Phương pháp Hồi quy với Biến Công cụ (Regression with Instrumental Variables)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 16 tháng 4 năm 2018

Ôn tập lý thuyết hồi quy tuyến tính cổ điển CLRM

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

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

- y gọi là biến phụ thuộc/biến được giải thích.
- x_1, x_2 là biến độc lập/biến giải thích.
- u là sai số, bao gồm tất cả những yếu tố khác ảnh hưởng đến y nhưng không nằm trong x_1, x_2.
- β_0, β_1, β_2 là các tham số trong mô hình.

Các giả định đối với hồi quy đa biến

Tương tự như các điều kiện của hồi quy đơn biến:

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.

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

Giả định phương sai của sai số không đổi (homoskedasticity)

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

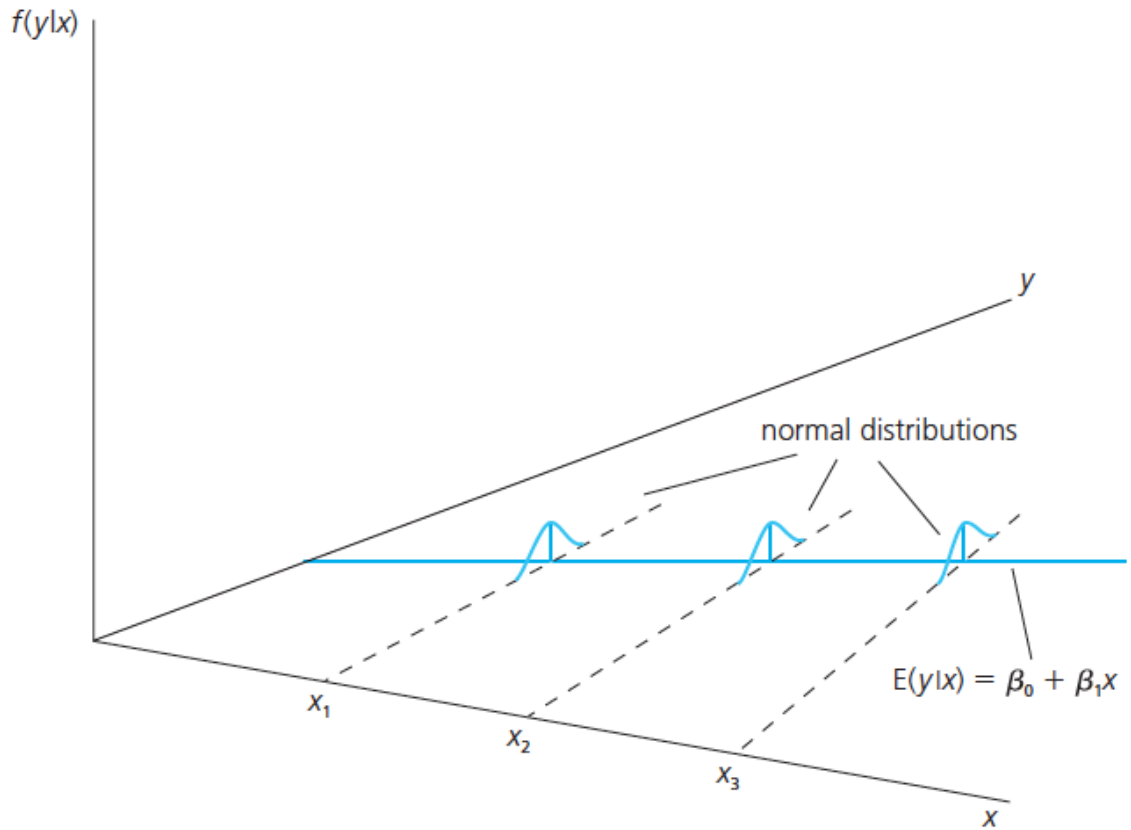
$$Var(u|x_1, ..., x_k) = \sigma^2$$

- 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**).
  - Ước lượng của β là hàm tuyến tính của biến phụ thuộc.
  - Trong tất cả các ước lượng tuyến tính, OLS có phương sai của ước lượng là nhỏ nhất.
  - Không chệch, $E(\hat{\beta}) = \beta$.

Giả định về phân phối mẫu của sai số

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:

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

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.

Hiệu lực nội tại trong mô hình OLS

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

- Ước lượng của β là không chệch và nhất quán:

$$E[\hat{\beta}] = \beta \tag{1}$$

$$plim(\hat{\beta}) \rightarrow \beta \tag{2}$$

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

Hiệu lực nội tại bị phá vỡ khi nào?

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

1. 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 $\beta_2 = 0$, khi biến x_2 không phải là biến quan trọng.
- Nếu $\sigma_{21} = 0$, khi x_1 và x_2 không tương quan, thì $\hat{\beta}_1$ cũng không chệch.
- Nếu không phải 2 trường hợp trên, 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.

