

**ÉCONOMÉTRIE  
(UGA, S2)  
CHAPITRE 3:  
ENDOGENÉITÉ ET VARIABLES INSTRUMENTALES**

Michal W. Urdanivia \*

\*Université de Grenoble Alpes, Faculté d'Économie, GAEL,  
e-mail: [michal.wong-urdanivia@univ-grenoble-alpes.fr](mailto:michal.wong-urdanivia@univ-grenoble-alpes.fr)

29 janvier 2024

# Contenu

1. Modèles à variables explicatives endogènes
2. La notion de variable instrumentale et l'estimateur des VI
3. L'estimateur des 2MC
4. Choix des VI et identification
5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée
6. Les variables de contrôle de l'hétérogénéité

# Outline

1. Modèles à variables explicatives endogènes
2. La notion de variable instrumentale et l'estimateur des VI
3. L'estimateur des 2MC
4. Choix des VI et identification
5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée
6. Les variables de contrôle de l'hétérogénéité

## **1. Modèles à variables explicatives endogènes**

# **1. Modèles à variables explicatives endogènes**

## **Les trois grands types d'endogénéité à partir d'exemples**

Variables explicatives pertinentes omises

Simultanéité

Erreurs de mesure sur les variables explicatives

**Les MCO ne sont pas convergents : biais d'endogénéité**

## Source d'endogénéité 1

### « Variables explicatives pertinentes omises »

#### L'exemple de la taille des classes (Angrist et Lavy ; Piketty)

**Objectif.** On veut estimer l'effet de la taille des classes de CE1 ( $tailclas_i$ ) sur les résultats des élèves de CE1 aux tests de math. ou de français ( $score_i$ ).

L'effet causal d'intérêt,  $taille_i \rightarrow score_i$ , est celui qui porte l'intérêt (ou non) de réduire la taille des classes.

**Données.** Un grand échantillon ( $i = 1, \dots, N$ ) d'élèves de CE1, avec leurs caractéristiques ( $c_i$ )

## Examen théorique de la question posée

Une manière simple de poser le problème consiste à écrire le modèle :

$$score_i = \alpha_0 + b_{1,0}tailclas_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0.$$

L'effet causal l'intérêt est  $b_{1,0}$  (même si  $\mathbf{b}_{-1,0}$  peut également être intéressant).

Si on augmente de 1 la taille de la classe, on « augmente » de  $b_{1,0}$  le score.

On s'attend à ce que  $b_{1,0} < 0$

Angrist et Lavy (Piketty) ont montré qu'on a de bonnes raisons de penser que :

$$Cov[tailclas_i; u_i] \neq 0$$

*Analyser ce problème passe par l'analyse PG des données.*

Le vecteur  $\mathbf{c}_i$  *contient* :

- âge, sexe, ... de l'élève  $i$
- éléments de description de sa famille (situation professionnelle, composition, ...)
- éléments de description de l'école (localisation, taille, ZEP, ...)

Ces données sont riches mais *ne contiennent pas* :

- l'origine des parents (maîtrise langue)
- éléments de description de l'attitude de  $i$  (turbulent ou calme, ...)
- éléments de description de ses aptitudes « scolaires », celles mesurées par le QI par exemple.

Tout ça joue sur  $score_i$  mais seul  $\mathbf{c}_i$  est observé.

Ce qui n'est pas mesuré (voire est difficilement mesurable) dans  $\mathbf{c}_i$  est regroupé dans  $\tilde{\mathbf{q}}_i$ .



Les effets de  $\tilde{\mathbf{q}}_i$  se retrouvent en partie dans  $\alpha_0$  et en partie dans le terme d'erreur  $u_i$ . Par exemple si le « vrai » modèle avec  $\tilde{\mathbf{q}}_i$  est donné par :

$$score_i = \delta_0 + b_{1,0}tailclas_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + \lambda'_0\tilde{\mathbf{q}}_i + v_i \quad \text{avec} \quad E[v_i] \equiv 0$$

Le lien avec le modèle utilisé :

$$score_i = \alpha_0 + b_{1,0}tailclas_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

se fait par :

$$\alpha_0 = \delta_0 + \lambda'_0 E[\tilde{\mathbf{q}}_i] \quad \text{et} \quad u_i = v_i + \lambda'_0 (\tilde{\mathbf{q}}_i - E[\tilde{\mathbf{q}}_i]).$$

- Les *effets moyens* de  $\tilde{\mathbf{q}}_i$  vont dans la constante du modèle,  $\alpha_0$ .
- Les *effets « hors-moyenne »* de  $\tilde{\mathbf{q}}_i$  vont dans les termes d'erreur,  $u_i$ .

**En conclusion** : le terme d'erreur  $u_i$  contient « beaucoup de choses ».

$$score_i = \alpha_0 + b_{1,0}tailclas_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

**Question de base :**  $tailclas_i$  et  $\mathbf{c}_i$  sont-elles exogènes ?

**Concrètement :** (i)  $score_i$  est-il un déterminant de  $tailclas_i$  et  $\mathbf{c}_i$  ?  
(ii)  $tailclas_i$  et  $\mathbf{c}_i$  sont-elles liées au contenu de  $u_i$  ?  
(iii) on néglige les erreurs de mesure sur  $tailclas_i$  et  $\mathbf{c}_i$ .

Pour  $\mathbf{c}_i$ , pas de problème en principe :

- Var. parfaitement « exogènes » par rapport à  $score_i$  : sexe, âge, ...
- Var. issues de décisions qui ont peu à voir avec  $score_i$  :  
composition du ménage, situation professionnelle des parents, ...
- Les parents ont peu de marge pour choisir l'école.

De fait on va chercher à estimer  $b_{1,0}$  à partir d'un *modèle conditionnel* en  $\mathbf{c}_i$ ,  
*i.e. en « contrôlant » les effets* de  $\mathbf{c}_i$  sur  $score_i$ .

Pour  $tailclas_i$ , l'analyse est plus « subtile »

- La taille de classe maximum est de 25 élèves
  - Si dans une école il y a 49 élèves de CE1, il y aura 2 classes
  - Si dans une école il y a 51 élèves de CE1, il y aura 3 classes
  - ...
- Les *élèves ne sont pas répartis aléatoirement* dans les classes d'une école
- Les *classes ne sont pas toutes de tailles égales* dans une école
  - Les élèves en difficulté dans les petites classes
  - Les élèves sans difficulté dans les grandes classes

Qu'est-ce que cela implique pour le modèle :

$$score_i = \alpha_0 + b_{1,0}tailclas_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0 ?$$

Dans :

$$score_i = \alpha_0 + b_{1,0}tailclas_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

on a vraisemblablement :

$$Cov[tailclas_i; u_i] > 0.$$

En effet, pour un élève  $j$  plutôt en difficulté, on a :

$u_j$  plutôt petit **et**  $tailclas_j$  plutôt petit.

***Problème : omission*** (forcée) *de variables explicatives pertinentes,  $\tilde{\mathbf{q}}_i$  ici*

**Implications :**  $tailclas_j$  est endogène dans le modèle

- Les MCO sont biaisés
- Les MCO sur-estiment  $b_{1,0}$  , empiriquement  $\hat{b}_{1,N}^{MCO} \simeq 0$

$\Rightarrow$  Il faudra estimer  $b_{1,0}$  autrement.

## Source d'endogénéité 2

### « Simultanéité »

#### L'exemple de l'effet de la maternité sur le salaire (Angrist et Evans)

**Objectif.** On veut estimer l'impact de la maternité ( $nbenf_i$ ) sur le salaire des mères ( $salaire_i$ ).

L'effet causal d'intérêt est  $nbenf_i \rightarrow salaire_i$ , est celui qui concerne l'effet de la maternité sur l'activité professionnelle des femmes.

**Données.** Un grand échantillon ( $i = 1, \dots, N$ ) de mères, avec des **variables de contrôle** ( $c_i$ )

## Examen théorique de la question posée

Une manière simple de poser le problème consiste à écrire le modèle :

$$\text{ salaire}_i = \alpha_0 + b_{1,0} \text{ nbenf}_i + \mathbf{b}'_{-1,0} \mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0.$$

L'effet causal l'intérêt est  $b_{1,0}$  (même si  $\mathbf{b}_{-1,0}$  peut également être intéressant).

### *Deux observations :*

- Les économistes expliquent le salaire par (entre autres) le nb d'enfants
- Les démographes expliquent le nb d'enfants par (entre autres) le salaire, l'inverse des économistes

*En fait, les uns et les autres n'ont ni complètement tort, ni complètement raison dans leurs raisonnements respectifs*

*Mais les deux auraient tort d'utiliser les MCO*

De fait, *les choix de maternité et les choix professionnels sont décidés plus ou moins conjointement*

Les deux propositions suivantes sont « vraies » :

- $salaire_i$  « cause »  $nbenf_i$  selon un double effet « ressources » et « temps disponible » pour les enfants
- $nbenf_i$  « cause »  $salaires_i$  selon double effet « ressources » et « temps disponible » pour la carrière professionnelle

***Les variables  $salaires_i$  et  $nbenf_i$  sont « simultanées » (i.e. des choix conjoints)***

Implications pour le modèle :

$$salaires_i = \alpha_0 + b_{1,0}nbenf_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0.$$

- $nbenf_i$  est **endogène** :  $nbenf_i$  est fonction de  $salaires_i$  et donc de  $u_i$
- Empiriquement  $\hat{b}_{1,N}^{MCO}$  est négatif proche de 0

### Source d'endogénéité 3

#### « Erreurs de mesure sur les variables explicatives »

Ce problème se distingue des deux autres dans le sens où il est essentiellement « technique », même s'il est également *lié au PG des données*.

**Objectif.** On veut estimer l'effet de  $\tilde{x}_i$  sur  $y_i$ .

**Données.** On ne dispose que d'une mesure bruitée de  $\tilde{x}_i$ ,  $\tilde{x}_i^e$ . Sinon, on dispose d'un grand échantillon ( $i = 1, \dots, N$ ) d'observations, *i.e.* de réalisations de  $(y_i, \tilde{x}_i^e)$ .

**Rmq.** Ce problème est, sûrement à tort, *peu traité en économétrie appliquée*, sauf pour ce qui concerne le revenu.

**Rmq.** Les erreurs de mesure *sur la variable à expliquer* « s'ajoutent » simplement aux termes d'erreur du modèle. Leur effet est de *détériorer la précision des estimateurs* des paramètres du modèle.



**Le modèle « d'intérêt »** est ici :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i / \tilde{x}_i] = E[u_i] \equiv 0,$$

*i.e.*, c'est un modèle de régression simple.

Ce modèle ne peut être utilisé car  $\tilde{x}_i$  n'est pas observée. Seule :

$$\tilde{x}_i^e = \tilde{x}_i + e_i$$

est observée,  $e_i$  étant une **erreur de mesure** de  $\tilde{x}_i$ . On suppose que :

$$E[e_i / \tilde{x}_i] = E[u_i / e_i] = E[e_i] = 0.$$

L'erreur de mesure  $e_i$  est donc *a priori* tout-à-fait anodine : elle est **sans biais** et n'est **liée ni à  $\tilde{x}_i$ , ni à  $u_i$** . Par la suite on utilisera le fait que :

$$V[\tilde{x}_i^e] = V[\tilde{x}_i] + V[e_i].$$

Par substitution de  $\tilde{x}_i$  par  $\tilde{x}_i^e - e_i$  dans le modèle d'intérêt on obtient le « **modèle observable** » :

$$y_i = \alpha_0 + b_0 \tilde{x}_i^e + v_i \text{ avec } v_i = u_i - b_0 e_i \text{ et } E[v_i] = 0.$$

On montre simplement que  $\tilde{x}_i^e$  est corrélée au terme d'erreur  $v_i$  :

$$\text{Cov}[\tilde{x}_i^e; v_i] = \text{Cov}[\tilde{x}_i + e_i; u_i - b_0 e_i] = -b_0 V[e_i].$$

***La variable explicative mesurée avec erreur est endogène dans le modèle observable, par « construction ».***

Les propriétés de l'estimateur des MCO indiquent que :

$$p \lim_{N \rightarrow +\infty} \hat{b}_N^{MCO} = V[\tilde{x}_i^e]^{-1} \text{Cov}[\tilde{x}_i^e; y_i]$$

par application de la régression linéaire dans le modèle observable.

Avec :

$$V[\tilde{x}_i^e] = V[\tilde{x}_i] + V[e_i]$$

et :

$$Cov[\tilde{x}_i^e; y_i] = Cov[\tilde{x}_i^e; b_0 \tilde{x}_i^e + v_i] = b_0 V[\tilde{x}_i^e] + Cov[\tilde{x}_i^e; v_i].$$

on obtient que :

$$p \lim_{N \rightarrow +\infty} \hat{b}_N^{MCO} = b_0 \times \frac{V[\tilde{x}_i]}{V[\tilde{x}_i] + V[e_i]}$$

et finalement :

$$\left| p \lim_{N \rightarrow +\infty} \hat{b}_N^{MCO} \right| < |b_0|.$$

On parle alors de ***biais d'atténuation***, car l'estimateur des MCO de  $b_0$  dans le modèle observable *sous-estime systématiquement  $b_0$  en valeur absolue*.

## Endogénéité des variables explicatives et Biais d'endogénéité de l'estimateur des MCO

Le biais d'atténuation calculé dans le cas du modèle avec erreurs de mesure sur les variables explicatives est un cas particulier de *biais d'endogénéité*.

Dans le modèle linéaire de forme générale :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0,$$

ce biais est celui affectant  $\hat{\mathbf{a}}_N^{MCO}$  pour l'estimation de  $\mathbf{a}_0$  lorsque  $E[u_i/\mathbf{x}_i] \neq 0$ .

