Tổng quan kinh tế lượng ứng dụng (Overview of applied econometrics)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 12 tháng 3 năm 2019

Hồi quy tuyến tính cổ điển CLRM và các giả định

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + ... + u_i$$

1. Tuyến tính theo tham số.
2. Chọn mẫu ngẫu nhiên.
3. Không có cộng tuyến hoàn hảo.
4. Trung bình có điều kiện của sai số bằng 0:

$$E(u|x^1, ..., x^k) = 0$$

$\Rightarrow$ Ước lượng của OLS là không chệch.

$$E(\hat{\beta}) = \beta$$

5. Với các giá trị của các biến giải thích cho trước, phương sai của sai số là một hằng số:

$$Var(u|x_1, ..., x_k) = \sigma^2$$

- Với các giả định 1-5, ước lượng của OLS là ước lượng tuyến tính, không chệch, và hiệu quả nhất (**Best Linear Unbiased Estimator - BLUE**).

6. Sai số u độc lập với các biến giải thích, có phân phối chuẩn với giá trị trung bình là 0 và phương sai σ^2.

$$u \sim N(0, \sigma^2)$$

Mô hình hồi quy tuyến tính cổ điển - CLRM

Nếu thỏa các giả định 1-6 thì mô hình được coi là mô hình hồi quy tuyến tính cổ điển.

- Ước lượng của β là BLUE.
- Phân phối mẫu của β là:

$$\hat{\beta} \sim N(\beta, Var(\beta))$$

- Viết dưới dạng chuẩn hóa:

$$\frac{\hat{\beta} - \beta}{sd(\hat{\beta})} \sim N(0, 1)$$

- Các kiểm định t, F có hiệu lực.

Khi các giả định của mô hình CLRM không thỏa

Mô hình có thể không có hiệu lực nội tại (internal validity)!

Khái niệm hiệu lực nội tại (internal validity) và hiệu lực ngoại vi (external validity) của mô hình ước lượng

- Hiệu lực nội tại: các giả thuyết thống kê đối với các tham số ước lượng được là hợp lý đối với mẫu hay quần thể dữ liệu và bối cảnh được nghiên cứu.
- Hiệu lực ngoại vi: các giả thuyết thống kê có thể được áp dụng đối với các bộ dữ liệu, quần thể hay bối cảnh khác so với bối cảnh nghiên cứu.
- Yêu cầu của mô hình là đảm bảo được hiệu lực nội tại. Một số mô hình có thể có hiệu lực nội tại nhưng hiệu lực ngoại vi yếu.

Hiệu lực nội tại trong mô hình OLS

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + u$$

Nếu các điều kiện 1-6 được thỏa:

- Ước lượng của β là không chệch (thiên lệch) và nhất quán:

$$E[\hat{\beta}] = \beta$$

$$plim(\hat{\beta}) \rightarrow \beta$$

- Các kiểm định có phân phối và mức ý nghĩa như dự báo.

Thiên lệch và nhất quán - Bias and Consistency

- Không thiên lệch: giá trị kỳ vọng của ước lượng bằng với giá trị thực – khi ước lượng mô hình với mẫu ngẫu nhiên lặp (repeated sampling) :

$$E(\hat{\beta}) = \beta$$

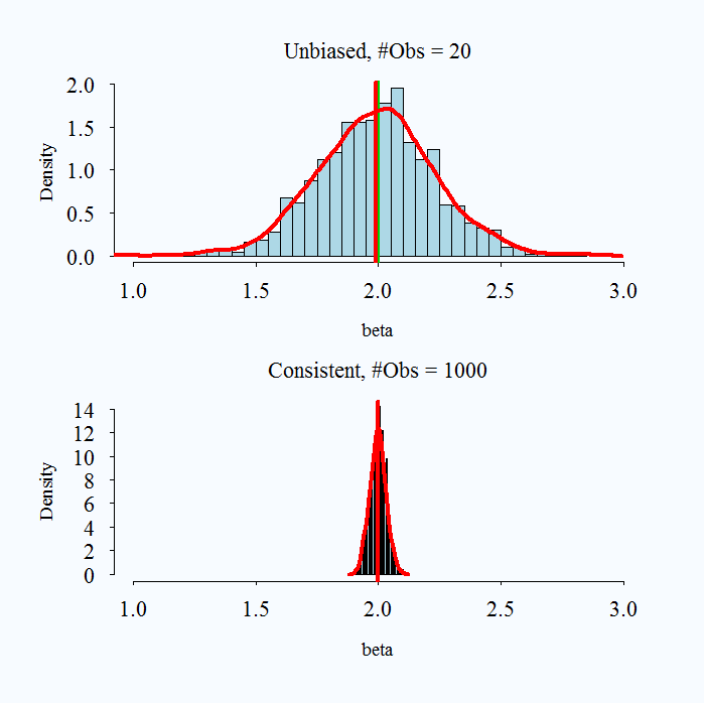
- Nhất quán: Phân phối của ước lượng của tham số hội tụ (còn gọi là tiệm cận - asymptotic) về giá trị thực khi cỡ mẫu tăng đến vô cùng:

$$plim(\hat{\beta}) \rightarrow \beta$$

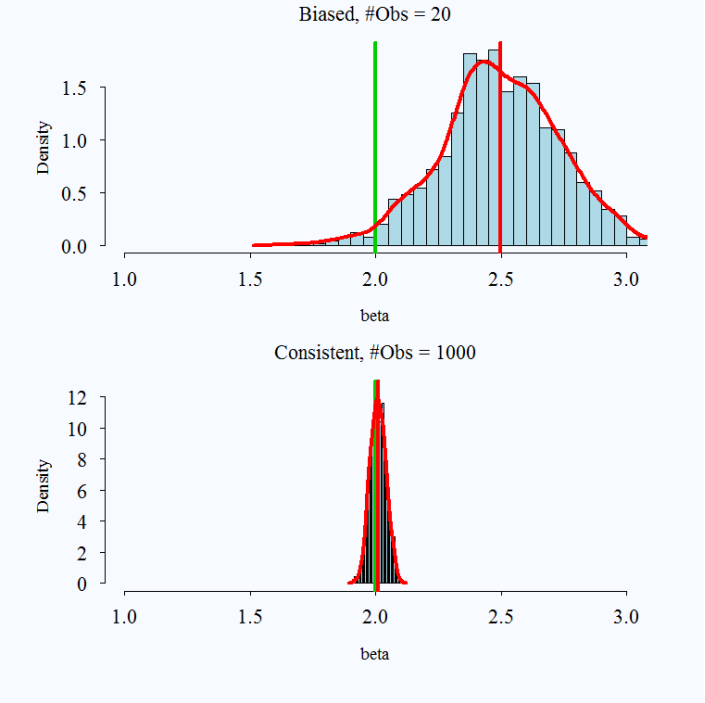
- Nếu ước lượng bị thiên lệch nhưng nhất quán, tăng cỡ mẫu có thể làm giảm mức độ thiên lệch.

Bias and Consistency

P1



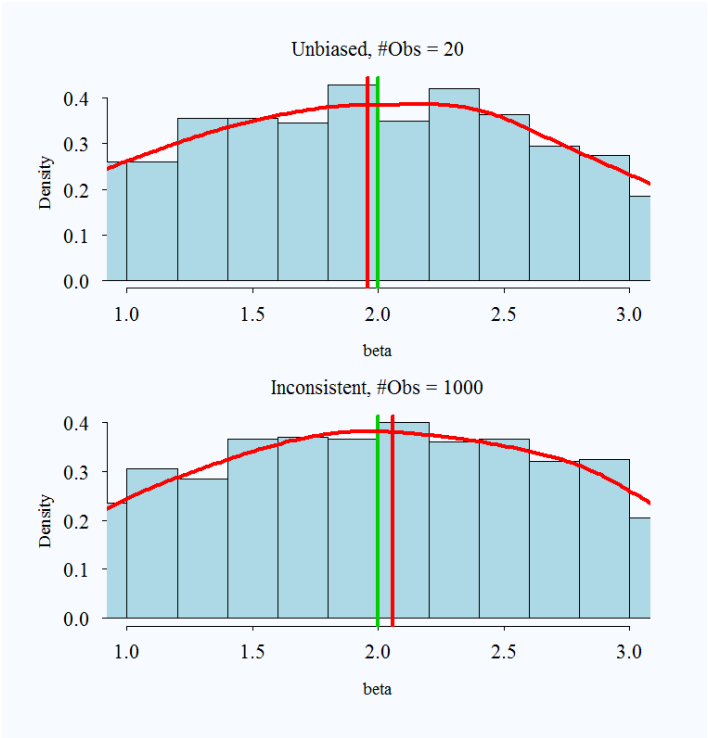
P2



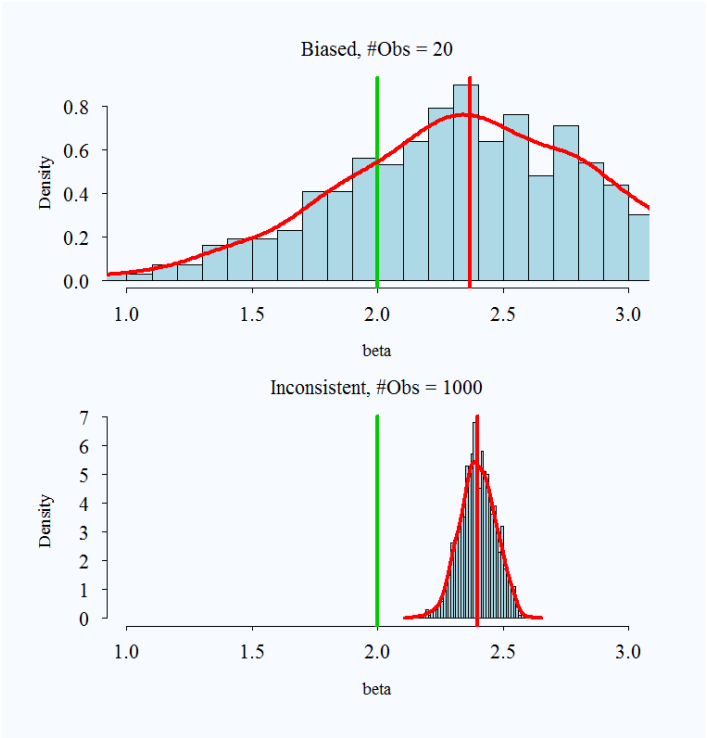
- P1: Ước lượng không chệch và nhất quán.
- P2: Ước lượng chệch nhưng nhất quán.

Bias and Consistency

P3



P4



- P3: Ước lượng không chệch và không nhất quán.
- P4: Ước lượng chệch và không nhất quán.

Hiệu lực nội tại bị phá vỡ khi nào và hậu quả gì xảy ra?

1. Phương sai của sai số thay đổi và tự tương quan (heteroskedasticy and autocorrelation)
2. Mô hình bị thiếu biến quan trọng (omitted variables bias)
3. Sai cấu trúc hàm (functional form misspecification)
4. Mẫu dữ liệu không ngẫu nhiên/hiện tượng tự lựa chọn mẫu (sample selection bias)
5. Quan hệ nhân quả đồng thời (simultaneous causality)
6. Sai số đo lường (measurement errors)

Hậu quả: ước lượng có thể không hiệu quả, bị thiên lệch, hoặc không nhất quán, và các kiểm định thống kê bị sai.

