụ thuộc vào ý muốn chủ quan của cá nhân hay hộ gia đình nhằm mô phỏng tình huống tham gia chương trình ngẫu nhiên. Ví dụ:

- Phân vùng cả nước làm 7 vùng để áp mức lương tối thiểu; hay áp thuế trước bạ khác nhau ở khu vực địa lý tỉnh thành khác nhau; hay độ tuổi nghỉ hưu theo luật định.
- Những hộ gia đình ở hai bên cận biên của tiêu chí được kỳ vọng giống nhau về mọi mặt, ngoại trừ việc tham gia chính sách $\Rightarrow$ Có thể so sánh các hộ này để tính ra tác động của việc tham gia chính sách.

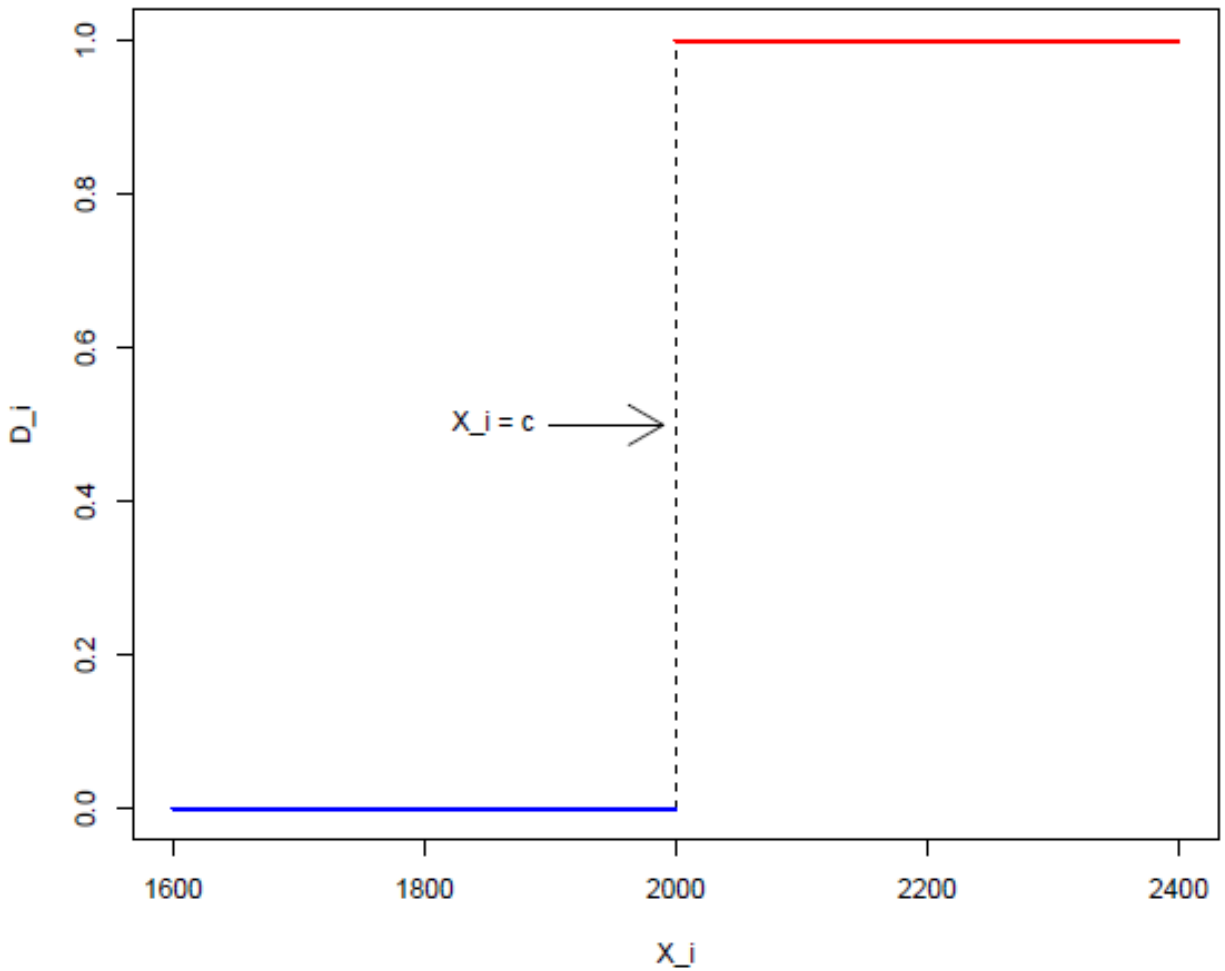
Thiết lập khung lý thuyết Sharp RDD (SRDD)

