Mô hình với dữ liệu không ngẫu nhiên (Models with sample selection)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 7 tháng 5 năm 2018

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

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:

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

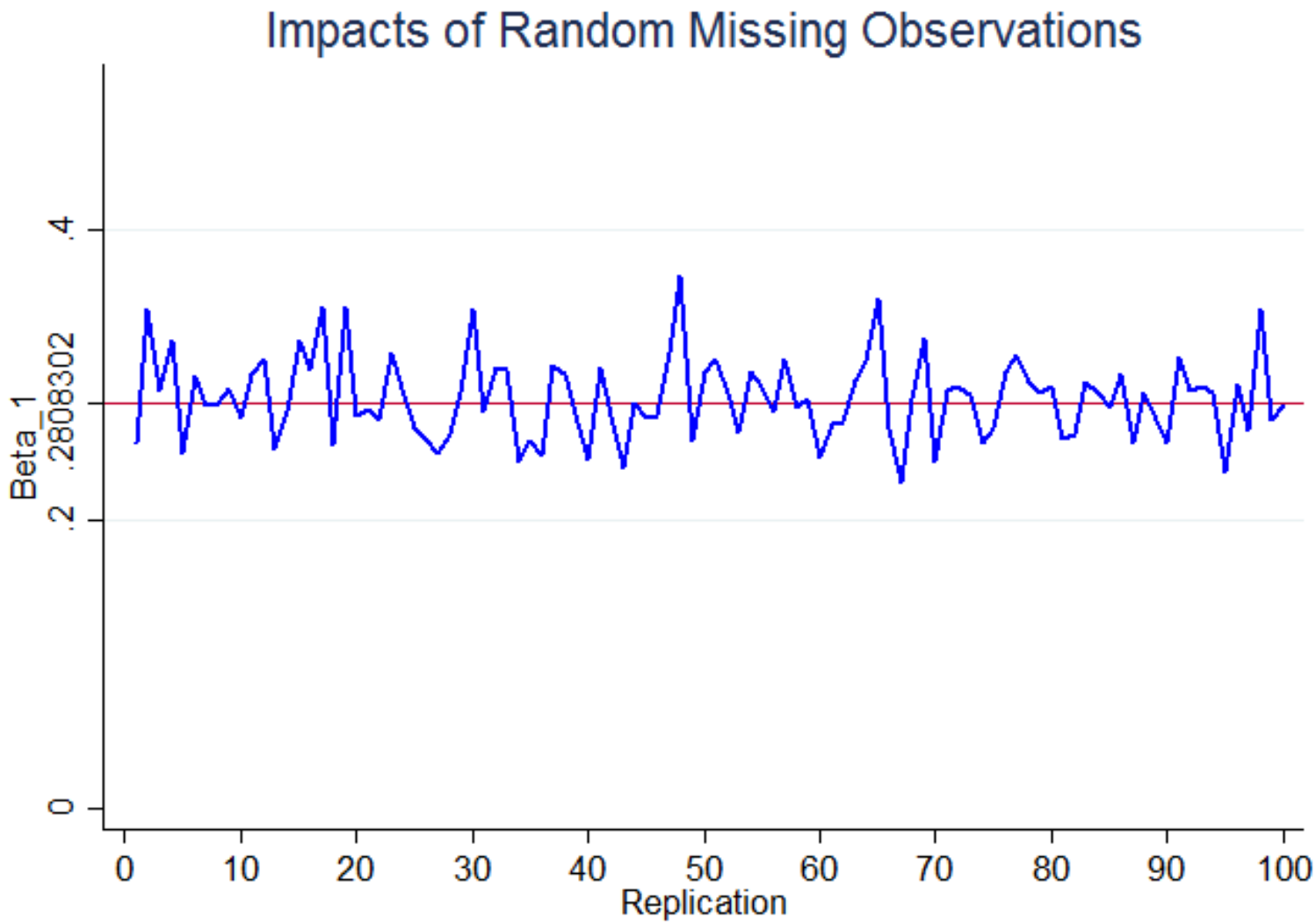
thỏa các điều kiện của mô hình CLRM, $E[u|x_1, ..., x_k] = 0$

- Nếu chúng ta quan sát được toàn bộ mẫu dữ liệu $\Rightarrow$ Ước lượng OLS của mô hình (1) không chệch và nhất quán.
- Nếu chúng ta chỉ có một số quan sát nhất định:
  - 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
- Tự lựa chọn mẫu theo các điều kiện quan sát được (selection on observables, exogenous sample selection): matching, heckman selection model
- Tự lựa chọn mẫu theo tiêu chí không quan sát được (selection on unobservales): OLS không có hiệu lực nội tại, sử dụng fixed effects/random effects, instrumental variables

Bootstrap tham số mô hình khi dữ liệu thiếu ngẫu nhiên

So sánh ước lượng OLS toàn bộ dữ liệu (4820 quan sát) với ước lượng chọn từ mẫu ngẫu nhiên của 4000 quan sát được lấy ngẫu nhiên từ bộ dữ liệu.