1. Phương sai của sai số thay đổi và tự tương quan

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$Var(u|x) \neq \sigma^2$$

hoặc

$$cov(u_i, u_j) \neq 0$$

- Ước lượng bằng OLS không bị chệch và vẫn nhất quán.
- Trị kiểm định sai, và khoảng tin cậy sai $\Rightarrow$ Ước lượng không có hiệu lực nội tại.

Chỉnh sửa bằng phương pháp White hoặc WLS/FGLS khi xảy ra hiện tượng phương sai thay đổi.

2. Mô hình thiếu biến quan trọng

- Ví dụ mô hình hồi quy chuẩn với hai biến giải thích:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

thỏa các điều kiện CLRM. Tuy nhiên không quan sát được x_2, do đó chúng ta sẽ ước lượng mô hình sau trên thực tế:

$$y = \beta_0 + \beta_1 x_1 + \underbrace{\beta_2 x_2 + u}_{v}$$

- Trong đó v là sai số gộp của cả sai số ngẫu nhiên u và biến không quan sát được x_2, $v = u + \beta_2 x_2$
- 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 x_2 lên x_1:

$$\sigma_{21} = \frac{cov(x_1, x_2)}{var(x_1)}$$

Đánh giá hướng chệch trong mô hình thiếu biến quan trọng

- Nếu $\beta_2 = 0$ (biến x_2 không phải là biến quan trọng) thì $\hat{\beta}_1$ không chệch.
- Nếu $\sigma_{21} = 0$ (x_1 và x_2 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 x_1 và biến không quan sát được x_2 thông qua hệ số σ_{21}.

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

Ví dụ trường hợp thiếu biến quan trọng trong mô hình tỷ suất thu nhập của đi học

Giả sử hàm hồi quy chuẩn là:

$$log(wage) = \beta_0 + \beta_1 educ + \underbrace{\beta_2 Ability + u}_{v}$$

- 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.

3. Sai cấu trúc hàm

Giả sử hàm hồi quy chuẩn là:

$$log(wage) = \beta_0 + \beta_1 educ + \underbrace{\beta_2 educ^2 + u}_{v}$$

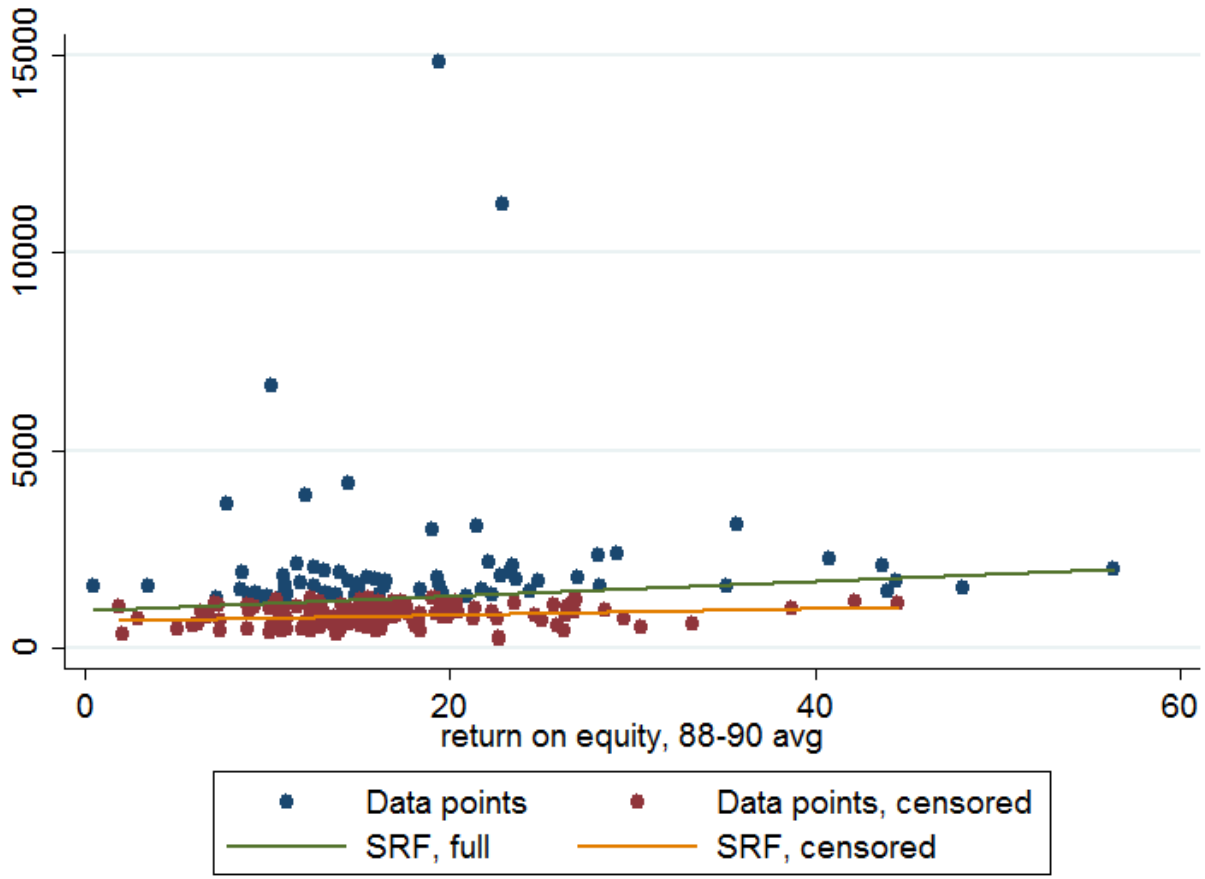
- Nếu nhà nghiên cứu bỏ sót biến $educ^2$ trong mô hình, ước lượng tỷ suất thu nhập khi đó là:

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{cov(educ, educ^2)}{var(educ)}$$

- Nếu đi học có quan hệ phi tuyến đến thu nhập (và kỳ vọng $\beta_2 < 0$), khi đó ước lượng của β_1 bị chệch xuống.
- Hậu quả giống trường hợp mô hình thiếu biến quan trọng.

