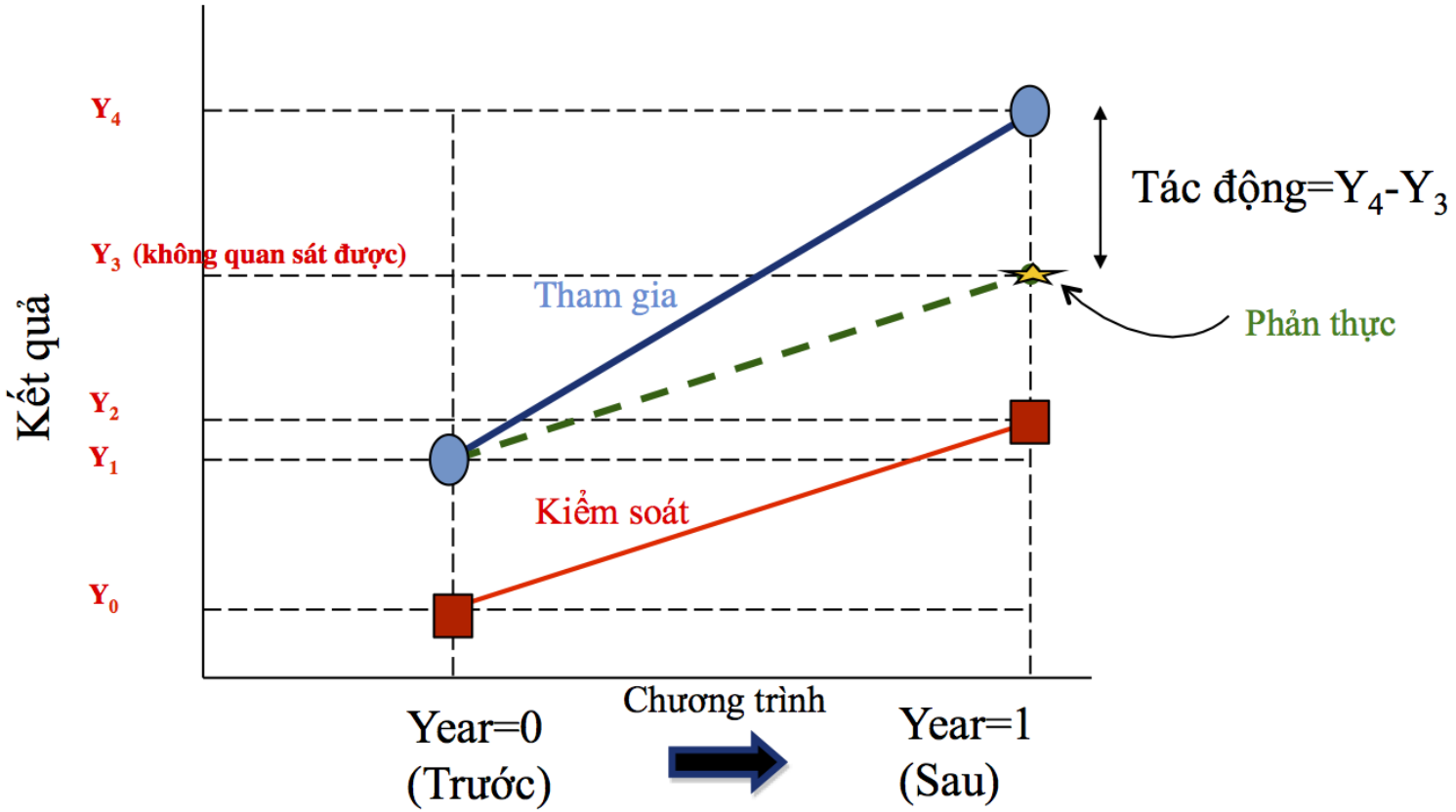
Mô hình Khác biệt Kép (Difference-in-Difference Method)

