Mô hình với dữ liệu không ngẫu nhiên (Models with non-random sample/ sample selection)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 14 tháng 4 năm 2019

Khái niệm dữ liệu không ngẫu nhiên/Vấn đề tự lựa chọn mẫu

Sample selection/non-random sample:

- Do cách thiết kế mẫu khiến dữ liệu bị mất.
- Do dữ liệu bị thiếu một số thông tin nhất định.
- Do cách thiết kế chính sách dẫn đến chỉ quan sát được những nhóm đối tượng nhất định.

Hiệu lực nội tại khi xảy ra vấn đề lựa chọn mẫu

Giả sử chúng ta có mô hình hồi quy của thu nhập y theo các biến giải thích x:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

thỏa các điều kiện của mô hình CLRM và $E[u|x_1, ..., x_k] = 0$

- Nếu chúng ta quan sát được toàn bộ mẫu dữ liệu $\Rightarrow$ Ước lượng OLS không chệch và nhất quán.
- Khi dữ liệu bị thiếu:
  - Dữ liệu bị thiếu ngẫu nhiên?
  - Dữ liệu bị thiếu không ngẫu nhiên?

- Thiếu ngẫu nhiên: Ước lượng OLS đảm bảo hiệu lực nội tại, nhưng độ tin cậy của ước lượng sẽ bị giảm.

- Thiếu không ngẫu nhiên: Ước lượng bằng OLS **có thể** bị chệch và không có hiệu lực nội tại. Cần hiểu rõ bản chất của dữ liệu!!

Dữ liệu không ngẫu nhiên

Dữ liệu bị thiếu không ngẫu nhiên do vấn đề chọn mẫu xảy ra trên biến giải thích, ví dụ chỉ điều tra những người làm việc ở HCM, hay có bằng cấp cao nhất không quá phổ thông trung học.

- Không ảnh hưởng đến hiệu lực nội tại, nhưng có thể ảnh hưởng đến hiệu lực ngoại vi.

- Ví dụ: Mô hình dựa trên điều tra thu nhập và tình trạng học vấn của nhóm cá nhân học không quá 12 năm sẽ không thể áp dụng cho nhóm học đại học hoặc cao hơn.

Dữ liệu không ngẫu nhiên

Dữ liệu bị thiếu do vấn đề lựa chọn mẫu dựa trên biến phụ thuộc. Ví dụ dữ liệu tiền lương quan sát bị chặn dưới bởi 0, hoặc không quan sát được tiền lương với những người không đi làm.

- Ảnh hưởng đến hiệu lực nội tại, và ước lượng bị chệch do vấn đề lựa chọn mẫu.
- Ví dụ: Ước lượng hàm tiền lương của người trong độ tuổi lao động. Những người không đi làm (do đó tiền lương bằng không hoặc không được ghi nhận) có thể do nhiều lý do (tiền lương thấp hơn kỳ vọng, hoặc có lựa chọn khác). Nếu không sử lý vấn đề chọn mẫu thì ước lượng sẽ bị sai lệch.

Xử lý khi dữ liệu không ngẫu nhiên

Cần hiểu rõ bản chất của dữ liệu mới nhận diện được vấn đề lựa chọn mẫu và đề xuất cách thức sử lý phù hợp!

- Nếu giả định những người không đi làm nhận mức lương bằng 0 $\Rightarrow$ Mô hình **Tobit** với biến phụ thuộc bị chặn dưới.

- Nếu giả định những người không đi làm là do có những lựa chọn khác tốt hơn (ví dụ làm tư, do đó không báo cáo thu nhập trong bảng câu hỏi tiền lương). Mặc dù những người này không được ghi nhận có thu nhập nhưng trên thực tế họ vẫn có thu nhập $\Rightarrow$ Dùng mô hình hồi quy điều chỉnh vấn đề lựa chọn mẫu **Heckman selection model/Heckit method**.

Mục đích của mô hình điều chỉnh vấn đề lựa chọn mẫu

Giả dụ chúng ta ước lượng hàm tỷ suất thu nhập của việc đi học. Mẫu dữ liệu của chúng ta có cả những người đang làm công ăn lương và những người trong độ tuổi lao động nhưng không báo cáo thu nhập do làm tư, kinh doanh tiểu thương.

- Nếu chỉ giới hạn ở mẫu dữ liệu những người đang đi làm và có thu nhập dương $\Rightarrow$ OLS có thể chệch và không nhất quán.

- Nếu chúng ta đưa toàn bộ dữ liệu (gồm cả những người không báo cáo thu nhập) vào mô hình thu nhập $\Rightarrow$ Xử lý thế nào với những người không báo cáo thu nhập?

