

Hiệu lực nội tại và hiệu lực ngoại vi (Internal validity and external validity)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

18/02/2020

Hồi quy tuyến tính cổ điển CLRM và các giả định

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

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:

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

⇒ 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:

$$\text{Var}(u|x_1, \dots, x_k) = \sigma^2$$

- ▶ 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:

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

- ▶ Viết dưới dạng chuẩn hóa:

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

- ▶ Các kiểm định t, F có hiệu lực.

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.

Khi các giả định của mô hình CLRM không thỏa

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

Hiệu lực nội tại trong mô hình OLS

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

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) :

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

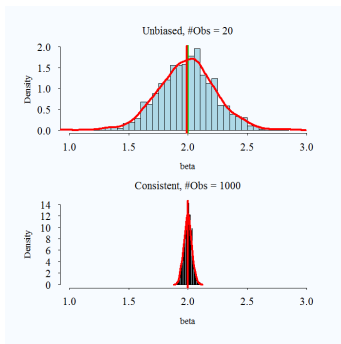
- ▶ Nhất quán: Phân phối của ước lượng của tham số hội tụ (còn gọi là tiệm cận - asymptotic) về giá trị thực khi cỡ mẫu tăng đến vô cùng:

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

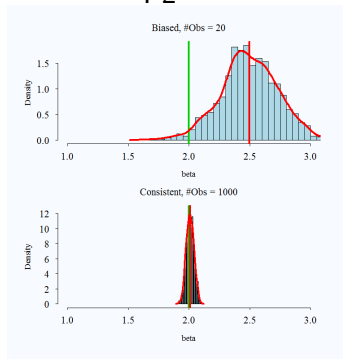
- ▶ Nếu ước lượng bị thiên lệch nhưng nhất quán, tăng cỡ mẫu có thể làm giảm mức độ thiên lệch.

Bias and Consistency

P1



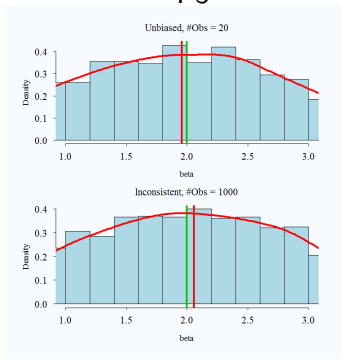
P2



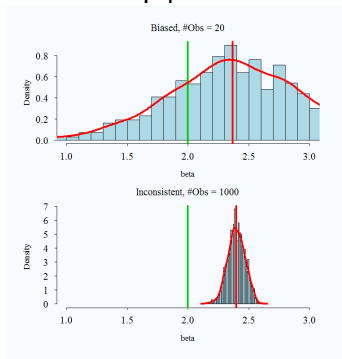
- ▶ P1: Ước lượng không chệch và nhất quán.
- ▶ P2: Ước lượng chệch nhưng nhất quán.

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 (heteroskedasticity 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

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

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

hoặc

$$\text{cov}(u_i, u_j) \neq 0$$

- ▶ 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 :

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

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} .

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

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:

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

- ▶ Tổ chất cá nhân *Ability* được kỳ vọng có tác động đến tiền lương.
- ▶ Tổ chất cá nhân tương quan với trình độ học vấn.
- ▶ Tổ 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:

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

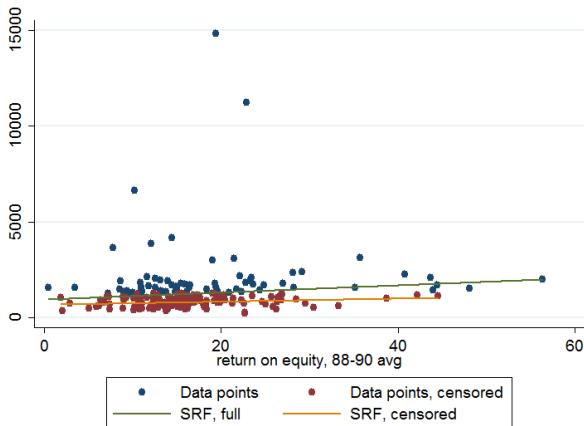
- ▶ Nếu nhà nghiên cứu bỏ sót biến educ^2 trong mô hình, ước lượng tỷ suất thu nhập khi đó là:

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

- ▶ 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.

