Hồi quy với Dữ liệu Bảng (Regression with Panel Data)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 24 tháng 3 năm 2019

Khái niệm 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)

Trường hợp mô hình hồi quy không có hiệu lực nội tại do thiếu biến quan trọng

- Ví dụ mô hình hồi quy tỷ suất thu nhập của đi học với hai biến giải thích số năm đi học (*educ*) và tố chất cá nhân (*Ability*):

$$log(income_i) = \beta_0 + \beta_1 educ_i + \beta_2 Ability_i + u_i$$

thỏa các điều kiện CLRM. i đại diện cho quan sát thứ i trong mẫu gồm có N quan sát.

- Tuy nhiên không quan sát được *Ability*, do đó chúng ta sẽ ước lượng mô hình sau trên thực tế:

$$log(income_i) = \beta_0 + \beta_1 educ_i + \underbrace{\beta_2 Ability_i + u_i}_{v_i}$$

Trong đó v_i là sai số gộp của cả sai số ngẫu nhiên u_i và biến không quan sát được $Ability_i$, $v_i = u_i + \beta_2 Ability_i$

Đánh giá hướng chệch trong mô hình thiếu biến quan trọng

Các đặc tính của ước lượng của $\hat{\beta}_1$:

$$\hat{\beta}_1 = \beta_1 + \beta_2 \sigma_{21}$$

σ_{21} là hệ số góc của hồi quy biến *Ability* lên *educ*:

$$\sigma_{21} = \frac{cov(educ, Ability)}{var(educ)}$$

- Nếu $\beta_2 = 0$ (biến *Ability* không phải là biến quan trọng) thì $\hat{\beta}_1$ không chệch.
- Nếu $\sigma_{21} = 0$ (*educ* và *Ability* không tương quan) thì $\hat{\beta}_1$ cũng không chệch.
- Nếu không phải 2 trường hợp trên thì β_1 chệch, với hướng và mức độ chệch tùy thuộc vào giá trị của β_2 và tương quan giữa biến *educ* và biến không quan sát được *Ability* thông qua hệ số σ_{21}.

Ước lượng bị thiên lệch do thiếu biến quan trọng - Omitted variables bias

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

- Tố chất cá nhân *Ability* được kỳ vọng có tác động đến tiền lương.
- Tố chất cá nhân tương quan với trình độ học vấn.
- Tố chất cá nhân không quan sát được.
- Kỳ vọng $\beta_2 > 0$ và $\sigma_{21} > 0 \Rightarrow$ Ước lượng tỷ suất thu nhập của đi học có khả năng bị chệch lên.

Sử dụng dữ liệu bảng để khắc phục vấn đề thiếu biến quan trọng không quan sát được

Với dữ liệu bảng, chúng ta có thể viết hàm hồi quy dữ liệu bảng như sau:

$$log(income_{it}) = \beta_0 + \beta_1 educ_{it} + \beta_2 Ability_{it} + \gamma t + u_{it}$$

với ký hiệu it đại diện cho quan sát thứ i tại năm quan sát t.

- γ là thay đổi thu nhập trung bình theo thời gian.

Trường hợp đơn giản nhất, ví dụ chúng ta có quan sát tại hai thời điểm, $t = 0$ và $t = 1$. Với giả định rằng tố chất cá nhân không thay đổi theo thời gian, khi đó hàm hồi quy có thể viết lại như sau:

$$log(income_{i0}) = \beta_0 + \beta_1 educ_{i0} + \beta_2 Ability_i + u_{i0} \quad (1)$$

$$log(income_{i1}) = \beta_0 + \beta_1 educ_{i1} + \beta_2 Ability_i + \gamma + u_{i1} \quad (2)$$

Lấy (2) trừ (1):

$$[log(income_{i1}) - log(income_{i0})] = \beta_1 [educ_{i1} - educ_{i0}] + \gamma + [u_{i1} - u_{i0}]$$

Khi đó, hàm hồi quy dựa trên sai phân của các biến giải thích có thể được viết dưới dạng sau:

$$\Delta log(income_i) = \gamma + \beta_1 \Delta educ_i + \Delta u_i \tag{3}$$

- Phương trình hồi quy sử dụng sai phân không còn biến *Ability*
- Giả sử $\Delta educ_i$ và Δu_i không tương quan, khi đó chúng ta có thể ước lượng β_1 bằng hồi quy OLS với phương trình (3).
- **Tên gọi: chuyển đổi sai phân bậc nhất với dữ liệu (first-differencing transformation) dùng để tạo ra ước lượng sai phân bậc nhất (first-differencing estimator) hoặc ước lượng khác biệt trong khác biệt (difference-in-difference, hoặc diff-in-diff estimator).**