Giả sử chúng ta có các thông tin sau về chương trình can thiệp như sau:

- Chúng ta có biến tiêu chí can thiệp chương trình X, gọi là forcing variable hay running variable. Chương trình can thiệp được thực hiện tại ngưỡng can thiệp c. Thiết kế Sharp RDD (SRDD) giả định xác suất xảy ra can thiệp thay đổi từ $0 \rightarrow 1$ tại ngưỡng can thiệp.
  - $D_i = 1$ nếu $X_i > c$
  - $D_i = 0$ nếu $X_i \leq c$
- Biến X_i có thể tương quan với kết quả thực hiện chương trình $Y_1(i)$ và $Y_0(i)$ trực tiếp hoặc gián tiếp thông qua các nhân tố không quan sát được.
- Chúng ta cần ước lượng tác động của chương trình can thiệp đối với các quan sát được chọn tham gia chương trình.

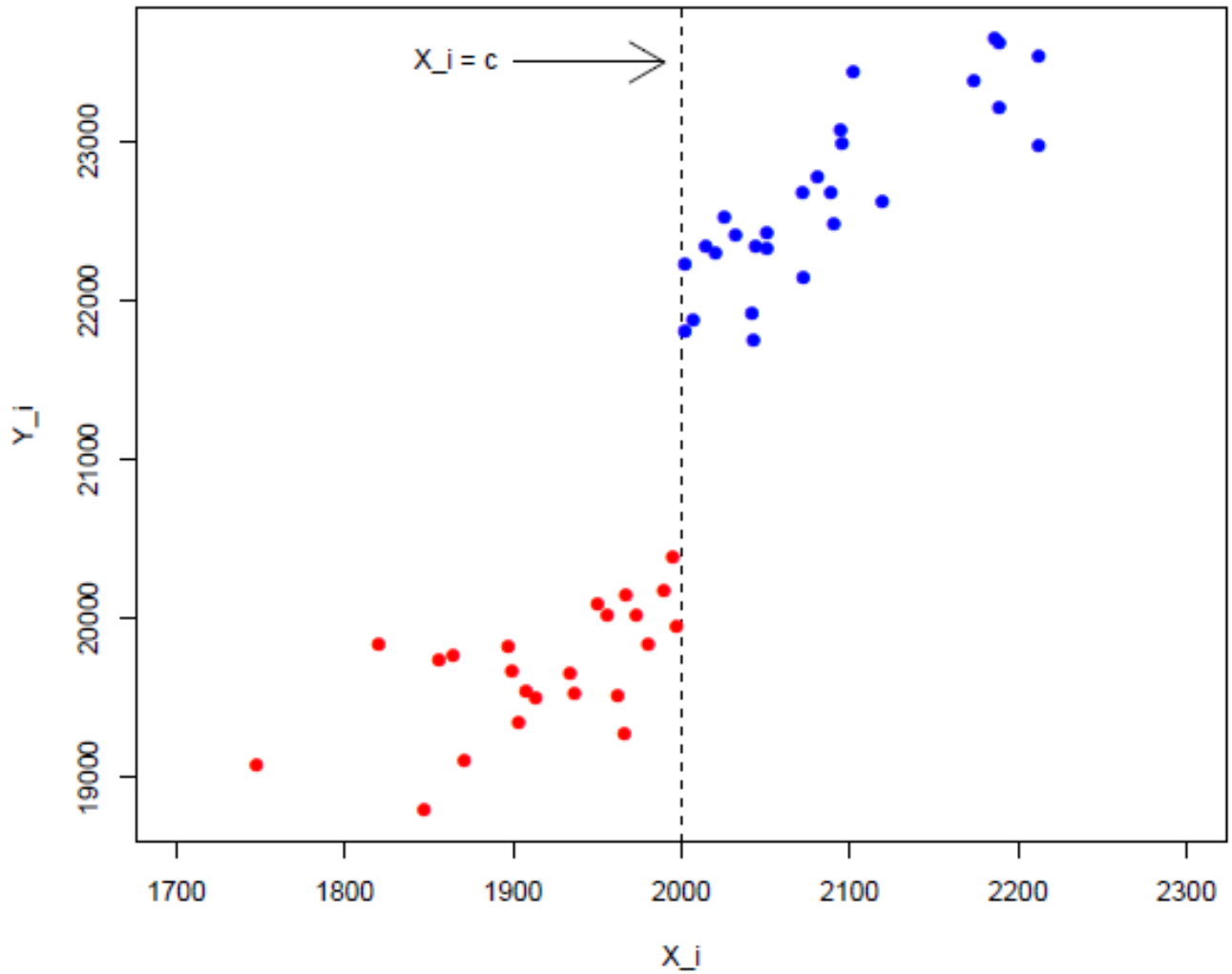
Thiết kế SRDD

Thay đổi trạng thái tham gia chương trình tại ngưỡng can thiệp chính sách.