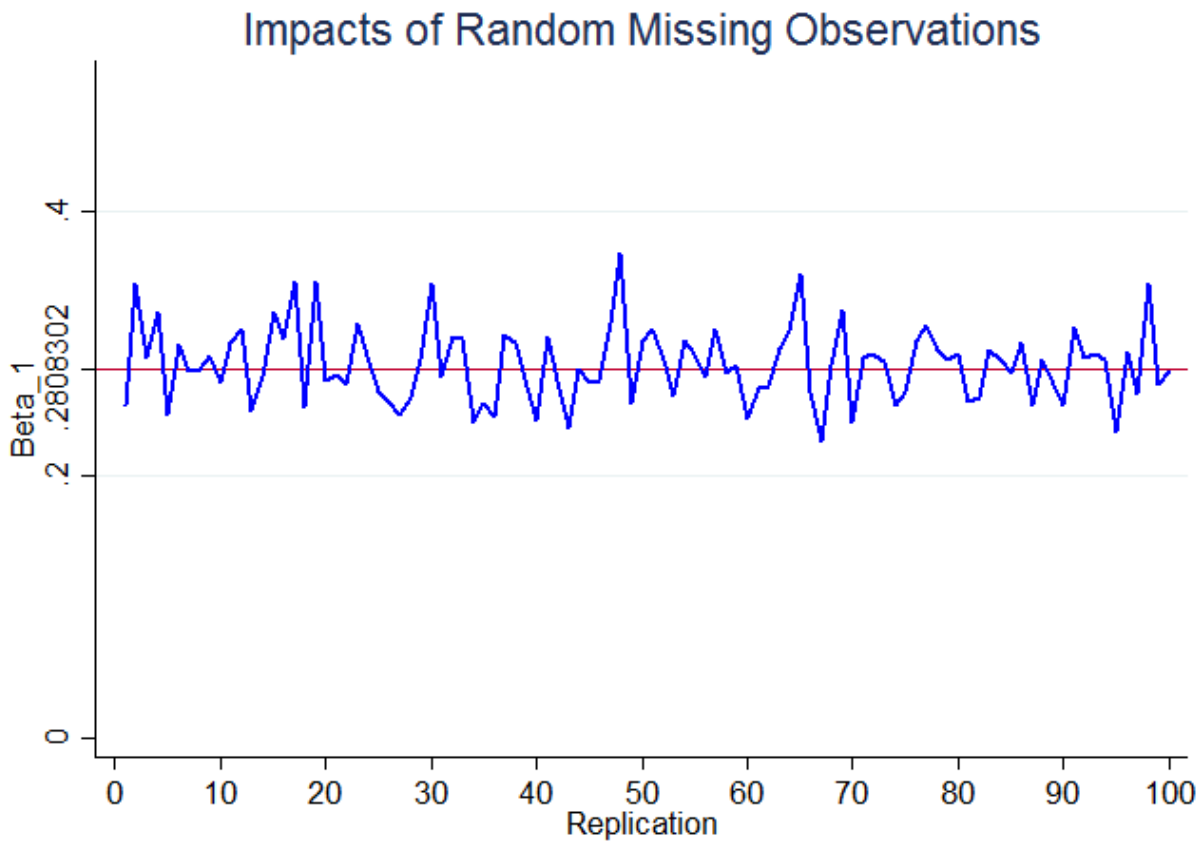
4. Dữ liệu không ngẫu nhiên và hiện tượng tự lựa chọn mẫu

Ảnh hưởng của vấn đề lựa chọn mẫu với biến phụ thuộc đến kết quả ước lượng:



Dữ liệu bị thiếu ngẫu nhiên

Không ảnh hưởng đến hiệu lực nội tại. 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.



Dữ liệu không ngẫu nhiên

- Dữ liệu bị thiếu không ngẫu nhiên dựa trên biến giải thích:
  - Không 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 có vấn đề lựa chọn mẫu dựa trên biến phụ thuộc:
  - 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.
  - Cần hiểu rõ bản chất của dữ liệu mới nhận diện được vấn đề lựa chọn mẫu!

5. Quan hệ nhân quả đồng thời

Ví dụ với giá cả và lượng tiêu thụ của hàng hóa quan sát được trên thị trường phụ thuộc đồng thời lẫn nhau:

$$Price = \beta_0 + \beta_1 Quantity + \beta_2 x + u$$

và

$$Quantity = \gamma_0 + \gamma_1 Price + \gamma_2 y + v$$

Ước lượng bằng OLS bị chệch và không có hiệu lực nội tại:

$$\hat{\beta}_1 = \beta_1 + \frac{\gamma_1 \sigma_u^2}{(1 - \gamma_1 \beta_1)\sigma_v^2} \neq \beta_1$$

6. Sai số đo lường

Giả sử hàm hồi quy chuẩn là:

$$wage = \beta_0 + \beta_1 educ + \beta_2 educ^2 + u$$

Thế nào là sai số đo lường?

- Sai số của biến phụ thuộc (ví dụ không ghi nhớ đủ các loại hình thu nhập ngoài tiền lương).
- Sai số của biến giải thích (ví dụ số năm đi học) có thể xảy ra do các loại hình học thêm bên ngoài học chính khóa.