Ví dụ ước lượng diff-in-diff

Sử dụng bộ dữ liệu energy.dta để ước lượng hàm sản xuất theo mô hình KLEM của 5,000 doanh nghiệp ở Việt Nam trong hai năm 2015-16.

$$logQ = \beta_0 + \beta_1 lnK + \beta_2 lnL + \beta_3 lnE + \beta_4 lnM + \gamma t + u$$

- Nếu mô hình trên bị thiếu biến quan trọng thì ước lượng của một hoặc tất cả các tham số bị chệch và không nhất quán.
- Nếu nhân tố không quan sát được không thay đổi theo thời gian (ví dụ đặc tính chủ doanh nghiệp, loại hình kinh doanh, vị trí địa lý, cơ sở hạ tầng...) thì chúng ta có thể sử dụng ước lượng với sai phân bậc nhất để sử lý vấn đề thiếu biến:

$$\Delta logQ = \gamma + \beta_1 \Delta lnK + \beta_2 \Delta lnL + \beta_3 \Delta lnE + \beta_4 \Delta lnM + v$$

Lưu ý với ước lượng diff-in-diff (DiD)

- Các biến không thay đổi theo thời gian sẽ bị loại bỏ khi thực hiện lấy sai phân bậc nhất. Do đó, không thể dùng mô hình Diff-in-Diff để ước lượng tác động của các nhân tố cố định đến biến phụ thuộc. Ví dụ giới tính, vị trí nơi ở, cơ sở hạ tầng (trong ngắn hạn), trình độ học vấn của những người đã kết thúc quá trình học hành...
- Phương pháp DiD dẫn đến giảm số lượng quan sát trong mô hình:
  - Biến sai phân làm giảm số lượng quan sát gốc.
  - Chỉ sử dụng quan sát có dữ liệu cả hai kỳ. Các quan sát chỉ có dữ liệu ở một kỳ sẽ bị loại bỏ.

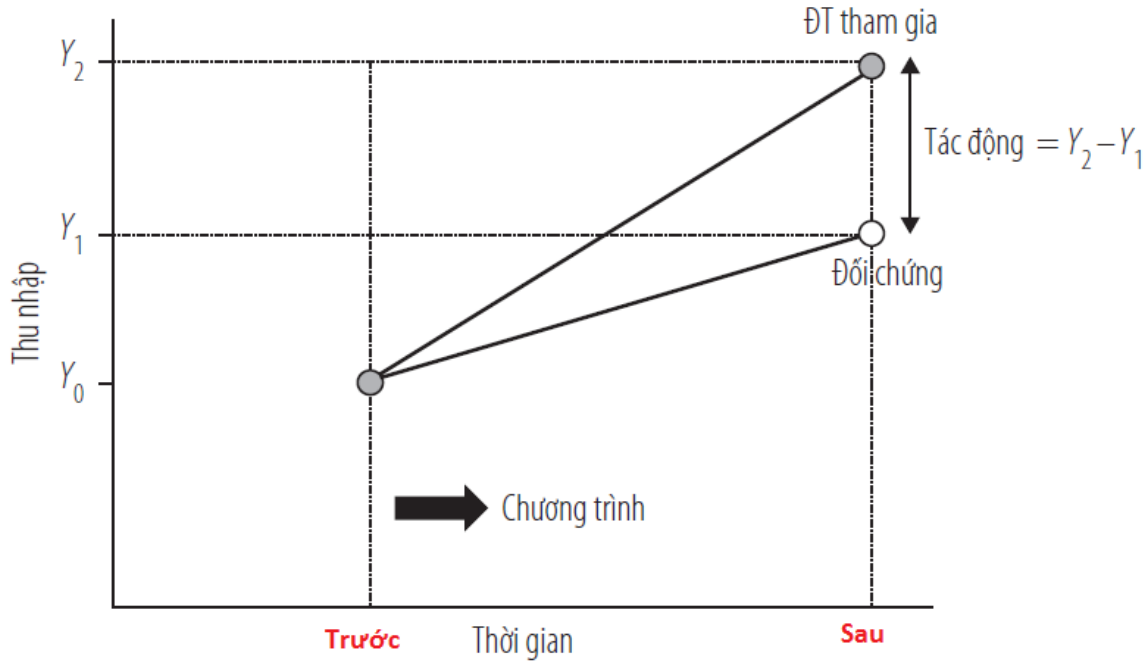
Ứng dụng phương pháp DiD trong phân tích tác động chính sách

