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

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

06/03/2020

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 xu hướng 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:

$$\Delta log(income_i) = \gamma + \beta_1 \Delta educ_i + \Delta u_i \qquad (3)$$

- 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).**

Thực hành ước lượng hàm sản xuất của doanh nghiệp với bốn yếu tố đầu vào trong mô hình KLEM

Sử dụng bộ dữ liệu energy.dta 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 xử lý vấn đề thiếu biến:

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

- So sánh kết quả ước lượng bằng pooled OLS và DiD.

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...
- Ước lượng tác động của các yếu tố ít thay đổi cũng thiếu chính xác.
- 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 $\Rightarrow$ Cảnh giác với dữ liệu bị mất/thiếu và quá trình lựa chọn mẫu có thể làm sai lệch kết quả!

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

- Mục tiêu của đánh giá tác động chính sách nhằm xác lập liệu chính sách can thiệp có tạo ra tác động hay không lên đối tượng hưởng lợi.
- Chính sách can thiệp được áp dụng lên một nhóm đối tượng tại một thời điểm.
  - Một nhóm bị ảnh hưởng hay được hưởng lợi từ chính sách, gọi là nhóm hưởng lợi (treatment group).
  - Một nhóm không bị ảnh hưởng bởi chính sách, được gọi là nhóm kiểm soát hoặc nhóm đối chứng (control group).

- Tác động của chính sách được định nghĩa là sự khác biệt giữa kết quả sau khi thực hiện chính sách so với kết quả **đáng lẽ đã xảy ra** nếu không có chính sách.

- Kết quả đáng lẽ đã xảy ra gọi là phản thực hay phản chứng (counterfactual). Chúng ta không bao giờ quan sát được phản chứng.

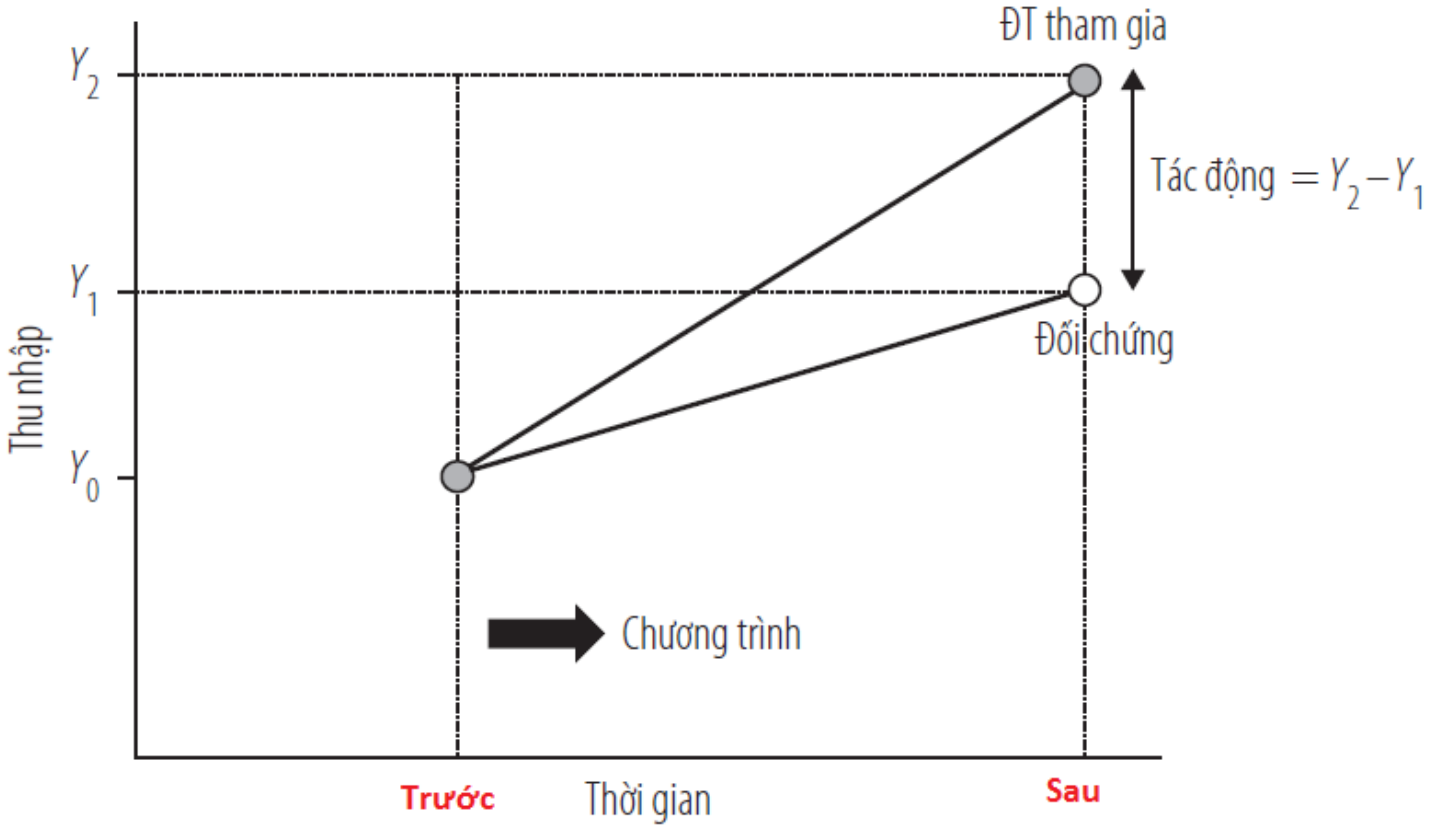
$$Impact = Y_{real} - Y_{counterfactual}$$