Tác động của sai số đo lường của biến phụ thuộc đến ước lượng OLS

Sai số đo lường của biến phụ thuộc:

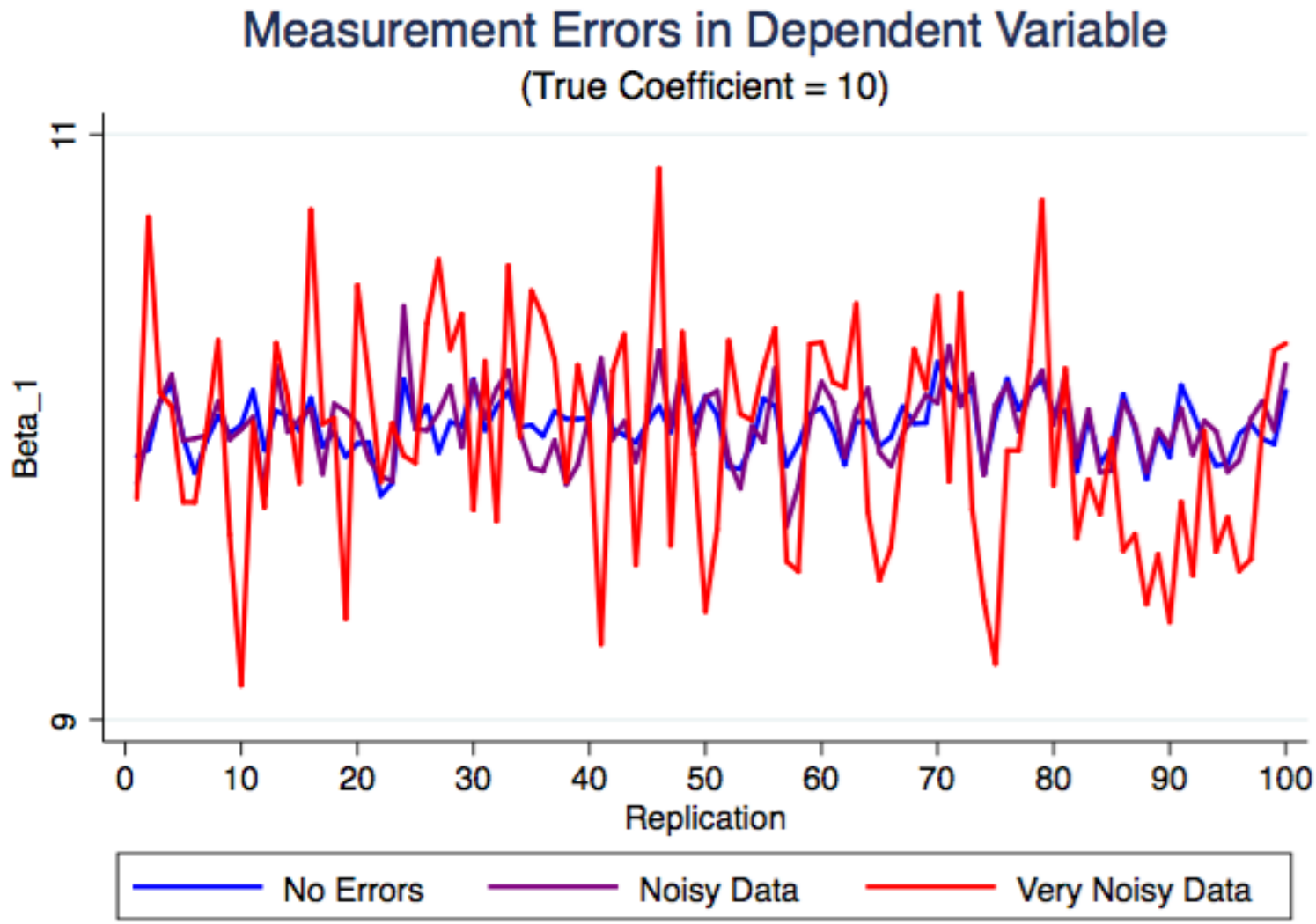
$$\widetilde{wage} = wage + v$$

với v là white noise. Khi đó chúng ta thực ước lượng mô hình:

$$\widetilde{wage} = \beta_0 + \beta_1 educ + \beta_2 educ^2 + (u + v)$$

- Mô hình vẫn thỏa các điều kiện CLRM, do đó ước lượng vẫn có hiệu lực nội tại.
- Tuy nhiên sai số càng lớn dẫn đến độ tin cậy của ước lượng càng giảm.

Mô phỏng Monte-Carlo trường hợp sai số đo lường đối với biến phụ thuộc



Sai số đo lường của biến giải thích có thể dẫn đến vi phạm các giả định CLRM và ước lượng sẽ không có hiệu lực nội tại

- Giả sử hàm hồi quy chuẩn là:

$$log(wage) = \beta_0 + \beta_1 educ + u$$

nhưng biến giải thích trong mô hình bị nhiễu thông tin, chúng ta quan sát được $educ^* = educ + \omega$.

- ω gọi là nhiễu sai số đo lường cổ điển:
$cov(educ, \omega) = 0$, $cov(\omega, u) = 0$, $E[\omega] = 0$, $var(\omega) = \sigma_\omega^2$