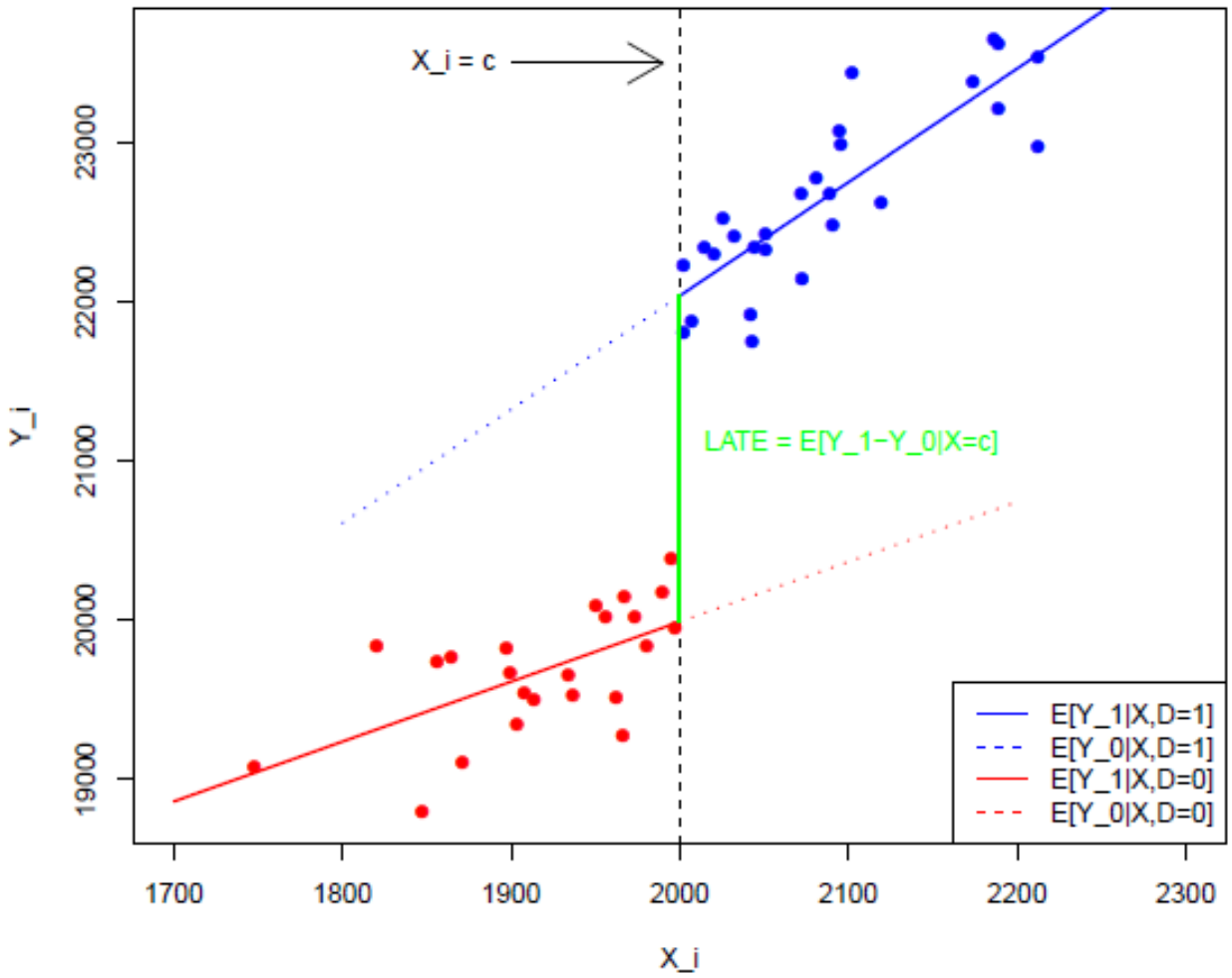
Thiết kế SRDD

Thay đổi biến kết quả tại ngưỡng can thiệp chính sách.



Thiết kế SRDD

Ước lượng tác động của chính sách bằng thiết kế SRDD.