2. Sai cấu trúc hàm

Giả sử hàm hồi quy chuẩn là:

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

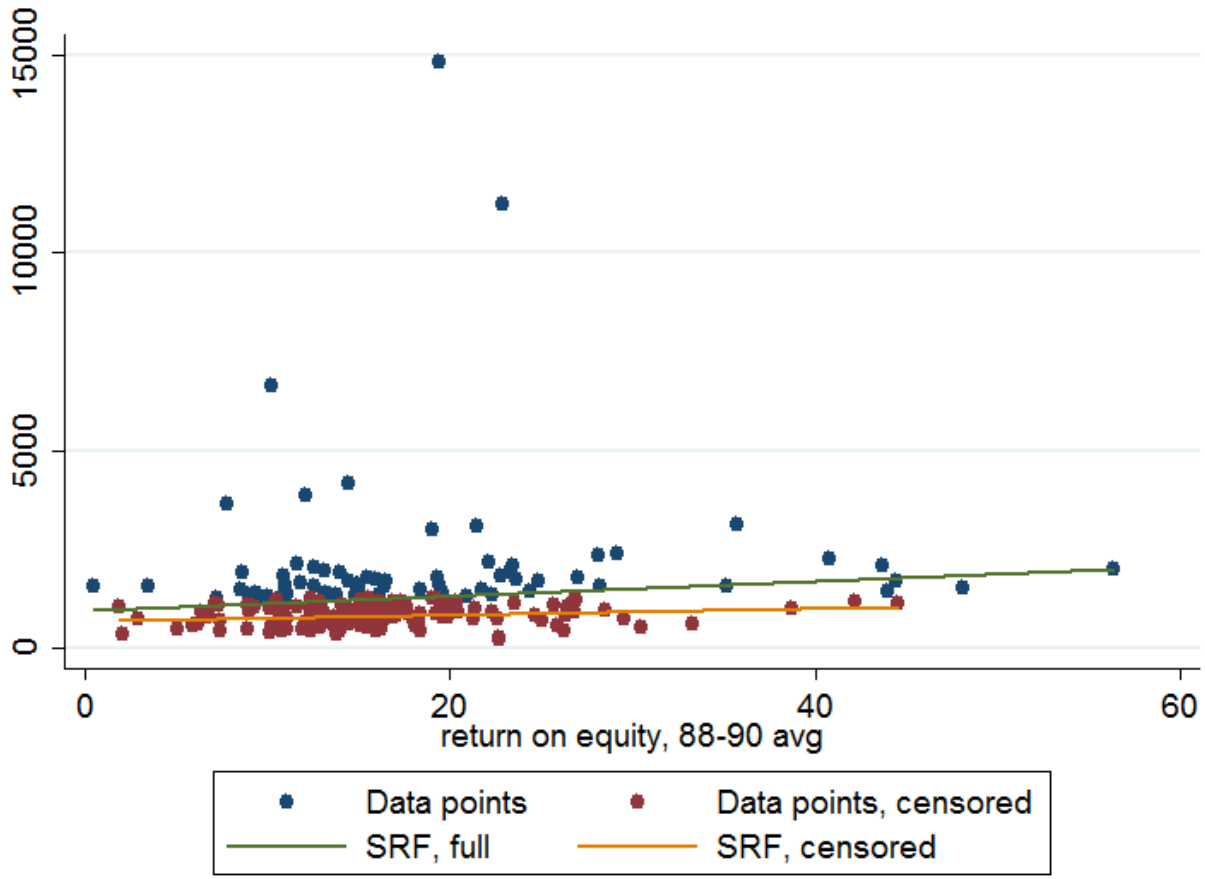
- Nếu nhà nghiên cứu bỏ sót biến $educ^2$ trong mô hình, ước lượng tỷ suất thu nhập khi đó là:

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{cov(educ, educ^2)}{var(educ)} \quad (3)$$

- 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.

3. Tự lựa chọn mẫu

Hàm hồi quy mẫu khi xảy ra vấn đề lựa chọn mẫu với biến phụ thuộc



- Nếu mẫu được quan sát không ngẫu nhiên $\Rightarrow$ không đảm bảo tính đại diện của ước lượng.

Tự lựa chọn mẫu

- Dữ liệu bị thiếu ngẫu nhiên: không ảnh hưởng đến hiệu lực nội tại
- 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: chỉ điều tra thu nhập và tình trạng học vấn của nhóm cá nhân học không quá 12 năm.
- 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.
  - Cần kỹ thuật cao cấp để xử lý.

4. Nhân quả đồng thời

Ví dụ với giá cả và lượng tiêu thụ của hàng hóa quan sát được trên thị trường:

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

và

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

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

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

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

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

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

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.

6. Sai số đo lường

Giả sử hàm hồi quy chuẩn là:

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

Thế nào là sai số đo lường?

- 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.
- Sai số của biến phụ thuộc (ví dụ không ghi nhớ đủ các loại hình thu nhập ngoài tiền lương).

Tác động của sai số đo lường đến ước lượng OLS

Sai số đo lường của biến phụ thuộc:

- Ít nghiêm trọng hơn sai số của biến giải thích
- Ước lượng vẫn có hiệu lực nội tại
- Sai số càng lớn dẫn đến độ tin cậy của ước lượng càng giảm.

Sai số đo lường của biến giải thích:

- Dẫn đến vi phạm các giả định CLRM và ước lượng sẽ không có hiệu lực nội tại.