Lê Việt Phú
Chương trình Giảng dạy Kinh tế Fulbright

Ngày 17 tháng 5 năm 2016

Khung phân tích của phương pháp DiD



	Trước	**Sau**	**Thay đổi**
Đối chứng	Y_0	Y_2	$Y_2 - Y_0 = a$
Tham gia	Y_1	Y_4	$Y_4 - Y_1 = b$
			$\mathbf{DiD = Y_4 - Y_3 = b - a}$

Điều kiện áp dụng phương pháp DiD

- Dữ liệu bảng (với mỗi quan sát có dữ liệu trước và sau khi có chính sách).
- Giả định song song (parallel assumption): Nếu không có chương trình thì xu hướng thay đổi của nhóm tham gia và nhóm đối chứng là như nhau. Khi này có thể kết hợp hai nhóm tham gia và đối chứng để xây dựng phản thực.
  - Điều kiện này nới lỏng hơn rất nhiều so với điều kiện nhóm đối chứng hoàn toàn tương đồng với nhóm tham gia trong phương pháp mẫu ngẫu nhiên.
  - Có thể sử dụng nhóm tham gia và nhóm đối chứng 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.

Mô hình ước lượng tác động bằng DiD

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 dummy (nhận giá trị 0 và 1 cho 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 chính sách:

	Trước	**Sau**	**$\triangle$Y**
Đối chứng	$Y = \beta_0$	$Y = \beta_0 + \beta_2$	β_2
Tham gia	$Y = \beta_0 + \beta_1$	$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
			DiD = β_3

Ước lượng mô hình DiD

- Hình thức ước lượng DiD đơn giản nhất là dùng hồi quy dữ liệu gộp (pooled regression): Gộp các quan sát qua nhiều năm của các hộ gia đình thành một bảng dữ liệu chéo. Có thể sử dụng với bảng dữ liệu không cân bằng (một số hộ chỉ có quan sát đầu kỳ, hoặc cuối kỳ).
- Hồi quy dữ liệu bảng với tác động cố định (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.
- 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.

Thực hành

- STATA data file hh_9198_2016.dta
- STATA program code did.do file

Thực hành

Nghiên cứu cấu trúc file hh_9198.dta

- Dữ liệu dạng bảng dọc (long format): 826 hộ gia đình, mỗi hộ có quan sát trước (Year=0) và sau (Year=1) khi thực hiện chương trình.
- Biến chính sách: Có phụ nữ tham gia vay vốn (**dmmfd=1**).
- Biến phụ thuộc: Tổng chi tiêu của hộ (exptot).

HHid	**Year**	**Village**	**Treatment (T)**	$\mathbf{Y_i}$	$\mathbf{X_i}$
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 dữ liệu rất quan trọng đối với dữ liệu bảng do các phương pháp khác nhau yêu cầu tổ chức cấu trúc dữ liệu khác nhau!**

Hồi quy dữ liệu gộp - Pooled regression

Để ước lượng được phương trình hồi quy (1) bằng phương pháp gộp dữ liệu, cần tạo biến chính sách $\mathbf{T = 1}$ (với hộ có tham gia) và biến tương tác $\mathbf{T \times Year}$:

HHid	**Year**	**Village**	**T**	**T × 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}
...	...	...	...	...	...	...

- *reg Y T Year (T ∗ Year) X_i* $\Rightarrow$ Tác động của chính sách là hệ số của biến tương tác.
- 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

```
****Data preparation
gen lexptot=ln(1+exptot)
gen lnland=ln(1+hhland/100)
egen dmmfd98=max(dmmfd), by(nh)
gen dmmfdyr=dmmfd98*year
****Basic model
reg lexptot year dmmfd98 dmmfdyr
****Full model
reg lexptot year dmmfd98 dmmfdyr sexhead agehead educhead lnland
vaccess pcirr rice wheat milk oil egg [pw=weight]
```