Giả định cần thiết để ước lượng mô hình với thiết kế SRDD

- $Y_1, Y_0 \perp D|X$
- $E[Y_1|X, D]$ và $E[Y_0|X, D]$ liên tục tại các giá trị xung quanh ngưỡng can thiệp chính sách.

Với các điều kiện trên thì chúng ta có ước lượng SRDD được tính như sau:

$$\begin{aligned}\beta_{SRDD} &= E[Y_1 - Y_0|X = c] \\ &= E[Y_1|X = c] - E[Y_0|X = c] \\ &= \lim_{x \to c+} E[Y_1|X = c] - \lim_{x \to c-} E[Y_0|X = c]\end{aligned}$$

β_{SRDD} là ước lượng LATE, chỉ có hiệu lực nội tại tại ngưỡng can thiệp chính sách.

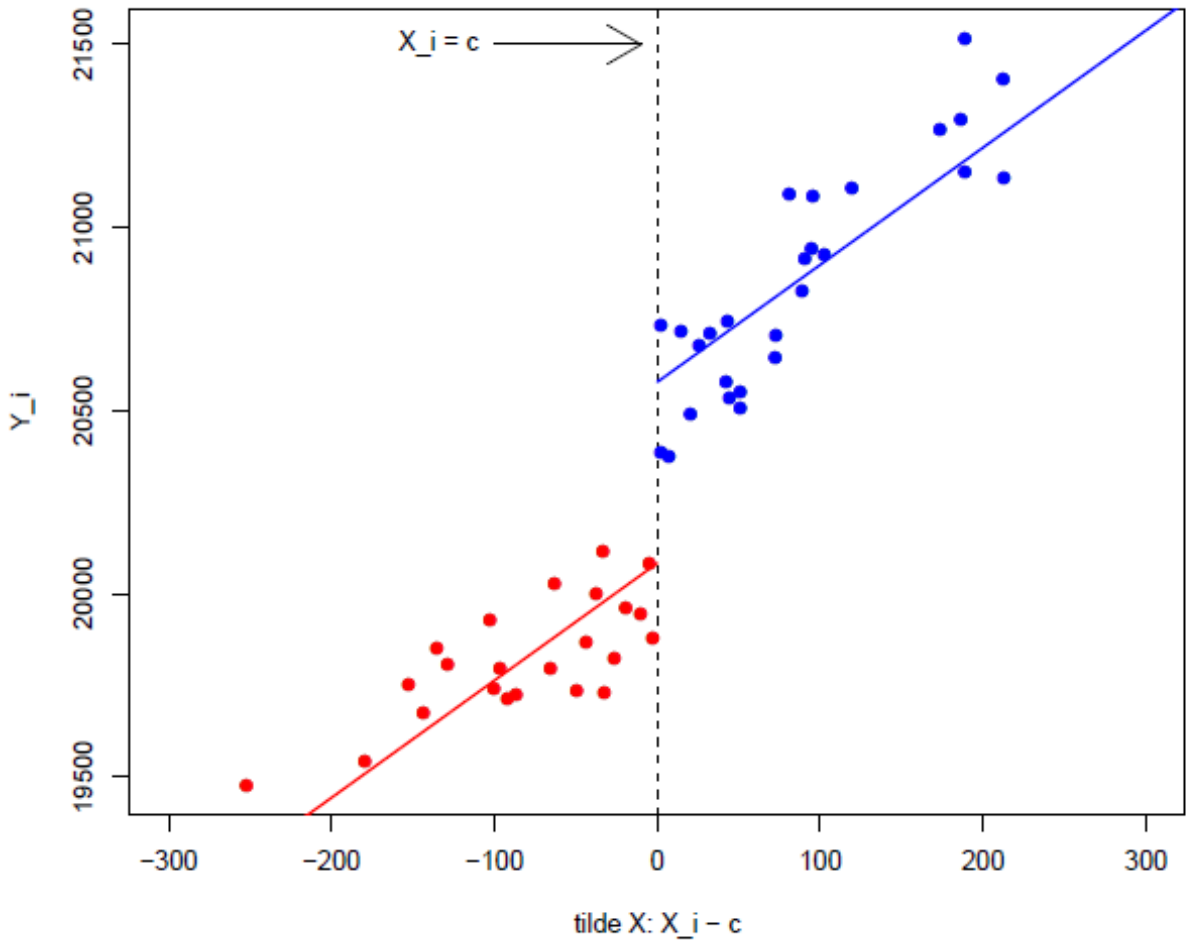
Cách thức ước lượng β_{SRDD}

- Giới hạn bộ dữ liệu nghiên cứu xung quanh ngưỡng can thiệp chính sách (bằng tiêu chí thực hiện/running variable) hay bằng khoảng cách địa lý.
  - $c - h \leq X_i \leq c + h$, h được gọi là bandwidth.