$\Rightarrow$ Chúng ta cần điều chỉnh hàm hồi quy để phản ánh vấn đề lựa chọn mẫu trong tham gia lực lượng lao động.

Cơ chế của mô hình điều chỉnh vấn đề lựa chọn mẫu

Mô hình lựa chọn mẫu được viết dưới dạng **hệ phương trình cấu trúc**, bao gồm một phương trình diễn giải hành vi và một phương trình diễn giải vấn đề lựa chọn mẫu:

$$\begin{cases} y &= X\beta + u \\ s &= 1[Z\gamma + v \geq 0] \end{cases}$$

trong đó $E[u|X] = 0$, X là các biến giải thích của phương trình hành vi y, Z là các biến giải thích trong phương trình lựa chọn mẫu s.

Ý nghĩa của phương trình lựa chọn mẫu s

Phương trình lựa chọn được biểu diễn dưới dạng **hàm chỉ số (index function)** của các biến giải thích Z, mục đích để giải thích tại sao một số quan sát nằm trong mẫu nghiên cứu (ví dụ có thu nhập) còn những người khác nằm ngoài mẫu (không có thu nhập).

$$s = \begin{cases} 1 & if \quad Z\gamma + v \geq 0 \\ \\ 0 & otherwise \end{cases}$$

- Nếu $Z_i\gamma + v \geq 0 \Rightarrow s_i = 1$ có nghĩa là chúng ta quan sát được cá nhân i trong phương trình hành vi (cá nhân i có thu nhập).
- Nếu $s_i = 0$ có nghĩa là chúng ta không có cá nhân i trong phương trình hành vi (cá nhân i không có thu nhập).

Ý nghĩa của phương trình hành vi y

Với điều kiện quan sát được cá nhân có thu nhập thì phương trình hành vi ước lượng tác động của các nhân tố X ảnh hưởng như thế nào đến thu nhập y.

Phương trình hành vi có điều kiện (conditional expectation function)

Chúng ta cần ước lượng phương trình hành vi y với điều kiện quan sát được các cá nhân nằm trong mẫu. Bỏ qua các bước biến đổi trung gian,

$$E[y|Z, s = 1] = X\beta + \rho\lambda(Z\gamma)$$

trong đó λ là tỷ số Mills nghịch đảo (Mills Inverse Ratio-IMR), được tính tại giá trị $Z\gamma$; β và ρ là tham số cần ước lượng của phương trình hành vi có điều kiện; X và $\lambda(Z\gamma)$ là các biến giải thích. $\lambda(Z\gamma)$ được tính như sau:

$$\lambda(Z\gamma) = \frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$$

$\phi(.)$ và $\Phi(.)$ là hàm mật độ và hàm tích lũy phân phối chuẩn.

Các bước xây dựng và ước lượng mô hình hồi quy điều chỉnh vấn đề lựa chọn mẫu

Bắt đầu bằng hệ phương trình cấu trúc:

$$\begin{cases} y &= X\beta + u \\ s &= 1[Z\gamma + v \geq 0] \end{cases}$$

Chúng ta cần ước lượng mô hình hành vi có điều kiện:

$$E[y|Z, s = 1] = X\beta + \rho\lambda(Z\gamma)$$

- Các tham số của mô hình hành vi có điều kiện là β và ρ.
- Các biến giải thích là X và tỷ số $\lambda(Z\gamma)$.

Do $\lambda(Z\gamma)$ phụ thuộc vào các tham số γ nên chúng ta phải ước lượng phương trình lựa chọn mẫu trước để tìm γ.

Ước lượng hồi quy điều chỉnh mẫu bằng hồi quy hai giai đoạn

1. Ước lượng mô hình lựa chọn (cá nhân nằm trong mẫu có thu nhập hay ngoài mẫu) bằng hồi quy Probit để ước lượng các tham số γ, và sử dụng toàn bộ bộ dữ liệu,

$$P(s = 1|Z) = \Phi(Z\gamma)$$

Tính giá trị $\widehat{\lambda(Z\gamma)}$ bằng công thức:

$$\widehat{\lambda(Z\gamma)} = \frac{\phi(Z\hat{\gamma})}{\Phi(Z\hat{\gamma})}$$

Tương tự như phương pháp 2SLS/IV, **phải có ít nhất một biến ngoại sinh trong Z nhưng không thuộc X** (biến chỉ ảnh hưởng đến việc lựa chọn vào mẫu có thu nhập chứ không ảnh hưởng đến thu nhập).

