Hồi quy với Dữ liệu Gộp và Dữ liệu Bảng (Regression with Pooled and Panel Data)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 26 tháng 3 năm 2018

Các loại cấu trúc dữ liệu

- Dữ liệu chéo (cross-sectional data)
- Dữ liệu chuỗi thời gian (time series data)
- Dữ liệu gộp (pooled cross-sectional data)
- Dữ liệu bảng (panel data)

Ứng dụng hồi quy dữ liệu bảng trong phân tích chính sách

- Có hai nhóm đối tượng nghiên cứu: một nhóm bị ảnh hưởng bởi chính sách (nhóm hưởng lợi - treatment group), một nhóm không (nhóm kiểm soát, nhóm đối chứng - control group).
- Cần ước lượng tác động của chính sách lên kết quả (thu nhập, chi tiêu) của nhóm hưởng lợi.

Giả sử mô hình cần ước lượng là:

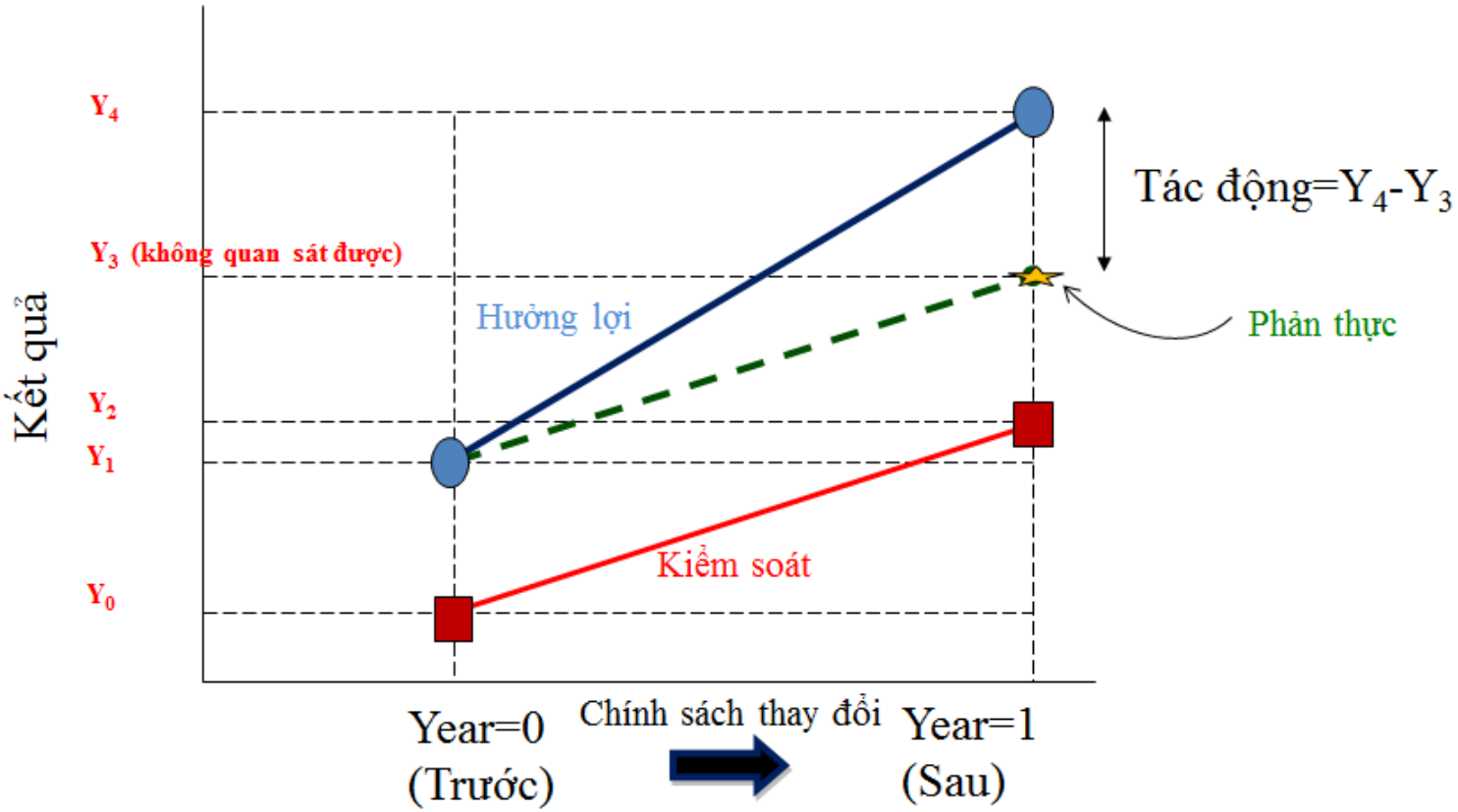
$$Y_{it} = \beta_0 + \beta_1 \times T_i + \gamma \times Year + \beta_j X_{it} + u_{it}$$