  - h được chọn dựa vào lý thuyết hay các thuật giải tối ưu. h có ảnh hưởng rất lớn đến kết quả ước lượng.
- Mã hóa lại biến tiêu chí can thiệp thành sai lệch từ ngưỡng can thiệp chính sách: $\tilde{X} = X - c$.
  - $\tilde{X} = 0$ nếu $X = c$
  - $\tilde{X} > 0$ nếu $X > c$ và do đó $D = 1$
  - $\tilde{X} < 0$ nếu $X < c$ và do đó $D = 0$
- Ước lượng β_{SRDD}
  - Sử dụng hàm hồi quy tuyến tính và có cùng độ dốc ở hai phía của ngưỡng can thiệp
  - Hàm hồi quy tuyến tính, khác độ dốc
  - Hàm phi tuyến
  - Kiểm tra bằng đồ thị hình cảnh để chọn mô hình phù hợp nhất.

Trường hợp đơn giản nhất

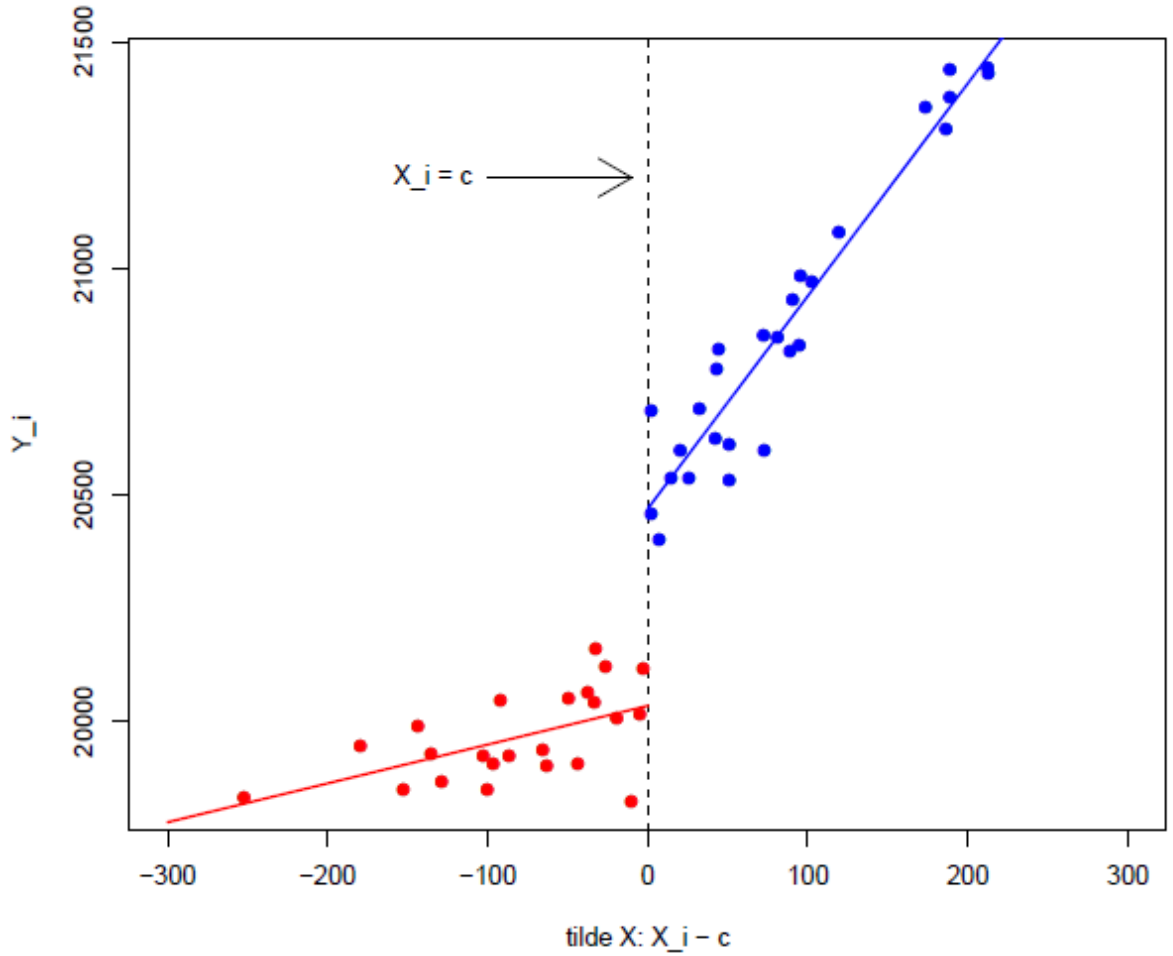
Chúng ta ước lượng β_{SRDD} bằng hồi quy tuyến tính với cùng độ dốc ở hai phía của ngưỡng can thiệp chính sách:

$$E[Y|X, D] = \beta_0 + \beta_{SRDD} D + \beta_2 \tilde{X}$$



Trường hợp hàm hồi quy tuyến tính có độ dốc khác nhau

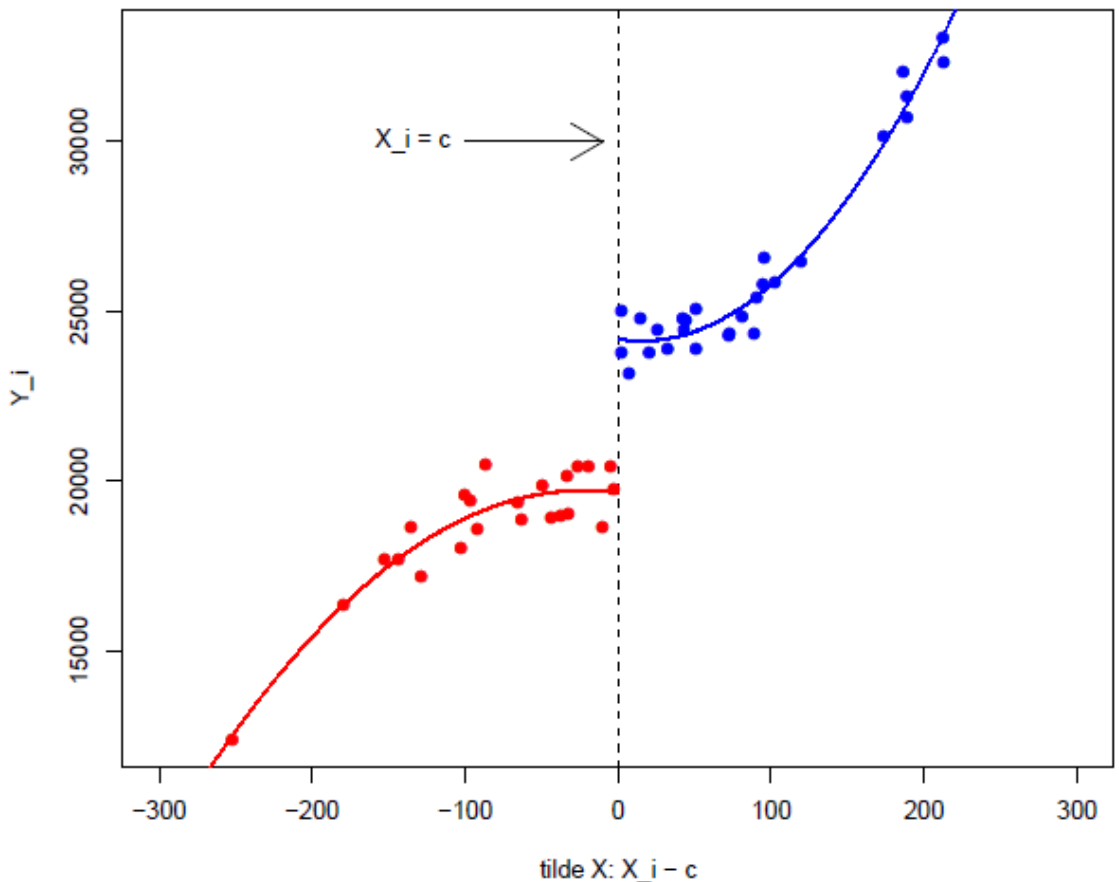
$$E[Y|X, D] = \beta_0 + \beta_{SRDD} D + \beta_2 \tilde{X} + \beta_3 D * \tilde{X}$$



Trường hợp hàm hồi quy phi tuyến

Ví dụ hàm hồi quy bậc 3

$$E[Y|X,D] = \beta_0 + \beta_{SRDD} D + \beta_2 \tilde{X} + \beta_3 \tilde{X}^2 + \beta_4 \tilde{X}^3 + \{\beta_5 \tilde{X} + \beta_6 \tilde{X}^2 + \beta_7 \tilde{X}^3\} * D$$



Tại sao SRDD lại xử lý được vấn đề lựa chọn mẫu?

