ÉCONOMÉTRIE
(UGA, S2)
CHAPITRE 3 :
ENDOGÉNÉITÉ ET VARIABLES INSTRUMENTALES(1)

Michal W. Urdanivia*

*UGA, Faculté d'Économie, GAEL,
e-mail : michal.wong-urdanivia@univ-grenoble-alpes.fr

12 avril 2022

1. Introduction à la notion de variable instrumentale

2. L'estimateur de VIs

1. Introduction à la notion de variable instrumentale

Endogénéité et (non-)identification dans un modèle linéaire simple

- On considère ici le modèle linéaire le plus simple :

$$y_i = \alpha_0 + b_0 x_i + u_i \text{ avec } \mathrm{E}[u_i] := 0, \tag{1}$$

  mais on considère que l'analyse du PGD indique que x_i est endogène dans ce modèle, i.e. que :

$$\mathrm{E}[u_i | x_i] \neq 0 \Rightarrow \mathrm{Cov}[x_i; u_i] \neq 0.$$

- Rappelons que lorsque x_i est exogène et que $\mathrm{Var}[x_i] \neq 0$, on a par l'exogénéité de x_i :

$$\mathrm{Cov}[x_i; u_i] = 0 \Leftrightarrow \mathrm{Cov}[x_i; y_i - \alpha_0 + b_0 x_i] = 0 \Rightarrow b_0 = \frac{\mathrm{Cov}[x_i; y_i]}{\mathrm{Var}[x_i]}.$$

- Autrement dit l'exogénéité de x_i permet d'identifier b_0(et aussi de α_0) comme une fonction de la distribution des variables observées y_i et x_i.
- Inversement dans la situation que nous considérons dans ce chapitre $\mathrm{Cov}[x_i; u_i] \neq 0$, rend impossible l'identification b_0(et aussi celle de α_0), et la construction d'un estimateur convergent.

Endogénéité et (non-)identification dans un modèle linéaire simple

- On peut représenter ce problème en utilisant un **graphe causal** :

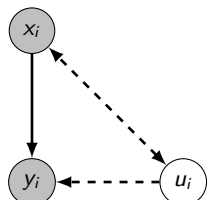

Figure 1 – Graphe causal du modèle : $y_i = \alpha_0 + b_0 x_i + u_i$, avec $\text{Cov}(x_i; u_i) \neq 0$. Les variables (x_i, y_i, u_i) sont les nœuds du graph et les nœuds foncés correspondent aux variables observées. Les arêtes représentent les relations entre les variables. Les relations observées sont en trait plein.

Identification avec une variable instrumentale

- L'intuition sous-jacente à la méthode des VIs consiste à répondre à la question de savoir si avec une variable, que nous notons z_i, il est possible d'obtenir une mesure de la relation causale entre x_i et y_i qui ne dépende pas de u_i.
- Autrement dit, z_i doit être exogène par rapport à u_i :

$$\mathrm{E}[u_i|z_i] = 0 \Rightarrow \mathrm{Cov}[z_i; u_i] = 0, \tag{2}$$

ce qui nous permets d'écrire :

$$\begin{aligned} \mathrm{Cov}[z_i; u_i] = 0 &\Leftrightarrow \mathrm{Cov}[z_i; y_i - \alpha_0 + b_0 x_i] = 0 \\ &\Leftrightarrow \mathrm{Cov}[z_i; y_i - \alpha_0 + b_0 \,\mathrm{Cov}[z_i; x_i] = 0 \end{aligned}$$

- Ceci indique que pour identifier b_0 on doit aussi supposer aussi que,

$$\mathrm{Cov}[z_i; x_i] \neq 0, \tag{3}$$

et b_0 est identifié par :

$$b_0 = \frac{\mathrm{Cov}[z_i; y_i]}{\mathrm{Cov}[z_i; x_i]}, \tag{4}$$

Identification avec une variable instrumentale

- On peut résumer les conditions (2)-(3) ainsi :

Définition 1 (Conditions de validité de VIs dans un modèle simple)

Dans le modèle $y_i = \alpha + b_0 x_i + u_i$ avec $\mathrm{E}[u_i] := 0$, une variable z_i est un instrument(de x_i) ssi :

(i) $\mathrm{Cov}[z_i; u_i] = 0$, i.e. z_i est exogène par rapport à u_i et :
(ii) $\mathrm{Cov}[z_i; x_i]$, i.e. z_i et x_i sont liées.