Il peut être calculé aisément à partir de l'équation (voir chapitre 2) :

$$\hat{\mathbf{a}}_N^{MCO} = \mathbf{a}_0 + \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{x}_i u_i.$$

L'application de la LGN et des propriétés des suites convergeant en probabilité permet de montrer que :

$$p \lim_{N \rightarrow +\infty} \hat{\mathbf{a}}_N^{MCO} = \mathbf{a}_0 + E[\mathbf{x}_i \mathbf{x}_i']^{-1} E[\mathbf{x}_i u_i].$$

*Le terme  $E[\mathbf{x}_i \mathbf{x}_i']^{-1} E[\mathbf{x}_i u_i]$  est le biais (asymptotique) d'endogénéité de  $\hat{\mathbf{a}}_N^{MCO}$  pour  $\mathbf{a}_0$  si  $E[\mathbf{x}_i u_i] \neq \mathbf{0}$*

Il est important de remarquer que *même si une seule variable explicative est endogène*, disons  $x_{Ki}$  :

$$E[x_{K,i} u_i] \neq 0 \text{ et } E[x_{k,i} u_i] = 0 \text{ pour } k = 1, \dots, K-1,$$

alors  $\hat{a}_{K,N}^{MCO}$  n'est pas le seul élément de  $\hat{\mathbf{a}}_N^{MCO}$  potentiellement biaisé. **Tous les éléments de  $\hat{\mathbf{a}}_N^{MCO}$  sont potentiellement biaisés.** Et ils le sont généralement en pratique, le problème d'endogénéité de  $x_{Ki}$  « **contaminant** » les  $\hat{a}_{k,N}^{MCO}$  pour  $k = 1, \dots, K-1$ .

Le biais d'endogénéité est donc potentiellement un problème très sérieux.

Dans le modèle :

$$y_i = \mathbf{a}'_{-K,0} \mathbf{x}_{-K,i} + a_{K,0} x_{K,i} + u_i \quad \text{avec} \quad E[u_i / \mathbf{x}_{-K,i}] = E[u_i] \equiv 0,$$

le biais de l'estimateur des MCO,  $\hat{\mathbf{a}}_N^{MCO} \equiv (\hat{\mathbf{a}}_{-K,N}^{MCO}, \hat{a}_{K,N}^{MCO})$ , de  $\mathbf{a}_0 \equiv (\mathbf{a}_{-K,0}, a_{K,0})$  dû à l'endogénéité de  $x_{K,i}$  est donné par l'équation :

$$p \lim_{N \rightarrow \infty} \hat{\mathbf{a}}_N^{MCO} = \mathbf{a}_0 + \underbrace{\text{Cov}[x_{K,i}; u_i] \times V[e_{K,i}] \times \begin{bmatrix} -\gamma_K \\ 1 \end{bmatrix}}_{\text{Biais as. de } \hat{\mathbf{a}}_N^{MCO}}$$

où :

$$e_{K,i} \equiv x_{K,i} - EL[x_{K,i} / \mathbf{x}_{-K,i}] \quad \text{et} \quad \gamma'_K \mathbf{x}_{-K,i} \equiv EL[x_{K,i} / \mathbf{x}_{-K,i}].$$

Seront donc biaisés (i) l'élément de  $\hat{\mathbf{a}}_N^{MCO}$  *correspondant à*  $x_{K,i}$  lui-même, (ii) les éléments de  $\hat{\mathbf{a}}_N^{MCO}$  *correspondant aux éléments de*  $\mathbf{x}_{-K,i}$  *liés à*  $x_{K,i}$  et (iii) la constante du modèle.

# Outline

1. Modèles à variables explicatives endogènes
2. La notion de variable instrumentale et l'estimateur des VI
3. L'estimateur des 2MC
4. Choix des VI et identification
5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée
6. Les variables de contrôle de l'hétérogénéité

## **2. La notion de variable instrumentale et l'estimateur des VI**



## 2. La notion de variable instrumentale et l'estimateur des VI

Cette section introduit une notion fondamentale pour la suite de cette partie : la notion de *variable instrumentale* (ou instrument, même si ce terme sera utilisé dans un sens précis par la suite).

Les *variables instrumentales* sont les « outils » privilégiés pour résoudre un problème d'identification lié à un problème d'endogénéité.

Dans cette section :

- La notion de VI est présentée à partir d'un exemple simple.
- Des exemples de VI sont donnés, dans les exemples de la section 1.
- L'estimateur dit des VI est présenté avec ses propriétés dans le cas général.

## 2.1. Introduction à la notion de variable instrumentale

On considère ici le modèle linéaire le plus simple :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

mais on considère que l'analyse du PGD indique que  $\tilde{x}_i$  est endogène dans ce modèle, *i.e.* que :

$$E[u_i / \tilde{x}_i] \neq 0 \quad \Rightarrow \quad \text{Cov}[\tilde{x}_i; u_i] \neq 0.$$

Le problème de l'identification de  $b_0$  (et par conséquent de  $\alpha_0$ ) provient de ce que la seule mesure potentielle de l'effet  $b_0$  qu'on sache estimer, *i.e.* la covariance de  $\tilde{x}_i$  et  $y_i$ , est rendue inutilisable par l'endogénéité de  $\tilde{x}_i$  :

$$\text{Cov}[\tilde{x}_i; y_i] = b_0 V[\tilde{x}_i] + \text{Cov}[\tilde{x}_i; u_i].$$

Il est impossible d'estimer  $b_0$ , i.e. d'identifier  $b_0$ , à partir de l'équation :

$$\text{Cov}[\tilde{x}_i; y_i] = b_0 V[\tilde{x}_i] + \text{Cov}[\tilde{x}_i; u_i]$$

puisque  $\text{Cov}[\tilde{x}_i; u_i]$  ne peut être estimée, les termes d'erreur  $u_i$  étant inconnus (ou alors il nous faudrait un estimateur convergent de  $b_0$ ).

Or : ***Problème d'identification = déficit d'information***

Ajouter de l'information au modèle peut prendre plusieurs formes :

**Approximation** : si  $\text{Cov}[\tilde{x}_i; u_i] \simeq 0$  alors  $b_0 \simeq V[\tilde{x}_i]^{-1} \text{Cov}[\tilde{x}_i; y_i]$

**Correction** : si  $E[u_i/\tilde{x}_i] = \delta$  alors  $y_i = (\alpha_0 + \delta) + b_0 \tilde{x}_i + v_i$  avec  $E[u_i/\tilde{x}_i] = 0$ ,  
et le modèle « corrigé » est un modèle de régression.

Ces approches sont utilisées en pratique mais sont peu satisfaisantes :

**Approximation** : comment savoir que  $Cov[\tilde{x}_i; u_i]$  est assez proche de 0 ?

**Correction** : comment savoir que  $E[u_i/\tilde{x}_i] = \delta$  plutôt que  $E[u_i/\tilde{x}_i] = \delta + \lambda\tilde{x}_i$  ?

***Le meilleur apport d'information*** est celui procuré par des ***variables qui permettent de contrôler ou de contourner de le problème posé par***  
 $Cov[\tilde{x}_i; u_i] \neq 0$ .

Il existe deux types de ces variables : les ***variables de contrôle*** et les ***variables instrumentales (VI)*** :

- Les ***variables instrumentales*** permettent de gérer les ***3 types d'endogénéité***
- Les ***variables de contrôle*** ne permettent de gérer que l'endogénéité dû à l'***omission de variables explicatives pertinentes*** (voir section 6)

***L'intuition liée à l'utilisation des Variables Instrumentales*** est la suivante :

- La covariance  $Cov[\tilde{x}_i; y_i]$  n'a pas d'intérêt lorsque  $\tilde{x}_i$  ***est endogène*** car si elle fournit une équation :

$$Cov[\tilde{x}_i; y_i] = b_0 V[\tilde{x}_i] + Cov[\tilde{x}_i; u_i]$$

qui contient  $b_0$ , cette équation contient également  $Cov[\tilde{x}_i; u_i]$  qui n'est pas identifiable.

- La question est dès lors la suivante :

Peut-on trouver ***une variable***, notée  $\tilde{z}_i$ , ***telle que***  $Cov[\tilde{z}_i; y_i]$  ***fournit une équation qui permette d'identifier  $b_0$  ?***

Avec  $Cov[\tilde{z}_i; y_i] = Cov[\tilde{z}_i; \alpha_0 + b_0 \tilde{x}_i + u_i]$ , on obtient :

$$Cov[\tilde{z}_i; y_i] = b_0 Cov[\tilde{z}_i; \tilde{x}_i] + Cov[\tilde{z}_i; u_i],$$

et on voit directement que  $\tilde{z}_i$  est *utile pour l'identification de*  $b_0$  si :

(i)  $Cov[\tilde{z}_i; u_i] = 0$ , i.e. si  $\tilde{z}_i$  est *exogène par rapport à*  $u_i$

et :

(ii)  $Cov[\tilde{z}_i; \tilde{x}_i] \neq 0$ , i.e. si  $\tilde{z}_i$  est *liée* (linéairement) à  $\tilde{x}_i$ .

En effet, on sait que :

$$p \lim_{N \rightarrow +\infty} N^{-1} \sum_{i=1}^N (\tilde{z}_i - \bar{z}_N) \tilde{x}_i = Cov[\tilde{z}_i; \tilde{x}_i]$$

et :

$$p \lim_{N \rightarrow +\infty} N^{-1} \sum_{i=1}^N (\tilde{z}_i - \bar{z}_N) y_i = Cov[\tilde{z}_i; y_i].$$

$b_0$  est donc le seul paramètre « réellement » inconnu de l'équation précédente  
dès lors que  $Cov[\tilde{z}_i; u_i] = 0$ .

Avec  $\tilde{z}_i$  qui vérifie :

(i)  $Cov[\tilde{z}_i; u_i] = 0$ , i.e.  $\tilde{z}_i$  est exogène par rapport à  $u_i$

et :

(ii)  $Cov[\tilde{z}_i; \tilde{x}_i] \neq 0$ , i.e.  $\tilde{z}_i$  liée (linéairement) à  $\tilde{x}_i$ ,

on a :

$$Cov[\tilde{z}_i; y_i] = b_0 Cov[\tilde{z}_i; \tilde{x}_i] \Leftrightarrow b_0 = Cov[\tilde{z}_i; \tilde{x}_i]^{-1} Cov[\tilde{z}_i; y_i].$$

En notant  $\bar{z}_N \equiv N^{-1} \sum_{i=1}^N \tilde{z}_i$  et en utilisant la LGN et les propriétés des suites convergeant en probabilité, on a également :

$$\hat{b}_N \equiv \left[ N^{-1} \sum_{i=1}^N (\tilde{z}_i - \bar{z}_N) \tilde{x}_i \right]^{-1} N^{-1} \sum_{i=1}^N (\tilde{z}_i - \bar{z}_N) y_i \xrightarrow[N \rightarrow +\infty]{p} b_0$$

**Pour résumer :** Le paramètre  $b_0$  est donc identifiable par  $\tilde{z}_i$ , puisque  $\hat{b}_N$  est un estimateur convergent de  $b_0$  construit à partir de  $\tilde{z}_i$ .

$\tilde{z}_i$  est une *variable instrumentale* de  $\tilde{x}_i$ .

$\hat{b}_N$  est l'*estimateur des VI* de  $b_0$  calculé avec  $\tilde{z}_i$  pour VI de  $\tilde{x}_i$ .

**Définition. Conditions de validité** de  $\tilde{z}_i$  en tant que VI de  $\tilde{x}_i$

Dans le modèle  $y_i = \alpha_0 + b_0 \tilde{x}_i + u_i$  avec  $E[u_i] \equiv 0$ ,  $\tilde{z}_i$  est une VI valide de  $\tilde{x}_i$  si et seulement si :

(i)  $Cov[\tilde{z}_i; u_i] = 0$ , i.e.  $\tilde{z}_i$  est exogène par rapport à  $u_i$

et :

(ii)  $Cov[\tilde{z}_i; \tilde{x}_i] \neq 0$ , i.e.  $\tilde{z}_i$  liée (linéairement) à  $\tilde{x}_i$ ,

*Ces conditions doivent être examinées par l'analyse théorique du PG des  $(y_i, \tilde{x}_i, \tilde{z}_i)$  et empiriquement, tout au moins en partie.*



L'équation :

$$Cov[\tilde{z}_i; y_i] = b_0 Cov[\tilde{z}_i; \tilde{x}_i]$$

montre que la VI  $\tilde{z}_i$  identifie  $b_0$  parce qu'elle n'influence  $y_i$  que *via*  $\tilde{x}_i$ , *i.e.* indirectement.

L'*estimateur des VI* est parfois appelé estimateur des *MC Indirects (MCI)*.

En fait le schéma de fonctionnement de l'estimateur des VI ou des MCI (puis des 2MC) peut être résumé par :

$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \text{ avec } E[u_i] \equiv 0 \quad ; \quad \tilde{z}_i \leftrightarrow \tilde{x}_i \rightarrow y_i \quad ; \quad Cov[\tilde{z}_i; u_i] = 0$
--

L'effet de  $\tilde{z}_i$  sur  $y_i$  ne « transite » que *via*  $\tilde{x}_i$ . La variable instrumentale  $\tilde{z}_i$  ***n'est pas une variable explicative dans le modèle de***  $y_i$ . On parle alors de ***relation d'exclusion*** (de la VI  $\tilde{z}_i$  vis-à-vis du modèle de  $y_i$ ).