trong đó T là biến chính sách ($T = 1$ nếu thuộc nhóm hưởng lợi, $T = 0$ với nhóm kiểm soát); $Year$ là biến thời gian; Y là biến kết quả; X là các biến kiểm soát khác trong mô hình.

Ứng dụng hồi quy dữ liệu bảng trong phân tích chính sách

- Nếu chỉ có dữ liệu chéo, có ước lượng được β_1 không?
- Điều gì xảy ra với ước lượng của β_1?
- Khái niệm đánh giá tác động chính sách (program valuation), điều tra mẫu ngẫu nhiên (randomized data), dữ liệu bán thử nghiệm/thử nghiệm tự nhiên (natural/quasi-experiment).

Phương pháp hồi quy Diff-in-Diff (DiD) với dữ liệu bảng



	Trước	**Sau**	**Thay đổi**
Kiểm soát	Y_0	Y_2	$Y_2 - Y_0 = a$
Hưởng lợi	Y_1	Y_4	$Y_4 - Y_1 = b$

Ước lượng DiD $= (\mathbf{Y_4} - \mathbf{Y_1}) - (\mathbf{Y_2} - \mathbf{Y_0}) = \mathbf{Y_4} - \mathbf{Y_3}$

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 chính sách can thiệp thì xu hướng thay đổi của nhóm hưởng lợi và nhóm kiểm soát là như nhau.
  - Điều kiện này nới lỏng hơn rất nhiều so với điều kiện nhóm kiểm soát hoàn toàn tương đồng với nhóm hưởng lợi trong điều tra ngẫu nhiên (RCT).
  - Có thể sử dụng nhóm hưởng lợi 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 chính sách (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 chính sách bằng DiD

$$Y_{it} = \beta_0 + \beta_1 * T_{it} + \beta_2 * Year_t + \beta_3 * (T \times Year) + \beta_j * X_{it} + u_{it}$$

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_j là các đặc tính của hộ gia đình (tạm thời bỏ qua).
- β_3 **là ước lượng tác động trung bình của việc tham gia chính sách (Average Treatment Effect - ATE).**

	Trước	**Sau**	**$\triangle$Y**
Kiểm soát	$Y = \beta_0$	$Y = \beta_0 + \beta_2$	β_2
Hưởng lợi	$Y = \beta_0 + \beta_1$	$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
			DiD $= \beta_3$

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.

Thực hành

Ước lượng tác động của chính sách cho vay tín dụng vi mô (microfinance) đến chi tiêu của hộ gia đình ở Bangladesh.

- STATA data file hh_9198_2018.dta
- STATA program code did.do file

Nghiên cứu cấu trúc dữ liệu:

- 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 (**dfmfd=1**) hoặc nam giới vay vốn (**dmmfd=1**).
- Biến kết quả: Tổng chi tiêu của hộ (exptot).
- Giả sử chúng ta muốn ước lượng mô hình hồi quy sau:

$$log(exptot_{it}) =$$

$$\beta_0 + \beta_1 * dmmfd_{it} + \beta_2 * Year_t + \beta_{\mathbf{3}} * (dmmfd_{it} \times Year_t) + \beta_j X_{it} + u_{it}$$

với X_{it} là đặc tính của hộ gia đình.

HHid	Year	Village	Treatment (T)	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 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!**

Phương pháp hồi quy dữ liệu gộp để ước lượng tác động DiD

Để ước lượng bằng phương pháp gộp dữ liệu, cần tạo biến chính sách $\mathbf{T = 1}$ (với hộ hưởng lợi) 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* ⇒ 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.

Kết quả ước lượng bằng hồi quy gộp

```
****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]
```

Nhận xét

- Bản chất của hồi quy dữ liệu gộp tương tự như hồi quy dữ liệu chéo.
- Các giả định của mô hình CLRM vẫn cần thiết. Nếu vi phạm $\Rightarrow$ ước lượng bị chệch hoặc không nhất quán.
- Chưa tận dụng tối đa khả năng của dữ liệu bảng (quan sát lặp qua thời gian) cho phép vi phạm giả định về tương quan giữa phần dư với biến chính sách.