- Dans la représentation en termes de graphe causal cela donne :

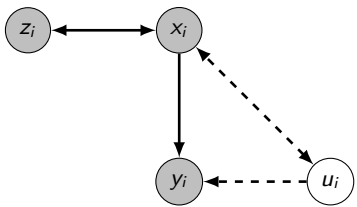

Figure 2 – Graphe causal du modèle : $y_i = \alpha_0 + b_0 x_i + u_i$, avec $\mathrm{Cov}(x_i; u_i) \neq 0$, $\mathrm{Cov}(z_i; u_i) = 0$. Les variables (z_i, x_i, y_i, u_i) sont les nœuds du graph et les nœuds foncés correspondent aux variables observées. Les arêtes représentent les relations entre les variables. Les relations observées sont en trait plein.

Estimateur des VIs dans le modèle simple

- L'identification de b_0 par (4) suggère l'estimateur :

$$\hat{b}_N^{VI} = \frac{N^{-1}\sum_{i=1}^{N}(z_i - \bar{z}_N)(y_i - \bar{y}_N)}{N^{-1}\sum_{i=1}^{N}(z_i - \bar{z}_N)(x_i - \bar{x}_N)} = \frac{N^{-1}\sum_{i=1}^{N}(z_- \bar{z}_N)y_i}{N^{-1}\sum_{i=1}^{N}(z_- \bar{z}_N)x_i}, \tag{5}$$

où $\bar{z}_N$, $\bar{x}_N$, et $\bar{y}_N$ sont le moyennes empiriques respectives de z_i, x_i, et y_i.

- De plus, $\hat{b}_N^{VI}$ est convergent. Nous avons en effet :

$$\operatorname*{plim}_{N\to+\infty} N^{-1}\sum_{i=1}^{N}(z_- \bar{z}_N)y_i \to \mathrm{Cov}[z_i; y_i],$$

$$\operatorname*{plim}_{N\to+\infty} N^{-1}\sum_{i=1}^{N}(z_- \bar{z}_N)x_i \to \mathrm{Cov}[z_i; x_i],$$

d'où :

$$\begin{aligned}\hat{b}_N^{VI} \underset{N\to+\infty}{\xrightarrow{P}} \frac{\mathrm{Cov}[z_i; y_i]}{\mathrm{Cov}[z_i; x_i]} &= \frac{\mathrm{Cov}[z_i; \alpha_0 + b_0 x_i + u_i]}{\mathrm{Cov}[z_i; x_i]}, \\ &= b_0 + \frac{\mathrm{Cov}[z_i; u_i]}{\mathrm{Cov}[z_i; x_i]}, \\ &= b_0.\end{aligned}$$

Estimateur des VIs dans le modèle simple

Remarque 1

- ⋆ On dit des variations de z_i qu'elles sont des variations exogènes : elles ne sont pas liées à u_i puisque $\text{Cov}[z_i; u_i] = 0$.
- ⋆ Ce sont les effets de ces variations exogènes sur x_i qui sont exploitées pour l'identification de b_0 grâce à $\text{Cov}[z_i; x_i] \neq 0$.
- ⋆ Noter qu'il n'est aucunement nécessaire que l'effet de z_i sur x_i soit causal.
- ⋆ L'effet de z_i sur y_i ne « transite » que via x_i. La variable instrumentale z_i n'est pas une variable explicative dans le modèle de y_i. On parle alors de relation d'exclusion (de la VI z_i vis-à-vis du modèle de y_i).
- ⋆ L'estimateur des VIs est parfois appelé estimateur des moindres carrés indirects. Cela provient de ce que b_0 dans (4) peut s'écrire :

$$b_0 = \frac{\text{Cov}[z_i; y_i] / \text{Var}[z_i]}{\text{Cov}[z_i; x_i] / \text{Var}[z_i]},$$

qui est le rapport entre le coefficient de z_i dans la projection de y_i sur z_i, et le coefficient de z_i dans la projection de x_i sur z_i.

