Mô hình Tobit với Biến Phụ thuộc bị chặn (Regression with Censored Data)

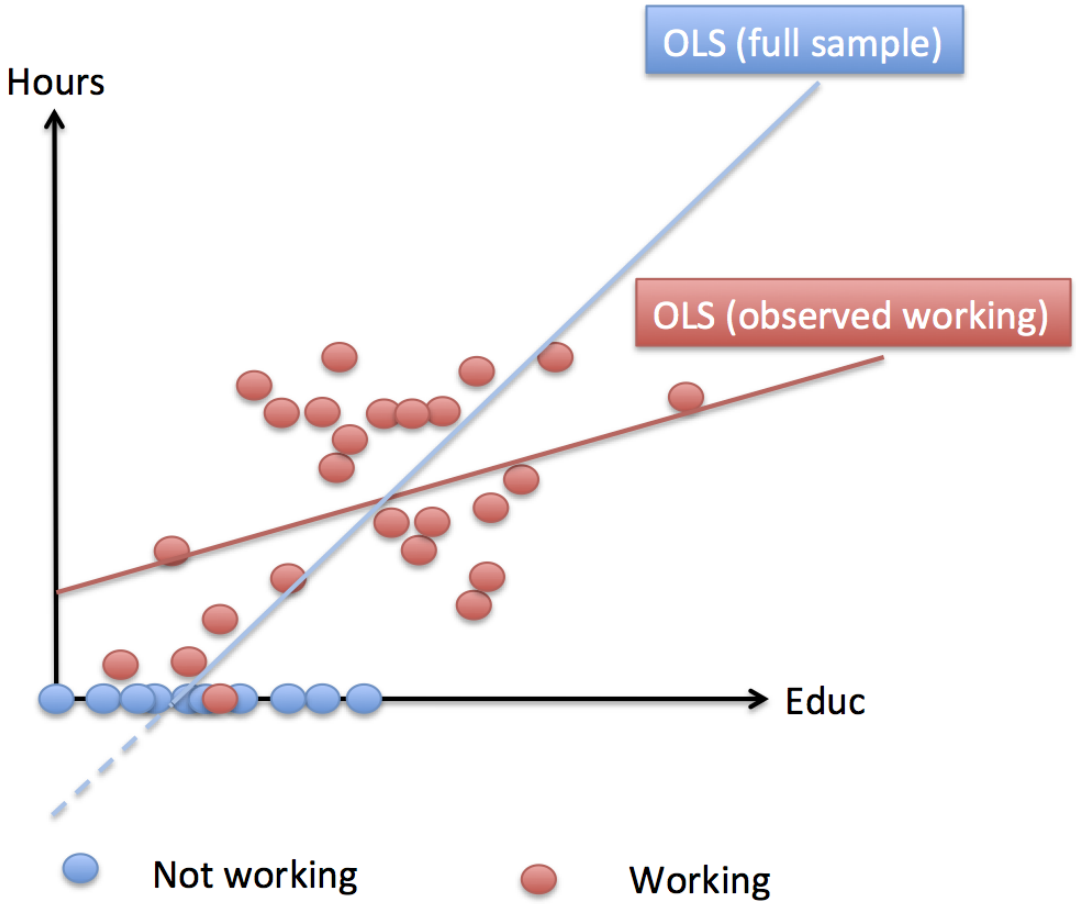
Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 14 tháng 1 năm 2018

Khái niệm biến phụ thuộc bị chặn/kiểm duyệt (censored data)

- Biến tiền lương bị chặn dưới bởi giá trị 0 đối với những người chưa đi làm, về hưu, hay đang thất nghiệp. Các giá trị quan sát được là dương.
- Rất nhiều biến số kinh tế bị chặn dưới bởi giá trị 0, ví dụ:
  - Số giờ lao động của phụ nữ đã có gia đình.
  - Số tiền làm từ thiện của một người trong một năm.
  - Số lít rượu bia một người uống trong một năm.
  - Chi tiêu cho hàng hoá xa xỉ của hộ gia đình trong dịp lễ tết.
  - Thời gian thất nghiệp của một người lao động.
- Dữ liệu có thể bị chặn trên hoặc chặn dưới do cách thức điều tra dữ liệu.

Hồi quy OLS với biến phụ thuộc bị chặn dưới tại 0



- Ước lượng số giờ làm việc bị thiên lệch giảm (downward bias) do bỏ qua nhóm không làm việc.
- Ước lượng OLS với toàn bộ dữ liệu gặp phải vấn đề dự báo biến phụ thuộc âm tương tự như mô hình xác suất tuyến tính LPM.

Các cách xử lý biến phụ thuộc bị chặn

- Cách 1: ước lượng mô hình Logit/Probit với biến phụ thuộc là có làm việc hay không. Tuy nhiên cách làm này chỉ ước lượng được xác suất có làm việc hay không (biến định tính rời rạc), nhưng không ước lượng được tác động của biến giải thích lên số giờ làm việc của những người đi làm như thế nào (biến định lượng liên tục).

- Cách 2: mô hình Tobit xử lý được cả hai vấn đề trên.