On dit des variations de  $\tilde{z}_i$  qu'elles sont des *variations exogènes* : elles ne sont pas liées à  $u_i$  puisque  $Cov[\tilde{z}_i; u_i] = 0$ .

Ce sont *les effets de ces variations exogènes sur  $\tilde{x}_i$*  qui sont exploitées pour *l'identification de  $b_0$*  grâce à  $Cov[\tilde{z}_i; \tilde{x}_i] \neq 0$ .

Noter qu'il n'est aucunement nécessaire que l'effet de  $\tilde{z}_i$  sur  $\tilde{x}_i$  soit causal :

*Une simple corrélation entre  $\tilde{z}_i$  et  $\tilde{x}_i$  suffit pour que  $\tilde{z}_i$  contienne de l'information sur  $\tilde{x}_i$  potentiellement exploitable pour l'inférence statistique.*

Cette *information est exploitable pour l'identification de  $b_0$*  car  $\tilde{z}_i$  *est exogène par rapport au modèle de  $y_i$ , i.e.* par rapport à  $u_i$ . Avec  $Cov[\tilde{z}_i; u_i] = 0$  on a :

$$Cov[\tilde{z}_i; y_i] = b_0 Cov[\tilde{z}_i; \tilde{x}_i] \Leftrightarrow b_0 = Cov[\tilde{z}_i; \tilde{x}_i]^{-1} Cov[\tilde{z}_i; y_i]$$

## 2.2. Variables instrumentales, instruments et conditions d'orthogonalité

**Idée ici :** Illustrer le rôle des conditions d'orthogonalité pour utiliser l'information apportée par les VI.

C'est comme avec la covariance comme ci-dessus mais sous une forme différente

On part modèle linéaire :

$$y_i = \alpha_0 + b_{1,0}\tilde{x}_i^x + b_{2,0}\tilde{x}_i^e + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

L'analyse du PGD indique que :

$$E[u_i / \tilde{x}_i^x] = 0 \quad \text{mais} \quad E[u_i / \tilde{x}_i^e] \neq 0 \Rightarrow E[\tilde{x}_i^e u_i] = \text{Cov}[\tilde{x}_i^e; u_i] \neq 0,$$

*i.e.* que  $\tilde{x}_i^x$  est exogène mais que  $\tilde{x}_i^e$  est endogène par rapport à  $u_i$ .

Après un nouvel examen du PGD, il existe une variable disponible telle que :

$$E[u_i/\tilde{x}_i^x, \tilde{z}_i^e] = 0 \text{ et } Cov[\tilde{x}_i^e; \tilde{z}_i^e] \neq 0,$$

i.e.  $\tilde{z}_i^e$  est une VI valide de  $\tilde{x}_i^e$  dans le modèle.

**Question** : Comment se servir de l'information apportée par  $\tilde{z}_i^e$  pour estimer  $\mathbf{a}_0 \equiv (\alpha_0, b_{1,0}, b_{2,0})$  ?

Les *covariances* entre  $\tilde{z}_i^e$ ,  $\tilde{x}_i^x$  d'une part, et  $y_i$  d'autre part *sont difficiles à manipuler directement* (surtout dans le cas de modèles non-linéaires). En fait, identifier  $\mathbf{a}_0$  à partir de l'« approche » par les covariances revient à résoudre le système :

$$\begin{cases} E[y_i] = \alpha_0 + b_{1,0}E[\tilde{x}_i^x] + b_{2,0}E[\tilde{x}_i^e] \\ Cov[\tilde{x}_i^x; y_i] = b_{1,0}V[\tilde{x}_i^x] + b_{2,0}Cov[\tilde{x}_i^x; \tilde{x}_i^e] \\ Cov[\tilde{z}_i^e; y_i] = b_{1,0}Cov[\tilde{z}_i^e; \tilde{x}_i^x] + b_{2,0}Cov[\tilde{z}_i^e; \tilde{x}_i^e] \end{cases}$$

**« Astuce »**

On fonde l'identification de  $\mathbf{a}_0$ , et le calcul de l'estimateur MM associé  $\hat{\mathbf{a}}_N^{MM}$ , sur un *système de conditions d'orthogonalité* et non sur le système ci-dessus

Le modèle :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \quad \text{avec} \quad E[u_i / \mathbf{z}_i] \equiv 0$$

où :

$$\mathbf{a}_0 \equiv \begin{bmatrix} \alpha_0 \\ b_{1,0} \\ b_{2,0} \end{bmatrix}, \quad \mathbf{x}_i \equiv \begin{bmatrix} 1 \\ \tilde{x}_i^x \\ \tilde{x}_i^e \end{bmatrix} \quad \text{et} \quad \mathbf{z}_i \equiv \begin{bmatrix} 1 \\ \tilde{x}_i^x \\ \tilde{z}_i^e \end{bmatrix}$$

donne :

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

**La condition d'orthogonalité dérivée de l'exogénéité de  $\mathbf{z}_i$  :**

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

sert alors de condition de moment estimante.

- Le vecteur  $\mathbf{z}_i$  est nommé ici ***vecteur d'instruments***. C'est le vecteur des variables utilisé pour construire des conditions de moment estimantes, *i.e.* des conditions d'orthogonalité avec le terme d'erreur du modèle.
- Le vecteur d'***instruments*** ne contient donc que des ***variables exogènes, explicatives et/ou instrumentales***.
- La variable 1 est exogène car  $E[u_i/1] = E[u_i] = E[1 \times u_i] = 0$ .

**Procédure.** Construction du vecteur d'instruments

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ \tilde{x}_i^x \\ \tilde{x}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{x}_i^e \end{bmatrix} \left\{ \begin{array}{l} \text{variables exogènes} \\ \text{variable endogène} \end{array} \right.$$

$$\mathbf{z}_i = \begin{bmatrix} 1 \\ \tilde{x}_i^x \\ \tilde{z}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{z}_i^e \end{bmatrix} \left\{ \begin{array}{l} \text{variables explicatives exogènes} \\ \text{variable instrumentale} \end{array} \right.$$

Pour construire  $\hat{\mathbf{a}}_N^{MM}$ , on utilise le *principe d'analogie* à partir de :

$$\mathbf{a}_0 \text{ solution en } \mathbf{a} \text{ de } E[\mathbf{z}_i(y_i - \mathbf{x}'_i \mathbf{a})] = \mathbf{0}$$

ce qui donne pour  $\hat{\mathbf{a}}_N^{MM}$  :

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

en supposant (ici) que  $\mathbf{a}_0$  est l'unique solution en  $\mathbf{a}$  de  $E[\mathbf{z}_i(y_i - \mathbf{x}'_i \mathbf{a})] = \mathbf{0}$ . On obtient alors :

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

et :

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

cette équation définissant ce qu'on appelle l'*estimateur des VI*.

**Rmq.** On peut avoir  $E[u_i/\tilde{x}_{1,i}] = 0$  et  $E[u_i/\tilde{x}_{2,i}] = 0$ , et  $E[u_i/\tilde{x}_{1,i}, \tilde{x}_{2,i}] \neq 0$ . Il convient donc d'être prudent en présence des conditions d'exogénéité et d'endogénéité des variables explicatives des modèles économétriques.

L'idée sous-jacente est qu'un vecteur de variables aléatoires apporte plus d'information que chacun des éléments du vecteur pris séparément.

Par exemple, les variables  $\tilde{x}_{1,i}$ ,  $\tilde{x}_{2,i}$  et  $u_i$  sont dichotomiques avec 1 = "*vrai*" et 0 = "*faux*", et :

$\tilde{x}_{1,i}$  :  $i$  préfère le candidat de qui il parle

$\tilde{x}_{2,i}$  :  $i$  parle du candidat A

$u_i$  :  $i$  vote pour A.

S'il n'y a que deux candidats, on a alors :

$$P[u_i/\tilde{x}_{1,i} = 1] = P[u_i/\tilde{x}_{2,i} = 1] = 1/2$$

mais :

$$P[u_i/\tilde{x}_{1,i} = 1, \tilde{x}_{2,i} = 1] = 1..$$



## 2.3. Exemples de variables instrumentales

### Exemple de l'effet la taille des classes sur les résultats des élèves

Pour  $tailclas_i$  dans :

$$score_i = \alpha_0 + b_{1,0}tailclas_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

Angrist et Lavy proposent  $tailclas\_moy_i$ , la taille de classe moyenne des classes (de CE1 pour Piketty) dans l'école de  $i$ .

- (i) exogène car dépend essentiellement du nombre d'élève de CE1 dans l'école et de la règle de la classe inférieure à 25 élèves
- (ii) bien corrélée empiriquement à  $tailclas_i$

Angrist et Lavy (1999) et Piketty (2004) obtiennent  $\hat{b}_{1,N} < 0$  ( $\hat{b}_{1,N}^{MCO} \approx 0$ ).

**Rmq.** Piketty montre également que l'effet de la taille de la classe est plus élevé en valeur absolue en ZEP qu'ailleurs.

## Exemple de l'effet de la maternité sur le salaire des femmes

Angrist et Evans considèrent un effet particulier : *l'effet du troisième enfant*.

$trois\_enf_i = 1$  si la femme  $i$  a 3 enfants

$trois\_enf_i = 0$  si la femme  $i$  a 2 enfants

Pour  $trois\_enf_i$  dans :

$$salaire_i = \alpha_0 + b_{1,0}trois\_enf_i + \mathbf{b}'_{-1,0}\mathbf{c}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

Angrist et Evans proposent  $meme\_sexe_i$  comme VI :

$meme\_sexe_i = 1$  si les premiers enfants de  $i$  ont le meme sexe

$meme\_sexe_i = 0$  si les premiers enfants de  $i$  n'ont pas le meme sexe

(i) exogène car le sexe des enfants est « purement » aléatoire

(ii) bien corrélée **positivement** empiriquement à  $trois\_enf_i$

Angrist et Evans (2000) obtiennent  $\hat{b}_{1,N} < 0$  ( $\hat{b}_{1,N}^{MCO} \approx 0$ ).

**Rmq.** Ils obtiennent un effet quasi-nul sur l'activité professionnelle des pères.

## Exemple de l'erreur de mesure sur la variable explicative

**Modèle « d'intérêt »:**

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \text{ avec } E[u_i / \tilde{x}_i] = E[u_i] \equiv 0,$$

**Mesure avec erreur de  $\tilde{x}_i$  :**  $\tilde{x}_i^e = \tilde{x}_i + e_i$ .

**Modèle « observable » :**

$$y_i = \alpha_0 + b_0 \tilde{x}_i^e + v_i \text{ avec } v_i = u_i - b_0 e_i \text{ et } E[v_i] = 0.$$

Une bonne VI de  $\tilde{x}_i^e$  est une autre mesure de  $\tilde{x}_i$  avec une erreur  $\varepsilon_i$  non corrélée à  $e_i$  et à  $u_i$  (*technique dite de la « double-proxy »*)

(i) exogène car  $\varepsilon_i$  exogène et non corrélée à  $e_i$

(ii) bien corrélée à  $\tilde{x}_i^e$  par  $\tilde{x}_i$

## 2.4. L'estimateur des VI

L'objectif est maintenant de généraliser l'approche présentée ici dans un cas simple et de présenter le *modèle linéaire à VI*.

Tout d'abord on considère un modèle linéaire général :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0,$$

dans lequel plusieurs variables explicatives sont potentiellement endogènes :

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} \quad \left. \begin{array}{l} \} \text{ variables exogènes} \quad (k = 1, \dots, M) \\ \} \text{ variables endogènes} \quad (k = M + 1, \dots, K) \end{array} \right\}$$

**Rmq.** La variable constante 1 est « exogène » :  $E[1 \times u_i] = E[u_i/1] = E[u_i] = 0$ .

Nous utiliserons ensuite la **Méthode des Moments** pour construire un estimateur convergent de  $\mathbf{a}_0$ , l'**estimateur des VI** « de forme générale ».

On considère ici que **chaque élément de  $\mathbf{x}_i^e = \tilde{\mathbf{x}}_i^e$  a une** variable instrumentale.

**Définition.**  $z_{k,i}$  est une **variable instrumentale** de  $x_{k,i}$  dans le modèle linéaire si

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

et :

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

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

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

où  $e_{k,i}$  est la partie spécifique de  $x_{k,i}$  dans  $\mathbf{x}_i$ , i.e. :

$$e_{k,i} \equiv x_{k,i} - EL[x_{k,i} / \mathbf{x}_{-k,i}].$$

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} \left\{ \begin{array}{l} \text{variables exogènes} \quad (k=1,\dots,M) \\ \text{variables endogènes} \quad (k=M+1,\dots,K) \end{array} \right.$$

**Rmq.** Si  $x_{k,i}$  est exogène, c'est une (*la meilleure*) VI d'elle-même.

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

$$\tilde{\mathbf{z}}_i^e \equiv \begin{bmatrix} \tilde{z}_{M+1,i} \\ \tilde{z}_{M+2,i} \\ \vdots \\ \tilde{z}_{K,i} \end{bmatrix} \text{ et } \mathbf{z}_i = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} \left\{ \begin{array}{l} \text{variables exogènes de } \mathbf{x}_i \quad (k=1,\dots,M) \\ \text{variables instrumentales} \quad (k=M+1,\dots,K) \end{array} \right.$$

Ce vecteur contient en fait *toutes les variables exogènes* du modèle. Ce sont *ces variables* qui *assurent l'identification* des paramètres du modèle.

Etant donnée son importance en matière d'inférence,  $\mathbf{z}_i$  est parfois nommé *ensemble d'information du modèle*.

**Définition.** Le modèle défini par :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \text{ avec } E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0$$

est un *modèle linéaire à variables instrumentales*.