Tác động của sai số đo lường đến ước lượng OLS: Trường hợp nhiễu thông tin

- 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

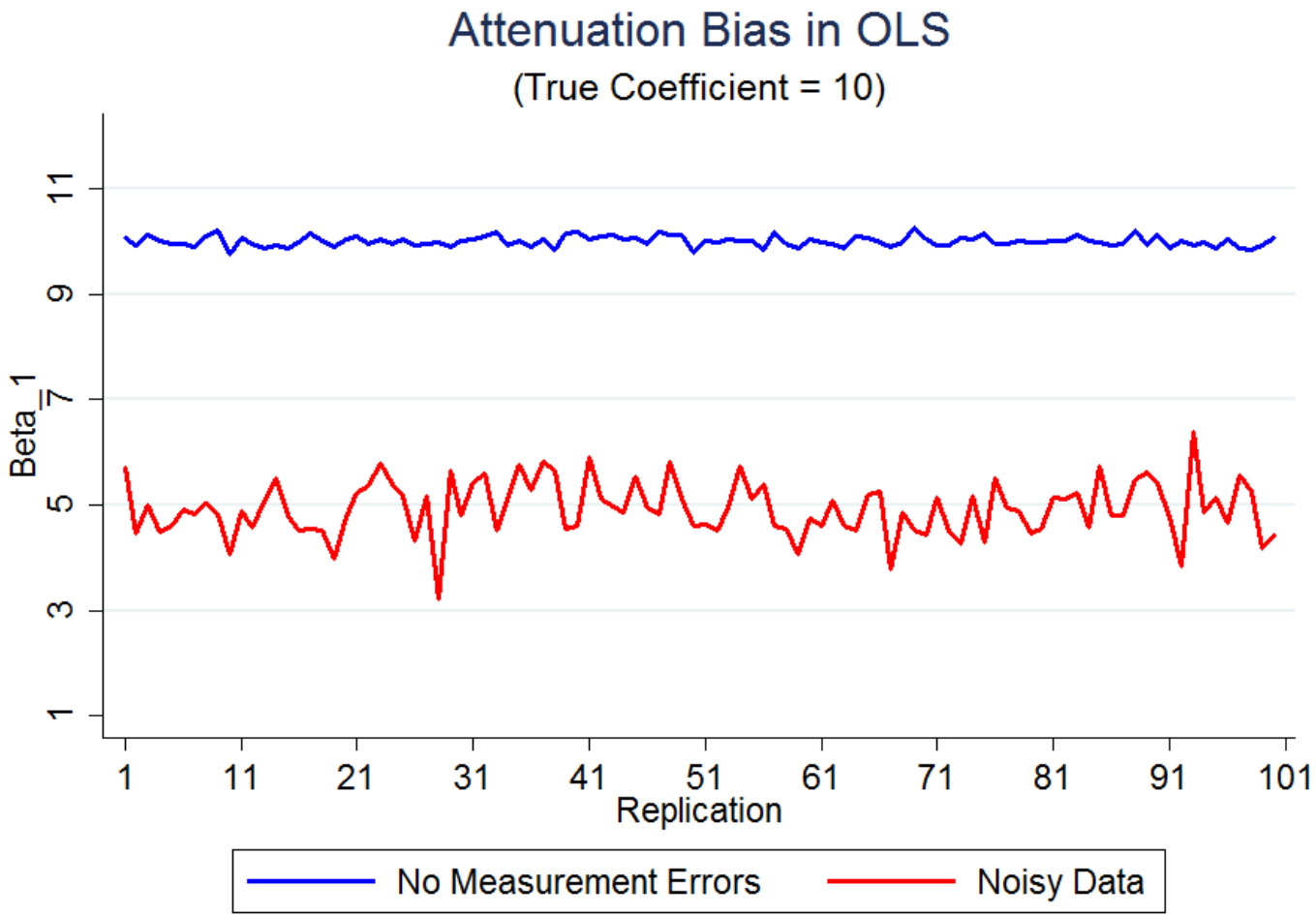
Nếu chúng ta ước lượng mô hình trên bằng OLS:

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

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%.

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

- Giả sử hàm hồi quy chuẩn là:

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

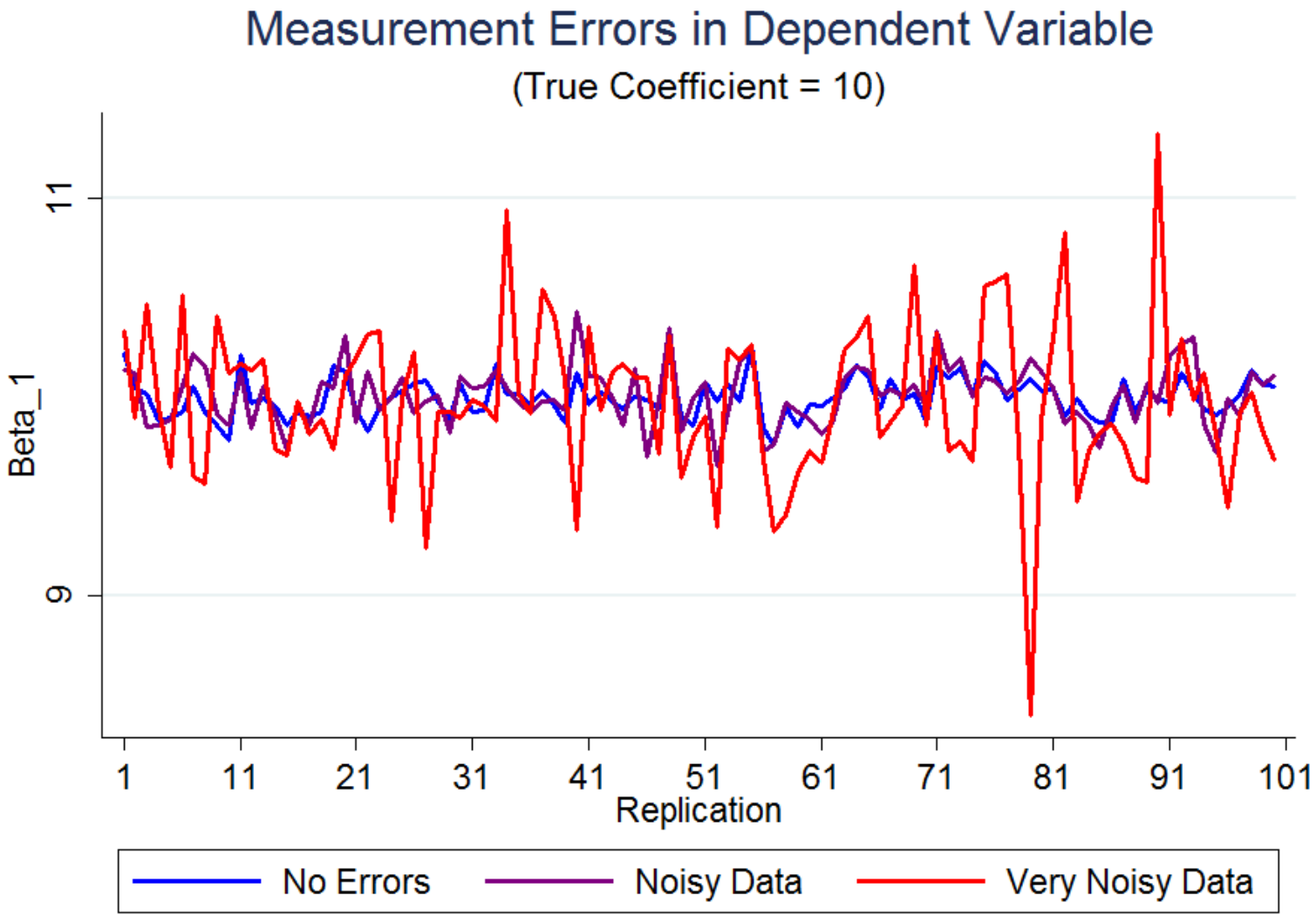
nhưng biến phụ thuộc trong mô hình bị nhiễu thông tin, chúng ta quan sát được $wage^* = wage + \omega$, với ω là white noise.