Linear regression

Number of obs	= 1652
F(14, 1637)	= 23.72
Prob > F	= 0.0000
R-squared	= 0.2816
Root MSE	= .42796

lexptot	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	.3044236	.0605892	5.02	0.000	.1855832	.423264
dmmfd98	.0169245	.0369018	0.46	0.647	-.0554553	.0893042
dmmfdyr	-.0417365	.0672451	-0.62	0.535	-.1736319	.0901589
sexhead	-.0441215	.0542985	-0.81	0.417	-.1506234	.0623805
agehead	.0017764	.0010926	1.63	0.104	-.0003666	.0039195
educhead	.0382679	.0049256	7.77	0.000	.0286067	.047929
lnland	.2206753	.0302448	7.30	0.000	.1613527	.2799979

Hồi quy dữ liệu bảng - Regression with panel data

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 tố chất 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: Hồi quy dữ liệu bảng với tác động cố định - Panel data regression with fixed effects

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

$$Y_i = \beta_0 + \beta_1 * T_i + \beta_2 * Year + \beta_3 * X_i + \eta_i + \varepsilon_i$$

η_i là tác động cố định của mỗi hộ gia đình i, không quan sát được. β_1 là tác động của việc tham gia chính sách.

Hồi quy dữ liệu bảng - Regression with panel data

Hình thức ước lượng thứ hai: Hồi quy với biến giả - Least Square Dummy Variables (LSDV)

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

Các lệnh này sẽ ước lượng mô hình OLS sau, với (N-1) biến giả D_i đại diện cho N quan sát:

$$Y_i = \beta_0 + \beta_1 * T_i + \beta_2 * Year + \beta_3 * X_i + \sum_i \sigma_i * D_i + \varepsilon_i$$

Hồi quy dữ liệu bảng - Regression with panel data

Hình thức ước lượng thứ ba: Hồi quy với sai phân bậc nhất của các biến số - Regression with first differences

- Lấy sai phân bậc nhất của các biến qua thời gian đối với từng quan sát (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:

$$\Delta Y_i = \beta_0 + \beta_1 * \Delta T_i + \beta_2 * \Delta X_i + \mu_i$$

- Sử dụng lệnh *reg dY dT* dX_i với sai phân bậc nhất của các biến số được tạo ra.

Thực hành

```
****Panel data with fixed effects
xtreg lexptot year dmmfd sexhead agehead educhead lnland vaccess
pcirr rice wheat milk oil egg, fe i(nh)
****Alternatives: LSDV
areg lexptot year dmmfd sexhead agehead educhead lnland vaccess
pcirr rice wheat milk oil egg, a(nh)
reg lexptot year dmmfd sexhead agehead educhead lnland vaccess
pcirr rice wheat milk oil egg i.nh
```

```
Fixed-effects (within) regression               Number of obs      =      1652
Group variable: nh                              Number of groups   =       826

R-sq:  within  = 0.1653                         Obs per group: min =         2
       between = 0.1898                                        avg =       2.0
       overall = 0.1726                                        max =         2

                                                F(13,813)          =     12.39
corr(u_i, Xb)  = 0.1165                         Prob > F           =    0.0000
```

lexptot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	.2775453	.0592425	4.68	0.000	.161259	.3938315
dmmfd	.0050621	.0449344	0.11	0.910	-.0831389	.0932632
sexhead	-.0492994	.0728122	-0.68	0.499	-.1922216	.0936227
agehead	.0000624	.0016969	0.04	0.971	-.0032685	.0033933
educhead	.0130055	.0083204	1.56	0.118	-.0033265	.0293375
lnland	.1402108	.0621941	2.25	0.024	.0181308	.2622908

Thực hành

Hồi quy với sai phân bậc nhất.

```
***Reorganize the data from long to wide format
reshape wide villid-lnland, i(nh) j(year)
***Create first-differencing variables
gen dlexptot = lexptot1 - lexptot0
gen ddmmfd = dmmfd1 - dmmfd0
...
reg dlexptot ddmmfd dsexhead dagehead deduchead dlnland dvaccess
dpcirr drice dwheat dmilk doil degg
```

