

Mô hình Khác biệt Kép (Difference-in-Difference Method)

Kinh tế lượng ứng dụng

Lê Việt Phú

Chương trình Giảng dạy Kinh tế Fulbright

Ngày 16 tháng 5 năm 2015

Table of contents

Ôn tập lý thuyết

Khung phân tích DiD

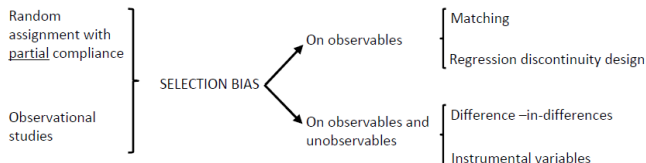
Phương pháp DiD - Mở rộng

Thực hành

Ôn tập lý thuyết

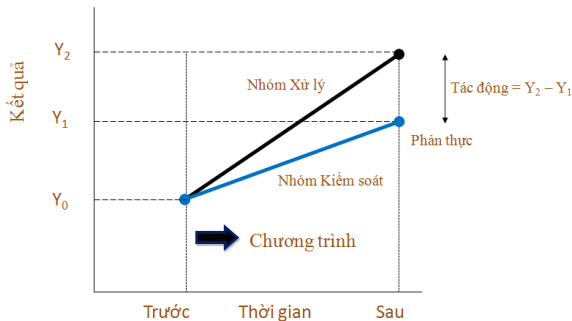
Các phương pháp với dữ liệu khác nhau:

Random assignment with average $\longrightarrow$ NO SELECTION BIAS $\longrightarrow$ ATT=(Average effect on treated) – (Average effect on non-treated)



Ôn tập lý thuyết (2)

- ▶ Bản chất của đánh giá tác động của chính sách là ước lượng phản thực $E[Y_0 | T = 1]$ bằng các phương pháp khác nhau.



- ▶ Dữ liệu điều tra ngẫu nhiên: phản thực chính là nhóm kiểm soát, $E[Y_0 | T = 1] = E[Y_0 | T = 0]$.
- ▶ Dữ liệu bán thực nghiệm: sử dụng thiết kế của chính sách để tìm cách ước lượng phản thực. Tùy thuộc vào việc có quan sát được nhân tố quyết định tham gia chính sách hay không.

Ước lượng điểm Xu hướng (Selection on Observables)

Mô hình việc tham gia chính sách để tạo nhóm đối chứng dựa vào những đặc tính quan sát được:

$$P(T = 1|X_i) = f(X_1, X_2, \dots, X_k)$$

Các giả định để đảm bảo ước lượng được nhóm đối chứng:

- ▶ Độc lập có điều kiện: $Y_i^1, Y_i^0 \perp T_i | X_i$ đảm bảo việc tham gia chương trình là ngẫu nhiên sau khi đã kiểm soát các khác biệt về đặc tính của hộ tham gia. Điều kiện này đảm bảo việc diễn giải tác động của tham gia chính sách là tác động nhân quả.
- ▶ Có vùng hỗ trợ chung: $0 < P(T_i = 1|X_i) < 1$ đảm bảo tìm được nhóm đối chứng có các đặc tính giống với nhóm tham gia.

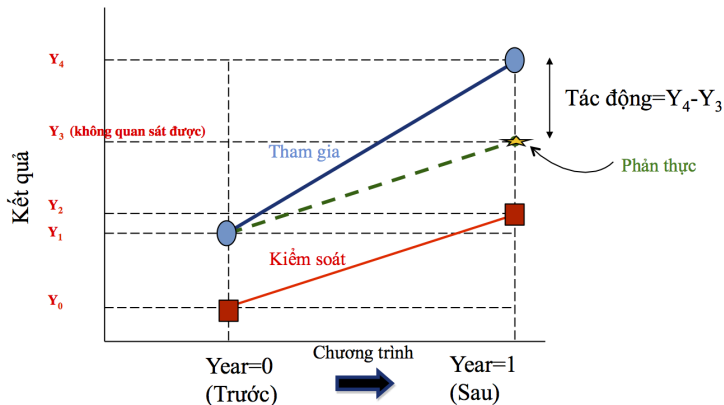
Kết hợp hai điều kiện trên thì PSM sẽ tìm ra nhóm đối chứng $E[Y_0|T = 1]$ dựa vào xác suất tham gia chương trình, để từ đó tính được tác động nhân quả của chính sách.