- Mô hình ước lượng khi này là:

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

- Ước lượng của β_1 vẫn không chệch và nhất quán nếu $cov(educ, v) = 0$, nhưng có thể không hiệu quả.

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



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

- Tìm biến đại diện cho tố chất cá nhân (IQ, điểm học...)
- Thêm biến lũy thừa/biến tương tác.
- Dùng phương pháp DiD khi có dữ liệu bảng để loại trừ nhân tố không quan sát được không thay đổi theo thời gian có tương quan với phần dư.
- Hồi quy với quyền số.
- **Phương pháp hồi quy với biến công cụ.**

Phương pháp hồi quy với biến công cụ

- Giả sử hàm hồi quy chuẩn là:

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

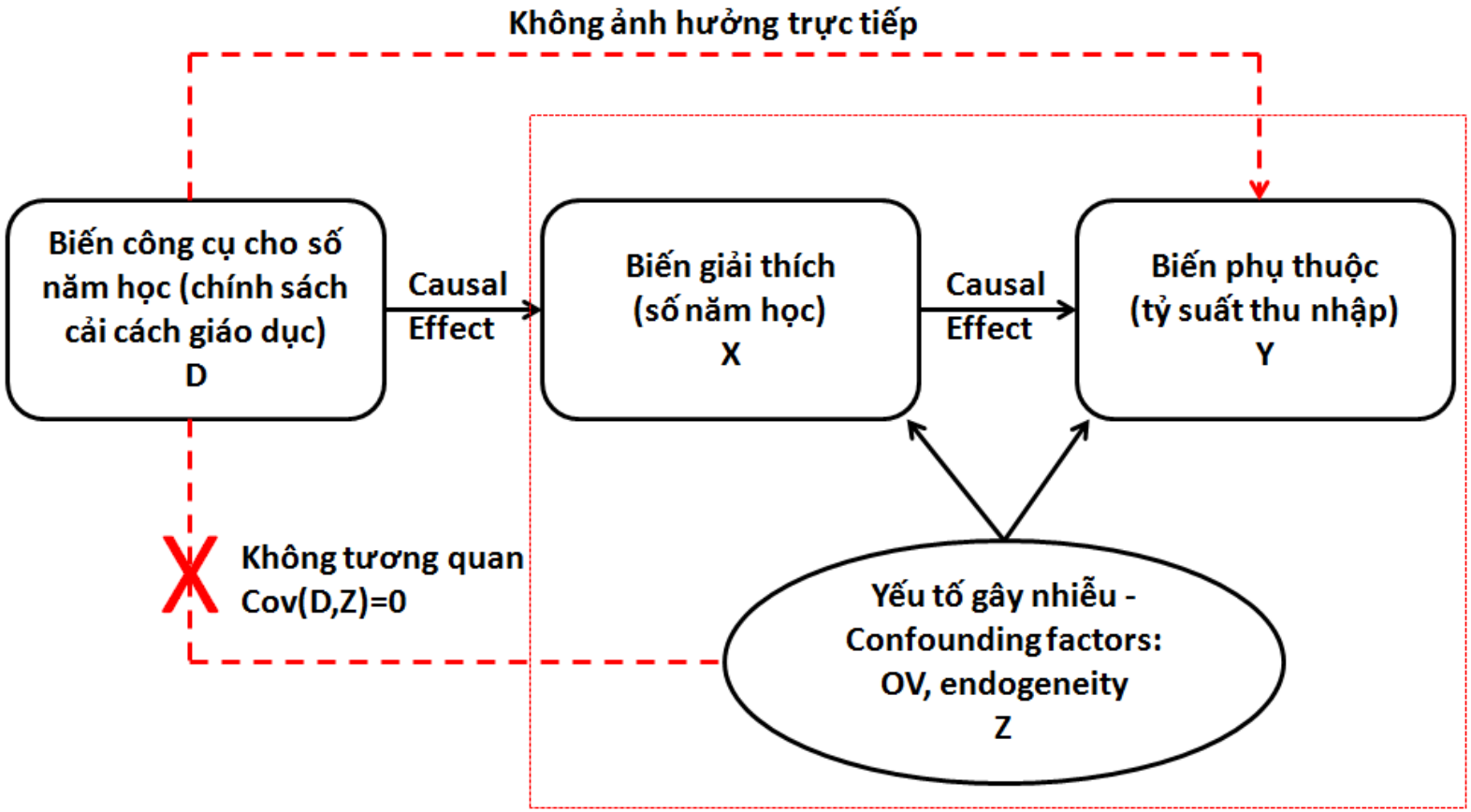
- Chúng ta biết giả định của CLRM bị vi phạm do biến quan trọng (tố chất cá nhân) trong mô hình không quan sát được có ảnh hưởng đến biến giải thích, $cov(educ, v) \neq 0$:

$$E[\beta_1] = \beta_1 + \frac{cov(educ, v)}{var(educ)}$$

- Biến *educ* được gọi là **biến nội sinh (endogenous variable)**, và mô hình trên gặp phải vấn đề biến nội sinh.
- Ước lượng OLS của mô hình bị vấn đề biến nội sinh không có hiệu lực nội tại.
- Vấn đề biến nội sinh là vấn đề nghiêm trọng nhất trong nghiên cứu định lượng!

Giả sử tồn tại một biến Z nào đó có thuộc tính sau:

- Z có tương quan với biến nội sinh *educ*, $cov(educ, Z) \neq 0$.
- Z không tương quan với phần dư của mô hình, $cov(Z, v) = 0$ (nói cách khác, Z không tác động trực tiếp lên biến phụ thuộc Y, nhưng Z có thể tác động lên biến phụ thuộc thông qua tác động lên biến nội sinh)



$$\begin{aligned} cov(Z, Y) &= cov(Z, \beta_0 + \beta_1 educ + v) \\ &= \beta_1 cov(Z, educ) + cov(Z, v) \end{aligned}$$

- Do $cov(Z, v) = 0 \Rightarrow$

$$\beta_1 = \frac{cov(Z, Y)}{cov(Z, educ)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(educ_i - \overline{educ})} \qquad (4)$$

- Ước lượng β_1 thông qua Z được gọi là ước lượng biến công cụ, khác với ước lượng bằng OLS.

Cơ chế của phương pháp biến công cụ

Chúng ta muốn ước lượng tác động của giáo dục lên thu nhập

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

Trong biến giáo dục *educ* có 2 phần:

- Phần ngẫu nhiên, không bị tác động bởi tố chất cá nhân (phần này bị ảnh hưởng bởi chính sách), là phần chúng ta muốn giữ lại.
- Phần nội sinh, do tố chất cá nhân quyết định. Phần này làm cho mô hình mất hiệu lực nội tại do tương quan với phần dư (bao gồm tố chất cá nhân trong đó).
  - Nếu chúng ta có biến Proxy cho Ability thì không cần phải sử dụng phương pháp hồi quy biến công cụ.
  - Nếu có dữ liệu bảng thì phần tố chất cá nhân cũng có thể bị loại bỏ bởi phương pháp DiD.

Cơ chế của phương pháp biến công cụ

- Chúng ta sử dụng biến công cụ Z tương quan với biến nội sinh *educ* nhưng không tương quan với phần dư v để lọc những thông tin cần giữ.
- Chúng ta sử dụng phương pháp hồi quy hai giai đoạn (Two-Stage Least Square-2SLS):
  - Bước 1: Hồi quy biến nội sinh educ theo biến công cụ, và thu được giá trị ước lượng $\widehat{educ}$.
  - Bước 2: Hồi quy Y theo $\widehat{educ}$ để tìm $\hat{\beta}_1$.

$$educ = \gamma_0 + \gamma_1 Z + \varepsilon$$

$$Y = \beta_0 + \beta_1 \widehat{educ} + v$$

- Do $\hat{Z}$ không tương quan với v nên $\hat{\beta}_1$ ước lượng được từ 2SLS nhất quán (nhưng luôn luôn chệch, đặc biệt với cỡ mẫu nhỏ!).

Ví dụ 1: Ước lượng tỷ suất thu nhập của đi học

Sử dụng bộ dữ liệu MROZ.dta, ước lượng mô hình sau:

$$Y = log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + \underbrace{\gamma Ability + u}_{v}$$

- Lý giải tại sao trình độ học vấn của cha/mẹ có thể sử dụng làm biến công cụ cho số năm đi học.
- Kiểm tra hồi quy bước 1.