Source	SS	df	MS
Model	7.80610462	12	.650508718
Residual	216.357698	813	.266122629
Total	224.163802	825	.2717137

Number of obs	= 826
F(12, 813)	= 2.44
Prob > F	= 0.0040
R-squared	= 0.0348
Adj R-squared	= 0.0206
Root MSE	= .51587

dlexptot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ddmmfd	.0050621	.0449344	0.11	0.910	-.0831389	.0932632
dsexhead	-.0492994	.0728122	-0.68	0.499	-.1922216	.0936227
dagehead	.0000624	.0016969	0.04	0.971	-.0032685	.0033933
deduchead	.0130055	.0083204	1.56	0.118	-.0033265	.0293375
dlnland	.1402108	.0621941	2.25	0.024	.0181308	.2622908

Nhận xét

- Hồi quy dữ liệu gộp đơn giản, dễ thực hiện, nhưng không tận dụng tối đa khả năng có thể có của dữ liệu bảng.

- Hồi quy dữ liệu bảng với tác động cố định **xtreg, fe** là hiệu quả nhất. Nhưng nếu bảng dữ liệu không cân bằng thì một số quan sát sẽ bị loại bỏ $\Rightarrow$ Giảm cỡ mẫu $\Rightarrow$ Giảm khả năng kiểm định các giả thuyết thống kê.

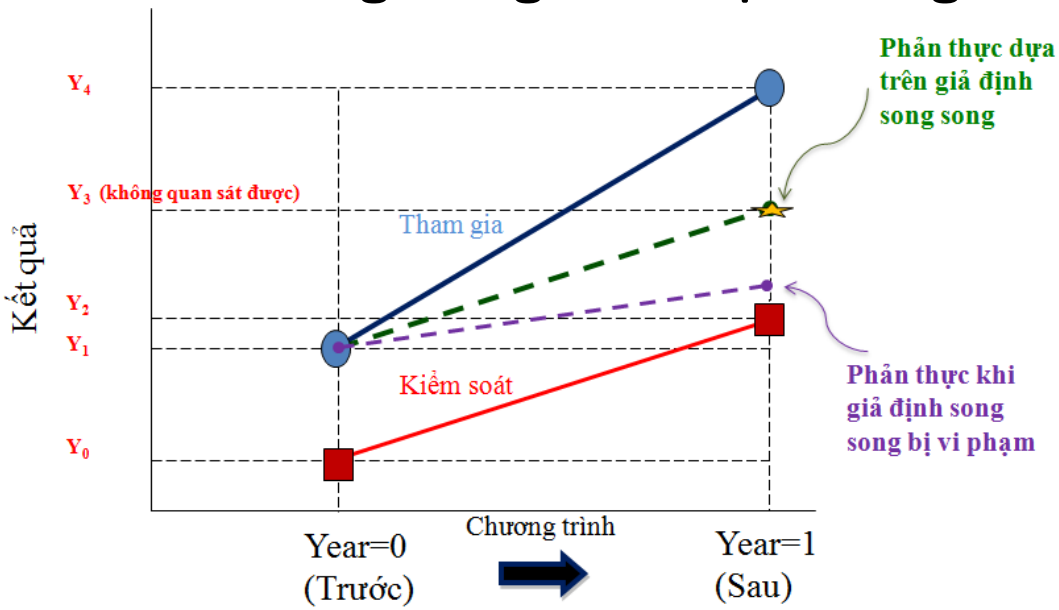
Mở rộng mô hình DiD

- Nếu giả định song song không đảm bảo $\Rightarrow$ sử dụng hồi quy DiD có tính đến điều kiện ban đầu.
- DiD kết hợp với PSM: Sử dụng PSM để lọc các quan sát có độ tương đồng cao (các quan sát nằm trong vùng hỗ trợ chung) trước khi chạy mô hình DiD có thể cải thiện kết quả của ước lượng.