  - Lưu ý tác động không phải là khác biệt giữa hai nhóm hưởng lợi và kiểm soát.
  - Không phải là sự khác biệt trước và sau khi thực hiện chính sách.

- Do đó, trọng tâm của việc đánh giá tác động chính sách là sử dụng các thiết kế nghiên cứu để **ước lượng phản thực**.

- Tùy vào cách thức thực hiện, độ phức tạp, khả năng thu thập dữ liệu, chi phí và yêu cầu về độ tin cậy mà dữ liệu có thể bao gồm cả dữ liệu trước và sau khi thực hiện chính sách, hoặc chỉ có dữ liệu sau khi thực hiện chính sách.

Tiêu chuẩn vàng: Đánh giá tác động chính sách bằng thiết kế mẫu ngẫu nhiên (Randomized Controlled Trial - RCT)



$$Impact = Y_{treatment} - Y_{control}$$

Đánh giá tác động chính sách bằng thiết kế mẫu ngẫu nhiên

- Dựa vào thiết kế đảm bảo nhóm đối chứng hoàn toàn tương đồng với nhóm hưởng lợi trước khi thực hiện chương trình.
- Khi này, sử dụng nhóm đối chứng làm counterfactual, và khác biệt về kết quả giữa hai nhóm sau khi thực hiện chính sách chính là tác động của chính sách can thiệp.
- Yêu cầu khắt khe việc thiết kế mẫu đảm các đặc tính của hai nhóm đối tượng hoàn toàn tương đồng (tham gia chính sách là hoàn toàn ngẫu nhiên, không có quá trình tự lựa chọn mẫu khi tham gia chương trình, hai nhóm tương đồng nhau về các đặc tính quan sát được và không quan sát được).

$\Rightarrow$ **Các nghiên cứu bằng RCT rất tốn kém, khó thực hiện, nhưng có hiệu lực nội tại cao nhất trong tất cả các thiết kế nghiên cứu.**

