

# Mô hình So sánh bằng Điểm Xu hướng (Propensity Score Matching Method)

Kinh tế lượng ứng dụng

Lê Việt Phú

Chương trình Giảng dạy Kinh tế Fulbright

Ngày 12 tháng 5 năm 2015

# Table of contents

Ôn tập lý thuyết

Phương pháp PSM

Thực hành

# Ôn tập lý thuyết

$$D = \mathbf{E}(Y_i | T = 1) - \mathbf{E}(Y_i^0 | T = 1) + \mathbf{E}(Y_i^0 | T = 1) - \mathbf{E}(Y_i | T = 0)$$

- ▶  $\mathbf{E}(Y_i | T = 1) - \mathbf{E}(Y_i^0 | T = 1)$  chính là tác động của chương trình đối với những hộ tham gia, so sánh với khi chính họ không tham gia. Tác động này được gọi là tác động trung bình với người tham gia (average treatment effect on the treated-ATT, hoặc TOT).
- ▶  $\mathbf{E}(Y_i^0 | T = 1) - \mathbf{E}(Y_i | T = 0)$  là tác động của lựa chọn mẫu (selection bias) lên tác động trung bình D. Đó là sự khác biệt giữa thu nhập của những hộ nếu như họ không tham gia, nhưng trên thực tế là có tham gia, với thu nhập của những hộ không tham gia (giải thích sau).  $Y_i^0 | T = 1$  không quan sát được nên không thể ước lượng D một cách chính xác tuyệt đối.

# Các phương pháp ước lượng

Bản chất của đánh giá chính sách là tìm cách ước lượng phản thực.

1. Thử nghiệm ngẫu nhiên (randomization) để loại bỏ vấn đề chọn mẫu. Phản thực chính là nhóm đối chứng (nhóm kiểm soát).

2. Thiết kế nghiên cứu sử dụng dữ liệu bán thực nghiệm hay thử nghiệm tự nhiên. Nhóm xử lý và nhóm kiểm soát không được phân bổ một cách ngẫu nhiên. Chính sách được áp dụng theo các tiêu chí độc lập, không phụ thuộc vào ý chí chủ quan của người tham gia. Là các tình huống chính sách thường gặp.

- ▶ **So sánh bằng điểm xu hướng - PSM**

- ▶ Khác biệt kép - DiD

- ▶ Biến công cụ/Hồi quy cắt (IV, RD)

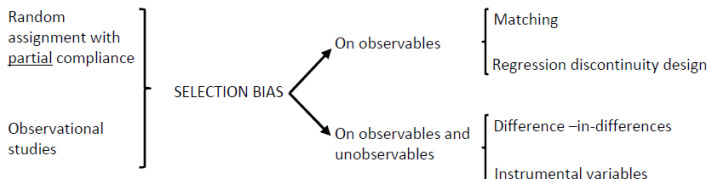
# Các phương pháp ước lượng

Dữ liệu bán thực nghiệm hay thực nghiệm tự nhiên là gì?

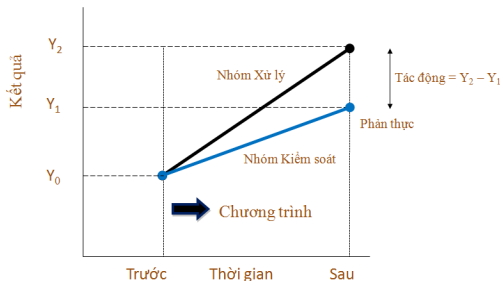
- ▶ Việc phân công vào nhóm xử lý hay nhóm kiểm soát là không ngẫu nhiên, tuy nhiên các quan sát không tự lựa chọn vào nhóm xử lý hay nhóm kiểm soát được. Các dữ liệu thu thập từ việc ban hành chính sách vĩ mô của chính phủ, hay các hiện tượng thời tiết bất thường, bất khả kháng thường được sử dụng làm dữ liệu bán thực nghiệm.
- ▶ Không đảm bảo tính ngẫu nhiên của nhóm xử lý và nhóm kiểm soát do nhóm bị ảnh hưởng và nhóm không bị ảnh hưởng có thể khác nhau  $\Rightarrow$  Luôn tồn tại vấn đề chọn mẫu.