So sánh kết quả ước lượng OLS so với IV

Regression Results

	OLS b/se	IV Estimates b/se
educ	0.1075***	0.0702*
	(0.0141)	(0.0343)
exper	0.0416**	0.0437**
	(0.0132)	(0.0133)
expersq	-0.0008*	-0.0009*
	(0.0004)	(0.0004)
Constant	-0.5220**	-0.0611
	(0.1986)	(0.4344)
Obs	428.0000	428.0000
R2	0.1568	0.1430
R2-adj	0.1509	0.1370
df(r)	424.0000	
SSR	188.3051	191.3867

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

Ví dụ 2: Sử dụng khoảng cách làm biến công cụ

Sử dụng bộ dữ liệu CARD.dta, ước lượng mô hình sau:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + \beta_4 black + \beta_5 smsa + \beta_6 south + \underbrace{\gamma Ability + u}_{v}$$

- trong đó các biến black, smsa, south là các biến giả đại diện cho người da đen, ở thành thị (Standard Metropolitan Statistical Area), và ở phía nam nước Mỹ.
- Biến công cụ được chọn là khu vực sinh sống có trường cao đẳng/đại học (chương trình 4 năm).

So sánh giữa OLS, OLS với Proxy cho biến Ability, và IV

Regression Results

	OLS b/se	OLS with P~y b/se	IV Estimates b/se
educ	0.0740***	0.0693***	0.1323**
	(0.0035)	(0.0049)	(0.0492)
exper	0.0836***	0.0935***	0.1075***
	(0.0066)	(0.0095)	(0.0213)
expersq	-0.0022***	-0.0027***	-0.0023***
	(0.0003)	(0.0005)	(0.0003)
black	-0.1896***	-0.1361***	-0.1308*
	(0.0176)	(0.0263)	(0.0528)
smsa	0.1614***	0.1534***	0.1313***
	(0.0156)	(0.0189)	(0.0301)
south	-0.1249***	-0.0791***	-0.1049***
	(0.0151)	(0.0180)	(0.0230)
IQ		0.0025***	
		(0.0007)	
Constant	4.7337***	4.4826***	3.7528***
	(0.0676)	(0.1036)	(0.8284)
Obs	3010.0000	2061.0000	3010.0000
R2	0.2905	0.2257	0.2252
R2-adj	0.2891	0.2231	0.2237
df(r)	3003.0000	2053.0000	
SSR	420.4760	278.4521	459.1785

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

Sử dụng phương pháp biến công cụ trong đánh giá tác động chính sách

- Chính sách luôn có mục tiêu cụ thể, ví dụ hướng vào đối tượng ưu tiên thay vì cho toàn bộ dân số (purposive placement).

- Tự lựa chọn mẫu (self selection): những hộ thực sự cần thiết tham gia chưa chắc đã là những hộ được tham gia chính sách, hoặc ngược lại, do những nguyên nhân không quan sát được.

- **Hiện tượng tham gia chính sách không ngẫu nhiên (nội sinh) cũng là vấn đề đặc biệt quan trọng bởi nếu không nhận diện được thì ước lượng không có hiệu lực nội tại và tham vấn chính sách có thể bị sai lệch.**

Hậu quả nếu việc tham gia chính sách là không ngẫu nhiên

$$plim\ \hat{\beta} = \beta + \frac{Cov(\varepsilon * T)}{Var(T)}$$

- Nếu biến chính sách không ngẫu nhiên $\Rightarrow cov(T, \varepsilon) \neq 0$, và ước lượng của β sẽ bị chệch và không nhất quán.

- Hướng chệch (lên hay xuống) phụ thuộc vào tương quan giữa phần dư với biến chính sách. Nếu chỉ hộ giàu được tham gia chính sách (ε_i lớn khi $T = 1$) thì ước lượng tác động chính sách sẽ bị chệch lên. Khi này kết luận chính sách có tác động tích cực bị phóng đại so với thực tế.

Một số tình huống nghiên cứu

- Giả sử chúng ta muốn đánh giá tác động của chính sách cho vay tín dụng ưu đãi đến phúc lợi (thu nhập của hộ). Có lý do để cho rằng việc tham gia chính sách là không ngẫu nhiên. Ví dụ gia đình nào có khả năng vay vốn là những hộ có quan hệ tốt với chính quyền, có phương án sử dụng vốn vay hiệu quả, có tài sản thế chấp... Nếu sử dụng hồi quy OLS thì khả năng ước lượng sẽ bị chệch lên do tương quan dương giữa biến chính sách và biến dư (thu nhập).

Vấn đề biến chính sách nội sinh trong các tình huống khác

- Đánh giá tác động của chính sách bảo hiểm y tế lên thu nhập của nông hộ.
- Chính sách hỗ trợ vốn cho doanh nghiệp trong giai đoạn khủng hoảng kinh tế.
- Chương trình cung cấp nước sạch đến cho người dân ảnh hưởng như thế nào đến thu nhập và phúc lợi của hộ sử dụng.
- Chương trình hỗ trợ đào tạo dạy nghề ảnh hưởng thế nào đến thu nhập của người lao động.

Lựa chọn biến công cụ như thế nào?

Biến công cụ phải thoả mãn 2 điều kiện:

- Tương quan với tình trạng tham gia chính sách.
- Không tương quan với phần dư của biến phụ thuộc (exclusion restriction).

Rất khó tìm được biến thoả mãn cả hai điều kiện trên. Các biến công cụ thường được sử dụng là các đặc tính địa lý như khoảng cách, hay các thay đổi có yếu tố bất ngờ như các hiện tượng thời tiết cực đoan, thiên tai, hay các chính sách vĩ mô của chính phủ.