2. L'estimateur de VIs

Variables endogènes, exogènes, instruments

- L'objectif est maintenant de généraliser l'approche présentée dans le cas simple précédent au modèle linéaire général :

$$y_i = \mathbf{x}_i' \mathbf{a}_0 + u_i, \text{ avec } \mathrm{E}[u_i] := 0. \tag{6}$$

- Plusieurs éléments du vecteur $\mathbf{x}_i$ peuvent être endogènes de sorte que dans l'estimateur des MCO de $\mathbf{a}_0$ plusieurs éléments sont potentiellement biaisés(c.f. cours précédent sur les VIs).
- Notons :

$$\mathbf{x}_i = \begin{bmatrix} \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \end{bmatrix} \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} \begin{matrix} \{\text{variables explicatives exogènes} & : \mathrm{E}[u_i | x_{k,i}^x] = 0 (k = 1, \ldots, M) \\ \{\text{variables explicatives endogènes} & : \mathrm{E}[u_i | x_{k,i}^x] \neq 0 (k = M+1, \ldots, K) \end{matrix}$$

Remarque 2

- ⋆ Il est clair que la variable constante 1 est «exogène» :$\mathrm{E}[1 \times u_i] = \mathrm{E}[ui|] = \mathrm{E}[u_i] = 0$.
- ⋆ Comme pour l'estimateur des MCO nous utiliserons la Méthode des Moments pour construire un estimateur convergent de $\mathbf{a}_0$, l'estimateur des VI du modèle (6).
- ⋆ On considère ici que chaque élément $\mathbf{x}_i^e$ a une variable instrumentale.

Variables endogènes, exogènes, instruments

Définition 2 (Variable instrumentale)

$z_{k,i}$ est une variable instrumentale de $x_{k,i}$ dans le modèle linéaire (6) si :

(i) $\text{Cov}[z_{k,i}; u_i] = 0$ i.e., $z_{k,i}$ est exogènes par rapport à u_i,

(ii) $z_{k,i}$ « suffisamment » liée à $x_{k,i}$.

Remarque 3

- On verra dans la suite (analyse des conditions de rang) que la condition (ii) doit en fait être définie comme :

$$\text{Cov}[z_{k,i}; e_{k,i}] \neq 0 \text{ pour } k > 1,$$

où $e_{k,i}$ est la partie spécifique de $x_{k,i}$ dans x_i , i.e. le résidu de la projection linéaire de $x_{k,i}$ sur les autres explicatives $\mathbf{x}_{-k,i}$.

$$e_{k,i} = x_{k,i} - \mathcal{EL}[x_{k,i}|\mathbf{x}_{-k,i}].$$

- Dans la définition d'un VI précédente, on voit que lorsqu'une variable explicative $x_{k,i}$ est exogène alors c'est aussi une variable instrumentale d'elle même. En ce sens que non seulement elle vérifie (i) mais elle vérifie forcément (ii) (car ayant une corrélation de 1 avec elle même)

Variables endogènes, exogènes, instruments

- On construit le vecteur des variables instrumentales $\mathbf{z}_i$ avec :

$$\tilde{\mathbf{z}}_i^e = \begin{bmatrix} \tilde{z}_{M+1,i} \\ \tilde{z}_{M+2,i} \\ \vdots \\ \tilde{z}_{K,i} \end{bmatrix} \text{ et } \mathbf{z}_i = \begin{bmatrix} \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \end{bmatrix} \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} \begin{matrix} \{\text{variables exogènes de } \mathbf{x}_i & : \mathrm{E}[u_i | x_{k,i}^x] = 0 (k = 1, \ldots, M) \\ \{\text{variables instrumentales} & : \mathrm{E}[u_i | z_{k,i}] = 0 (k = M+1, \ldots, K) \end{matrix}$$

- Ce vecteur contient en fait toutes les variables exogènes du modèle. Ce sont ces variables qui assurent l'identification des paramètres du modèle.
- $\mathbf{z}_i$ est parfois nommé ensemble d'information du modèle.