Mô hình Tobit với biến phụ thuộc bị chặn

Mô hình Tobit gồm có 2 bước theo thứ tự:

- Bước 1: Ước lượng xác suất quan sát được một người có tham gia lao động hay không.

- Bước 2: Ước lượng các nhân tố ảnh hưởng đến số giờ lao động, và điều chỉnh hệ số ước lượng để tính đến xác suất có đi làm hay không.

Xây dựng mô hình Tobit

Giả sử số giờ làm việc của một người được diễn giải bởi hàm:

$$y = X * \beta + u$$

với $u \sim N(0, \sigma^2)$.

- Chúng ta quan sát được $y > 0$ đối với những người đi làm.
- Với những người không đi làm, $y = 0$.

Ước lượng mô hình Tobit

- Bước 1 ước lượng xác suất có tham gia lao động hay không:

$$P(y > 0|x) = \Phi(\frac{X * \beta}{\sigma})$$

với $\Phi(.)$ là hàm phân phối tích luỹ chuẩn.

- Bước 2 ước lượng hàm số giờ làm việc theo các biến giải thích:

$$E[y|x] = \Phi(\frac{X * \beta}{\sigma}) * \left[X * \beta + \sigma\lambda(\frac{X * \beta}{\sigma})\right]$$

  - $E[y|x]$ là kỳ vọng không điều kiện hay hàm hồi quy mẫu với tất cả dữ liệu (gồm cả có và không đi làm).
  - $\lambda(c) = \frac{\phi(c)}{\Phi(c)}$, còn được gọi là tỷ số Mills nghịch đảo (inverse Mills ratio), là tỷ lệ giữa hàm mật độ và hàm tích lũy của phân phối chuẩn được tính tại giá trị c.

Tác động biên trong mô hình Tobit

Tác động biên của biến giải thích lên biến phụ thuộc bằng đạo hàm bậc nhất của hàm hồi quy theo biến giải thích:

$$\frac{\partial E[y|x]}{\partial x_j} = \beta_j * \Phi(\frac{X * \beta}{\sigma})$$

- Tác động biên của mô hình Tobit được tính gián tiếp bằng β có điều chỉnh giảm theo hệ số $P(y > 0|x) = \Phi(\frac{X*\beta}{\sigma}) < 1$.
- Nếu $P(y > 0|x) = \Phi(\frac{X*\beta}{\sigma}) = 1$ thì biến phụ thuộc nhận giá trị dương cho toàn bộ mẫu quan sát (tất cả các quan sát đều tham gia lao động). Khi đó OLS và Tobit là đồng nhất.
- Tương tự như phương pháp MLE, $\Phi(\frac{X*\beta}{\sigma})$ được tính tại các giá trị đặc trưng như trung bình, các tứ phân vị... của các biến giải thích.

Thực hành: Sử dụng bộ dữ liệu Labor.dta và ước lượng hàm cung lao động của phụ nữ đã có gia đình.

	OLS b/se	Tobit b/se
main		
netincome	-3.4466	-8.8142*
	(2.5440)	(4.4591)
educ	28.7611*	80.6456***
	(12.9546)	(21.5832)
exper	65.6725***	131.5643***
	(9.9630)	(17.2794)
expersq	-0.7005*	-1.8642***
	(0.3246)	(0.5377)
age	-30.5116***	-54.4050***
	(4.3639)	(7.4185)
KIDS6	-442.0899***	-894.0217***
	(58.8466)	(111.8779)
KIDS7	-32.7792	-16.2180
	(23.1762)	(38.6414)
Constant	1330.4824***	965.3053*
	(270.7846)	(446.4358)
sigma		
Constant		1122.0217***
		(41.5790)
N	753	753

* p<0.05, ** p<0.01, *** p<0.001

So sánh ước lượng OLS và Tobit thế nào?

- Với OLS, β là tác động biên trực tiếp từ hàm hồi quy.
- Với Tobit,

$$\frac{\partial E[y|x]}{\partial x_j} = \beta_j * \Phi(\frac{X * \beta}{\sigma})$$

Cần tính $\Phi(\frac{X*\beta}{\sigma})$ tại giá trị cho trước của các biến giải thích.

Ví dụ tại giá trị trung bình của các biến giải thích:
$\Phi(20.12 * -8.81 + 12.28 * 80.65 + 10.63 * 131.6 + 10.63^2 * -1.86 + 42.53 * -54.41 + .24 * -894.0 + 1.35 * -16.22 + 965.3)/1122.02) = \Phi(.3727) = .645.$

- Tác động biên của việc học thêm một năm lên số giờ lao động của phụ nữ, tại giá trị trung bình của các biến giải thích, tính cho toàn bộ mẫu gồm cả những người đang tham gia lao động và không lao động, là 80.65*.645 = 52 giờ.
- Ước lượng OLS là 28.76 giờ. Ước lượng nào hợp lý hơn, OLS hay Tobit?
- Chỉ giới hạn vào 428 phụ nữ đang tham gia lao động, ước lượng OLS và Tobit cho kết quả giống nhau.