# Các phương pháp ước lượng

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



# Phương pháp PSM



- ▶ Bản chất của PSM là xây dựng nhóm phản chứng (counterfactual) bằng các phương pháp thống kê.
- ▶ Dựa vào các đặc tính quan sát được giữa nhóm xử lý (treatment group) và nhóm kiểm soát (control group), chúng ta xây dựng một chỉ số gọi là điểm xu hướng - propensity score - là xác suất một quan sát được chọn vào nhóm xử lý hay nhóm kiểm soát.
- ▶ Các quan sát có cùng điểm xu hướng có thể được dùng làm nhóm đối chứng.

# Phương pháp PSM

Các giả định của phương pháp PSM:

- ▶ Chỉ số xu hướng hoàn toàn được xây dựng dựa trên các đặc tính quan sát được, do đó phương pháp PSM yêu cầu việc lựa chọn mẫu chỉ phụ thuộc vào các đặc tính có thể thấy được (selection on observables). Các đặc tính không quan sát được không ảnh hưởng đến quá trình chọn nhóm xử lý hay nhóm kiểm soát.
- ▶ Giả định độc lập có điều kiện (*conditional independence* hay *unconfoundedness*): sau khi đã kiểm soát các yếu tố khác quan sát được, sự khác biệt về tác động chính sách lên nhóm xử lý hay nhóm kiểm soát không phụ thuộc vào việc phân bổ chính sách. Nói cách khác là sau khi đã kiểm soát sự khác biệt về đặc tính (i.e. *ceteris paribus*), thì việc phân bổ vào nhóm xử lý hay nhóm kiểm soát là hoàn toàn ngẫu nhiên. Từ đó chúng ta mới có thể sử dụng nhóm kiểm soát để làm đối chứng cho nhóm xử lý.

$$Y_i^0, Y_i^1 \perp T_i | X_i$$



## Phương pháp PSM

Giả định có vùng hỗ trợ chung (*common support*) hoặc điều kiện trùng lặp (*overlap condition*):

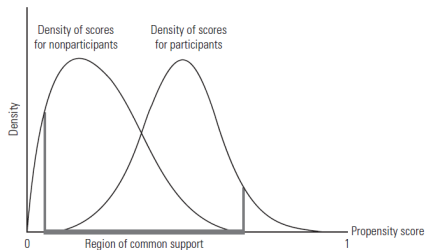
- ▶ Vùng hỗ trợ chung là vùng có ước lượng điểm xu hướng của cả nhóm xử lý và nhóm kiểm soát. Điều kiện của vùng hỗ trợ chung là:

$$0 < P(T_i = 1|X_i) < 1$$

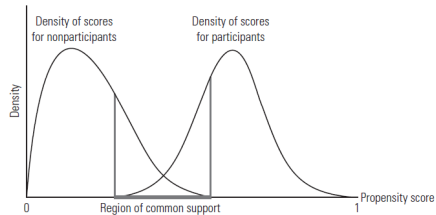
- ▶ Đối với mỗi khoảng giá trị của điểm xu hướng, có thể tìm được các quan sát đối chứng để đảm bảo việc so sánh giữa nhóm xử lý và nhóm kiểm soát là hợp lệ. Nếu các quan sát có điểm xu hướng khác biệt nhau quá thì không thể sử dụng làm nhóm đối chứng tốt.
- ▶ Có càng nhiều quan sát của cả nhóm xử lý và nhóm kiểm soát trong vùng hỗ trợ chung càng tốt.
- ▶ Quan sát không nằm trong vùng hỗ trợ chung sẽ bị loại.