La condition d'identification de  $\mathbf{a}_0$  dans ce modèle est donnée par :

$$\text{rang} E[\mathbf{z}_i \mathbf{x}'_i] = K = \dim \mathbf{x}_i.$$

**Rmq.** Noter que la condition d'exogénéité de  $\mathbf{z}_i$  est définie par  $E[u_i / \mathbf{z}_i] = 0$ , et non par  $\text{Cov}[\mathbf{z}_i; u_i] = \mathbf{0}$ .

Ce n'est pas nécessaire pour un modèle linéaire où  $\text{Cov}[\mathbf{z}_i; u_i] = \mathbf{0}$  suffit mais c'est standard et cela simplifie la présentation des hypothèses d'homoscédasticité.

Comme dans le cas où on a construit l'estimateur des MCO de  $\mathbf{a}_0$  par la MM, 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 :

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

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

$$\mathbf{a}_0 \text{ solution en } \mathbf{a} \text{ de } E[\mathbf{z}_i(y_i - \mathbf{x}'_i \mathbf{a})]=\mathbf{0}$$

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

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



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 une solution sous forme explicite. On a :

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

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

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

Ceci donne finalement :

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

Cette équation définit ce qu'on appelle l'***estimateur des VI***.

**Définition.** Soit le modèle à variables instrumentales :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \quad \text{avec} \quad E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0.$$

L'*estimateur des variables instrumentales (des VI)* de  $\mathbf{a}_0$  dans ce modèle est défini par :

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

Cet estimateur est convergent pour  $\mathbf{a}_0$  est asymptotiquement normal, ce que nous allons démontrer (rapidement car « on fait toujours les mêmes démonstrations »).

### Propriété 29.

#### *Convergence de $\hat{\mathbf{a}}_N^{VI}$ dans un modèle à VI*

Soit  $\{(y_i, \mathbf{x}_i, \mathbf{z}_i); i = 1, 2, \dots, N\}$  un échantillon de variables aléatoires telles que :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \quad \text{avec} \quad E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0.$$

L'estimateur des VI de  $\mathbf{a}_0$  :

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

(i) existe avec une probabilité approchant 1

et :

(ii) est convergent, i.e. :  $\hat{\mathbf{a}}_N^{VI} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{a}_0$ .

Le modèle de  $y_i$  nous donne que  $y_i = \mathbf{x}_i' \mathbf{a}_0 + u_i$ , on a donc :

$$\hat{\mathbf{a}}_N^{VI} = \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i y_i = \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i (\mathbf{x}_i' \mathbf{a}_0 + u_i).$$

Après développement, on obtient :

$$\hat{\mathbf{a}}_N^{VI} = \mathbf{a}_0 + \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i.$$

On sait, par la loi LGN, que :

$$N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i \mathbf{x}_i'] \text{ et } N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i u_i]$$

En combinant ces résultats et on obtient :

$$\left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[\mathbf{z}_i u_i] = E[\mathbf{z}_i \mathbf{x}_i']^{-1} \times \mathbf{0} = \mathbf{0}$$

en utilisant  $E[\mathbf{z}_i u_i] = \mathbf{0}$ . Finalement on obtient :  $\boxed{\hat{\mathbf{a}}_N^{VI} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{a}_0 + \mathbf{0} = \mathbf{a}_0.}$

### Propriété 30.

#### *Normalité asymptotique de $\hat{\mathbf{a}}_N^{VI}$ dans un modèle à VI*

Soit  $\{(y_i, \mathbf{x}_i, \mathbf{z}_i); i = 1, 2, \dots, N\}$  un échantillon de variables aléatoires telles que  $y_i = \mathbf{a}_0' \mathbf{x}_i + u_i$  avec  $E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0$ . L'estimateur des VI de  $\mathbf{a}_0$  :

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

vérifie :

$$\sqrt{N}(\hat{\mathbf{a}}_N^{VI} - \mathbf{a}_0) \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N}(\mathbf{0}, \Sigma_0)$$

avec :

$$\Sigma_0 = E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1}$$

**Rmq.** La condition d'homoscédasticité éventuelle des  $u_i$  est définie en référence à  $\mathbf{z}_i$  pas à  $\mathbf{x}_i$ .

On utilise ici directement le fait que :

$$\hat{\mathbf{a}}_N^{VI} = \mathbf{a}_0 + \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i$$

qui implique que :

$$\sqrt{N} (\hat{\mathbf{a}}_N^{VI} - \mathbf{a}_0) = \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \sqrt{N} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \right].$$

La LGN et les propriétés des suites convergeant en probabilité donnent :

$$\left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i \mathbf{x}_i']^{-1}.$$

Le TCL implique que :

$$\sqrt{N} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \right] \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N} \left( E[\mathbf{z}_i u_i], E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] \right)$$

et avec  $E[\mathbf{z}_i u_i] = \mathbf{0}$  que :

$$\sqrt{N} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \right] \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N} \left( \mathbf{0}, E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] \right).$$

On a :

$$\sqrt{N}(\hat{\mathbf{a}}_N^{VI} - \mathbf{a}_0) \xrightarrow[N \rightarrow +\infty]{L} E[\mathbf{z}_i \mathbf{x}_i']^{-1} \times \mathcal{N}(\mathbf{0}, E[u_i^2 \mathbf{z}_i \mathbf{z}_i'])$$

et donc :

$$\sqrt{N}(\hat{\mathbf{a}}_N^{VI} - \mathbf{a}_0) \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N}(\mathbf{0}, E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1})$$

On applique ici des propriétés classiques :

$$\begin{aligned} V[\mathbf{A}_0 \mathbf{m}_i] &= \mathbf{A}_0 V[\mathbf{m}_i] \mathbf{A}_0' \Rightarrow \boldsymbol{\Sigma}_0 = E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] (E[\mathbf{z}_i \mathbf{x}_i']^{-1})' \\ (\mathbf{A}_0^{-1})' &= (\mathbf{A}_0')^{-1} \Rightarrow (E[\mathbf{z}_i \mathbf{x}_i']^{-1})' = (E[\mathbf{z}_i \mathbf{x}_i'])'^{-1} \end{aligned}$$

et :

$$(\mathbf{AB})' = \mathbf{B}' \mathbf{A}' \text{ et } E[\mathbf{M}_i]' = E[\mathbf{M}_i'] \Rightarrow E[\mathbf{z}_i \mathbf{x}_i']' = E[\mathbf{x}_i \mathbf{z}_i']$$

On a besoin d'un estimateur de  $\boldsymbol{\Sigma}_0 = E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1}$ .

Pour calculer un estimateur de :

$$\Sigma_0 = E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1}.$$

*On distingue deux cas.*

Dans le *cas homoscédastique* on a  $V[u_i/\mathbf{z}_i] = E[u_i^2/\mathbf{z}_i] = \sigma_0^2$ . La loi des conditionnements successifs donne alors :

$$E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] = \sigma_0^2 E[\mathbf{z}_i \mathbf{z}_i'] = E[\mathbf{z}_i \mathbf{z}_i']$$

et donc :

$$\Sigma_0 = \sigma_0^2 E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[\mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1}.$$

***Dans tous les cas, on utilise toujours les mêmes techniques :***

On remplace les espérances mathématiques par des moyennes et les paramètres inconnus par des estimateurs convergents.



## Cas homoscédastique

On a :

$$\Sigma_0 = \sigma_0^2 E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[\mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1}$$

et on sait :

$$\left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right]^{-1} \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{x}_i \mathbf{z}_i']^{-1} \text{ et } \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i \mathbf{x}_i']^{-1},$$

et :

$$N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i \mathbf{z}_i'].$$

Avec  $u_i = y_i - \mathbf{x}_i' \mathbf{a}_0$  et  $\hat{\mathbf{a}}_N^{VI} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{a}_0$  on a :

$$\hat{\sigma}_N^2 \equiv N^{-1} \sum_{i=1}^N (y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{VI})^2 \xrightarrow[N \rightarrow +\infty]{p} \sigma_0^2 = E[(y_i - \mathbf{x}_i' \mathbf{a}_0)^2].$$

et :

$$\begin{aligned} \hat{\Sigma}_N &\equiv \hat{\sigma}_N^2 \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right]^{-1} \\ &\xrightarrow[N \rightarrow +\infty]{p} \Sigma_0 = \sigma_0^2 E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[\mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1}. \end{aligned}$$

## Cas hétéroscédastique

On a :

$$\Sigma_0 = E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1}$$

et on sait :

$$\left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right]^{-1} \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{x}_i \mathbf{z}_i']^{-1} \text{ et } \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i \mathbf{x}_i']^{-1} .$$

Avec  $u_i = y_i - \mathbf{x}_i' \mathbf{a}_0$  on a  $E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] = E[(y_i - \mathbf{x}_i' \mathbf{a}_0)^2 \mathbf{z}_i \mathbf{z}_i']$ . Par conséquent, avec  $\hat{\mathbf{a}}_N^{VI} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{a}_0$  on a :

$$N^{-1} \sum_{i=1}^N (y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{VI})^2 \mathbf{z}_i \mathbf{z}_i' \xrightarrow[N \rightarrow +\infty]{p} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] = E[(y_i - \mathbf{x}_i' \mathbf{a}_0)^2 \mathbf{z}_i \mathbf{z}_i'] .$$

et :

$$\begin{aligned} \hat{\Sigma}_N &\equiv \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N (y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{VI})^2 \mathbf{z}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right]^{-1} \\ &\xrightarrow[N \rightarrow +\infty]{p} \Sigma_0 = \sigma_0^2 E[\mathbf{z}_i \mathbf{x}_i']^{-1} E[\mathbf{z}_i \mathbf{z}_i'] E[\mathbf{x}_i \mathbf{z}_i']^{-1} . \end{aligned}$$

**Propriété 31. Estimateurs de la variance asymptotique de  $\hat{\mathbf{a}}_N^{VI}$  dans un modèle à VI**

La variance asymptotique,  $\Sigma_0$ , de l'estimateur des VI de  $\mathbf{a}_0$ ,  $\hat{\mathbf{a}}_N^{VI}$ , peut être estimée par  $\hat{\Sigma}_N$  si  $V[u_i/\mathbf{x}_i] = \sigma_0^2$  (cas homoscédastique) ou par  $\hat{\Sigma}_N^W$  (cas général, hétéroscédastique). On a :

$$(i) \quad \hat{\Sigma}_N^W \equiv \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \hat{u}_{i,N}^2 \mathbf{z}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right]^{-1}$$

et :

$$(ii) \quad \hat{\Sigma}_N \equiv \hat{\sigma}_N^2 \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right]^{-1}$$

avec  $\hat{\sigma}_N^2 \equiv N^{-1} \sum_{i=1}^N \hat{u}_{i,N}^2$  et en notant,  $\hat{u}_{i,N} \equiv y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{VI}$  le résidu d'estimation.

**Rmq.** L'estimateur  $\hat{\Sigma}_N^W$  est dit « *robuste à l'hétéroscédasticité* » ou « *de White* ».

# Outline

1. Modèles à variables explicatives endogènes
2. La notion de variable instrumentale et l'estimateur des VI
- 3. L'estimateur des 2MC**
4. Choix des VI et identification
5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée
6. Les variables de contrôle de l'hétérogénéité

### **3. L'estimateur des 2MC**

### 3. L'estimateur des 2MC

On continue à se placer dans le cadre du modèle *modèle linéaire à VI* de forme générale :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \quad \text{avec} \quad E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0,$$

dans lequel plusieurs variables explicatives sont potentiellement en endogènes :

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \mathbf{x}_i^e \end{bmatrix} \quad \left. \begin{array}{l} \} \text{ variables exogènes} \\ \} \text{ variables endogènes} \end{array} \right\} \quad \begin{array}{l} (k = 1, \dots, M) \\ (k = M + 1, \dots, K) \end{array}$$

En fait, quand on cherche à « *instrumenter* » une *variable explicative endogène*, disons  $\tilde{x}_{k,i}$  (avec  $k > M$ ), on cherche des variables exogènes par rapport à  $u_i$  susceptibles d'être corrélées à  $\tilde{x}_{k,i}$  (ce qu'on peut vérifier).

### Technique de recherche de VI pour la variable endogène $\tilde{x}_{k,i}$

Si je devais *prédire au mieux* (pas nécessairement construire un modèle causal de)  $\tilde{x}_{k,i}$  :

(i) quelle variables « explicatives » j'utiliserais ?

sachant que :

(ii) ces variables explicatives de  $\tilde{x}_{k,i}$  doivent être exogènes dans mon modèle d'intérêt, *i.e.* par rapport à  $u_i$ .

En pratique on peut donc disposer de *plusieurs VI pour une seule* variable explicative endogène.

- C'est bon en matière d'inférence : *plusieurs VI contiennent plus d'information qu'une seule* ...
- ... mais cela rend impossible l'utilisation *directe* de l'estimateur des VI.

- C'est pour ces cas qu'a été défini l'estimateur des *Doubles Moindres Carrés (2MC)*

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

$$\mathbf{z}_i = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} \left\{ \begin{array}{l} \text{variables exogènes} \quad (k=1, \dots, M) \\ \text{variables instrumentales} \quad (k=M+1, \dots, L) \end{array} \right.$$

où  $\tilde{\mathbf{z}}_i^e$  contient toutes les VI « externes » disponibles pour instrumenter  $\tilde{\mathbf{x}}_i^e$  et :

$$\dim \tilde{\mathbf{x}}_i^e = K - M \leq L - M = \dim \tilde{\mathbf{z}}_i^e$$



Nb de var. explicatives endogènes  $\leq$  Nb de var. instrumentales "externes"

ou, de manière équivalente :

$$\dim \mathbf{x}_i = K \leq L = \dim \mathbf{z}_i$$



Nb de var. explicatives  $\leq$  Nb de var. instrumentales



Le problème posé par l'utilisation de l'estimateur des VI de  $\mathbf{a}_0$  est simple.