- Mô hình ước lượng khi này là:

$$log(wage) = \beta_0 + \beta_1 educ^* + \underbrace{u - \beta_1 \omega}_{v}$$

Tác động của sai số đo lường đến ước lượng OLS

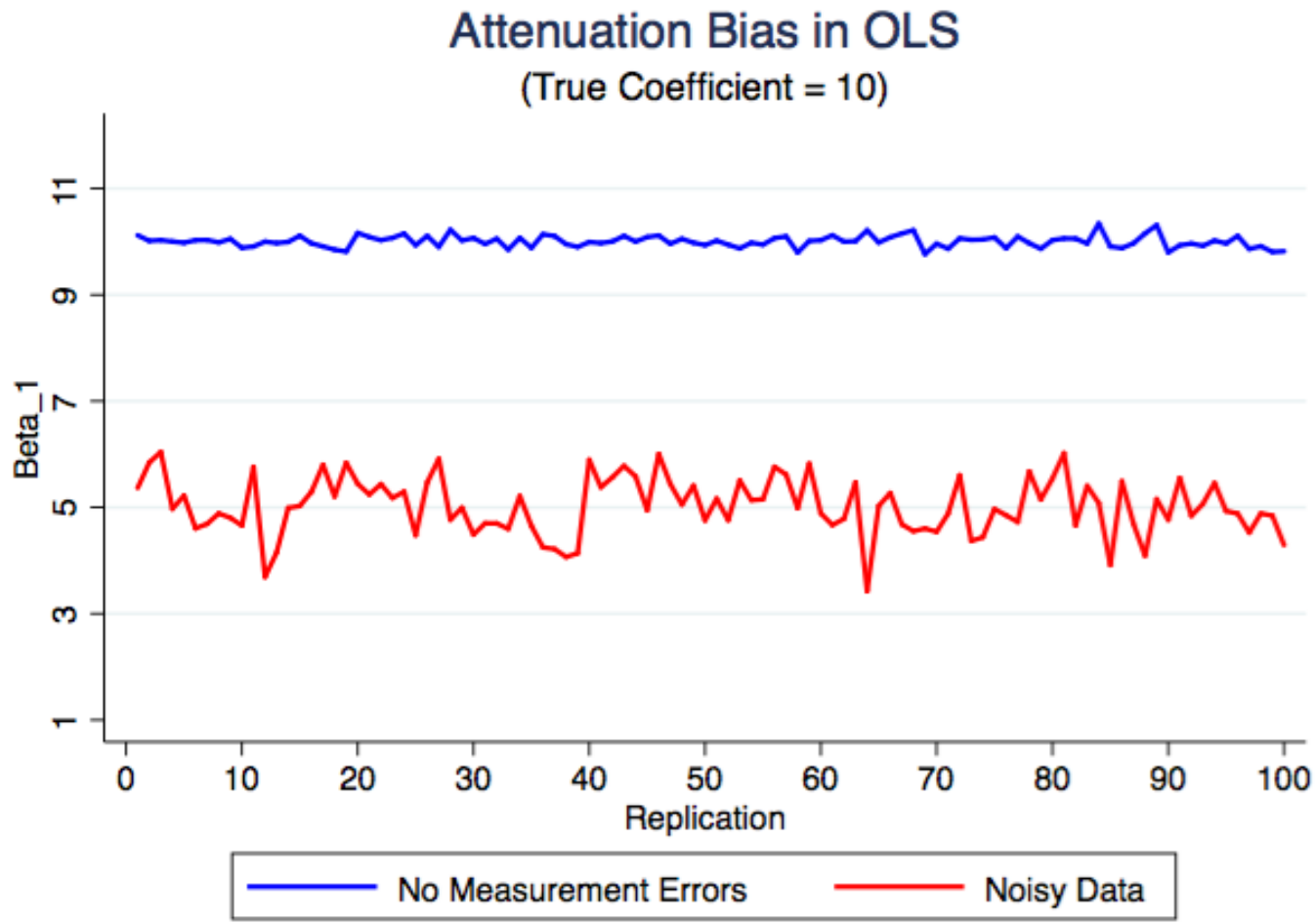
Nếu chúng ta ước lượng mô hình trên bằng OLS:

$$
\begin{aligned}
plim(\hat{\beta}_1) &= \beta_1 + \frac{cov(educ^*, v)}{var(educ^*)} \\
&= \beta_1 + \frac{cov(educ + \omega, u - \beta_1\omega)}{var(educ + \omega)} \\
&= \beta_1 - \beta_1 \frac{cov(\omega, \omega)}{var(educ) + var(\omega)} \\
&= \beta_1 \frac{\mathbf{var(educ)}}{\mathbf{var(educ)} + \sigma_\omega^2}
\end{aligned}
$$

Do $\frac{\mathbf{var(educ)}}{\mathbf{var(educ)}+\sigma_\omega^2} < 1$ nên ước lượng của $|\hat{\beta}_1| < |\beta_1|$. Đây gọi là vấn đề chệch hướng giảm thiểu (attenuation bias) khi xảy ra vấn đề sai số đo lường.

Mô phỏng Monte-Carlo để chứng minh đặc tính thống kê của các ước lượng dựa trên dữ liệu mô phỏng

- Tạo bộ dữ liệu mô phỏng
- Tạo biến giải thích có sai số đo lường
- Chứng minh tham số ước lượng bị thiên lệch suy giảm.



Trường hợp sai số đo lường có tính hệ thống

- Giả sử hàm hồi quy chuẩn là:

$$log(consumption) = \beta_0 + \beta_1 wage + u$$

nhưng biến giải thích trong mô hình bị báo cáo thiếu, chúng ta quan sát được $wage^* = wage - \omega$, với $\omega > 0$.

- Mô hình ước lượng khi này là:

$$log(consumption) = \beta_0 + \beta_1 wage^* + \underbrace{u + \beta_1 \omega}_{v}$$

$$plim(\hat{\beta}_1) = \beta_1 + \frac{cov(wage^*, u + \beta_1 \omega)}{var(wage^*)}$$

- Giả sử thu nhập báo cáo thấp hơn 10% thu nhập thực, $\omega = .1 * wage$. Khi đó ước lượng của β_1 sẽ bị phóng đại 10%.

Hình thức sử lý khi ước lượng không có hiệu lực nội tại?

Đã học kỳ trước...

- Khi mô hình thiếu biến quan trọng: Tìm biến đại diện (proxy) cho tố chất cá nhân (IQ, điểm học...) trong mô hình tỷ suất thu nhập của đi học.
- Cấu trúc hàm: Thêm biến lũy thừa/biến tương tác và kiểm định RESET.
- Phương sai thay đổi: Sử dụng robust standard errors hoặc hồi quy với quyền số.