# Phương pháp PSM

## Vùng hỗ trợ tốt



## Vùng hỗ trợ kém



## Phương pháp PSM

Nếu các điều kiện trên được thỏa mãn, thì tác động tham gia trung bình đối với những người tham gia (ATT hay TOT) có thể ước lượng được bằng PSM khi sử dụng dữ liệu chéo (cross-sectional data):

$$TOT_{PSM} = \frac{1}{N_T} \left[ \sum_{i \in T} Y_i^T - \sum_{j \in C} \omega(i, j) Y_j^C \right]$$

- ▶ Các nhóm đối chứng có cùng điểm xu hướng (hay xác suất) với các nhóm xử lý sẽ so sánh - match - với nhau để tìm tác động trung bình là sai biệt giữa kết quả của hai nhóm.
- ▶  $TOT_{PSM}$  là sai lệch trung bình giữa nhóm xử lý và nhóm đối chứng, với trọng số  $\omega(i, j)$  được tính bằng các phương pháp so sánh khác nhau.
- ▶ Do tồn tại vấn đề chọn mẫu nên  $ATE \neq ATT(TOT)$  trong phương pháp PSM.

# Phương pháp PSM

Các cách thức chọn nhóm đối chứng dựa vào điểm xu hướng:

- ▶ Với mỗi khoảng giá trị, lựa chọn  $n$  quan sát gần nhất (nearest-neighbor matching).
- ▶ Chọn quan sát trong khoảng giá trị cho trước (caliper or radius matching).
- ▶ So sánh trong từng khoảng giá trị cho trước (stratification or interval matching).
- ▶ So sánh bằng hàm hồi quy nội tại sử dụng phương pháp phi tham số (kernel and local linear matching).

# Phương pháp PSM

Các bước thực hiện đánh giá theo phương pháp PSM:

- ▶ Bước 1: Ước lượng mô hình xác suất tham gia chương trình bằng hàm logit hay probit, với các biến giải thích là các đặc tính có ảnh hưởng đến khả năng tham gia chương trình:

$$\hat{P}(T = 1) = F(X_1, \dots, X_m)$$

- ▶ Bước 2: Xác định vùng hỗ trợ chung và thực hiện các kiểm định đảm bảo các điều kiện cân bằng được thỏa mãn. Các nhóm có cùng một giá trị điểm xu hướng (hay khoảng giá trị) cần có các thuộc tính không quá khác biệt nhau.
- ▶ Bước 3: So sánh nhóm xử lý với nhóm kiểm soát trong vùng hỗ trợ chung, dựa vào các cách thức tạo nhóm đối chứng như đã trình bày ở trên.

# Thực hành

- ▶ Cài đặt chương trình cho STATA ("findit pscore")
- ▶ Sử dụng dữ liệu hh\_98.dta
- ▶ STATA dofile psm.do
- ▶ Lập trình phương pháp PSM trong STATA

## Thực hành (2)

Sử dụng lệnh:

- ▶ Chạy chương trình pscore để tính điểm xu hướng và kiểm tra tính cân bằng giữa nhóm đối chứng và nhóm xử lý:

```
pscore T X1 X2... [pw=weight], pscore(myscore)  
blockid(myblock) numblo(k) comsup
```

- ▶ T là biến chính sách, có tham gia hay không
- ▶ X1 X2 ... là các biến giải thích cho việc tham gia chính sách
- ▶ weight là biến trọng số, thường sử dụng trong các bộ dữ liệu điều tra dân số
- ▶ myscore là tên biến điểm xu hướng
- ▶ blockid là tên khoảng điểm xu hướng
- ▶ numblo là số lượng block, tùy chọn
- ▶ comsup là biến trạng thái chỉ dẫn số liệu có nằm trong vùng hỗ trợ hay không

## Thực hành (3)

- ▶ Chương trình pscore sẽ tính xác suất tham gia chương trình dựa vào các biến giải thích  $[X_1, X_2, \dots]$  sử dụng hàm logit hay probit. Mặc định của pscore là probit.
- ▶ Kiểm tra điều kiện cân bằng: các biến giải thích phải cân bằng (giá trị trung bình bằng nhau) trong mỗi khoảng giá trị của điểm xu hướng. Điều kiện này đảm bảo nhóm đối chứng tương đồng với nhóm xử lý. Nếu không đạt được thì phải chạy lại chương trình pscore và lựa chọn biến giải thích khác. Thêm option **detail** để biết thêm chi tiết biến nào và khoảng giá trị nào không cân bằng.