- Các bài toán đánh giá tác động của chính sách thường bắt đầu bằng 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).
- Chính sách hay một can thiệp nào đó chỉ được thực hiện với nhóm hưởng lợi.
- Sau khi chính sách được thực hiện, chính phủ cần đánh giá tác động của chính sách để biết liệu chính sách có đạt hiệu quả kinh tế xã hội hay không so với chi phí bỏ ra.

- Tác động của chính sách được định nghĩa là sự khác biệt giữa kết quả thực so với kết quả đáng lẽ đã xảy ra nếu không có chính sách.
  - Không phải là khác biệt của biến phụ thuộc giữa hai nhóm hưởng lợi và không hưởng lợi!
- Kết quả đáng lẽ đã xảy ra gọi là phản thực hay phản chứng (counterfactual). Chúng ta không quan sát được phản chứng.
- Cách thức đánh giá tùy thuộc vào thiết kế của chính sách trước khi thực hiện và mức độ thu thập dữ liệu. Dữ liệu có thể bao gồm cả dữ liệu trước khi thực hiện chính sách và sau khi hoàn thành, hoặc chỉ có dữ liệu sau khi hoàn thành.

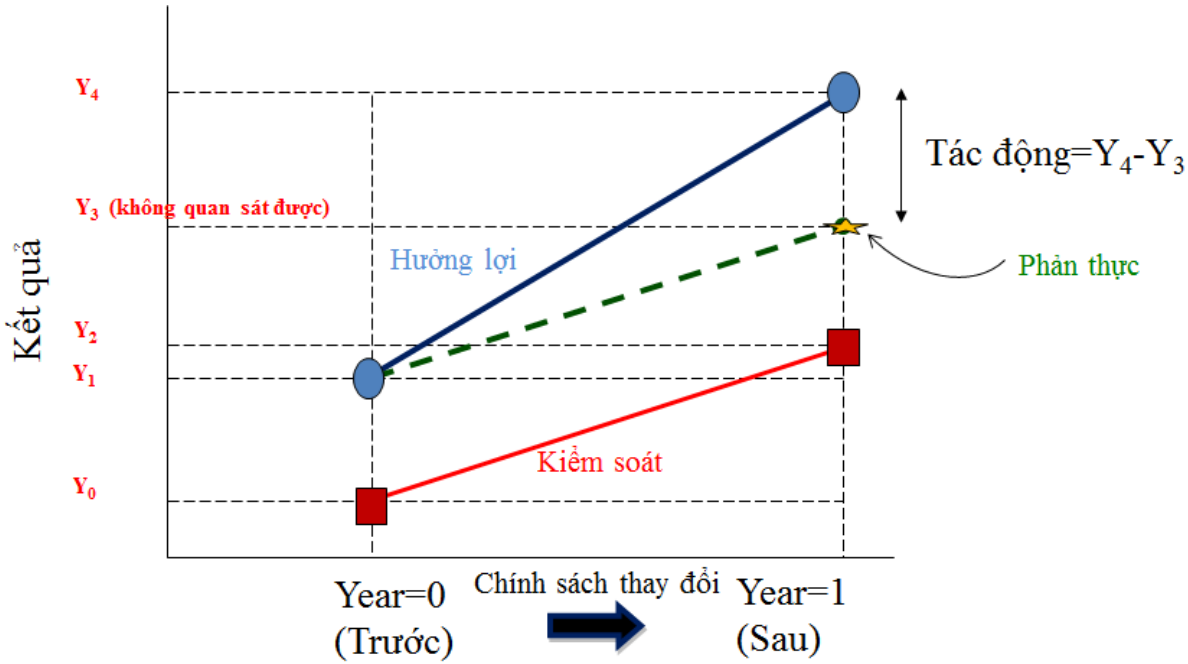
$$Impact = Y_{real} - Y_{counterfactual}$$

Trường hợp chuẩn - Thiết kế mẫu ngẫu nhiên trước khi thực hiện chương trình (RCT)



- Nếu nhóm đối chứng hoàn toàn tương đồng với nhóm hưởng lợi thì khác biệt về kết quả giữa 2 nhóm sau khi thực hiện chính sách là tác động của chính sách can thiệp.
- Yêu cầu thiết kế mẫu đảm bảo việc tham gia chính sách là hoàn toàn ngẫu nhiên và các đặc tính của hai nhóm đối tượng hoàn toàn giống nhau.