Khung phân tích của phương pháp DiD

Nếu nhân tố quyết định tham gia chính sách không quan sát được $\Rightarrow$ lựa chọn dựa vào biến không quan sát được (selection on unobservables). Không sử dụng được phương pháp PSM.

- ▶ Giả định song song (parallel assumption): Nếu không có chương trình thì thay đổi kết quả của nhóm tham gia (T) và nhóm kiểm soát (C) là như nhau. Khi này có thể sử dụng nhóm kiểm soát $E[Y^C | T = 1]$ để xây dựng nhóm đối chứng $E[Y_0^T | T = 1]$.
 - ▶ Giả định này nói lỏng hơn rất nhiều so với giả định nhóm kiểm soát tương đồng với nhóm tham gia trong phương pháp mẫu ngẫu nhiên bởi nó cho phép sử dụng nhóm tham gia và nhóm kiểm soát có khác biệt về các thuộc tính, kể cả các thuộc tính không quan sát được có thể ảnh hưởng đến lựa chọn tham gia chương trình (unobserved heterogeneity).
 - ▶ Chúng ta sẽ nghiên cứu tình huống phức tạp hơn khi giả định song song bị vi phạm.

Khung phân tích của phương pháp DiD



	Trước	Sau	
Kiểm soát	Y_0	Y_2	$Y_2 - Y_0 = a$
Tham gia	Y_1	Y_4	$Y_4 - Y_1 = b$
		DID	$= b - a = Y_4 - Y_3$

Khung phân tích của phương pháp DiD (2)

Ước lượng tác động bằng hồi quy:

$$Y_i = \beta_0 + \beta_1 * T_i + \beta_2 * Year + \beta_3 * (T \times Year) + \beta_4 * X_i + \varepsilon_i \quad (1)$$

Trong đó:

- ▶ T là biến trạng thái tham gia chính sách
- ▶ $Year$ là biến thời gian trước và sau khi thực hiện chính sách
- ▶ X_i là các đặc tính của hộ gia đình (tạm thời bỏ qua)

β_3 là ước lượng ATT của việc tham gia chương trình:

	Year = 0	Year = 1	ΔY
T=0	$Y = \beta_0$	$Y = \beta_0 + \beta_2$	β_2
T=1	$Y = \beta_0 + \beta_1$	$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
			DiD = β_3

Khung phân tích của phương pháp DiD (3)

Hình thức ước lượng DiD đơn giản nhất:

- ▶ Dùng dữ liệu gộp (pooled regression): gộp các quan sát qua nhiều năm của các hộ gia đình thành một bảng dữ liệu chéo (repeated cross-sectional data).
- ▶ DiD cũng có thể áp dụng với dữ liệu chéo (chỉ có một năm quan sát duy nhất đối với tất cả các hộ gia đình), tuy nhiên rất hiếm khi được sử dụng do thiếu tính tin cậy.

Phương pháp DiD - Mở rộng

- ▶ DiD có tính đến điều kiện ban đầu:
Nếu điều kiện ban đầu ảnh hưởng đến giả định song song $\Rightarrow$ hồi quy với tung độ gốc và hệ số góc thay đổi.
- ▶ Ước lượng tác động cố định sử dụng dữ liệu bảng (panel data with fixed effects)
Sử dụng dữ liệu bảng có thể kiểm soát được các yếu tố không quan sát được (ví dụ như IQ, tố chất cá nhân) không thay đổi theo thời gian nhưng có ảnh hưởng đến kết quả.
- ▶ FE vs RE:
Dữ liệu bảng cho phép sử dụng nhiều giả định khác nhau về yếu tố không quan sát được có thể ảnh hưởng đến chương trình.
- ▶ DiD kết hợp với PSM:
Sử dụng PSM để lọc các quan sát có độ tương đồng cao trước khi sử dụng DiD để tính tác động của chương trình.