Variables endogènes, exogènes, instruments

Définition 3 (Modèle linéaire à variables instrumentales)

Le modèle défini par :

$$y_i = \mathbf{x}_i'\mathbf{a}_0 + u_i, \text{ avec } \mathrm{E}[u_i|\mathbf{z}_i] = \mathrm{E}[u_i] := 0,$$

est un modèle linéaire à variables instrumentales. La condition d'identification de $\mathbf{a}_0$ dans ce modèle est donnée par :

$$\mathrm{Rang}\left(\mathrm{E}[\mathbf{z}\mathbf{x}']\right) = K = \dim(\mathbf{x}).$$

Remarque 4

La condition d'exogénéité de $\mathbf{z}_i$ est définie par $\mathrm{E}[u_i|\mathbf{z}_i] = 0$, et non par $\mathrm{Cov}[\mathbf{z}_i; u_i] = \mathbf{0}$. Ce n'est pas nécessaire pour un modèle linéaire où $\mathrm{Cov}[\mathbf{z}_i; u_i] = \mathbf{0}$ suffit mais c'est standard et cela simplifie la présentation des hypothèses d'homoscédasticité.

Variables endogènes, exogènes, instruments

- Comme dans le cas où on a construit l'estimateur des MCO de a on part de la condition d'exogénéité des $\mathbf{z}_i$ (et non des $\mathbf{x}_i$ comme dans le cas des MCO), i.e. la condition d'orthogonalité donnée par :

$$\mathrm{E}[u_i \mathbf{z}_i] = 0 \Rightarrow \mathrm{E}[\mathbf{z}_i u_i] = 0 \Leftrightarrow \mathrm{E}[z_i(y_i - \mathbf{x}_i' \mathbf{a}_0)] = \mathbf{0}.$$

- On a ici la condition de moment estimante pour $\mathbf{a}_0$ est $\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i \mathbf{a}_0)] = \mathbf{0}$. Et on a alors :

$$\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i' \mathbf{a}_0)] = \mathbf{0} \Leftrightarrow \mathbf{a} = \mathbf{a}_0.$$

- On suppose ici que $\mathbf{a}_0$ est l'unique solution en $\mathbf{a}$ de $\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}'\mathbf{a})] = 0$.
- Le principe d'analogie définit l'estimateur de la MM de $\mathbf{a}_0$ par :

$$N^{-1} \sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i' \mathbf{a}) = \mathbf{0}_{K \times 1} \Leftrightarrow \mathbf{a} = \hat{\mathbf{a}}_N^{MM}.$$

- L'équation dont $\hat{\mathbf{a}}_N^{MM}$ est définie comme la solution en $\mathbf{a}$ est en fait un système de K équations linéaires à K inconnues (les éléments de $\hat{\mathbf{a}}_N^{MM}$). Il a solution sous forme explicite. On a :

$$N^{-1} \sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{MM}) = \mathbf{0}_{K \times 1}.$$

Variables endogènes, exogènes, instruments

- Il est aisé de définir la forme de $\hat{\mathbf{a}}_N^{MM}$,

$$N^{-1}\sum_{i=1}^{N}\mathbf{z}_i y_i - \left[N^{-1}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right]\hat{\mathbf{a}}_N^{MM} = \mathbf{0}_{K\times 1},$$

qui donne finalement :

$$\hat{\mathbf{a}}_N^{MM} = \left[N^{-1}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right]^{-1} N^{-1}\sum_{i=1}^{N}\mathbf{z}_i y_i,$$

qui définit ce qu'on appelle l'estimateur des VI.

Références