Sử dụng DiD khi hai nhóm có sự khác biệt



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

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

Mô hình ước lượng tác động chính sách bằng DiD

Tác động của chính sách có thể được ước lượng bằng mô hình sau:

$$Y = \beta_0 + \beta_1 * T + \beta_2 * Year + \beta_{\mathbf{3}} * (T \times Year) + \beta_j * X + u$$

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 ($Year = 0$ trước khi thực hiện chính sách và $Year = 1$ sau khi kết thúc).
- Y là biến kết quả; X là các biến giải thích khác trong mô hình (tạm thời bỏ qua).

$$Y = \beta_0 + \beta_1 * T + \beta_2 * Year + \beta_3 * (T \times Year) + u$$

	Trước ($Year = 0$)	**Sau** ($Year = 1$)	$\triangle$**Y**
Đối chứng ($T = 0$)	$Y = \beta_0$	$Y = \beta_0 + \beta_2$	β_2
Hưởng lợi ($T = 1$)	$Y = \beta_0 + \beta_1$	$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
			DiD $= \beta_3$

β_3 là ước lượng tác động trung bình của việc tham gia chính sách (Average Treatment Effect - ATE).

Điều kiện áp dụng phương pháp DiD để đánh giá tác động chính sách

- Dữ liệu bảng – nhưng không nhất thiết phải cân bằng!
- 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.

Các hình thức ước lượng mô hình DiD

- Hình thức ước lượng DiD đơn giản nhất là dùng hồi quy OLS với dữ liệu gộp (pooled regression). Tác động của chính sách là tham số của biến tương tác $T * Year$.

$$reg\ Y\ T\ Year\ (T * Year)\ X$$

- 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 do vấn đề lựa chọn mẫu.

Thực hành

Sử dụng bộ dữ liệu microcredit.dta để ước lượng tác động của chính sách cho vay tín dụng vi mô (microfinance) đến chi tiêu của hộ gia đình ở Bangladesh.

- 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 $treat = 1$ nếu hộ có tham gia vay vốn.
- Biến kết quả: Tổng chi tiêu của hộ (exptot).

Chúng ta cần ước lượng mô hình hồi quy sau:

$$log(exptot_{it}) = \beta_0 + \beta_1 * treat_{it} + \beta_2 * Year_t + \beta_3 * (treat_{it} \times Year_t) + \beta_j X_{it} + u_{it}$$

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

Cách thức tổ chức dữ liệu bảng

Bảng dọc (long format)

HHid	**Year**	**Treatment (T)**	$\mathbf{Y_i}$	$\mathbf{X_i}$
1	0	1	y_{10}	x_{10}
1	1	1	y_{11}	x_{11}
2	0	0	y_{20}	x_{20}
2	1	0	y_{21}	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!

Nhận xét với hồi quy dữ liệu gộp

- 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 với tác động cố định - Panel data regression with fixed effects

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

$$Y_{it} = \beta_0 + \beta_1 * T_{it} + \beta_2 * Year_t + \beta_j * X_{it} + \underbrace{a_i + u_{it}}_{v_{it}} \quad (4)$$

a_i không thay đổi qua thời gian đối với các quan sát trong cùng một hộ gia đình i (time invariant unobserved heterogeneity), ví dụ tính cách, quan hệ xã hội, tố chất cá nhân, giới tính chủ hộ không thay đổi theo thời gian.

- Do a_i không quan sát được nên a_i sẽ bị gom chung vào phần dư gộp của mô hình ($v_{it} = a_i + u_{it}$).
- Nếu a_i tương quan dương 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.