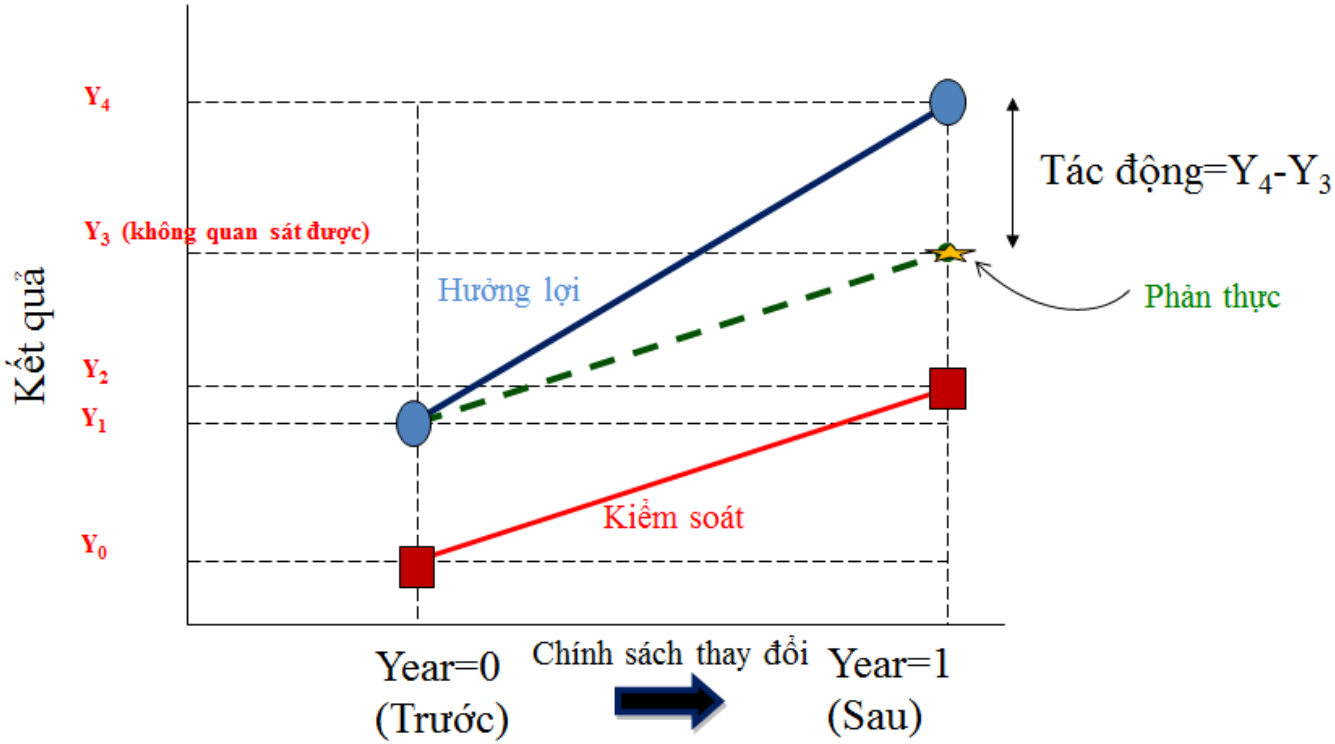
Các phương pháp khác đánh giá tác động chính sách

Bản chất của đánh giá tác động chính sách là ước lượng counterfactual.

- **Dữ liệu quan sát lặp (dữ liệu bảng) có thể được sử dụng để ước lượng counterfactual.**
- Các thiết kế nghiên cứu đặc biệt như hồi quy gián đoạn (hồi quy cắt - regression discontinuity design) hay hồi quy biến công cụ.
- Các hiện tượng ngẫu nhiên xảy ra (natural experiments) cho phép ước lượng phản thực từ nhóm không bị ảnh hưởng.
- Ước lượng phản thực bằng các thuật toán thống kê (matching, synthetic controls)

Sử dụng phương pháp DiD để đánh giá tác động 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.



	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:

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

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 can thiệp trung bình của 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 thiết kế đánh giá 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 (unobserved heterogeneity).

Lưu ý về giả định song song và hiệu lực của phương pháp DiD

- Nếu giả định song song bị vi phạm thì phản chứng là không hợp lệ (invalid counterfactual) $\Rightarrow$ Ước lượng bị chệch!
  - Khi xu hướng thay đổi của hai nhóm không tương đồng (ví dụ tốc độ tăng lương của nhóm rất nghèo so với nhóm rất giàu có thể khác nhau).
  - Khi thời gian thực hiện chương trình quá dài dẫn đến những thay đổi mang tính cấu trúc giữa các nhóm.

- Nếu có dữ liệu từ 3 kỳ quan sát trở lên thì có thể kiểm định giả định song song (falsification test).