Giả sử với hàm hồi quy đơn giản,

$$Y = \beta_0 + \beta_{SRDD} D + \beta_2 \tilde{X} + u$$

$$E[Y_i|D=1] - E[Y_i|D=0] =$$

$$\Big\{\beta_0 + \beta_{SRDD} + \beta_2 E[\tilde{X}|D=1] + E[u|D=1]\Big\} -$$

$$\Big\{\beta_0 + 0 + \beta_2 E[\tilde{X}|D=0] + E[u|D=0]\Big\}$$

$$= \beta_{SRDD} + \beta_2 \Big\{E[\tilde{X}|D=1] - E[\tilde{X}|D=0]\Big\}$$

$$= \beta_{SRDD} + bias$$

Nếu chúng ta giới hạn mẫu ở xung quanh ngưỡng can thiệp c và hàm $\tilde{X}$ không có hiện tượng dứt quãng hay nhảy vọt quanh giá trị c thì $bias = 0$. β_{SRDD} chính là tác động can thiệp trung bình (ATE), có hiệu lực quanh ngưỡng c, do đó ước lượng RDD cho tác động Local ATE (LATE).

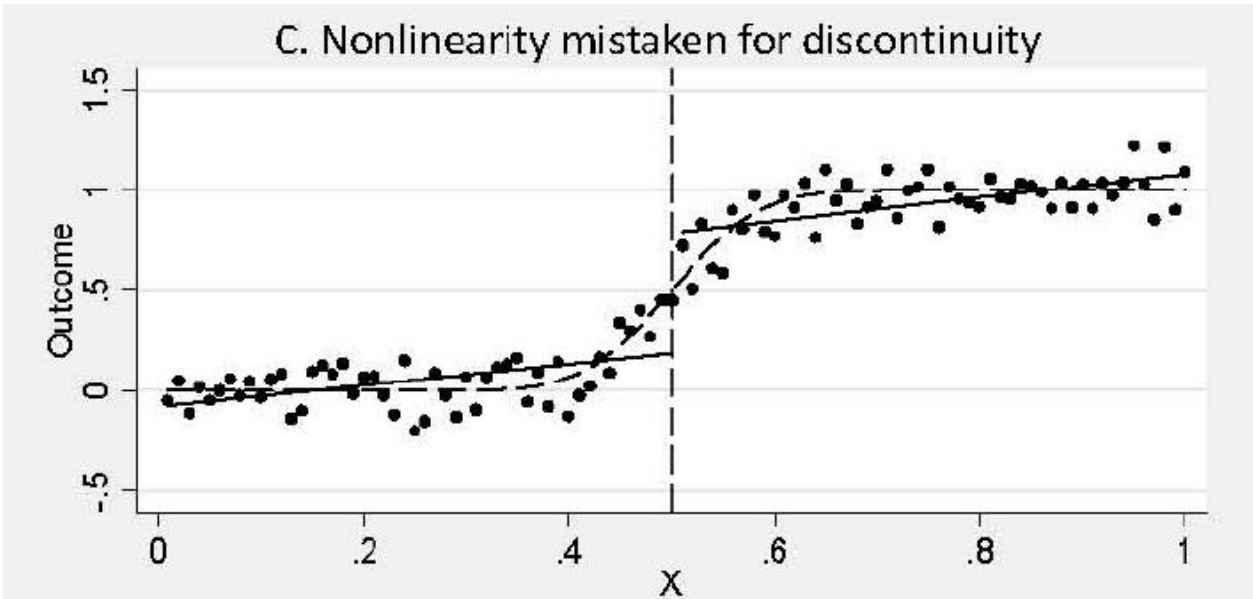
Những vấn đề phải lưu ý khi sử dụng SRDD

- Kết quả có thay đổi khi thay đổi cấu trúc hàm?
- Kiểm tra điều kiện cân bằng: các biến giải thích của mô hình có liên tục hay ngắt quãng ở ngưỡng can thiệp?
- Kiểm tra nếu có hiện tượng ngắt quãng ở ngưỡng can thiệp giả (placebo c^*)?
- Kiểm tra nếu có hiện tượng tự lựa chọn xunh quanh ngưỡng can thiệp?