La condition de moment estimante de  $\mathbf{a}_0$  considérée par l'estimateur des VI est la condition d'orthogonalité issue de l'exogénéité de  $\mathbf{z}_i$  :

$$E[\mathbf{z}_i(y_i - \mathbf{a}'_0 \mathbf{x}_i)] = \mathbf{0}_{L \times 1}.$$

*Lorsque*  $L = K$ , le cas considéré pour la construction de  $\hat{\mathbf{a}}_N^{VI}$ , cette équation définit un système de  $L = K = \dim \mathbf{z}_i$  équations à  $K = \dim \mathbf{a}_0$  inconnues. Elle est donc aisée à résoudre, en particulier :

$$\mathbf{a}_0 = E[\mathbf{z}_i \mathbf{x}'_i]^{-1} E[\mathbf{z}_i y_i] \quad \text{si } L = K.$$

On retrouve ici la structure de base de l'estimateur  $\hat{\mathbf{a}}_N^{VI}$  qui est la contre-partie empirique de  $E[\mathbf{z}_i \mathbf{x}'_i]^{-1} E[\mathbf{z}_i y_i]$ .

**Lorsque  $L > K$** , le cas considéré pour la construction de  $\hat{\mathbf{a}}_N^{VI}$ , cette équation définit un système de  $L = \dim \mathbf{z}_i$  équations à  $K = \dim \mathbf{a}_0$  inconnues.

Il y a donc ***plus d'équations que d'inconnues***. On dit alors que :

**Définition.** Dans un modèle à VI, le vecteur d'instruments  $\mathbf{z}_i$

- (i) ***juste-identifie***  $\mathbf{a}_0$  si  $\dim \mathbf{a}_0 = \dim \mathbf{x}_i = K = L = \dim \mathbf{z}_i$
- (ii) ***sur-identifie***  $\mathbf{a}_0$  si  $\dim \mathbf{a}_0 = \dim \mathbf{x}_i = K < L = \dim \mathbf{z}_i$
- (iii) ***sous-identifie*** (n'identifie pas)  $\mathbf{a}_0$  si  $\dim \mathbf{a}_0 = \dim \mathbf{x}_i = K > L = \dim \mathbf{z}_i$

En cas de ***sur-identification***, l'équation :

$$E[\mathbf{z}_i(y_i - \mathbf{a}_0' \mathbf{x}_i)] = \mathbf{0}_{L \times 1}$$

ne permet pas l'application ***directe*** du principe d'analogie.

La contre-partie empirique de  $E[\mathbf{z}_i(y_i - \mathbf{a}'\mathbf{x}_i)] = \mathbf{0}_{L \times 1}$  :

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

n'admet pas, en général, de solution en  $\mathbf{a}$ . En quelque sorte, il y a « *trop* » d'équations.

## Il y a plusieurs manières de résoudre ce problème :

- (i) *Éliminer des éléments de  $\mathbf{z}_i$*  pour se ramener au cas juste-identifié et pouvoir employer  $\hat{\mathbf{a}}_N^{VI}$ . *Pas satisfaisant* car :

Élimination d'éléments de $\mathbf{z}_i$ = perte d'information = perte d'efficacité
---

- (ii) *Utiliser la Méthode des Moments Généralisée* (MM généralisée pour ça)
- (iii) *Utiliser une astuce pour « réduire » la dimension* du vecteur de VI utilisé *sans perdre d'information (ou tout au moins un minimum)*

Dans le cas (iii) on cherche à réduire la dimension de la condition d'orthogonalité estimante pour $\mathbf{a}_0$ , $E[\mathbf{z}_i(y_i - \mathbf{a}_0' \mathbf{x}_i)] = \mathbf{0}_{L \times 1}$ .
---

On veut en fait définir à partir de  $\mathbf{z}_i$ , *un vecteur d'instruments*, noté  $\mathbf{w}(\mathbf{z}_i)$ , de dimension  $K$  tel que :

$$E[\mathbf{w}(\mathbf{z}_i)(y_i - \mathbf{a}'_0 \mathbf{x}_i)] = \mathbf{0}_{K \times 1}.$$

Or, un bon vecteur d'instruments doit (i) être exogène par rapport à  $u_i$  et (ii) permettre de « bien prédire »  $\mathbf{x}_i$  (ou avoir une « forte » corrélation avec  $\mathbf{x}_i$ ).

*La projection linéaire* de  $\mathbf{x}_i$  sur  $\mathbf{z}_i$ ,  $EL[\mathbf{x}_i/\mathbf{z}_i]$ , est  
un *excellent candidat* pour  $\mathbf{w}(\mathbf{z}_i)$ .

**Rmq.**  $EL[\mathbf{x}_i/\mathbf{z}_i]$  est un *excellent* candidat pour  $\mathbf{w}(\mathbf{z}_i)$  car c'est la *meilleure combinaison linéaire des éléments* de  $\mathbf{z}_i$  en matière de prédiction de  $\mathbf{x}_i$ .

**Rmq.** Le *meilleur* (Chamberlain) est  $E[\mathbf{x}_i/\mathbf{z}_i]V[u_i/\mathbf{z}_i]^{-1}$  pour l'efficacité as. de l'estimation de  $\mathbf{a}_0$  à partir du modèle de  $\mathbf{x}_i$  et des VI  $\mathbf{z}_i$ .

## Rappel sur les projections linéaires

La projection linéaire de  $x_i$  sur  $\mathbf{z}_i$ , notée  $EL[x_i/\mathbf{z}_i]$ , est le *meilleur prédicteur linéaire de  $x_i$  par  $\mathbf{z}_i$*  au sens de l'*erreur quadratique moyenne* :

$$EL[x_i/\mathbf{z}_i] \equiv \boldsymbol{\gamma}'\mathbf{z}_i \text{ où } \boldsymbol{\gamma} \equiv \arg \min_{\mathbf{g}} E[(x_i - \mathbf{g}'\mathbf{z}_i)^2].$$

Elle s'interprète de deux manières :

- *Espérance conditionnelle linéaire*
- Prédiction par une *régression asymptotique*

C'est un *outil mathématique très pratique*. Il permet de *décomposer*  $x_i$  en la somme de la *projection* de  $x_i$  sur  $\mathbf{z}_i$  ( $\boldsymbol{\gamma}'\mathbf{z}_i$ ) et le *résidu de cette projection* ( $e_i$ ) :

$$x_i = \boldsymbol{\gamma}'\mathbf{z}_i + e_i \text{ avec } e_i \equiv x_i - \boldsymbol{\gamma}'\mathbf{z}_i$$

telle que :

$$E[\mathbf{z}_i e_i] = \mathbf{0}, \text{ par construction de } e_i.$$

La notion de projection est très liée à la notion de régression. En particulier, si  $E[\mathbf{z}_i \mathbf{z}_i']$  est inversible on a :

$$\boldsymbol{\gamma} = E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i x_i].$$

On peut alors obtenir très simplement un *estimateur convergent de  $\boldsymbol{\gamma}$* , le paramètre de la projection :

$$\hat{\boldsymbol{\gamma}}_N^{MCO} \equiv \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i x_i \xrightarrow[N \rightarrow +\infty]{p} \boldsymbol{\gamma} = E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i x_i],$$

*i.e.* il suffit d'utiliser l'*estimateur des MCO de  $\boldsymbol{\gamma}$*  dans l'équation  $x_i = \boldsymbol{\gamma}' \mathbf{z}_i + e_i$ .

On a également :

$$\mathbf{z}_i' \hat{\boldsymbol{\gamma}}_N^{MCO} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{z}_i' \boldsymbol{\gamma} = EL[x_i / \mathbf{z}_i].$$

**Rmq.** L'équation  $x_i = \gamma' \mathbf{z}_i + e_i$  *ne définit pas un modèle de*  $x_i$  mais une simple décomposition de  $x_i$ . *Cette équation n'est qu'un outil mathématique.*

**Rmq.** « Régresser »  $x_i$  sur  $\mathbf{z}_i$  fournit, par construction, un estimateur convergent du meilleur prédicteur linéaire de  $x_i$  sur  $\mathbf{z}_i$  au sens de l'erreur quadratique moyenne, *i.e.* des « Moindres Carrés ».

Cette prédiction est obtenue dans une *logique d'ajustement*, *i.e.* sans référence à un modèle causal.



Dans le cas multi-varié on a :

$$EL[\mathbf{x}_i/\mathbf{z}_i] \equiv \begin{bmatrix} EL[x_{1,i}/\mathbf{z}_i] \\ EL[x_{2,i}/\mathbf{z}_i] \\ \vdots \\ EL[x_{K,i}/\mathbf{z}_i] \end{bmatrix} \equiv \begin{bmatrix} \gamma'_1 \mathbf{z}_i \\ \gamma'_2 \mathbf{z}_i \\ \vdots \\ \gamma'_K \mathbf{z}_i \end{bmatrix} \equiv \Gamma \mathbf{z}_i = \begin{bmatrix} EL[1/\mathbf{z}_i] \\ EL[x_{2,i}/\mathbf{z}_i] \\ \vdots \\ EL[x_{M,i}/\mathbf{z}_i] \\ EL[x_{M+1,i}/\mathbf{z}_i] \\ \vdots \\ EL[x_{K,i}/\mathbf{z}_i] \end{bmatrix} = \begin{bmatrix} 1 \\ x_{2,i} \\ \vdots \\ x_{M,i} \\ \gamma'_{M+1} \mathbf{z}_i \\ \vdots \\ \gamma'_K \mathbf{z}_i \end{bmatrix}$$

- (i)  $\dim EL[\mathbf{x}_i/\mathbf{z}_i] = K$ , ce qu'on voulait pour utiliser la MM.
- (ii)  $EL[\mathbf{x}_i/\mathbf{z}_i] = \Gamma \mathbf{z}_i$  est **exogène** car c'est une fonction de variables exogènes.
- (iii)  $EL[\mathbf{x}_i/\mathbf{z}_i]$  est **bien corrélée** à  $\mathbf{x}_i$  puisque c'est, par construction, son meilleur prédicteur linéaire par  $\mathbf{z}_i$ .
- (iv) Il est facile de construire un estimateur convergent de  $EL[\mathbf{x}_i/\mathbf{z}_i]$ .

### Propriété 32. *Projection linéaire de $\mathbf{x}_i$ sur $\mathbf{z}_i$*

La projection linéaire de  $\mathbf{x}_i$  sur  $\mathbf{z}_i$ , notée  $EL[\mathbf{x}_i/\mathbf{z}_i] \equiv \Gamma \mathbf{z}_i$  est donnée par :

$$EL[\mathbf{x}_i/\mathbf{z}_i] \equiv \Gamma \mathbf{z}_i = E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} \mathbf{z}_i$$

(si  $E[\mathbf{z}_i \mathbf{z}_i']$  est inversible), et on a :

$$\hat{\Gamma}_N \equiv \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \xrightarrow[N \rightarrow +\infty]{p} \Gamma$$

et donc :

$$\hat{\Gamma}_N \mathbf{z}_i \xrightarrow[N \rightarrow +\infty]{p} \Gamma \mathbf{z}_i.$$

(si la LGN et les propriétés des suites convergent en probabilité s'appliquent).

**Rmq.**  $\hat{\Gamma}_N$  est en fait l'empilement des estimateurs des MCO des  $\gamma'_k$ .

**Propriété 32bis. *Projection linéaire de  $\mathbf{x}_i$  sur  $\mathbf{z}_i$  et variances***

La projection linéaire de  $\mathbf{x}_i$  sur  $\mathbf{z}_i$ , notée  $EL[\mathbf{x}_i/\mathbf{z}_i] \equiv \Gamma \mathbf{z}_i$  est donnée par :

$$EL[\mathbf{x}_i/\mathbf{z}_i] \equiv \Gamma \mathbf{z}_i = E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} \mathbf{z}_i$$

(si  $E[\mathbf{z}_i \mathbf{z}_i']$  est inversible). En notant :

$$\mathbf{e}_i \equiv \mathbf{x}_i - EL[\mathbf{x}_i/\mathbf{z}_i] = \mathbf{x}_i - \Gamma \mathbf{z}_i$$

on a :

$$(i) \quad \mathbf{x}_i = EL[\mathbf{x}_i/\mathbf{z}_i] + \mathbf{e}_i = \Gamma \mathbf{z}_i + \mathbf{e}_i \text{ avec } E[\mathbf{z}_i \mathbf{e}_i'] = \mathbf{0}$$

et :

$$(ii) \quad E[\mathbf{x}_i \mathbf{x}_i'] = E \left[ EL[\mathbf{x}_i/\mathbf{z}_i] EL[\mathbf{x}_i/\mathbf{z}_i]' \right] + E[\mathbf{e}_i \mathbf{e}_i'] = \Gamma E[\mathbf{z}_i \mathbf{z}_i'] \Gamma' + E[\mathbf{e}_i \mathbf{e}_i']$$

Avec :

$$\mathbf{w}(\mathbf{z}_i) \equiv EL[\mathbf{x}_i/\mathbf{z}_i] = \Gamma \mathbf{z}_i = E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} \mathbf{z}_i,$$

la condition d'orthogonalité estimante « modifiée » de  $\mathbf{a}_0$  :

$$E[\mathbf{w}(\mathbf{z}_i)(y_i - \mathbf{a}_0' \mathbf{x}_i)] = \mathbf{0}_{K \times 1}$$

s'écrit :

$$E\left[E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} \mathbf{z}_i (y_i - \mathbf{a}_0' \mathbf{x}_i)\right] = \mathbf{0}_{K \times 1},$$

équation qui montre que  $\mathbf{z}_i$  identifie  $\mathbf{a}_0$  parce qu'elle n'influence  $y_i$  que *via* son effet sur  $\mathbf{x}_i$  ou encore :

$$E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i (y_i - \mathbf{a}_0' \mathbf{x}_i)] = \mathbf{0}_{K \times 1}$$

équation plus facile à manipuler par la suite.