Các phương pháp sẽ học trong học phần này để đảm bảo hiệu lực nội tại của ước lượng

- Hồi quy Tobit và Heckman selection để sử lý vấn đề dữ liệu bị chặn hoặc dữ liệu không ngẫu nhiên.
- Dùng dữ liệu bảng với tác động cố định (fixed effects) để sử lý trường hợp thiếu biến quan trọng trong mô hình bằng giả định nhân tố không quan sát được không thay đổi theo thời gian.
- Phương pháp hồi quy hai bước với biến công cụ để sử lý trường hợp thiếu biến quan trọng trong mô hình/biến nội sinh.
- Phương pháp hồi quy hệ phương trình trong trường hợp các biến có quan hệ nhân quả đồng thời.

Sau cùng, học viên sẽ học cách ứng dụng các phương pháp trên vào thiết kế nghiên cứu đánh giá tác động chính sách.

Mô hình với biến phụ thuộc bị giới hạn (Regression with limited dependent variables)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 12 tháng 3 năm 2019

Các loại hình biến phụ thuộc bị giới hạn

- Đơn giản nhất là biến phụ thuộc là biến xác suất xảy ra một sự kiện, có hoặc không xảy ra.
  - Doanh nghiệp có bị phá sản hay không; có vay tiền ngân hàng không.
- Biến phụ thuộc thể hiện hành vi lựa chọn trong mô hình đa lựa chọn:
  - Lựa chọn smartphone thương hiệu gì trong số các mặt hàng bán trên thị trường: Apple, Samsung, LG, Xiaomi, Oppo...
- Biến phụ thuộc là biến xếp hạng/thứ tự:
  - Xếp hạng một bộ phim từ: rất kém, kém, trung bình, hay, rất hay.
- Biến phụ thuộc là số lần xảy ra một sự kiện:
  - Số lần một người vi phạm hành vi bạo lực gia đình, số lần đi khám bệnh một năm.
- Biến phụ thuộc có giá trị bị giới hạn:
  - Tiền lương từ các điều tra thu nhập bị chặn dưới ở 0 đồng; số giờ làm việc một tuần không vượt quá $7 * 24 = 168$ giờ.

Tại sao kiểm soát vấn đề biến phụ thuộc bị giới hạn rất quan trọng?

- Không thỏa các giả định của mô hình hồi quy tuyến tính cổ điển CLRM $\Rightarrow$ Ước lượng có thể gặp một hoặc nhiều các vấn đề sau:
  - Phương sai của sai số thay đổi
  - Ước lượng bị chệch
  - Ước lượng không nhất quán
  - Ước lượng không hiệu quả
- Để hiểu xảy ra vấn đề gì thì phải dựa vào hiểu biết của dữ liệu và lý thuyết để giải thích.
- Lựa chọn khi phải đối phó với biến phụ thuộc bị giới hạn:
  - Tiếp tục sử dụng OLS và chấp nhận các vấn đề có thể gặp phải.
  - Sử dụng phương pháp phù hợp với dữ liệu.

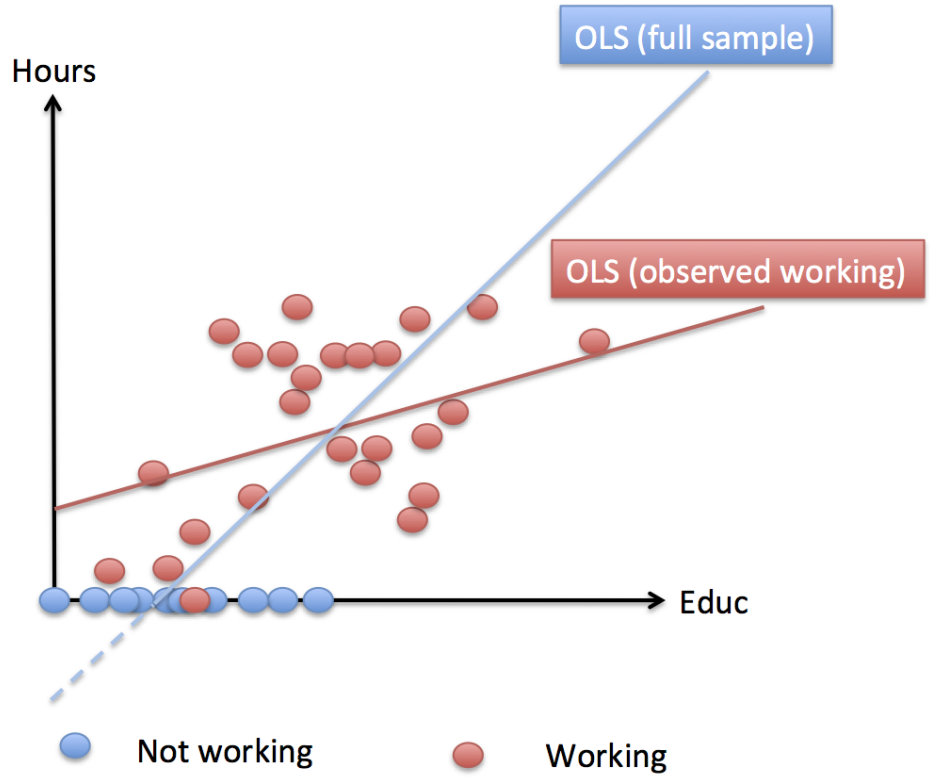
Các mô hình tương ứng với các loại biến phụ thuộc bị giới hạn

- Mô hình xác suất: **LPM, Logit, Probit**
- Mô hình đa lựa chọn: Multinomial logit/probit, conditional logit
- Mô hình biến xếp hạng: Ordered logit/probit
- Mô hình số lần xảy ra một sự kiện: Poisson count model
- Mô hình biến phụ thuộc bị chặn: **Tobit**, censored/truncated model
- Mô hình với mẫu dữ liệu bị vấn đề tự lựa chọn: **Sample selection model/Heckman correction model**

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 của số giờ đi làm trong năm



- Biến phụ thuộc bị chặn dưới tại 0.
- Ước lượng bằng OLS với nhóm làm việc có thể 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 số giờ làm việc 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

Bản chất của mô hình Tobit là hồi quy hai bước theo tuần tự:

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

- Bước 2: Ước lượng các nhân tố ảnh hưởng đến số giờ lao động bằng OLS, và điều chỉnh hệ số ước lượng để tính đến xác suất có đi làm hay không đã thực hiện ở bước 1.