Ước lượng hồi quy điều chỉnh mẫu bằng hồi quy hai giai đoạn

2. Ước lượng mô hình hành vi có điều kiện bằng OLS, với dữ liệu trong mẫu (chỉ những cá nhân có thu nhập), với các biến giải thích X và $\widehat{\lambda(Z\gamma)}$ được tính ở bước 1:

$$y = X\beta + \rho\widehat{\lambda(Z\gamma)} + u$$

   - Bản chất của phương pháp Heckit là chúng ta đưa thêm một biến giải thích là tỷ số IMR được tính từ phương trình chọn mẫu vào hồi quy OLS.
   - Do các biến giải thích ảnh hưởng đến cả phương trình chọn mẫu (thông qua $\widehat{\lambda(Z\gamma)}$) lẫn phương trình hành vi (thông qua X) nên phải giải thích sự khác biệt giữa ước lượng bằng OLS với Heckman.

Ước lượng giá trị của tưới tiêu đến năng suất lúa và ngô bằng phương pháp đánh giá thụ hưởng (hedonic valuation)

- Sử dụng bộ dữ liệu IrrigationValuation.dta
- Chúng ta quan sát được sản lượng lúa và ngô trên từng mảnh đất, các đặc tính đất đai thổ nhưỡng của các khoảnh ruộng, biến nhân khẩu học... Biến chính sách là tình trạng tưới tiêu (đất có được tưới tiêu bằng thủy lợi hay không).
- Mảnh đất được tưới tiêu được kỳ vọng có sản lượng cao hơn. Chênh lệch sản lượng giữa các mảnh đất có và không có tưới tiêu sẽ cho phép ước lượng giá trị của thủy lợi.

Giả sử hàm sản xuất dạng logarithm như sau:

$$log(Q_i) = \alpha_0 + \alpha_1 \times D_{IRRI_i} + \sum_j INPUT^j{}_i \times \alpha_j + \sum_k LAND^k{}_i \times \alpha_k + \sum_l DEMO^n{}_i \times \alpha_n + \varepsilon_i$$

trong đó:

- Q là tổng sản lượng trên một công (kg/1000m2).
- D_{IRRI} là biến mảnh ruộng có được tưới tiêu hay không.
- $INPUT$, $LAND$, $DEMO$ là các biến đầu vào, đặc tính đất đai, và nhân khẩu học của hộ gia đình.

Mô hình 1: Ước lượng hàm sản xuất bằng OLS

Việc lựa chọn loại cây trồng trên mỗi mảnh đất bị ảnh hưởng bởi nhiều nhân tố, bao gồm chính sách của chính phủ (một số loại đất chỉ được trồng lúa), đặc tính đất, đặc tính thủy lợi... $\Rightarrow$ Dữ liệu bị ảnh hưởng bởi vấn đề chọn mẫu.

Mô hình 2: Hàm hồi quy có điều chỉnh vấn đề chọn mẫu bằng phương pháp Heckit. Ví dụ với đất lúa:

$$\begin{cases} log(Q_i^{rice}) & = & \alpha_0 + \alpha_1 \times D_{IRRI_i} + ... + \rho\lambda(Z_i\gamma) + \varepsilon_i \\ P(Rice_i|R_i) & = & \Phi(Z_i\gamma + u_i) \end{cases}$$

trong đó Z là các đặc tính đất đai và chính sách có thể ảnh hưởng đến việc chọn loại cây trồng. Biến ngoại sinh là quy định mảnh đất đó chỉ được trồng lúa hay có thể trồng cây khác.

So sánh và kiểm định mô hình lựa chọn mẫu

- So sánh kết quả giữa mô hình OLS và Heckit.
- Kiểm tra các tham số ước lượng trong mô hình lựa chọn mẫu.
- Kiểm định có vấn đề tự lựa chọn mẫu: $H_o : \rho = 0$. Nếu bác bỏ H_o thì cần sử dụng mô hình lựa chọn mẫu.

RICE MODEL

	OLS b/se	Heckman Co~d b/se
main		
plotIrriga~n	**0.5300*** **	**0.2752*** **
	(0.0883)	**(0.0323)**
plotArea	**-0.0515**	**-0.0488*** **
	(0.0251)	**(0.0024)**

MAIZE MODEL

	OLS b/se	Heckman Co~d b/se
main		
plotIrriga~n	**0.1385***	**0.2700** **
	(0.0565)	**(0.0862)**
plotArea	**-0.0449** **	**-0.0399*** **
	(0.0125)	**(0.0066)**