On a alors :

$$\mathbf{a}_0 \text{ solution en } \mathbf{a} \text{ de } E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i (y_i - \mathbf{a}' \mathbf{x}_i)] = \mathbf{0}_{K \times 1},$$

En supposant que  $\mathbf{a}_0$  est l'unique solution de système de  $K$  équations à  $K$  inconnues, l'utilisation du principe d'analogie donne :

$$\hat{\mathbf{a}}_N^{MM} \text{ solution en } \mathbf{a} \text{ de } N^{-1} \sum_{i=1}^N \hat{\Gamma}_N \mathbf{z}_i (y_i - \mathbf{x}_i' \mathbf{a}) = \mathbf{0}_{K \times 1},$$

ou, en remplaçant  $\hat{\Gamma}_N$  par  $\left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1}$  :

$$\hat{\mathbf{a}}_N^{MM} \text{ solution en } \mathbf{a} \text{ de}$$

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

On a alors :

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

ou encore :

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

On obtient ainsi :

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

ou, en développant :

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

L'estimateur  $\hat{\mathbf{a}}_N^{MM}$  défini ci-dessus est l'*estimateur des 2MC* :

$$\hat{\mathbf{a}}_N^{2MC} = \left[ N^{-1} \sum_{i=1}^N \hat{\mathbf{\Gamma}}_N \mathbf{z}_i \mathbf{x}'_i \right]^{-1} N^{-1} \sum_{i=1}^N \hat{\mathbf{\Gamma}}_N \mathbf{z}_i y_i$$

avec :

$$\hat{\mathbf{\Gamma}}_N \equiv \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1}$$

ou :

$$\begin{aligned} \hat{\mathbf{a}}_N^{2MC} = & \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}'_i \right] \right\}^{-1} \\ & \times \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i y_i \end{aligned}$$

Cet estimateur s'écrit sous *trois formes équivalentes* :

$$\hat{\mathbf{a}}_N^{2MC} = \left[ N^{-1} \sum_{i=1}^N \hat{\mathbf{\Gamma}}_N \mathbf{z}_i \mathbf{x}'_i \right]^{-1} N^{-1} \sum_{i=1}^N \hat{\mathbf{\Gamma}}_N \mathbf{z}_i y_i$$

avec :

$$\hat{\mathbf{\Gamma}}_N \equiv N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1}$$

ou :

$$\begin{aligned} \hat{\mathbf{a}}_N^{2MC} = & \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}'_i \right] \right\}^{-1} \\ & \times \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i y_i \end{aligned}$$

ou :

$$\hat{\mathbf{a}}_N^{2MC} \equiv \left\{ \mathbf{X}' \mathbf{Z} [\mathbf{Z}' \mathbf{Z}]^{-1} \mathbf{Z}' \mathbf{X} \right\}^{-1} \mathbf{X}' \mathbf{Z} [\mathbf{Z}' \mathbf{Z}]^{-1} \mathbf{Z}' \mathbf{y}$$

Les notations  $\mathbf{X}$  et  $\mathbf{y}$  ayant déjà été introduites, la matrice  $\mathbf{Z}$ , similaire à  $\mathbf{Z}$  est détaillée plus bas.



- L'estimateur  $\hat{\mathbf{a}}_N^{2MC}$  a la structure d'un estimateur des VI :

$$\hat{\mathbf{a}}_N^{2MC} = \left[ N^{-1} \sum_{i=1}^N \hat{\mathbf{\Gamma}}_N \mathbf{z}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \hat{\mathbf{\Gamma}}_N \mathbf{z}_i y_i$$

avec un vecteur d'instruments « estimés »,  $\hat{\mathbf{\Gamma}}_N \mathbf{z}_i$ , qui est de fait un estimateur (convergent) de  $\mathbf{w}(\mathbf{z}_i) \equiv EL[\mathbf{x}_i / \mathbf{z}_i]$

- *On peut également montrer* ( $\hat{\mathbf{\Gamma}}_N$  est une matrice de projection, et donc idempotente) que l'estimateur  $\hat{\mathbf{a}}_N^{2MC}$  a la structure d'un estimateur des MCO :

$$\hat{\mathbf{a}}_N^{2MC} = \left[ N^{-1} \sum_{i=1}^N (\hat{\mathbf{\Gamma}}_N \mathbf{z}_i)(\hat{\mathbf{\Gamma}}_N \mathbf{z}_i)' \right]^{-1} N^{-1} \sum_{i=1}^N (\hat{\mathbf{\Gamma}}_N \mathbf{z}_i) y_i$$

avec un vecteur de variables explicatives « estimées » (régresseurs estimés),  $\hat{\mathbf{w}}_N(\mathbf{z}_i) \equiv \hat{\mathbf{\Gamma}}_N \mathbf{z}_i$ , l'estimateur (convergent) de  $\mathbf{w}(\mathbf{z}_i) \equiv EL[\mathbf{x}_i / \mathbf{z}_i]$ .

- Il tient son nom « Doubles Moindres Carrés » de cette propriété. Il peut en fait être calculé en deux étapes de MCO, technique dite « des **MCO successifs** » :

1. Le calcul de  $\hat{\Gamma}_N$  est en fait un calcul d'estimateurs des MCO

$$EL[\mathbf{x}_i/\mathbf{z}_i] \equiv \Gamma \mathbf{z}_i = \begin{bmatrix} \mathbf{x}_i^x \\ \gamma'_{M+1} \mathbf{z}_i \\ \vdots \\ \gamma'_K \mathbf{z}_i \end{bmatrix} \text{ et } \hat{\Gamma}_N = \begin{bmatrix} [\mathbf{I}_M & \mathbf{0}] \\ \hat{\gamma}_{M+1,N}^{MCO'} \\ \hat{\gamma}_{K,N}^{MCO'} \end{bmatrix}$$

avec :

$$\hat{\gamma}_{k,N}^{MCO} \equiv \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i x_{k,i} .$$

$$2. \hat{\mathbf{a}}_N^{2MC} = \left[ N^{-1} \sum_{i=1}^N (\hat{\Gamma}_N \mathbf{z}_i)(\hat{\Gamma}_N \mathbf{z}_i)' \right]^{-1} N^{-1} \sum_{i=1}^N (\hat{\Gamma}_N \mathbf{z}_i) y_i .$$

- ***Cette propriété été importante*** lorsque les moyens de calculs étaient limités. ***Elle n'a plus beaucoup d'intérêt maintenant.***
- Il est conseillé d'utiliser cette propriété avec précaution (et je changerais bien le nom de cet estimateur) car elle peut être « dangereuse ».
  - (i) Par exemple, un estimateur des 2MC non linéaires a été défini par extension au cas linéaire, mais il ne possède pas cette propriété des « MCO successifs » !
  - (ii) De même, les calculs par estimations successives sont à employer avec précaution, surtout lorsqu'on « estime » des variables explicatives. Cela sera discuté plus en détail dans le cadre de la présentation du « test de la régression augmentée ».

La technique des « MCO successifs » repose (en seconde étape) sur l'utilisation de « ***régresseurs estimés*** » (*i.e.* de variables explicatives estimées), ces derniers devant être utilisés avec précaution.

**Rmq.** L'utilisation d'« ***instruments estimés*** » ne pose pas de problème.

En utilisant la matrice dite « matrice d'instruments »:

$$\mathbf{Z} \equiv \begin{bmatrix} \mathbf{z}'_1 \\ \mathbf{z}'_2 \\ \vdots \\ \mathbf{z}'_N \end{bmatrix}_{N \times L} = \begin{bmatrix} z_{1,1} & z_{2,1} & \cdots & z_{L,1} \\ z_{1,2} & z_{2,2} & \cdots & z_{L,2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1,N} & z_{2,N} & \cdots & z_{L,N} \end{bmatrix}$$

on peut écrire les estimateurs  $\hat{\mathbf{a}}_N^{VI}$  et  $\hat{\mathbf{a}}_N^{2MC}$  sous forme compacte :

$$\hat{\mathbf{a}}_N^{VI} \equiv [\mathbf{Z}'\mathbf{X}]^{-1} \mathbf{Z}'\mathbf{y}$$

et :

$$\hat{\mathbf{a}}_N^{2MC} \equiv \left\{ \mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1} \mathbf{Z}'\mathbf{X} \right\}^{-1} \mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1} \mathbf{Z}'\mathbf{y}$$

tout comme on a celle de l'estimateur  $\hat{\mathbf{a}}_N^{MCO}$  :

$$\hat{\mathbf{a}}_N^{MCO} \equiv [\mathbf{X}'\mathbf{X}]^{-1} \mathbf{X}'\mathbf{y}.$$

Cette écriture compacte permet de montrer que l'estimateur des VI est un cas particulier d'estimateur des 2MC. Si  $K = L$  alors les matrices  $\mathbf{X}'\mathbf{Z}$ ,  $\mathbf{Z}'\mathbf{Z}$  et  $\mathbf{Z}'\mathbf{X}$  sont carrées et de même dimension :  $K \times K$ .

En utilisant les propriétés des inverses de produits de matrices inversibles,  $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ . On obtient :

$$\begin{aligned}\hat{\mathbf{a}}_N^{2MC} &\equiv \left\{ \mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{X} \right\}^{-1} \mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y} \\ &= [\mathbf{Z}'\mathbf{X}]^{-1} \mathbf{Z}'\mathbf{Z}[\mathbf{X}'\mathbf{Z}]^{-1} \mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1} \mathbf{Z}'\mathbf{y} \\ &= [\mathbf{Z}'\mathbf{X}]^{-1} \mathbf{Z}'\mathbf{y} = \hat{\mathbf{a}}_N^{VI}\end{aligned}$$

De même, l'estimateur des MCO est un cas particulier d'estimateur des VI, celui où les **variables explicatives peuvent être utilisées comme VI**. **On a alors  $\mathbf{Z} = \mathbf{X}$  et :**

$$\hat{\mathbf{a}}_N^{VI} = [\mathbf{Z}'\mathbf{X}]^{-1} \mathbf{Z}'\mathbf{y} = [\mathbf{X}'\mathbf{X}]^{-1} \mathbf{X}'\mathbf{y} = \hat{\mathbf{a}}_N^{MCO}.$$

***En économétrie : l'estimateur des 2MC est une référence essentielle.***

### Propriété 33.

#### *Convergence de $\hat{\mathbf{a}}_N^{2MC}$ dans un modèle à VI*

Soit  $\{(y_i, \mathbf{x}_i, \mathbf{z}_i); i = 1, 2, \dots, N\}$  un échantillon de variables aléatoires telles que :

$$y_i = \mathbf{a}_0' \mathbf{x}_i + u_i \quad \text{avec} \quad E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0.$$

L'estimateur des 2MC de  $\mathbf{a}_0$

$$\begin{aligned} \hat{\mathbf{a}}_N^{2MC} = & \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right] \right\}^{-1} \\ & \times \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i y_i \end{aligned}$$

(i) existe avec une probabilité approchant 1

et :

(ii) est convergent, i.e. :  $\hat{\mathbf{a}}_N^{2MC} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{a}_0$ .

Cette propriété est facile à démontrer en utilisant les techniques habituelles à partir de l'équation :

$$\hat{\mathbf{a}}_N^{2MC} = \mathbf{a}_0 + \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}'_i \right] \right\}^{-1} \\ \times \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}'_i \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}'_i \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i$$

obtenue en remplaçant  $y_i$  par  $\mathbf{a}'_0 \mathbf{x}_i + u_i$  dans l'expression de  $\hat{\mathbf{a}}_N^{2MC}$ . Les résultats cruciaux pour la convergence de  $\hat{\mathbf{a}}_N^{2MC}$  sont comme toujours :

$$N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \xrightarrow[N \rightarrow +\infty]{p} E[\mathbf{z}_i u_i] \text{ et } E[\mathbf{z}_i u_i] = \mathbf{0}$$

*i.e.* résultent de l'application de la LGN (et de ses variantes) et des conditions d'exogénéité, du vecteur de VI  $\mathbf{z}_i$  ici.

On démontre également aisément que  $\hat{\mathbf{a}}_N^{2MC}$  est asymptotiquement normal dans le cadre d'un modèle à VI.

Cependant, les expressions de la variance asymptotique de  $\hat{\mathbf{a}}_N^{2MC}$  est particulièrement « affreuse » dans le cas général, *i.e.* à termes d'erreur hétéroscédastiques.

On se restreint donc ici au cas d'un modèle à VI à termes d'erreur homoscédastiques, *i.e.* on se place dans le cas où :

$$V[u_i/\mathbf{z}_i] = E[u_i^2/\mathbf{z}_i] = E[u_i^2] = \sigma_0^2.$$

Une seconde raison de ce choix, la plus importante, est que  $\hat{\mathbf{a}}_N^{2MC}$  *est asymptotiquement efficace dans le cas « homoscédastique »*, *i.e.* il n'existe pas de meilleur estimateur de la MM fondé sur la condition estimante

$E\left[E[\mathbf{x}_i\mathbf{z}_i']E[\mathbf{z}_i\mathbf{z}_i']^{-1}\mathbf{z}_i(y_i - \mathbf{a}_0'\mathbf{x}_i)\right] = \mathbf{0}_{K \times 1}$ . Ce n'est pas le cas si les  $u_i$  sont hétéroscédastiques.