So sánh các phương pháp đã học

	Randomization	PSM	DiD
Giả định			
Dữ liệu			
Tác động			
Ưu điểm			
Nhược điểm			

Thực hành

- ▶ data file STATA hh_9198.dta
- ▶ STATA dd.do file

Nghiên cứu cấu trúc file hh_9198.dta

- ▶ Dữ liệu dạng bảng $N \times t = 826 \times 2$: 826 hộ gia đình, mỗi hộ có quan sát trước ($t=0$) và sau ($t=1$) khi thực hiện chương trình.
- ▶ Bảng dọc (long format)

HHid	Year	Village	Treatment	Y_i	X_i
1	0	...	0	y_0^T	x_{10}
1	1	...	1	y_1^T	x_{11}
2	0	...	0	y_0^C	x_{20}
2	1	...	0	y_1^C	x_{21}
...	...	...	...	...	...

- ▶ Các kỹ thuật xử lý và chuyển đổi rất quan trọng đối với dữ liệu bảng!

Thực hành (2) - Hồi quy dữ liệu gộp (pooled regression, repeated cross-sectional data)

Để ước lượng được phương trình hồi quy (1) bằng phương pháp gộp dữ liệu, cần tạo biến tương tác $\mathbf{T} \times \mathbf{Year}$:

HHid	Year	Village	Treatment	$\mathbf{T} \times \mathbf{Year}$	$\mathbf{Y}_i$	$\mathbf{X}_i$
1	0	...	1	0	y_0^T	x_{10}
1	1	...	1	1	y_1^T	x_{11}
2	0	...	0	0	y_0^C	x_{20}
2	1	...	0	0	y_1^C	x_{21}
...	...	...	...	...	...	...

- ▶ $\text{reg } Y \text{ Year } T \text{ (Year} * T) X_i \Rightarrow$ Ước lượng chênh lệch trung bình giữa nhóm có và không tham gia chính sách, có kiểm soát sự ảnh hưởng của các nhân tố khác.
- ▶ Lợi ích của hồi quy dữ liệu gộp là thực hiện đơn giản, không yêu cầu dữ liệu bảng phải cân bằng (mỗi hộ gia đình đều có quan sát ở tất cả các thời kỳ). Tuy nhiên, nếu dữ liệu bị thiếu một cách hệ thống (non-random missing values) thì việc ước lượng có thể bị chệch.

Thực hành (3) - Hồi quy dữ liệu bảng với tác động cố định

- Panel data with fixed effects

Khác với hồi quy dữ liệu gộp, hồi quy dữ liệu bảng cho phép tách được ảnh hưởng của khác biệt không quan sát được nhưng không thay đổi theo thời gian (time invariant unobserved heterogeneity). Ví dụ tổ chức cá nhân không thay đổi theo thời gian, và có thể có ảnh hưởng đến quyết định tham gia chương trình cũng như kết quả chương trình.

Hình thức ước lượng thứ nhất:

- ▶ *xtreg Y Year T (Year * T) X_i, fe i(id)* id là mã hộ gia đình.
Mô hình FE ước lượng phương trình hồi quy sau:

$$Y_i = \beta_0 + \beta_1 * T_i + \beta_2 * Year + \beta_3 * (T_i \times Year_i) + \beta_4 * X_i + \eta_i + \varepsilon_i \quad (2)$$

η_i là tác động cố định của mỗi hộ gia đình i , không quan sát được.

Thực hành (4) - Hồi quy dữ liệu bảng với tác động cố định

- Panel data with fixed effects

Các lệnh khác có kết quả tương tự như FE bằng cách tạo một biến giả cho mỗi hộ gia đình để ước lượng tác động cố định và dùng phương pháp LSDV:

- ▶ *areg Y Year T (Year * T) X_i, a(id)*
- ▶ *reg Y Year T (Year * T) X_i i.id*

Hình thức ước lượng thứ hai:

- ▶ Hồi quy sử dụng sai biệt của các biến số của phương trình (2):
Lấy sai biệt của các biến trong phương trình (2) qua thời gian đối với từng hộ gia đình (lấy dữ liệu năm sau trừ đi dữ liệu năm trước), khi đó tác động cố định và tung độ gốc sẽ bị trừ khử, và bản chất là chúng ta ước lượng mô hình sau bằng OLS:

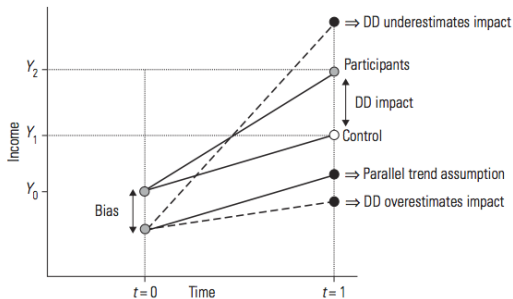
$$\Delta Y_i = \beta_2 + \beta_1 * \Delta T_i + \beta_3 * \Delta(T_i \times Year_i) + \beta_4 * \Delta X_i + \mu_i \quad (3)$$

- ▶ Sử dụng lệnh *reg dY dT dInteraction dX_i* với các sai biệt được tạo ra.

Thực hành (5) - DiD có tính đến điều kiện ban đầu

Giả định song song hàm ý các nhân tố phi chính sách không quan sát được ảnh hưởng đến nhóm tham gia và nhóm kiểm soát không thay đổi theo thời gian (time-invariant selection bias). Nếu giả định song song bị vi phạm $\Rightarrow$ có thể dẫn đến ước lượng chệch trên hoặc dưới:

Time-Varying Unobserved Heterogeneity



- ▶ Ví dụ tốc độ tăng trưởng thu nhập của nhóm tham gia lớn hơn nhóm kiểm soát $\Rightarrow$ DiD chệch dưới (ước lượng thấp hơn thực tế) và ngược lại.

Thực hành (6) - DiD có tính đến điều kiện ban đầu

Tương tự như ước lượng sử dụng sai biệt của các biến số, nhưng kiểm soát thêm điều kiện ban đầu $\mathbf{X}_i$ trong cùng một phương trình hồi quy:

$$\Delta Y_i = \beta_2 + \beta_1 * \Delta T_i + \beta_3 * \Delta(T_i \times Year_i) + \beta_4 * \Delta X_i + \beta_5 * \mathbf{X}_i + \mu_i \quad (4)$$

Sử dụng lệnh *reg dY dT dInteraction dX_i X_i* với các sai biệt được tạo ra và điều kiện ban đầu (quan sát X_i tại thời điểm $Year = 0$).

Thực hành (7) - DiD kết hợp với PSM

- ▶ Ôn tập: PSM tìm ra nhóm phản chứng dựa vào các đặc tính quan sát được và loại bỏ những quan sát nằm ngoài vùng hỗ trợ.
- ▶ Kết hợp PSM và DiD sẽ mạnh hơn là chỉ sử dụng DiD.

Các bước thực hiện:

- ▶ Bước 1: Lọc các hộ gia đình nằm trong vùng hỗ trợ chung bằng cách ước lượng xác suất tham gia chương trình (điểm xu hướng) dựa trên điều kiện ban đầu. Lưu ý là nếu dữ liệu là dạng dài (long format) thì cần phải chuyển đổi thành dạng rộng (wide format).
 - ▶ Chạy chương trình pscore
 - ▶ Kiểm tra cho đến khi điều kiện cân bằng được đảm bảo
 - ▶ Lọc các quan sát đảm bảo điều kiện cân bằng và bỏ các quan sát nằm ngoài vùng hỗ trợ chung
- ▶ Bước 2: Tạo bộ dữ liệu chỉ với các quan sát nằm trong vùng hỗ trợ chung và sử dụng các phương pháp ước lượng DiD trên mẫu dữ liệu đã chọn lọc này.

Thực hành (8) - Thực hành DiD kết hợp với PSM

Bước 1:

Cần thiết tạo dữ liệu chéo (wide format) từ dữ liệu bảng (long format) sao cho với mỗi quan sát thì biến tham gia chính sách T được tính tại thời điểm $Year = 1$, X_i là đặc tính hộ gia đình thời điểm $Year = 0$:

```
pscore T Xi, pscore(myscore) blockid(myblock) comsup  
keep if myblock! = .  
keep id  
sort id
```

Bước 2:

```
merge id using master.dta  
keep if _merge == 3  
drop merge  
gen Interact = T * Year  
xtreg income T Year Interact Xi, fe i(id)
```