Các hình thức ước lượng mô hình DiD

Cách 1: OLS với dữ liệu gộp (pooled regression) và biến tương tác (interaction effect).

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

- Tác động của chính sách là tham số của biến tương tác $T \times Year$.
- 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/sample attrition) thì việc ước lượng có thể bị chệch do vấn đề lựa chọn mẫu.
  - Những quan sát rớt rụng khỏi mẫu có đặc tính khác biệt so với phần còn lại (ví dụ hộ gia đình vay vốn không có khả năng trả nợ thì bỏ trốn hay không trả lời phỏng vấn).

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

- Tìm hiểu bộ dữ liệu.
- Cấu trúc dữ liệu dạng bảng dọc (long format): 826 hộ gia đình, trong đó có 468 hộ hưởng lợi, 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).

Cách thức tổ chức dữ liệu bảng

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!

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}
...	...	...	...	...

Với cấu trúc trên, mô hình ước lượng được viết như sau:

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

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

Nhận xét với hồi quy dữ liệu gộp

- Bản chất của hồi quy dữ liệu gộp tương tự như hồi quy dữ liệu chéo. Dễ thực hiện, không yêu cầu dữ liệu cân bằng.
- 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.
- Dữ liệu bị thiếu có hệ thống có thể làm mất hiệu lực nội tại của mô hình.

Cách 2: Hồi quy dữ liệu bảng - Regression with panel data

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

a_i là tác động cố định không quan sát được, 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 hồi quy dữ liệu bảng với tác động cố định

1. **Hồi quy dữ liệu bảng với tác động cố định (Fixed Effects Panel Regression):**
   xtreg Y T Year X, fe i(id)
   với id là mã hộ gia đình.

   - Phương pháp tối ưu với dữ liệu bảng.
   - Chỉ sử dụng các quan sát lặp $\Rightarrow$ Cảnh giác với vấn đề mẫu bị rớt rụng (attrition) có thể làm giảm hiệu lực ngoại vi của kết quả.
   - Nếu mẫu bị rớt rụng có hệ thống thì kết quả có thể bị sai lệch.

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

   Bản chất của phương pháp này là 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.

$$Y_{it} = \beta_0 + \beta_1 * T_{it} + \beta_2 * Year_t + \beta_k * X_{it} + \sum_{j=1}^{N-1} \sigma_j * D_j + u_{it}$$

3. **Hồi quy với dữ liệu sai phân bậc nhất - Regression with First Differences**

Lấy sai phân bậc nhất của các biến số qua thời gian (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:

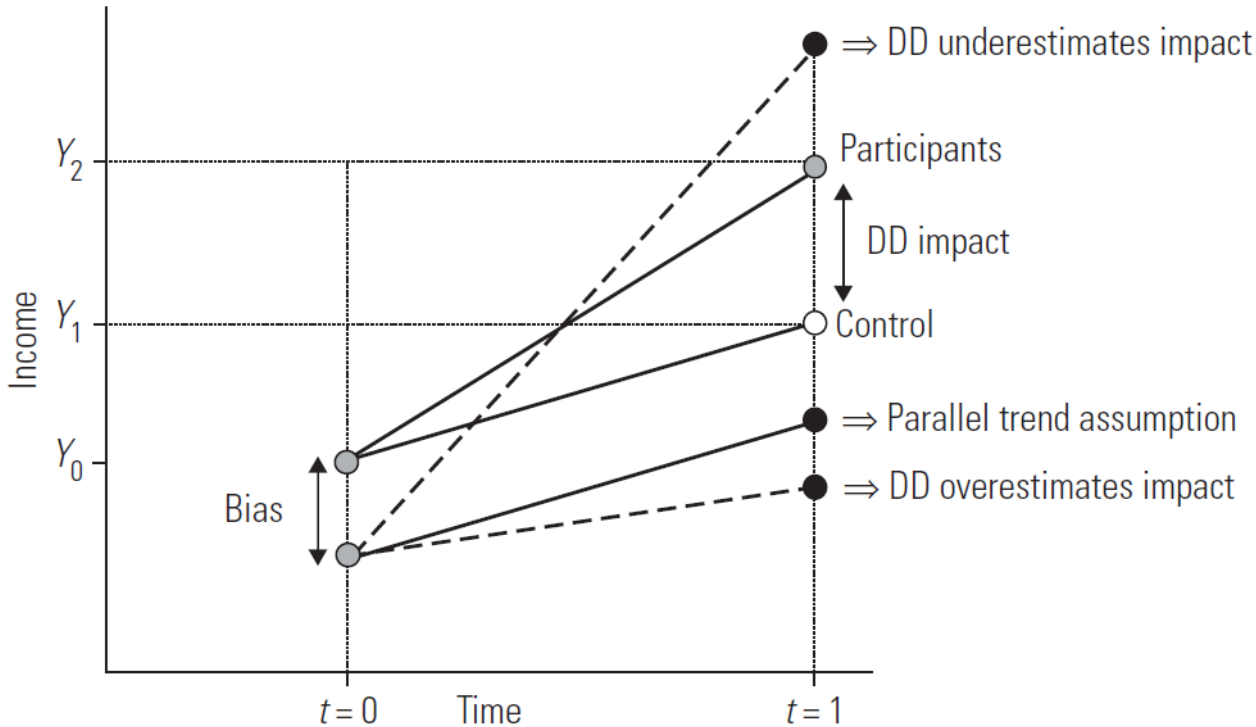
$$\Delta Y_i = \beta_2 + \beta_1 * \Delta T_i + \beta_k * \Delta X_i + u_i$$