Hồi quy dữ liệu bảng với tác động cố định 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):

$$\ddot{Y}_{it} = \beta_1 * \ddot{T}_{it} + \beta_2 * \ddot{Year}_t + \beta_j * \ddot{X}_{it} + \ddot{u}_{it} \qquad (5)$$

  trong đó $\ddot{Y}_{it} = Y_{it} - \bar{Y}_i$... (lấy giá trị quan sát được trừ đi giá trị trung bình của từng hộ gia đình).
- Tác động cố định a_i sẽ bị loại khỏi mô hình (5).
- Ước lượng mô hình (5) bằng OLS sẽ cho kết quả β_1 không chệch.

Các hình thức thực hiện

1. **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.
2. **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.

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

3. **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 $\Delta Y_i = Y_{i1} - Y_{i0}$...

reg dY dT dX_i với sai phân bậc nhất của các biến số được tạo ra.

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 X_i^0 (quan sát X_i tại thời điểm $Year = 0$).
- Có thể áp dụng để kiểm định tính vững của giả định song song.
- Cần tổ chức dữ liệu để ghép dữ liệu sai phân với điều kiện ban đầu.

Thực hành với bộ dữ liệu microcredit.dta

- Viết phương trình hồi quy.
- So sánh các loại ước lượng.
- Diễn giải ý nghĩa.

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 do vấn đề lựa chọn mẫu.

- Có thể sử dụng **hồi quy 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.

- Các phương pháp trên không nhất thiết ra kết quả giống nhau. Khi dữ liệu chỉ có hai kỳ quan sát, và cân bằng, thì các phương pháp sẽ cho kết quả tương đồng.

Hồi quy dữ liệu bảng - Nâng cao

Mô hình tổng quát của hồi quy dữ liệu bảng

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

- 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:

$$\overline{Y_i} = \beta_j * \overline{X_i} + a_i + \overline{u_i} \tag{7}$$

- Ước lượng các tham số dựa trên mô hình (7) được gọi là **between estimator** (ước lượng dựa vào sự khác biệt giữa các hộ gia đình với nhau về mặt trung bình).

Lấy phương trình (6) trừ đi phương trình (7), do nhân tố cố định a_i không đổi nên nó sẽ bị loại:

$$Y_{it} - \overline{Y_i} = \beta_j * (X_{it} - \overline{X_i}) + (u_{it} - \overline{u_i}) \tag{8}$$

viết gọn lại thành:

$$\ddot{Y}_{it} = \beta_j * \ddot{X}_{it} + \ddot{u}_{it} \tag{9}$$

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 hộ gia đình (còn gọi là chuyển đổi bên trong - within transformation, time-demeaned transformation).

- Ước lượng của mô hình (9) được gọi là **ước lượng tác động cố định**, **within estimator/fixed-effects estimator** (ước lượng dựa vào biến động nội tại cùng một hộ gia đình).

Hồi quy tác động ngẫu nhiên - random effects regression

- Giả sử tác động cố định không quan sát được a_i không tương quan với biến chính sách và các biến giải thích X_i trong mô hình (6):

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

khi này, ước lượng bằng fixed-effects là không tối ưu do chuyển đổi dữ liệu làm mất thông tin và giảm số bậc tự do.

- Áp dụng mô hình random-effects trong trường hợp này:

$$Y_{it} = \beta_j * X_{it} + v_{it} \tag{10}$$

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

- Ước lượng (10) bằng OLS 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 hồi quy với quyền số GLS (generalized least square) để sử lý vấn đề tương quan chuỗi:

- Chuyển đổi bộ dữ liệu bằng hệ số θ,

$$\theta = 1 - \sqrt{\frac{\sigma_u^2}{(\sigma_u^2 + T\sigma_a^2)}}$$

  - 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:

$$Y_{it} - \theta\overline{Y_i} = \beta_j * (X_{it} - \theta\overline{X_i}) + (v_{it} - \theta\overline{v_i}) \qquad (11)$$

Thực hành

- Ước lượng và so sánh mô hình pooled OLS, fixed effects, và random effects với bộ dữ liệu microcredit.
- Kiểm định Hausman kiểm tra sự khác biệt mang tính hệ thống giữa hai ước lượng FE/RE và lựa chọn mô hình phù hợp nhất.
  - 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 sẽ dẫn đến hậu quả nghiêm trọng: Áp dụng FE sai dẫn đến ước lượng không hiệu quả; Áp dụng RE sai dẫn đến ước lượng không nhất quán.