### Propriété 34.

*Normalité asymptotique de  $\hat{\mathbf{a}}_N^{2MC}$  dans un*

*modèle à VI à termes d'erreur homoscédastiques*

Soit  $\{(y_i, \mathbf{x}_i, \mathbf{z}_i); i = 1, 2, \dots, N\}$  un échantillon de variables aléatoires telles que :

$$y_i = \mathbf{a}_0' \mathbf{x}_i + u_i \text{ avec } E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0 \text{ et } E[u_i^2 / \mathbf{z}_i] = \sigma_0^2.$$

L'estimateur des 2MC de  $\mathbf{a}_0$  :

$$\begin{aligned} \hat{\mathbf{a}}_N^{2MC} = & \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right] \right\}^{-1} \\ & \times \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i y_i \end{aligned}$$

vérifie :

$$\sqrt{N} (\hat{\mathbf{a}}_N^{2MC} - \mathbf{a}_0) \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N}(\mathbf{0}, \Sigma_0)$$

avec :

$$\Sigma_0 = \sigma_0^2 \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1}.$$

Cette propriété est facile à démontrer en utilisant les techniques habituelles à partir de l'équation :

$$\begin{aligned}\hat{\mathbf{a}}_N^{2MC} = \mathbf{a}_0 + & \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right] \right\}^{-1} \\ & \times \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i\end{aligned}$$

qui donne :

$$\begin{aligned}\sqrt{N}(\hat{\mathbf{a}}_N^{2MC} - \mathbf{a}_0) = & \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right] \right\}^{-1} \\ & \times \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \sqrt{N} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \right]\end{aligned}$$

Les résultats cruciaux pour la normalité asymptotique de  $\hat{\mathbf{a}}_N^{2MC}$  sont comme toujours :

$$\sqrt{N} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i u_i \right] \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N} \left( E[\mathbf{z}_i u_i], E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] \right) \text{ et } E[\mathbf{z}_i u_i] = \mathbf{0}$$

*i.e.* résultent de l'application du TCL (et de la LGN et de ses variantes) et des conditions d'exogénéité, du vecteur de VI  $\mathbf{z}_i$  ici. Avec l'homoscédasticité des  $u_i$  on a également :

$$E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] = \sigma_0^2 E[\mathbf{z}_i \mathbf{z}_i']$$

On obtient ici :

$$\begin{aligned} & \sqrt{N} (\hat{\mathbf{a}}_N^{2MC} - \mathbf{a}_0) \\ & \xrightarrow[N \rightarrow +\infty]{L} \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1} \\ & \quad \times E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} \times \mathcal{N}(\mathbf{0}, \sigma_0^2 E[\mathbf{z}_i \mathbf{z}_i']) \end{aligned}$$

La variance de la loi limite de  $\sqrt{N}(\hat{\mathbf{a}}_N^{2MC} - \mathbf{a}_0)$  est donnée par :

$$\begin{aligned} & \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1} E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} \\ & \times \left( \sigma_0^2 E[\mathbf{z}_i \mathbf{z}_i'] \right) \times E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1} \end{aligned}$$

et cette expression se simplifie en :

$$\sigma_0^2 \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1} = \Sigma_0.$$

On utilise ici les propriétés  $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$  et  $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ .

Ce type de simplification n'existe pas si les  $u_i$  sont potentiellement hétéroscédastiques.

Pour information, la variance de la loi limite de  $\sqrt{N}(\hat{\mathbf{a}}_N^{2MC} - \mathbf{a}_0)$  est donnée par :

$$\begin{aligned} & \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1} \\ & \times E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[u_i^2 \mathbf{z}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \\ & \times \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1} \end{aligned}$$

dans le cas où les  $u_i$  sont potentiellement hétéroscédastiques.

### Propriété 35.

*Estimateurs de la variance asymptotique de  $\hat{\mathbf{a}}_N^{2MC}$  dans un modèle à VI à termes d'erreur homoscedastiques*

La variance asymptotique,  $\Sigma_0$ , de l'estimateur des 2MC de  $\mathbf{a}_0$ ,  $\hat{\mathbf{a}}_N^{2MC}$ , peut être estimée par :

$$\hat{\Sigma}_N \equiv \hat{\sigma}_N^2 \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right] \right\}^{-1}$$

avec :

$$\hat{\sigma}_N^2 \equiv N^{-1} \sum_{i=1}^N \hat{u}_{i,N}^2 \quad \text{et} \quad \hat{u}_{i,N} \equiv y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{2MC}.$$

dans le cas d'un modèle à VI à termes d'erreur homoscedastiques.

L'estimateur :

$$\hat{\Sigma}_N \equiv \hat{\sigma}_N^2 \left\{ \left[ N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{z}_i' \right] \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \left[ N^{-1} \sum_{i=1}^N \mathbf{z}_i \mathbf{x}_i' \right] \right\}^{-1}$$

avec :

$$\hat{\sigma}_N^2 \equiv N^{-1} \sum_{i=1}^N (y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{2MC})^2$$

de  $\Sigma_0$ , est simplement la contre-partie empirique de :

$$\Sigma_0 = \sigma_0^2 \left\{ E[\mathbf{x}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{x}_i'] \right\}^{-1}$$

en notant que :

$$\sigma_0^2 = E[(y_i - \mathbf{x}_i' \mathbf{a}_0)^2]$$

On montre sa convergence par la LGN (est ses variance) et par la propriété de convergence de  $\hat{\mathbf{a}}_N^{2MC}$ , i.e.  $\hat{\mathbf{a}}_N^{2MC} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{a}_0$ .

## Remarques importantes pour la « pratique des VI »

- Le nom de modèle à VI est peu utilisé. Je l'utilise parce qu'il suggère que *les VI font partie de la spécification du modèle économétrique*, ce qui est le cas. *Le choix des VI ( $z_i$ ) est aussi important que celui des variables explicatives ( $x_i$ ) du modèle.*
- La section suivante donne quelques « clés » pour le choix des VI et vous éviter certaines pièges. Le choix des VI doit résulter d'une *analyse fine du PGD*. En particulier, l'exogénéité des VI est quasiment impossible à tester dans l'absolu.
- *Le choix des VI est délicat*, il vaut mieux s'inspirer de la littérature. Dans 99% des cas vous aurez à traiter des cas déjà analysés, ou des cas analogues à des cas analysés.  
Toutefois il convient de noter que les VI d'Angrist sont exceptionnelles de par leur simplicité et leur pertinence (les « mauvaises langues » disent qu'il détermine les problèmes en fonction des VI qu'il a).



- ***L'estimateur des VI « n'existe pas » dans les logiciels d'économétrie*** (et il a tendance à disparaître des manuels d'économétrie). Ceci provient de ce que ***l'estimateur des VI est un cas particulier de l'estimateur des 2MC*** dont le calcul est programmé dans tous les logiciels d'économétrie.
- ***L'efficacité (as.) de l'estimateur des 2MC s'accroît en théorie avec le nombre de VI utilisées, de manière « mécanique »***. Il est donc tentant de chercher à utiliser de nombreuses VI. Il convient néanmoins de rester prudent en pratique :
  - Plus on utilise de VI, plus on risque d'en utiliser de « mauvaises », *i.e.* de variables non exogènes.
  - Utiliser de trop nombreuses conditions estimantes peut finir par biaiser l'estimateur des 2MC lorsque  $N$  n'est pas suffisamment grand.
  - Des VI ayant une faible corrélation avec les variables qu'elles sont censées instrumenter peuvent être à l'origine de biais importants lorsqu'elles sont « légèrement endogènes ». C'est ce qu'on appelle le problème dit des « ***instruments faibles*** ».

## Biais de l'estimateur des 2MC à distance finie et dimension de $\tilde{\mathbf{z}}_i^e$

L'estimateur des 2MC est convergent :

$$\hat{\mathbf{a}}_N^{2MC} \xrightarrow[N \rightarrow +\infty]{p} \mathbf{a}_0$$

mais biaisé à distance finie :

$$E[\hat{\mathbf{a}}_N^{2MC} / \mathbf{Z}] \neq \mathbf{a}_0 \text{ (en général).}$$

- Schématiquement, lorsque  $L = \dim \mathbf{z}_i$  croît, le ratio  $N/L$  diminue et il devient de plus en plus difficile de justifier l'utilisation des propriétés asymptotiques de  $\hat{\mathbf{a}}_N^{2MC}$  pour approximer ses propriétés « réelles », *i.e.* avec  $N$  grand mais pas infini.
- De fait, c'est surtout le nombre de VI « externes », *i.e.*  $\dim \tilde{\mathbf{z}}_i^e$ , qui importe. Il est préférable de se contenter de quelques VI « dont on est sûr » par variable explicative endogène que de tenter d'en utiliser un grand nombre, surtout si le ratio  $N/K$  n'est pas très grand.

## Le problème des instruments faibles

Dans le modèle  $y_i = \alpha_0 + b_0 \tilde{x}_i + u_i$  avec  $E[u_i] \equiv 0$  avec  $\tilde{z}_i$  pour VI de  $\tilde{x}_i$  on a :

$$\hat{b}_N^{2MC} = \hat{b}_N^{VI} \xrightarrow[N \rightarrow +\infty]{P} b_0 + Cov[\tilde{z}_i; \tilde{x}_i]^{-1} Cov[\tilde{z}_i; u_i].$$

La variable  $\tilde{z}_i$  est un *instrument faible* de  $\tilde{x}_i$  si  $Cov[\tilde{z}_i; \tilde{x}_i] \equiv \varepsilon \simeq 0$ .

- L'utilisation d'instruments faibles ne pose pas de problème si  $Cov[\tilde{z}_i; u_i] = 0$ , ce qui ne peut réellement être assuré qu'« en théorie ».
- Le problème des instruments faibles est que si on a  $Cov[\tilde{z}_i; u_i] = \varsigma$  avec  $\varsigma \simeq 0$  mais  $\varsigma \neq 0$ , alors  $Cov[\tilde{z}_i; \tilde{x}_i]^{-1} Cov[\tilde{z}_i; u_i] = \varsigma / \varepsilon$  peut être grand en valeur absolue, ce qui implique un biais asymptotique significatif de  $\hat{b}_N^{2MC}$ .
- Si  $\tilde{z}_i$  est un instrument « fort » de  $\tilde{x}_i$ , i.e. si  $|Cov[\tilde{z}_i; \tilde{x}_i]|$  est grand alors, le biais as. de  $\hat{b}_N^{2MC}$  est nécessairement « petit » et peut être « négligé ».

# Outline

1. Modèles à variables explicatives endogènes
2. La notion de variable instrumentale et l'estimateur des VI
3. L'estimateur des 2MC
- 4. Choix des VI et identification**
5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée
6. Les variables de contrôle de l'hétérogénéité

## **4. Choix des VI et identification**

## 4. Choix des VI et identification

Que ce soit pour l'estimateur des VI ou des 2MC on a utilisé l'exogénéité des VI, *i.e.*  $E[u_i/\mathbf{z}_i] = 0$ , pour définir la condition d'orthogonalité :

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

comme base pour la construction d'un estimateur de  $\mathbf{a}_0$ . L'exogénéité des  $\mathbf{z}_i$  est une condition nécessaire pour l'identification de  $\mathbf{a}_0$ .

Etant entendu que  $E[\mathbf{z}_i(y_i - \mathbf{x}_i'\mathbf{a}_0)] = \mathbf{0}$  est valide, il est également nécessaire que cette équation caractérise  $\mathbf{a}_0$  de manière unique, *i.e.* que :

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

L'objectif de cette section est double, il s'agit :

- d'une part de montrer que :

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

*si et seulement si*  $\text{rang}E[\mathbf{z}_i\mathbf{x}_i'] = K$  et :

- d'autre part de d'interpréter cette condition de rang,  $\text{rang}E[\mathbf{z}_i\mathbf{x}_i'] = K$ , sachant que cette condition a des *implications importantes pour le choix des VI*.

Ici on a :

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

et un résultat standard d'algèbre linéaire donne que :

l'équation  $E[\mathbf{z}_i\mathbf{x}_i']\mathbf{a} = E[\mathbf{z}_iy_i]$  a une solution unique en  $\mathbf{a}$  si et seulement si  $\text{rang}E[\mathbf{z}_i\mathbf{x}_i'] = \dim \mathbf{a} = K$ .

### Propriété 36.

#### *Identification de $\mathbf{a}_0$ dans un modèle linéaire à VI*

Soit  $\{(y_i, \mathbf{x}_i, \mathbf{z}_i); i = 1, 2, \dots, N\}$  un échantillon de variables aléatoires réelles telles que  $y_i = \mathbf{x}_i' \mathbf{a}_0 + u_i$  avec  $E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0$ . Le vecteur de paramètres  $\mathbf{a}_0$  est identifiable par la condition  $E[\mathbf{z}_i(y_i - \mathbf{x}_i' \mathbf{a}_0)] = \mathbf{0}$  si et seulement si :

$$\text{rang} E[\mathbf{z}_i \mathbf{x}_i'] = K.$$

On appelle cette condition d'identification, *la condition de rang sur les VI* ou sur le  $\text{rang} E[\mathbf{z}_i \mathbf{x}_i']$ .

