Hồi quy Hai Giai đoạn với Biến Công cụ (Two-staged Regression with Instrumental Variables)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 29 tháng 3 năm 2019

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.

Hiệu lực nội tại của ước lượng bằng OLS khi mô hình thiếu biến quan trọng

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_j X_j + \gamma Ability + u$$

- Khi mô hình bị thiếu biến quan trọng (Tố chất cá nhân *Ability* không quan sát được) thì ước lượng của β_1 bị chệch và không nhất quán.
- Trường hợp tổng quát: khi biến chính sách tương quan với phần dư (hiện tượng nội sinh - endogeneity).

Chúng ta có thể sử dụng mô hình hồi quy dữ liệu bảng với tác động cố định để loại trừ nhân tố *Ability*.

Hiệu lực nội tại khi xảy ra 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:

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

Hiệu lực nội tại khi có 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 của biến giải thích đến ước lượng OLS

- 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

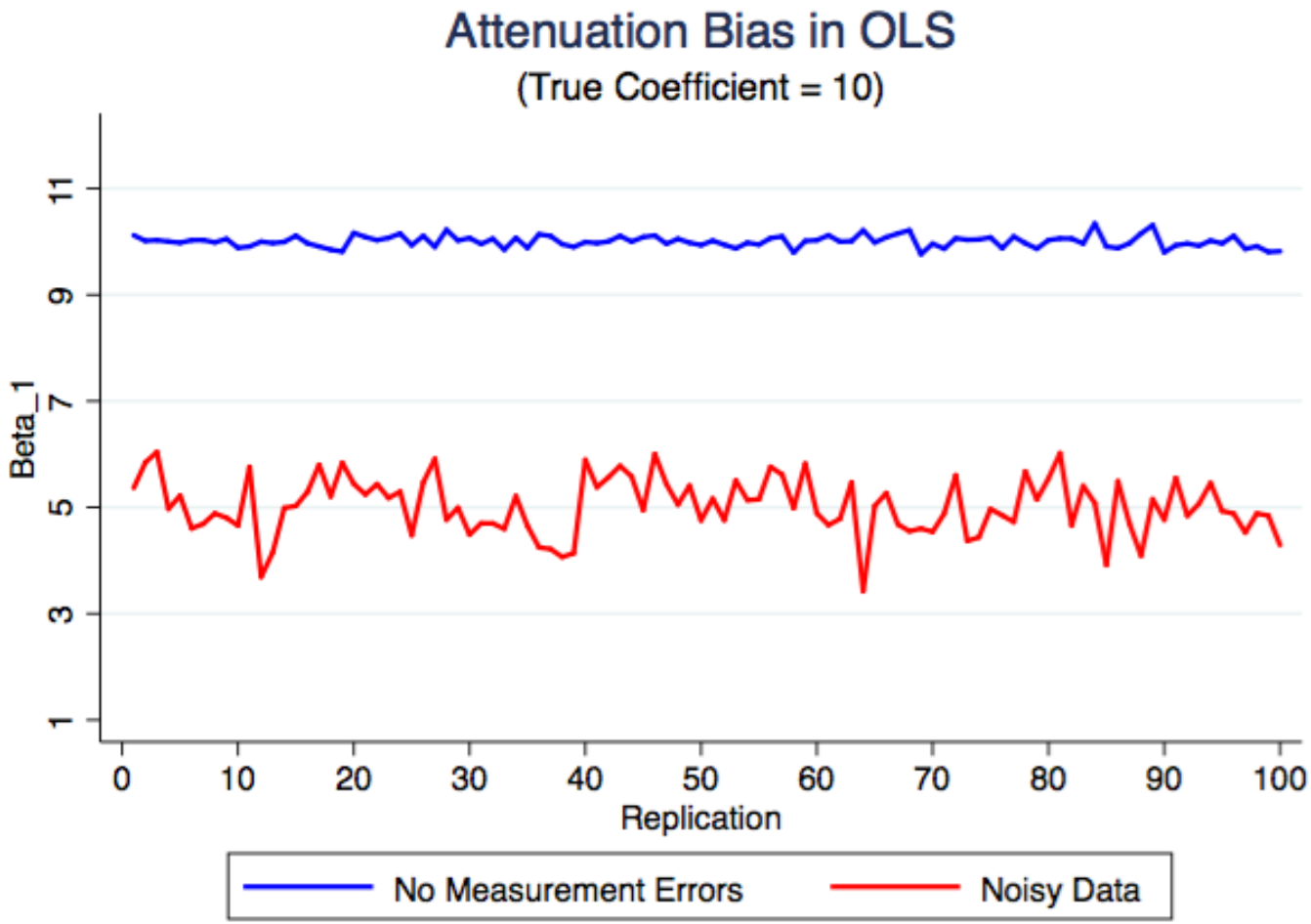
Nếu chúng ta ước lượng mô hình trên bằng OLS:

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

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.