DiD có tính đến điều kiện ban đầu

Phản chứng được xây dựng dựa trên giả định song song. Nếu giả định song song bị vi phạm ⇒ ước lượng có thể bị chệch trên hoặc dưới.

Ước lượng bị chệch dưới khi xu hướng tăng của nhóm tham gia thấp hơn nhóm đối chứng trong điều kiện không có chính sách.



- Ví dụ nếu không tham gia chính sách, tốc độ tăng trưởng thu nhập của nhóm tham gia (chủ yếu người đã có thu nhập cao) thấp hơn nhóm không tham gia (chủ yếu là người nghèo có thu nhập thấp) ⇒ DiD chệch dưới (ước lượng thấp hơn thực tế).

Thực hành

Mô hình hồi quy với sai phân bậc nhất của các biến số, có kiểm soát thêm điều kiện ban đầu $\mathbf{X_i}$:

$$\Delta Y_i = \beta_0 + \beta_1 * \Delta T_i + \beta_2 * \Delta X_i + \beta_{\mathbf{3}} * \mathbf{X_i} + \mu_i$$

Sử dụng lệnh *reg dY dT dX$_i$ X$_i$* với sai phân bậc nhất của các biến số được tạo ra và điều kiện ban đầu (quan sát X_i tại thời điểm $Year = 0$).

DiD kết hợp với PSM

- Ôn tập: PSM tìm ra nhóm đối 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 cải thiện ước lượng so với 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 (thời điểm $Year = 0$).
  - 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 hộ trợ chung và sử dụng các phương pháp ước lượng DiD trên mẫu dữ liệu đã chọn lọc này.

Thực hành

Bước 1: Ước lượng mô hình pscore với biến tham gia chính sách T (tại thời điểm $Year = 1$) dựa trên các điều kiện ban đầu ($Year = 0$).

```
*Reorganize the data from long to wide format
reshape wide villid-dmmfd, i(nh) j(year)
pscore dmmfd1 sexhead0 agehead0 educhead0 hhland0 vaccess0 pcirr0
rice0 wheat0 milk0 oil0 egg0, pscore(score) blockid(block) comsup
level(0.001)

*keep observations in common support
keep if comsup==1

*keep observation ID only
keep nh

*merge to the original dataset
merge nh using hh_9198_2016.dta

*keep only observations which matched the ID identified above
tab _merge
keep if _merge==3
drop _merge
```

Thực hành

Bước 2: Ước lượng mô hình DiD trên bộ dữ liệu đã lọc.

```
*Estimate DiD model with panel data and fixed effects
gen lexptot=ln(1+exptot)
gen lnland=ln(1+hhland/100)

xtreg lexptot year dmmfd sexhead agehead educhead lnland vaccess
pcirr rice wheat milk oil egg, fe i(nh)
```

```
Fixed-effects (within) regression               Number of obs      =      1500
Group variable: nh                              Number of groups   =       750

R-sq:  within  = 0.1751                         Obs per group: min =         2
       between = 0.1352                                        avg =       2.0
       overall = 0.1477                                        max =         2

                                                F(13,737)          =     12.04
corr(u_i, Xb)  = 0.0580                         Prob > F           =    0.0000
```

lexptot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	.2999704	.0616634	4.86	0.000	.1789135	.4210273
dmmfd	.015489	.0450185	0.34	0.731	-.0728908	.1038687
sexhead	-.143559	.0828466	-1.73	0.084	-.3062024	.0190844
agehead	-.0012272	.0018241	-0.67	0.501	-.0048083	.002354
educhead	.0092429	.0087175	1.06	0.289	-.0078712	.026357
lnland	.1828212	.0818227	2.23	0.026	.0221878	.3434546

So sánh các phương pháp đã học

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