Chọn cấu trúc hàm sai

RDD yêu cầu chọn cấu trúc hàm và bandwidth phù hợp. Chọn sai dẫn đến kết luận sai. Ví dụ:

- Nhầm lẫn hàm phi tuyến với gián đoạn tại ngưỡng can thiệp.
- Tăng độ phức tạp của hàm hồi quy (polynomials) làm giảm mức độ chệch (bias), nhưng phải đánh đổi với hiệu quả (variance).
- Thay đổi bandwidth có thể hưởng đến kết quả.



Kiểm tra điều kiện cân bằng

Kiểm tra điều kiện cân bằng của các biến kiểm soát xung quanh ngưỡng can thiệp.

- Kiểm tra bằng đồ thị: Đồ thị scatter plot của biến giải thích Z theo biến can thiệp X phải trơn (smooth) tại ngưỡng can thiệp c.
- Kiểm định bằng thống kê: Sử dụng Z làm biến phụ thuộc giả (placebo outcome) và ước lượng mô hình sau:

$$E[Z|X, D] = \beta_0 + \beta_{SRDD} D + \beta_2 \tilde{X} + \beta_3 D * \tilde{X}$$

  sau đó kiểm định nếu $\beta_{SRDD} = 0$.
  - Nếu đảm bảo thì Z cân bằng tại ngưỡng can thiệp.
  - Nếu không đảm bảo điều kiện cân bằng thì có thể kèm biến Z vào mô hình hồi quy để kiểm soát vấn đề thiếu cân bằng.
- Chỉ kiểm định cân bằng với các nhân tố quan sát được. Vẫn có thể tồn tại nhân tố không quan sát được không cân bằng.

Kiểm định can thiệp giả/Falsification test/Placebo threshold test

Để kiểm tra cấu trúc hàm và các giả định của mô hình là phù hợp, giả sử chúng ta kiểm định liệu mô hình có phát hiện ra tác động tại một ngưỡng can thiệp giả c^* nào đó khác với ngưỡng can thiệp thực c.

- Nếu mô hình và giả định là đúng thì sẽ không phát hiện được tác động nào ở ngưỡng can thiệp giả.
- Chúng ta sẽ ước lượng mô hình:

$$E[Y|X, D] = \beta_0 + \beta_{SRDD} D + \beta_2 \tilde{X}^* + \beta_3 D * \tilde{X}^*$$

với $\tilde{X}^* = X - c^*$. Lưu ý chỉ sử dụng dữ liệu của một phía đối với ngưỡng can thiệp thực, $X > c$ hoặc $X < c$.

- Kiểm tra nếu $\beta_{SRDD} = 0$ thỏa. Nếu vi phạm không có nghĩa là thiết kế *RDD* sai. Có thể có những nguyên nhân khác như hàm hồi quy bị gián đoạn tại nhiều ngưỡng, dẫn đến ước lượng thực bị nhiễu bởi các nhân tố đó, dẫn đến kết luận kém chính xác.

Kiểm tra vấn đề tự lựa chọn mẫu xung quanh ngưỡng can thiệp/Sorting around the threshold

Đây là vấn đề nghiêm trọng nhất trong thiết kế SRDD. Liệu có xảy ra hiện tượng cá nhân thay đổi hành vi để tự lựa chọn vào nhóm can thiệp hay đối chứng không?

- Cá nhân có thể cố ý thay đổi đặc tính liên quan đến biến can thiệp X để được lựa chọn vào nhóm hưởng lợi.
- Nhà hoạch định chính sách có thể lựa chọn loại chỉ số can thiệp X_k hay ngưỡng giá trị can thiệp c nhằm một mục đích nào đó.
- Hiện tượng lựa chọn mẫu là một thách thức cho việc nhận diện tác động đối với thiết kế SRDD do vấn đề không tuân thủ. Khi này có thể phải sử dụng thiết kế Fuzzy RDD.
- Khi hiện tượng chọn mẫu xảy ra với quy mô lớn có thể vô hiệu hóa thiết kế RDD.

Kiểm tra vấn đề sorting

- Sử dụng đồ thị phân phối ở hai phía của ngưỡng can thiệp c.
- Nếu đồ thị phân phối bị ngắt quãng hoặc nhảy vọt ở ngưỡng c chứng tỏ có hiện tượng sorting.
- Ví dụ vấn đề sorting trong chương trình đào tạo nghề cho những người thu nhập thấp. Đồ thị bên trái khi không xảy ra hiện tượng sorting. Nếu một số người chủ động hạn chế số giờ làm để hạn chế thu nhập và do đó được phân bổ vào nhóm hưởng lợi, dẫn đến đồ thị phân phối bị ngắt quãng tại ngưỡng thu nhập c (đồ thị trái).