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:

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

- Chúng ta biết giả định của CLRM bị vi phạm do mô hình thiếu biến quan trọng (tố chất cá nhân *Ability*), dẫn đến phần dư có tương quan với biến chính sách, $cov(educ, v) \neq 0$:

$$E[\hat{\beta}_1] = \beta_1 + \frac{cov(educ, v)}{var(educ)}$$

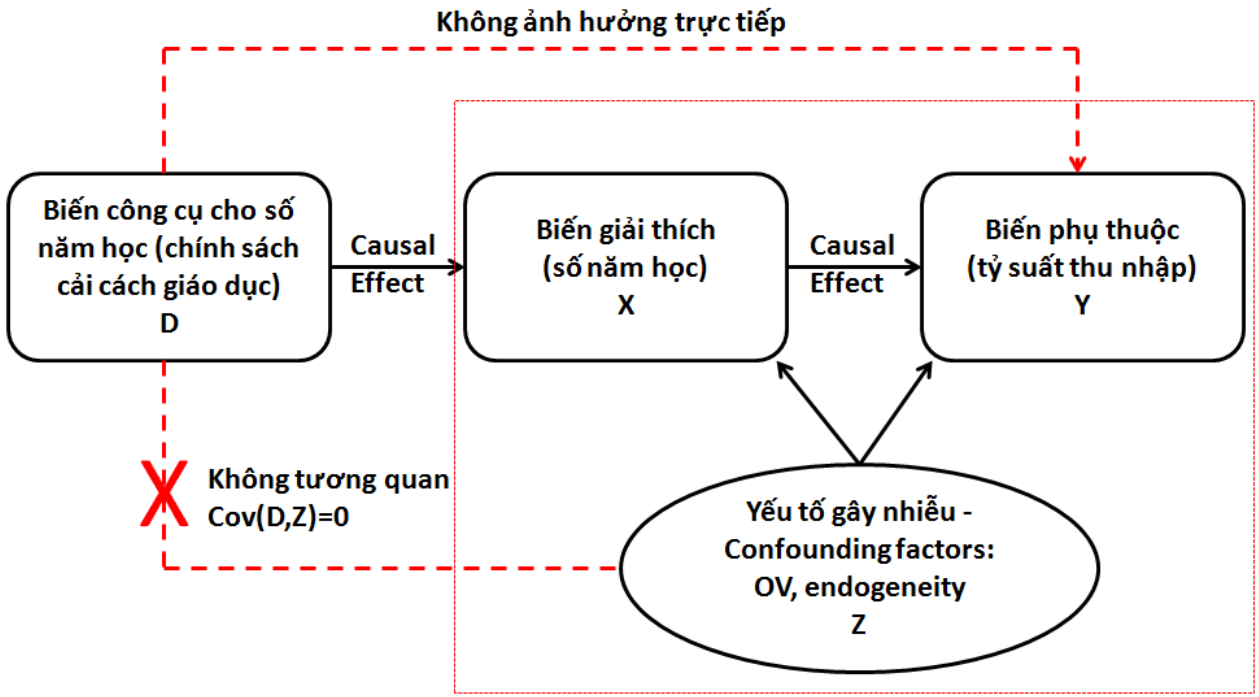
- Biến chính sách tương quan với phần dư được gọi là **hiện tượng nội sinh (endogeneity)**, và biến bị ảnh hưởng được gọi là biến nội sinh (endogenous variable).
- Ước lượng OLS của mô hình bị vấn đề biến nội sinh không có hiệu lực nội tại.

- **Vấn đề biến nội sinh là vấn đề nghiêm trọng nhất trong nghiên cứu định lượng!**
- Nếu có biến Proxy cho *Ability* như điểm số hay chỉ số IQ thì có thể sử lý được vấn đề thiếu biến quan trọng.
- Nếu có dữ liệu bảng thì phần tố chất cá nhân cũng có thể bị loại bỏ bởi phương pháp DiD.

Nếu không có biến proxy hay dữ liệu bảng, có thể sử dụng phương pháp biến công cụ để sử lý vấn đề biến nội sinh.

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

- **D** có tương quan với biến nội sinh *educ*, $cov(D, educ) \neq 0$.
- **D** không tương quan với phần dư của mô hình, $cov(D, v) = 0$ (nói cách khác, **D** không tác động trực tiếp lên biến phụ thuộc Y, nhưng **D** có thể tác động gián tiếp lên biến phụ thuộc thông qua tác động lên biến nội sinh).
- **D** được gọi là biến công cụ cho biến nội sinh số năm đi học.