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.

Mở rộng: DiD có tính đến điều kiện ban đầu

- Sử dụng để kiểm tra tính vững của kết quả khi nghi ngờ điều kiện trước khi thực hiện chính sách ảnh hưởng đến tốc độ thay đổi của kết quả (độ dốc của giả định song song).



- Không kiểm soát điều kiện ban đầu có thể dẫn đến sai lầm khi xây dựng phản thực, dẫn đến ước lượng bị chệch.

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

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

- 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 $Year = 0$).

Thực hành đánh giá tác động của chương trình tín dụng vi mô đến tổng chi tiêu của hộ gia đình với phương pháp hồi quy dữ liệu bảng

Sử dụng bộ dữ liệu microcredit.dta của 826 hộ gia đình thu thập qua hai năm.

- Viết phương trình hồi quy với các phương pháp đã học (pooled regression, panel data with FE, LSDV, OLS with first differences with and without the initial condition).
- Ước lượng và so sánh các mô hình.
- Diễn giải ý nghĩa.

1. Pooled regression with an interaction term:

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

2. Fixed-effects panel regression:

$$log(exptot_{it}) = \beta_0 + \beta_1 T_{it} + \beta_2 Year_t + \beta_k X_{it} + a_i + u_{it}$$

3. LSDV:

$$log(exptot_{it}) = \beta_0 + \beta_1 T_{it} + \beta_2 Year_t + \beta_k X_{it} + \sum_{j=1}^{N-1} \sigma_j D_j + u_{it}$$

4. OLS with first differencing data (with and without the initial condition $\gamma_{\mathbf{k}} * \mathbf{X_i^0}$):

$$\Delta log(exptot_i) = \beta_2 + \beta_1 \Delta T_i + \beta_k \Delta X_i + \gamma_{\mathbf{k}} * \mathbf{X_i^0} + u_i$$

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 ưu điểm điều tra lặp của dữ liệu bảng.

- Hồi quy dữ liệu bảng với tác động cố định **xtreg fe** là hiệu quả nhất. Cũng có thể sử dụng **hồi quy sai phân bậc nhất** để loại bỏ những nhấn tố không thay đổi theo thời gian. 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.

- **Hồi quy với biến giả** cũng có thể được sử dụng để kiểm soát các nhân tố không thay đổi theo thời gian. Tuy nhiên đưa nhiều biến giả làm giảm bậc tự do và giảm sức mạnh của kiểm định thống kê.

- 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 pooled, xtreg fe, lsdv và first differencing đều cho kết quả tương đồng.