Cơ chế của mô hình tự lựa chọn mẫu - Model of Sample Selection

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:

$$y = X\beta + u \tag{1}$$
$$s = 1[Z\gamma + v \geq 0] \tag{2}$$

trong đó $E[u|X] = 0$, X là các biến giải thích của mô hình hành vi. Phương trình lựa chọn được biểu diễn dưới dạng hàm chỉ số (index function) của các biến giải thích Z (**Lưu ý Z phải có ít nhất một biến khác với X**).

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

Ý nghĩa của hàm chỉ số - Index function

- Nếu $Z\gamma + v \geq 0 \Rightarrow s_i = 1$ có nghĩa là chúng ta quan sát được hộ gia đình i trong phương trình hành vi (1).
- Nếu $s_i = 0$ có nghĩa là chúng ta không có hộ gia đình i trong phương trình (1).

Diễn giải ý nghĩa của hệ phương trình 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 đi làm ($s_i = 1$) và những người trong độ tuổi lao động nhưng không làm việc ($s_i = 0$) vì nhiều lý do (lương quá thấp, nghỉ hưu, làm việc khác không tạo ra thu nhập...)

- 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 vào mô hình thu nhập $\Rightarrow$ Sử lý thế nào với những người không có thu nhập?

$\Rightarrow$ Chúng ta cần phương trình lựa chọn mẫu để diễn giải hành vi tham gia lực lượng lao động.

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

Chúng ta muốn ước lượng mô hình hành vi (1), áp dụng cho những cá nhân quan sát được trong mô hình lựa chọn mẫu (2). 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), và ρ là tham số của biến số IMR mới đưa vào phương trình trên.

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

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

Phương pháp hồi quy điều chỉnh mẫu - Heckman selection model, Heckit method

Bắt đầu bằng hệ phương trình cấu trúc:

$$y = X\beta + u$$
$$s = 1[Z\gamma + v \geq 0]$$

Chúng ta cần ước lượng mô hình hành vi có điều chỉnh vấn đề lựa chọn mẫu:

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

- Các tham số của mô hình hành vi có điều kiện là β và ρ.
- Các biến giải thích là X và tỷ số IMR (λ) được tính tại các giá trị $Z\gamma$.

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

Phương pháp hồi quy điều chỉnh mẫu bằng hồi quy 2 giai đoạn

1. Ước lượng mô hình lựa chọn 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 IMR từ các tham số $\hat{\gamma}$ cho các dữ liệu được lựa chọn $(s_i = 1)$ bằng công thức:

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

2. Ước lượng mô hình hành vi có điều kiện bằng OLS, với dữ liệu được lựa chọn $(s_i = 1)$, đồng thời đưa thêm một biến giải thích mới là $\hat{\lambda}$ được tính ở bước 1 vào mô hình:

$$y = X\beta + \rho\hat{\lambda} + u$$

Ước lượng mô hình năng suất gạo và ngô trong nông nghiệp để ước tính giá trị của thủy lợi

Đánh giá tác động của tưới tiêu đến năng suất lúa và ngô sử dụng bộ dữ liệu IrrigationValuation.dta.

Mô hình 1: 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 (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.

Vấn đề đối với ước lượng hàm sản xuất bằng OLS:

- Việc lựa chọn loại cây trồng bị ảnh hưởng bởi nhiều nhân tố, bao gồm chính sách của chỉnh phủ (một số loại đất chỉ được trồng lúa), đặc tính đất, đặc tính thủy lợi, lợi nhuận...

  $\Rightarrow$ 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:

$$\begin{cases} log(Q_i^{rice}) = \alpha_0 + \alpha_1 \times D_{IRRI_i} + ... + \rho\sigma_\varepsilon\lambda(Z\gamma) + \varepsilon_i & (2) \\ P(Rice_i|R_i) = \Phi(R_i\gamma + u_i) & (1) \end{cases}$$

trong đó R là các đặc tính đất đai và chính sách có thể ảnh hưởng đến việc chọn loại cây trồng.

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.