Hồi quy dữ liệu bảng - Regression with panel data

Giả sử mô hình hồi quy với tác động cố định không quan sát được a_i được viết dưới dạng:

$$Y_{it} = \beta_0 + \beta_1 * T_{it} + \beta_2 * Year_t + \beta_j * X_{it} + a_i + u_{it}$$

a_i không thay đổi qua thời gian đối với quan sát i (time invariant unobserved heterogeneity), ví dụ tính cách, quan hệ xã hội, tố chất cá nhân không thay đổi theo thời gian, và có thể có ảnh hưởng đến quyết định tham gia chính sách cũng như tác động của chính sách.

- Do a_i không quan sát được nên a_i sẽ bị gom chung vào phần dư gộp của mô hình $(a_i + u_{it})$.
- Nếu a_i tương quan với biến chính sách T_i (người có quan hệ tốt có khả năng vay vốn tốt hơn) $\Rightarrow$ ước lượng của β_1 sẽ bị chệch lên do tương quan dương giữa phần dư gộp với biến chính sách.

$\Rightarrow$ Hồi quy dữ liệu bảng với tác động cố định - Panel data regression with fixed effects - có thể xử lý được vấn đề tác động cố định tương quan với biến chính sách.

- Thực hiện chuyển đổi loại trừ giá trị trung bình (time-demeaned tranformation) và ước lượng dựa trên bộ dữ liệu chuyển đổi:

$$\ddot{Y}_{it} = \beta_1 * \ddot{T}_{it} + \beta_2 * \ddot{Year}_t + \beta_j * \ddot{X}_{it} + \ddot{u}_{it} \qquad (1)$$

  trong đó $\ddot{Y}_{it} = Y_{it} - \bar{Y}_i$... (lấy giá trị quan sát được trừ đi giá trị trung bình, áp dụng đối với từng hộ gia đình).
- Tác động cố định a_i sẽ bị loại khỏi mô hình (1).
- Ước lượng β_1 bằng OLS.

Cách thực hiện

- **Hồi quy với tác động cố định (Fixed Effects Regression):**
  xtreg Y T Year X, fe i(id)
  với id là mã hộ gia đình.
- **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 dữ liệu gộp OLS với (N-1) biến giả D_j đại diện cho N hộ gia đình. β_1 là tác động của chính sách.

$$Y_{it} = \beta_0 + \beta_1 * T_{it} + \beta_2 * Year_t + \beta_j * X_{it} + \sum_j \sigma_j * D_j + u_{it}$$

- **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:

$$\Delta Y_i = \beta_2 + \beta_1 * \Delta T_i + \beta_j * \Delta X_i + u_i$$

  với $\Delta Y_i = Y_1 - Y_0$...