Hồi quy dữ liệu bảng - Nâng cao

Mô hình tổng quát của hồi quy dữ liệu bảng

$$Y_{it} = \beta_k * X_{it} + \underbrace{a_i + u_{it}}_{v_{it}} \tag{6}$$

- 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:

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

- 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 trừ:

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

viết gọn lại thành:

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

với các giá trị $\ddot{Y}_{it}$, $\ddot{X}_{it}$ được tính bằng cách lấy 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 (FE) estimator** (ước lượng dựa vào biến động nội tại cùng một hộ gia đình).

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

- Giả sử tác động cố định không quan sát được a_i không tương quan với các biến giải thích X_{it} khác trong mô hình (6):

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

khi này, mô hình (6) vẫn thỏa điều kiện 4.2 ($cov(X_{it}, v_{it}) = 0$) và ước lượng bằng OLS vẫn không chệch.

- Nếu ước lượng bằng fixed-effects trong trường hợp này 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:

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

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

- Ước lượng (10) bằng OLS không chệch (unbiased) nhưng không hiệu quả nhất do các phần dư v_{it} tương quan chuỗi với nhau (vi phạm điều kiện *iid*):

$$cov(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} \neq 0$$

Ước lượng mô hình tác động ngẫu nhiên

Tương tự như phương pháp hồi quy với quyền số (generalized least square-GLS) để xử lý vấn đề tương quan chuỗi:

1. Ước lượng quyền số chuyển đổi dữ liệu θ,

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

   - T là số kỳ quan sát, và thỏa điều kiện số quan sát lớn hơn nhiều số kỳ quan sát, $N \gg T$.
   - θ luôn dương và nhỏ hơn 1.
   - Chuyển đổi bộ dữ liệu theo công thức: $Y_{it} - \theta\overline{Y_i}$, $X_{it} - \theta\overline{X_i}$, và $v_{it} - \theta\overline{v_i}$.

2. Và ước lượng mô hình OLS với dữ liệu đã chuyển đổi:

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

Stata: *xtreg Y T Year X, re i(id)*

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

- θ phản ánh mức độ quan trọng tương đối của tác động cố định a_i so với phần dư u_{it} của mô hình thông qua phương sai σ_a^2 và σ_u^2.
- Nếu tác động cố định không quan trọng trong mô hình, $\sigma_a^2 \ll \sigma_u^2 \Rightarrow \theta \to 0$. Khi này ước lượng RE tương tự như pooled OLS.
- Nếu tác động cố định rất quan trọng trong mô hình, $\sigma_a^2 \gg \sigma_u^2 \Rightarrow \theta \to 1$. Khi này ước lượng RE sẽ tiệm cận ước lượng FE.

Khi nào thì sử dụng pooled OLS, fixed-effects và random-effects model?

Lựa chọn mô hình nào tùy thuộc vào lý thuyết nền tảng, lập luận bối cảnh nghiên cứu, dữ liệu và kiểm định.

- Luôn sử dụng pooled OLS làm mô hình tham chiếu trước khi ước lượng các mô hình khác phức tạp hơn.
- Nếu tác động cố định tương quan với biến giải thích thì mô hình FE sẽ xử lý được vấn đề thiếu biến quan trọng. Nếu tác động cố định không tương quan với biến giải thích thì mô hình RE sẽ hiệu quả hơn FE.
- Áp dụng sai 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.

Kiểm định Hausman để lựa chọn FE hoặc RE model

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.

$$H_0 : \beta_{FE} = \beta_{RE}$$
$$H_1 : \beta_{FE} \neq \beta_{RE}$$

- Trị kiểm định χ^2 được tính với giả định các tham số ước lượng được theo phương pháp FE thì nhất quán (consistent), và phương phương pháp RE thì hiệu quả (efficient).
- Nếu hai tham số ước lượng tương đương nhau thì chọn ước lượng có hiệu quả hơn (tham số ước lượng có sai số chuẩn thấp nhất).
- Nếu có sự khác biệt giữa hai tham số ước lượng, khi này giả định sử dụng trong ước lượng RE có thể không hợp lý.
- Nguyên tắc chọn mô hình với Hausman test:
  - 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.

Thực hành

Ước lượng hàm sản xuất KLEM với dữ liệu energy bằng cả bốn mô hình pooled OLS, between efects, fixed effects, và random effects. So sánh kết quả và giải thích.