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

Do giả định $cov(D, v) = 0 \Rightarrow$

$$\beta_1 = \frac{cov(D, Y)}{cov(D, educ)} = \frac{\sum_{i=1}^{n}(D_i - \bar{D})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(D_i - \bar{D})(educ_i - \overline{educ})}$$

Ước lượng β_1 thông qua D được gọi là ước lượng sử dụng phương pháp biến công cụ, khác với ước lượng bằng OLS.

Phương pháp hồi quy hai giai đoạn với biến công cụ (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 $\hat{\beta}_1$.

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

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

Ước lượng sử dụng biến công cụ được gọi là ước lượng 2SLS, IV, 2SLS/IV.

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

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

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \underbrace{\gamma Ability + u}_{v}$$

- 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 ước lượng OLS và 2SLS.

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:

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

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 (chường trình 4 năm).

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

Regression Results

	OLS b/se	OLS with Proxy 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

Khác biệt giữa 2SLS/IV với hồi quy rút gọn (reduced-form regression)

Tại sao không sử dụng trực tiếp biến công cụ **D** thay cho biến nội sinh *educ* và ước lượng phương trình hồi quy tỷ suất thu nhập như sau:

$$log(wage) = \beta_0 + \beta_1 * D + v$$

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:

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

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

- Phương sai xấp xỉ (asymptotic variance) của tham số ước lượng β_1 bằng 2SLS có công thức:

$$Var(\hat{\beta}_1)_{IV} = \frac{\hat{\sigma}^2}{SST_X R^2_{X,D}}$$

  - SST_X là tổng biến thiên của biến giải thích X.
  - $R^2_{X,D}$ là hệ số thích hợp của hồi quy X lên D.

- Trong QM-I, 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.

- Để đơn giản hóa, giả định hàm hồi quy có một biến giải thích, khi đó $R_X^2 = 0$. Ta có thể so sánh sai số của ước lượng OLS và IV trực tiếp:

$$Var(\hat{\beta}_1)_{IV} = \frac{\hat{\sigma}^2}{SST_X R_{X,D}^2} > Var(\hat{\beta}_1)_{OLS} = \frac{\hat{\sigma}^2}{SST_X}$$

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

- Phương sai của ước lượng bằng IV luôn lớn hơn OLS (giả sử khi sử dụng OLS là đúng) $\Rightarrow$ Khoảng tin cậy tăng và ước lượng kém chính xác.

- Khi biến công cụ tương quan yếu với biến nội sinh (weak instruments), $R^2_{X,D}$ nhỏ, thì phương sai của ước lượng sử dụng phương pháp IV bị thổi phồng $\Rightarrow$ Ước lượng càng kém chính xác.

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

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

$$plim\hat{\beta}_{1,IV} = \beta_1 + \frac{corr(D,v)}{corr(D,X)}.\frac{\sigma_v}{\sigma_X}$$

$$plim\hat{\beta}_{1,OLS} = \beta_1 + corr(X,v).\frac{\sigma_v}{\sigma_X}$$

- Ước lượng OLS bị thiên lệch và không nhất quán khi $corr(X,v) \neq 0$.
- Ước lượng IV nhất quán khi tìm được biến công cụ tốt, $corr(D,X) \neq 0$ và $corr(D,v) = 0$.
- Nếu $corr(D,X)$ nhỏ (biến công cụ yếu) thì ước lượng IV có thể rất không nhất quán (và hậu quả xấu hơn là sử dụng OLS).
- Ước lượng IV luôn bị thiên lệch.

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 (tình trạng hưởng lợi là 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

Giả sử chúng ta muốn đánh giá tác động của chính sách cho vay vốn đến thu nhập hộ gia đình bằng hàm hồi quy đơn giản hóa như sau:

$$Y = \beta_0 + \beta_1 T + v$$

trong đó T là tình trạng tham gia chính sách (có hoặc không).

$$plim\ \hat{\beta}_1 = \beta_1 + \frac{Cov(T, v)}{Var(T)}$$

- Nếu T tương quan với v thông qua nhân tố không quan sát được (ví dụ quan hệ tốt thì dễ được vay vốn), $cov(T, v) \neq 0 \Rightarrow$ ước lượng bằng OLS của β_1 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 có nhiều quan hệ được tham gia chính sách (v 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.

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

$$Y = \beta_0 + \beta_1 T + \beta_2 X + v$$

Biến công cụ cho biến chính sách T 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.

- Le (2017) sử dụng tình trạng hộ khẩu làm biến công cụ giải thích cho giá điện trong ước lượng hàm cầu điện tiêu thụ ở hộ gia đình.
- 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.

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 ($Cov(D, v) = 0$, 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.
- Kiểm định nhận diện mô hình quá lỏng (underidentification test): Kiểm định tương quan giữa biến công cụ với biến nội sinh.

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.