Một số ví dụ về biến công cụ

- Kinh điển: Nghiên cứu về tỷ suất thu nhập của số năm đi học của Angrist và Krueger (1991). Sử dụng thời gian sinh theo quý để làm biến công cụ cho biến chính sách là số năm đi học.
- Nghiên cứu về tác động lâu dài của bom Mỹ đến tăng trưởng kinh tế ở VN (Miguel, JDS). Cường độ ném bom là biến nội sinh, và tăng ở những điểm gần vĩ tuyến 17. Do đó dùng khoảng cách từ các tỉnh đến vĩ tuyến 17 làm biến công cụ.
- Le (2014) sử dụng vĩ tuyến 17 làm biến công cụ để giải thích sự thay đổi của số năm đi học do cải cách giáo dục xóa bỏ lớp 9 và hợp nhất hệ thống giáo dục Bắc-Nam theo hệ 12 năm khi ước lượng tỷ suất thu nhập cho việc đi học.

Một số ví dụ về biến công cụ

- Le (2017) sử dụng tình trạng hộ khẩu để làm biến công cụ cho giá điện sinh hoạt khi ước lượng hàm cầu điện sinh hoạt.
- Đánh giá tác động của chương trình đào tạo để giúp người thất nghiệp. Việc tham gia chương trình là không ngẫu nhiên. Cần biến công cụ tương quan với việc tham gia, nhưng không trực tiếp tương quan với xác suất xin được việc. Dùng khoảng cách quan sát được giữa nhà với trung tâm đào tạo làm biến công cụ.
- Nghiên cứu về thu nhập và nội chiến (Miguel et al 2005, JPE). Thu nhập ảnh hưởng đến cạnh tranh tài nguyên và xung đột. Tuy nhiên thu nhập là biến nội sinh. Dùng thay đổi lượng mưa bất thường làm biến công cụ.

Khác biệt giữa IV với hồi quy rút gọn (reduced-form regression)

Tại sao không sử dụng biến công cụ Z thay cho biến nội sinh *educ* và ước lượng phương trình hồi quy một giai đoạn như sau:

$$Y = \beta_0 + \beta_1 * Z + u$$

mà phải dùng hồi quy 2SLS?

Các đặc tính thống kê của ước lượng sử dụng biến công cụ

- Giả sử hàm hồi quy chuẩn là:

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

Chúng ta sử dụng biến Z làm biến công cụ cho biến X, và giả định $Var(u|Z) = \sigma^2$.

- Phương sai tiệm cận (asymptotic variance) của tham số ước lượng β_1 có công thức:

$$Var(\hat{\beta}_1)_{IV} = \frac{\hat{\sigma}^2}{SST_x R^2_{x,z}}$$

Trong đó SST_x là tổng biến thiên của biến X, $R^2_{x,z}$ là hệ số thích hợp của hồi quy X lên Z.

Các đặc tính thống kê của ước lượng sử dụng biến công cụ

- Trong QMI, chúng ta đã biết phương sai của β_1 đối với ước lượng OLS là:

$$Var(\hat{\beta}_1)_{OLS} = \frac{\hat{\sigma}^2}{SST_x(1 - R_x^2)}$$

Trong đó R_x^2 là hệ số thích hợp của hồi quy biến X lên tất cả các biến giải thích còn lại trong mô hình.

- Đối với hồi quy có một biến giải thích (đơn giản hóa), $R_x^2 = 0$, khi đó ta có thể so sánh sai số của ước lượng OLS và IV trực tiếp:

$$Var(\hat{\beta}_1)_{IV} = \frac{\hat{\sigma}^2}{SST_x R_{x,z}^2} > Var(\hat{\beta}_1)_{OLS} = \frac{\hat{\sigma}^2}{SST_x}$$

Các đặc tính thống kê của ước lượng sử dụng biến công cụ

- Sử dụng biến công cụ yếu (weak instruments), tương quan yếu với biến nội sinh, dẫn đến phương sai của ước lượng sử dụng phương pháp IV bị thổi phồng $\Rightarrow$ Ước lượng kém chính xác và khoảng tin cậy tăng.

- Nếu Z trùng lặp với X thì ước lượng IV trùng với ước lượng OLS.

Các đặc tính thống kê của ước lượng sử dụng biến công cụ

- Tính nhất quán và thiên lệch của ước lượng IV và OLS khi có biến nội sinh:

$$plim\hat{\beta}_{1,IV} = \beta_1 + \frac{corr(z,u)}{corr(z,x)}.\frac{\sigma_u}{\sigma_x}$$