## Thực hành (4)

Sau khi chạy lệnh `pscore` và điều kiện cân bằng đã được đảm bảo, tính tác động tham gia trung bình đối với người tham gia (*TOT*) theo các hình thức so sánh khác nhau:

Command	Method
<code>attnd</code>	nearest-neighbor matching
<code>atts</code>	stratification matching
<code>attr</code>	radius matching
<code>atrk</code>	kernel matching

- ▶ `attnd Y T [pw=weight], pscore(myscore) comsup`
- ▶ `atts Y T, pscore(myscore) blockid(myblock) comsup`
- ▶ `attr Y T, pscore(myscore) radius(xxxx) comsup`
- ▶ `atrk Y T, pscore(myscore) comsup bootstrap reps(50)`

## Thực hành (5) - Advanced level

- ▶ Tự viết chương trình để tính ra tác động thay vì sử dụng các phần mềm hay các chương trình viết sẵn (canned commands, e.g. `reg Y X`) là phương pháp tốt nhất để học kinh tế lượng.
- ▶ Pseudocode của phương pháp PSM
  - ▶ Bước 1: ước lượng propensity score bằng logit hoặc probit
  - ▶ Bước 2: kiểm tra vùng hỗ trợ, phân khoảng (block), kiểm tra điều kiện cân bằng trong mỗi block. Nếu thỏa mãn chuyển sang bước 3, nếu không thì hoặc là phân lại block, hoặc là sửa hàm ước lượng ở bước 1. Lặp lại cho đến khi điều kiện cân bằng được đảm bảo.
  - ▶ Bước 3: tính *TOT* trung bình của các block.

## Thực hành (6) - Advanced level

This code was adapted from the Cameron and Trivedi (2005) book.

```
1  *This code was simplified from Cameron and Trivedi (2005)'s
2  *Microeconometrics: Methods and Applications for demonstration
3
4  *step 1: estimate propensity score
5  logit dmmfd sexhead agehead educhead vaccss pciirr rice wheat milk oil [pw=weight]
6  predict PSCORE
7
8  *step 2: create common support
9  sum PSCORE if dmmfd==1
10 scalar PTMIN = r(min)
11 scalar PTMAX = r(max)
12
13 sum PSCORE if dmmfd==0
14 scalar PCMIN = r(min)
15 scalar PCMAX = r(max)
16
17 drop if PSCORE < PTMIN
18 drop if PSCORE < PCMIN
19 drop if PSCORE > PTMAX
20 drop if PSCORE > PCMAX
21
22 sum PSCORE
23
24 *Check the common support
25 graph twoway (kdensity PSCORE if dmmfd==1, lpattern(longdash)) ///
26 (kdensity PSCORE if dmmfd==0, lpattern(shortdash)), ///
27 legend(label(1 "Treatment") label(2 "Control")) ///
28 bgcolor(white) graphregion(color(white)) ///
29 ytitle("Density") xtitle("Estimated Propensity Score")
30
31 *divide PSCORE into blocks and check balances within blocks
32 global cut1 = 0.1
33 global cut2 = 0.2
34 global cut3 = 0.3
```

## Thực hành (7) - Advanced level

```
35
36 gen myblock = 1
37 replace myblock = 2 if PSCORE > $cut1 & PSCORE < $cut2
38 replace myblock = 3 if PSCORE > $cut2 & PSCORE < $cut3
39 replace myblock = 4 if PSCORE > $cut3
40
41 tab myblock
42 tab myblock dmmfd, sum(agehead) nostand nofreq
43
44 *check balances in covariates
45 foreach i of numlist 1 2 3 4 {
46     ttest agehead if myblock==`i', by(dmmfd)
47 }
48
49 *step 3: calculate AIT by stratification weighing
50 keep exptot dmmfd PSCORE myblock
51 sort myblock PSCORE dmmfd
52
53 local i=1
54 while `i' <= 4 {
55     reg exptot dmmfd PSCORE if myblock==`i'
56     di "BLOCK " `i'
57     di "ATT" `i' " = " _b[dmmfd]
58     di "SE " _se[dmmfd]
59     di "t-stat = " _b[dmmfd]/_se[dmmfd]
60     local ++i
61 }
62
63 *weighing ATT by block size
64 di 1522*33/1102
65
```