- 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
```

Thực hành

```
***regression with first differences
***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
```

DiD có tính đến điều kiện ban đầu

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

$$\Delta Y_i = \beta_2 + \beta_1 * \Delta T_i + \beta_j * \Delta X_i + \beta_\mathbf{k} * \mathbf{X_i^0} + u_i$$

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 $Year = 0$).

Nhận xét ưu nhược điểm của các hình thức ước lượng

- 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. Nếu dữ liệu bị thiếu một cách hệ thống (systematic attrition) $\Rightarrow$ mô hình có thể bị chệch.

- Có thể sử dụng sai phân bậc nhất để loại bỏ những nhấn tố không thay đổi theo thời gian, hoặc hồi quy với biến giả để kiểm soát các tác động cố định.

- Hình thức ước lượng ảnh hưởng đến tính chính xác của kết quả (độ lệch chuẩn của ước lượng).

Hồi quy dữ liệu bảng - Nâng cao

Mô hình tổng quát

$$Y_{it} = \beta_j * X_{it} + a_i + u_{it} \tag{2}$$

- với a_i là tác động cố định, đặc trưng cho từng quan sát i, và không quan sát được. a_i khác nhau giữa các hộ/cá nhân nhưng trong cùng một hộ/cá nhân, đặc trưng này không thay đổi theo thời gian.
- Lấy trung bình đối với từng quan sát theo thời gian, ta có phương trình:

$$\overline{Y_i} = \beta_j * \overline{X_i} + a_i + \overline{u_i} \tag{3}$$

- Ước lượng các tham số dựa trên mô hình (3) được gọi là **between estimator** (so sánh giữa các quan sát với nhau).

Lấy phương trình (2) trừ đi phương trình (3), do nhân tố cố định không đổi nên nó sẽ bị loại:

$$Y_{it} - \overline{Y_i} = \beta_j * (X_{it} - \overline{X_i}) + (u_{it} - \overline{u_i}) \tag{4}$$

viết gọn lại thành:

$$\ddot{Y}_{it} = \beta_j * \ddot{X}_{it} + \ddot{u}_{it} \tag{5}$$

với các giá trị $\ddot{Y}_{it}$, $\ddot{X}_{it}$ bằng giá trị quan sát được trừ đi giá trị trung bình đối với từng quan sát (còn gọi là chuyển đổi bên trong - within transformation, time-demeaned transformation).

- Ước lượng của mô hình (5) được gọi là **ước lượng tác động cố định**, **within estimator**, hay fixed-effects estimator (thu được thông qua so sánh nội tại cùng một quan sát).
- *xtreg Y T Year X, fe i(id)*

Hồi quy tác động ngẫu nhiên - random effects (RE) regression

- Giả sử tác động không quan sát a_i được không tương quan với biến chính sách và các biến giải thích X_i trong mô hình (2):

$$cov(a_i, X_{it}) = 0,$$

khi này, ước lượng bằng FE là không tối ưu (làm mất thông tin và giảm số bậc tự do).

- Áp dụng mô hình RE trong trường hợp này:

$$Y_{it} = \beta_j * X_{it} + v_{it} \tag{6}$$

với $v_{it} = a_i + u_{it}$ là phần dư gộp (composite error term).

- Ước lượng OLS của mô hình (5) sẽ không là BLUE do các phần dư tương quan chuỗi với nhau:

$$cov(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2},$$

Ước lượng mô hình tác động ngẫu nhiên

Sử dụng phương pháp GLS (generalized least square) để sử lý vấn đề tương quan chuỗi:

- Chuyển đổi bộ dữ liệu bằng hệ số θ,

$$\theta = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)]^{1/2}$$

  θ luôn dương và nhỏ hơn 1. θ phản ánh mức độ quan trọng tương đối của tác động cố định so với phần dư của mô hình thông qua phương sai σ_a^2 và σ_u^2.

- Và ước lượng mô hình sau bằng OLS:

$$Y_{it} - \theta\overline{Y_i} = \beta_j * (X_{it} - \theta\overline{X_i}) + (v_{it} - \overline{v_i}) \qquad (7)$$

- Stata: *xtreg Y T Year X, re i(id)*
  với id là mã hộ gia đình.

Thực hành

- Ước lượng mô hình random effects với bộ dữ liệu microfinance.
- So sánh kết quả với hồi quy pooled OLS và fixed effects.
- Kiểm định Hausman để lựa chọn mô hình. Kiểm định Hausman kiểm tra sự khác biệt mang tính hệ thống giữa hai ước lượng.
  - Bác bỏ $H_0 \Rightarrow$ ước lượng RE khác với ước lượng FE $\Rightarrow$ sử dụng ước lượng FE.
  - Không bác bỏ $H_0 \Rightarrow$ sử dụng ước lượng RE.

So sánh pooled OLS, fixed effects và random effects

Bản chất của ước lượng RE là kết hợp giữa pooled OLS với FE thông qua quyền số θ:

- Nếu $\theta \to 0$ (ảnh hưởng của tác động cố định nhỏ hơn nhiều so với phần dư) thì ước lượng RE tương tự như pooled OLS.
- Nếu $\theta \to 1$ (ảnh hưởng của tác động cố định lớn hơn nhiều so với phần dư) thì ước lượng RE sẽ tiệm cận ước lượng FE.
- Lựa chọn mô hình nào tùy thuộc vào lý thuyết nền tảng, dữ liệu và kiểm định.
  - Nếu tác động cố định tương quan với biến giải thích thì chọn mô hình FE. Nếu không thì chọn mô hình RE.
  - Áp dụng sai trường hợp sẽ dẫn đến hậu quả: áp dụng FE sai dẫn đến ước lượng không hiệu quả. Áp dụng RE sai dẫn đến ước lượng không nhất quán.