Xây dựng mô hình Tobit

Thông thường hành vi làm việc của một người được diễn giải bởi hàm ẩn:

$$y^* = X * \beta + u, \qquad u \sim N(0, \sigma^2)$$

trong đó y^* là biến phụ thuộc ẩn (latent variable), không quan sát được. Chúng ta quan sát được biến y là số giờ làm việc trong năm:

$$y = max(0, y^*)$$

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

Xây dựng mô hình Tobit

Chúng ta có thể tìm được phương trình ước lượng của biến phụ thuộc là trung bình có quyền số của xác xuất đi làm và số giờ đi làm:

$$E[y|x] = \underbrace{P(y=0|x) * E[y=0|x]}_{=0} + \underbrace{P(y>0|x) * E[y|y>0,x]}_{>0}$$

trong đó:

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

với $\Phi(.)$ là hàm tích lũy phân phối chuẩn, được tính tại giá trị $\frac{X*\beta}{\sigma}$

Xây dựng mô hình Tobit

Ngoài ra, chúng ta có biểu thức sau (bài tập 3):

$$E[y|y > 0, x] = X * \beta + \sigma\lambda(\frac{X * \beta}{\sigma})$$

với $\lambda(c) = \frac{\phi(c)}{\Phi(c)}$, còn được gọi là tỷ số Mills nghịch đảo (inverse Mills ratio - IMR), là tỷ lệ giữa hàm mật độ và hàm tích lũy của phân phối chuẩn được tính tại giá trị c.

Xây dựng mô hình Tobit

Từ các công thức trên, chúng ta có phương trình hàm hồi quy Tobit như sau:

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

So sánh với hồi quy OLS:

$$E[y|x] = X * \beta$$

- Hồi quy Tobit là hàm phi tuyến của các tham số và biến giải thích thông qua hàm tích lũy và phân phối xác suất.
- Có thể chứng minh (!) là giá trị dự báo của biến phụ thuộc của hàm Tobit là dương với mọi giá trị của X, khác so với hồi quy OLS có thể nhận giá trị dự báo âm.

Ước lượng mô hình Tobit và diễn giải ý nghĩa

- Mô hình Tobit được ước lượng bằng phương pháp MLE thay vì OLS.
- Diễn giải các hệ số ước lượng:
  - Với OLS thì β là tác động biên của các biến giải thích lên biến phụ thuộc và không đổi.
  - Với Tobit thì chúng ta phải tính tác động biên từ phương trình hàm hồi quy bằng đạo hàm bậc nhất của biến phụ thuộc theo biến giải thích.

Tác động biên trong mô hình Tobit

- Nếu biến giải thích là biến liên tục:

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

- Nếu biến giải thích là biến rời rạc:

$$\Delta y = E[y|x_1] - E[y|x_0]$$

- Tác động biên của mô hình Tobit sẽ phụ thuộc vào giá trị tham chiếu, loại biến (liên tục hay rời rạc).
- Tương tự như hồi quy Logit/Probit, $\Phi(\frac{X*\beta}{\sigma})$ 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

Giả sử chúng ta muốn ước lượng mô hình hàm cung số giờ lao động như sau:

$$\begin{aligned} hours =& \beta_0 + \beta_1 netincome + \beta_2 educ + \beta_3 exper + \beta_4 expersq + \beta_5 age \\ &+ \beta_6 KIDS6 + \beta_7 KIDS7 + u \end{aligned}$$

trong đó có 325/753 quan sát có số giờ làm việc bằng 0.

So sánh ước lượng OLS và Tobit thế nào?

	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

Ước tính tác động biên

- 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, là 80.65*.645 = 52 giờ. Ước lượng OLS là 28.76 giờ.
- Tác động biên lên số giờ lao động của phụ nữ chưa có con nhỏ dưới 6 tuổi so với có một con dưới 6 tuổi, tại giá trị trung bình của các biến giải thích khác, là 503.5 giờ.
- Chỉ giới hạn vào 428 phụ nữ đang tham gia lao động, ước lượng OLS và Tobit cho kết quả giống nhau.

Tổng kết mô hình Tobit

- Khi dữ liệu quan sát được bị chặn tại một ngưỡng giá trị nào đó thì ước lượng OLS có thể bị chệch hoặc gặp phải vấn đề dự báo không chính xác.
- Sử dụng mô hình Tobit và phương pháp MLE có thể sửa được lỗi của mô hình OLS.
- Diễn giải ý nghĩa của các tham số của mô hình Tobit phức tạp hơn mô hình OLS do giá trị dự báo là hàm phi tuyến của các biến giải thích và tham số ước lượng – tương tự như hàm hồi quy xác suất Logit hoặc Probit.
- Trường hợp liên quan: Khi dữ liệu gặp phải vấn đề tự lựa chọn mẫu (ví dụ không quan sát được một số cá nhân có các thuộc tính nhất định) thì cần sử dụng hàm hồi quy điều chỉnh mẫu - Heckman selection model (cuối môn học).