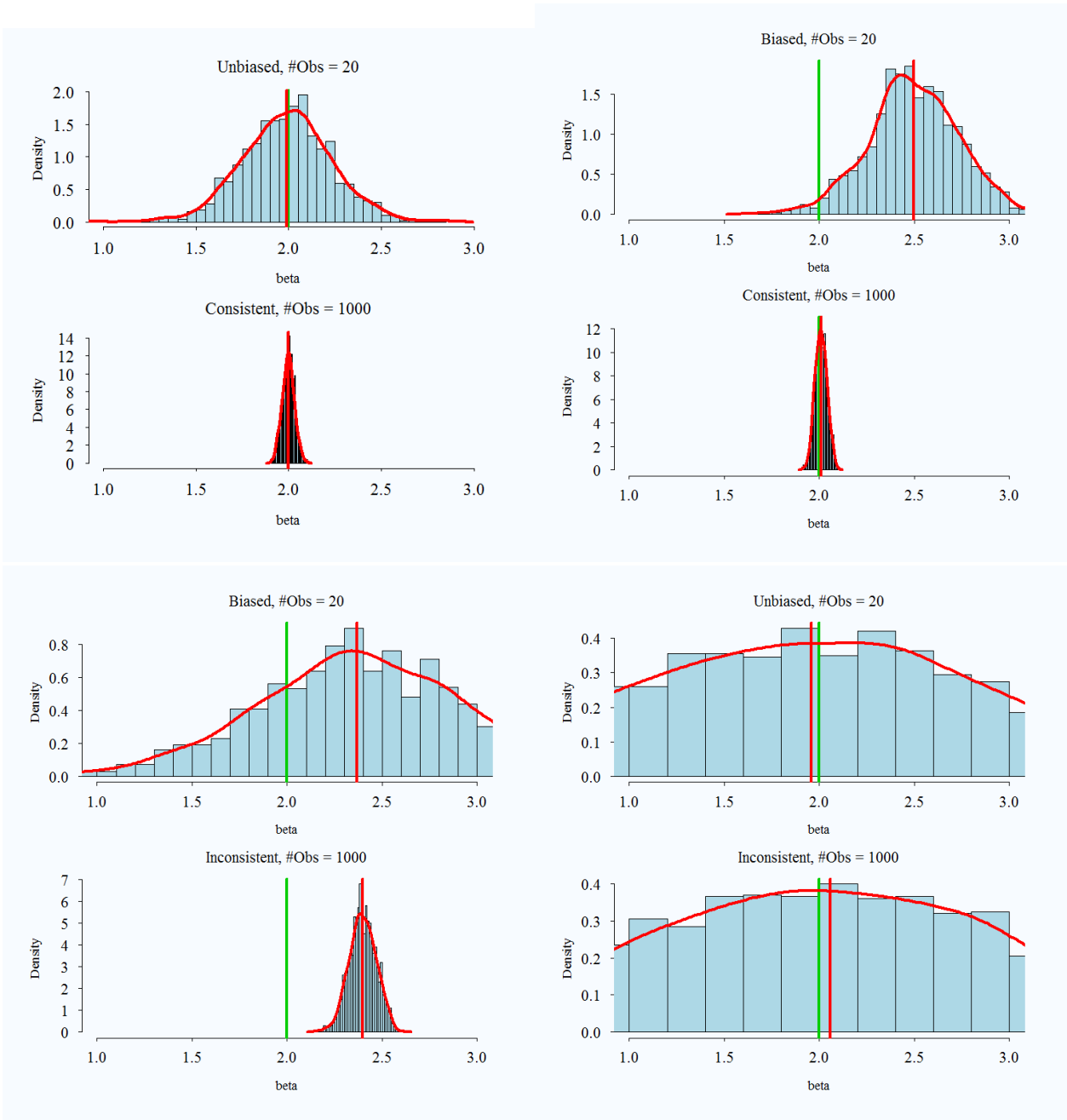
$$plim\hat{\beta}_{1,OLS} = \beta_1 + corr(x,u).\frac{\sigma_u}{\sigma_x}$$

- Ước lượng OLS luôn bị thiên lệch và không nhất quán khi $corr(x,u) \neq 0$.
- Ước lượng IV nhất quán khi tìm được biến công cụ tốt (Z tương quan với X và không tương quan với u.)
- Với cỡ mẫu nhỏ, nếu $corr(z,x)$ nhỏ thì ước lượng IV có thể rất không nhất quán (và hậu quả xấu hơn là sử dụng OLS).
- Ước lượng IV hầu như luôn thiên lệch.

Thiên lệch và nhất quán - Bias and Consistency

- Không thiên lệch: giá trị kỳ vọng của ước lượng bằng với giá trị thực, $E(\hat{\beta}_1) = \beta_1$
- 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, $plim\hat{\beta}_1 \rightarrow \beta_1$
- Nếu ước lượng bị thiên lệch nhưng nhất quán, tăng cỡ mẫu có thể làm giảm mức độ thiên lệch.

Bias and Consistency



Các kiểm định đối với phương pháp biến công cụ

- Kiểm định Wu-Hausman về sự hiện diện của biến nội sinh.
- Kiểm định biến công cụ yếu (weak instruments): Nếu 1st-stage F-stat > 10 với trường hợp 1 biến công cụ thì chấp nhận biến công cụ (Stock and Yogo, 2005).
- Điều kiện loại trừ ($Cov(Z, v) = 0$, exclusion restriction) không thể kiểm định được đối với trường hợp số biến công cụ bằng với số biến nội sinh, do đó cần giải thích dựa trên kiến thức và bối cảnh của mô hình.
- Kiểm định ràng buộc chặt (overidentification): Khi có nhiều biến công cụ hơn biến nội sinh thì có thể kiểm định điều kiện loại trừ bằng kiểm định ràng buộc chặt.

Nhận xét đối với phương pháp biến công cụ

- Là một trong những phương pháp mạnh nhất để ước lượng quan hệ nhân quả trong đánh giá tác động chính sách, đặc biệt đối với dữ liệu thử nghiệm tự nhiên. Nhưng đồng thời cũng là một trong những phương pháp khó hiểu nhất đối với cả các chuyên gia nghiên cứu kinh tế.
- Có thể sử dụng nhiều biến công cụ, nhiều biến nội sinh đồng thời.
- Rất khó tìm biến công cụ hoàn hảo.
- Nếu tìm được biến công cụ tốt thì ước lượng IV có hiệu lực nội tại. Nếu không thì ước lượng IV có thể còn tệ hơn ước lượng OLS.

Tham khảo

- Lecture notes, Monique de Haan, Department of Economics, University of Oslo. http://www.sv.uio.no/econ/english/people/aca/moniqued/
- https://eranraviv.com/bias-vs-consistency/