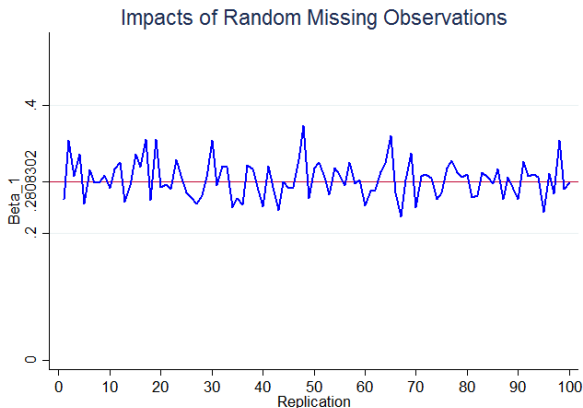
4. Dữ liệu không ngẫu nhiên và hiện tượng tự lựa chọn mẫu

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 xử 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:

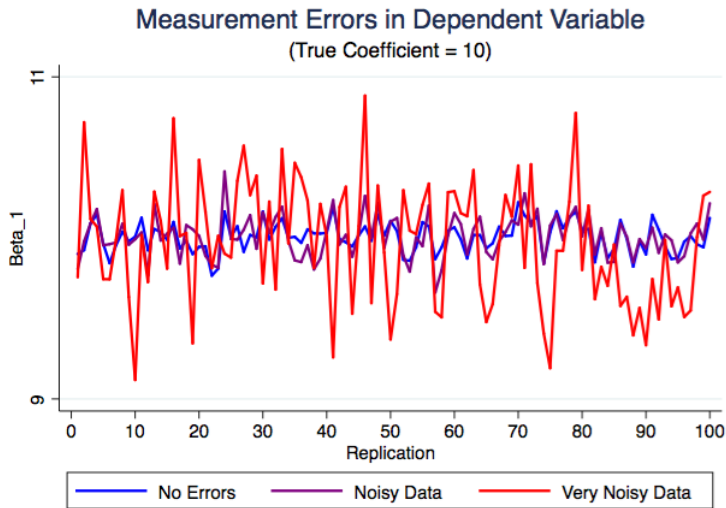
$$\widetilde{wage} = wage + v$$

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

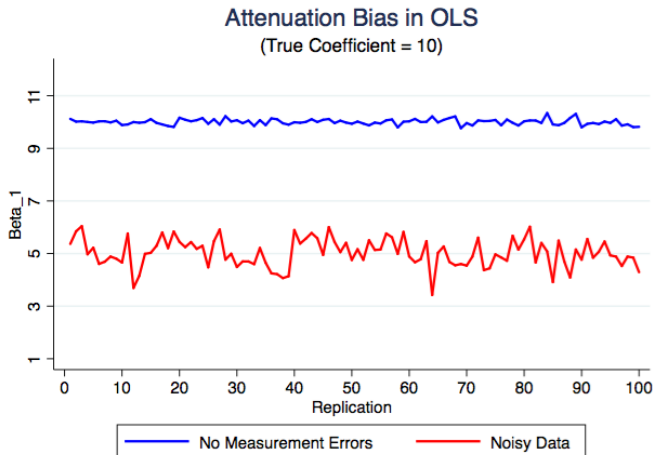
Nếu chúng ta ước lượng mô hình trên bằng OLS:

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

Do $\frac{\text{var}(\text{educ})}{\text{var}(\text{educ}) + \sigma_\omega^2} < 1$ 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:

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

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

- ▶ Mô hình ước lượng khi này là:

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

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

- ▶ Giả sử thu nhập báo cáo thấp hơn 10% thu nhập thực, $\omega = .1 * \text{wage}$. 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 tổ 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 xử 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 xử 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 xử 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.