L'interprétation de la condition de rang sur  $E[\mathbf{z}_i \mathbf{x}_i']$  n'est pas immédiate. Mais certaines conditions nécessaires pour  $\text{rang} E[\mathbf{z}_i \mathbf{x}_i'] = K$  sont immédiates.



### Propriété 37.

**Conditions nécessaires pour  $\text{rang}E[\mathbf{z}_i\mathbf{x}_i'] = K$**

Les conditions suivantes sont nécessaires pour la condition  $\text{rang}E[\mathbf{z}_i\mathbf{x}_i'] = K$  :

- (i)  $\dim \mathbf{z}_i \geq \dim \mathbf{x}_i$ , **condition d'ordre**
- (ii)  $\text{rang}E[\mathbf{x}_i\mathbf{x}_i'] = K$ , condition d'identification des effets des  $\mathbf{x}_i$
- (iii)  $\text{rang}E[\mathbf{z}_i\mathbf{z}_i'] \geq K$ , condition sur l'indépendance des éléments de  $\mathbf{z}_i$

**Interprétations.** (i) Il doit y avoir *au moins autant de VI que de variables explicatives* ou, de manière équivalente, *au moins autant de conditions de moment que de paramètres à estimer*.  
(ii) Les éléments de  $\mathbf{x}_i$  doivent être tels que  $\mathbf{a}_0$  serait identifié si  $y_i = \mathbf{x}_i'\mathbf{a}_0 + u_i$  était un modèle de régression.  
(iii) Au moins  $K = \dim \mathbf{a}_0 = \dim \mathbf{x}_i$  éléments de  $\mathbf{z}_i$  doivent être linéairement indépendants, *i.e.* non redondants des autres quant à leur apport d'information.

On utilise ici les partitions habituelles :

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i \end{bmatrix} = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \mathbf{x}_i^e \end{bmatrix} \left. \begin{array}{l} \} \text{ variables exogènes} \\ \} \text{ variables endogènes} \end{array} \right\} \begin{array}{l} (k = 1, \dots, M) \\ (k = M + 1, \dots, K) \end{array}$$

$$\mathbf{z}_i = \begin{bmatrix} 1 \\ \tilde{\mathbf{z}}_i \end{bmatrix} = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} \left. \begin{array}{l} \} \text{ variables exogènes} \\ \} \text{ VI "externes"} \end{array} \right\} \begin{array}{l} (k = 1, \dots, M) \\ (k = M + 1, \dots, L) \end{array}$$

Et on utilise les projections linéaires pour examiner le contenu « concret » de la condition  $\text{rang} E[\mathbf{z}_i \mathbf{x}_i'] = K$ .

On note ici  $\tilde{e}_{\ell,i}$  le résidu de la projection du  $\ell^{\text{ième}}$  élément de  $\tilde{\mathbf{x}}_i$  sur  $(1, \tilde{\mathbf{x}}_{-\ell,i})$ , *i.e.* la partie spécifique de  $\tilde{x}_{\ell,i}$  dans  $\tilde{\mathbf{x}}_i$  :

$$\tilde{e}_{\ell,i} \equiv \tilde{x}_{\ell,i} - EL[\tilde{x}_{\ell,i} / 1, \tilde{\mathbf{x}}_{-\ell,i}] \text{ pour } \ell = 1, \dots, K - 1.$$

On définit  $\tilde{\mathbf{e}}_i$  comme l'empilement des  $\tilde{e}_{\ell,i}$  pour  $\ell = 1, \dots, K-1$  :

$$\tilde{\mathbf{e}}_i \equiv \begin{bmatrix} \tilde{e}_{1,i} \\ \tilde{e}_{2,i} \\ \vdots \\ \tilde{e}_{K-1,i} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{e}}_i^x \\ \tilde{\mathbf{e}}_i^e \end{bmatrix} \left\{ \begin{array}{l} \text{parties spécifiques des exogènes } (\ell = 1, \dots, M-1) \\ \text{parties spécifiques des endogènes } (\ell = M, \dots, K-1) \end{array} \right\}.$$

$$\tilde{\mathbf{x}}_i = \begin{bmatrix} \tilde{\mathbf{x}}_i^x \\ \tilde{\mathbf{x}}_i^e \end{bmatrix} \left\{ \begin{array}{l} \text{variables exogènes } (\ell = 1, \dots, M-1) \\ \text{variables endogènes } (\ell = M, \dots, K-1) \end{array} \right\}$$

**Rmq.** Les «  $M-1$  » et «  $K-1$  » sont dus à ce qu'on ne calcule pas la partie spécifique de la variable constante ( $x_{1i} = 1$ ).

**Propriété 38. Conditions de rang sur  $E[\mathbf{z}_i \mathbf{x}_i']$ .**

Les trois conditions suivantes sont *équivalentes* :

$$(i) \text{ rang} E[\mathbf{z}_i \mathbf{x}_i'] = K$$

$$(ii) \text{ rangCov}[\tilde{\mathbf{z}}_i; \tilde{\mathbf{x}}_i] = K - 1$$

$$(iii) \text{ rang} V[\tilde{\mathbf{e}}_i^x] = M - 1 \text{ et } \text{rangCov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] = K - M$$

Il est difficile pousser plus avant l'analyse de la condition  $\text{rang} E[\mathbf{z}_i \mathbf{x}_i'] = K$ .

- La matrice  $V[\tilde{\mathbf{e}}_i^x]$  étant diagonale l'analyse de son rang est immédiate.
- Le problème est qu'il est difficile de déterminer des conditions facilement interprétables et d'examen aisé de la condition  $\text{rangCov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] = K - M$ , et ce même si les éléments de  $\tilde{\mathbf{e}}_i^e$  ne sont pas corrélés entre eux.

## Eléments de démonstration et d'interprétation

La condition (ii) est une formulation alternative de la condition (i).

La condition (iii) utilise :

$$\text{rangCov}[\tilde{\mathbf{z}}_i; \tilde{\mathbf{x}}_i] = \text{rangCov}[\tilde{\mathbf{z}}_i; \tilde{\mathbf{e}}_i]$$

et :

$$\text{Cov}[\tilde{\mathbf{z}}_i; \tilde{\mathbf{e}}_i] = \begin{bmatrix} V[\tilde{\mathbf{e}}_i^x] & \mathbf{0} \\ \text{Cov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^x] & \text{Cov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] \end{bmatrix}.$$

Ces équations indiquent  $\text{Cov}[\tilde{\mathbf{z}}_i; \tilde{\mathbf{x}}_i]$  est de plein rang colonne si et seulement si  $V[\tilde{\mathbf{e}}_i^x]$  et  $\text{Cov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e]$  le sont également.

Il est nécessaire que  $V[\tilde{\mathbf{e}}_i^e]$  soit inversible pour  $\text{rangCov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] = K - M$ .

Lorsque  $V[\tilde{\mathbf{x}}_i]$  est inversible, le rang de  $E[\mathbf{z}_i \mathbf{x}'_i]$  ne concerne que les relations entre  $\tilde{\mathbf{e}}_i^e$  et  $\tilde{\mathbf{z}}_i^e$ .

La matrice  $V[\tilde{\mathbf{z}}_i^e]$  est inversible si et seulement les éléments de  $\tilde{\mathbf{z}}_i^e$  ne sont pas linéairement redondants, *i.e.* s'ils n'apportent pas d'informations redondantes. Dans ce cas, elle est de rang  $L - M$  et il est nécessaire qu'elle soit au moins de rang  $K - M$  pour  $\text{rangCov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] = K - M$

Si  $V[\tilde{\mathbf{z}}_i^e]$  est inversible, alors  $\text{rangCov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] = K - M$  si et seulement si les éléments de :

$$EL[\tilde{\mathbf{e}}_i^e/1, \tilde{\mathbf{z}}_i^e] = \text{Cov}[\tilde{\mathbf{e}}_i^e; \tilde{\mathbf{z}}_i^e] V[\tilde{\mathbf{z}}_i^e]^{-1} (\tilde{\mathbf{z}}_i^e - E[\tilde{\mathbf{z}}_i^e])$$

sont linéairement indépendants.

Il existe une condition suffisante « simple » pour  $\text{rangCov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] = K - M$ .

En notant :

$$EL\left[\tilde{e}_{k,i}^e/1,\tilde{\mathbf{z}}_i^e\right]=\sum_{m=1}^{L-M}\gamma_{k,\ell}\left(\tilde{z}_{\ell,i}^e-E\left[\tilde{z}_{\ell,i}^e\right]\right),$$

on dira que  $\tilde{z}_{\ell,i}^e$  *est une VI spécifique de*  $\tilde{x}_{k,i}^e$  si et seulement si :

$$\gamma_{k,\ell} \neq 0 \text{ et } \gamma_{m,\ell} = 0 \text{ pour } m \in \{1, \dots, K-M\} \text{ et } m \neq k.$$

En désignant par :

$$\tilde{\zeta}_{\ell,i}^e \equiv \tilde{z}_{\ell,i}^e - EL\left[\tilde{z}_{\ell,i}^e/1,\tilde{\mathbf{z}}_{-\ell,i}^e\right]$$

la partie spécifique de  $\tilde{z}_{\ell,i}^e$  dans  $\tilde{\mathbf{z}}_i^e$ ,  $\tilde{\zeta}_{\ell,i}^e$  *est spécifique de*  $\tilde{x}_{k,i}^e$  si et seulement si :

$$Cov\left[\tilde{\zeta}_{\ell,i}^e;\tilde{e}_{k,i}^e\right] \neq 0 \text{ et } Cov\left[\tilde{\zeta}_{\ell,i}^e;\tilde{e}_{m,i}^e\right] = 0 \text{ pour } m \in \{1, \dots, K-M\} \text{ et } m \neq k.$$

Si chaque élément de  $\tilde{\mathbf{e}}_i^e$  a une VI externe, *i.e.* un élément de  $\tilde{\mathbf{z}}_i^e$ , qui lui est spécifique alors  $rangCov\left[\tilde{\mathbf{z}}_i^e;\tilde{\mathbf{e}}_i^e\right] = K-M$ . Cette condition suffisante de  $rangCov\left[\tilde{\mathbf{z}}_i^e;\tilde{\mathbf{e}}_i^e\right] = K-M$  peut être examinée empiriquement.

**Pour  $\text{rang}E[\mathbf{z}_i\mathbf{x}'_i] = K$  il faut :**

- (i) Que les variables explicatives, *i.e.* les éléments de  $\tilde{\mathbf{x}}_i$ , soient suffisamment variables et suffisamment indépendantes entre elles (Elles ont alors de « grosses » parties spécifiques).

C'est la ***condition de rang du modèle de régression***. Si le modèle linéaire utilisé est problématique, de très bonnes VI n'y changeront rien.

et :

- (ii) Que ***chaque variable explicative endogène***, *i.e.* que chaque élément de  $\tilde{\mathbf{x}}_i^e$ , soit « ***instrumentée*** » ***par au moins une VI « externe »***, *i.e.* par au moins un élément de  $\tilde{\mathbf{z}}_i^e$ . Ca ne sert à rien d'avoir beaucoup de VI si une variable explicative endogène est « abandonnée » en termes d'instrumentation.

L'examen de cette condition doit passer par les parties spécifiques des  $\tilde{\mathbf{x}}_i^e$ . Plus que les  $\tilde{\mathbf{x}}_i^e$ , ce sont les  $\tilde{\mathbf{e}}_i^e$  qui doivent être « instrumentées » par les  $\tilde{\mathbf{z}}_i^e$ .



**En pratique, pour ce qui concerne  $\text{rangCov}[\tilde{\mathbf{z}}_i^e; \tilde{\mathbf{e}}_i^e] = K - M$**

- (i) Lorsqu'on cherche des VI pour  $\tilde{\mathbf{x}}_i^e$  alors on cherche des **VI aussi spécifique que possible de chacun des éléments de  $\tilde{\mathbf{x}}_i^e$** . Cette pratique se réfère à la condition suffisante pour la condition de rang.
- (ii) La condition  $\text{rangE}[\mathbf{z}_i, \mathbf{x}_i'] = K$  étant difficile à examiner *ex ante*, on calcule ensuite directement l'estimateur des 2MC d'intérêt en utilisant le vecteur de VI  $\mathbf{z}_i$ .
- (iii) On **diagnostique un problème relatif à  $\text{rangE}[\mathbf{z}_i, \mathbf{x}_i']$**  si :
  - (a) Le calcul de l'estimateur est impossible parce que **certaines matrices ne peuvent être inversées**, ce qui est relativement rare.
  - (b) Les **écarts-types estimés** de l'estimateur sont « **énormes** » pour certains paramètres.  
Il s'agit des paramètres associés à des **var. explicatives peu variables**, à des **var. explicatives très liées** ou/et à des **var. explicatives endogènes « mal » instrumentées**.

**En résumé**, la propriété suivante montre que ce sont les projections des parties spécifiques des variables explicatives endogènes sur le vecteur des instruments qui identifient les paramètres du modèle associées à ces variables.

**Propriété 39. « Anatomie » de l'estimateur des 2MC**

Dans le modèle à VI considéré ici :

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i + u_i \text{ avec } E[u_i / \mathbf{z}_i] \equiv 0,$$

si  $\text{rangCov}[\tilde{\mathbf{z}}_i; \tilde{\mathbf{x}}_i] = K - 1$  alors :

$$b_{m,0} = V[\tilde{e}_{m,i}^x]^{-1} \text{Cov}[\tilde{e}_{m,i}^x; y_i] \text{ pour } m = 1, \dots, M - 1$$

et :

$$b_{\ell,0} = V[EL[\tilde{e}_{\ell,i}^e / 1, \tilde{\mathbf{z}}_i^e]]^{-1} \text{Cov}[EL[\tilde{e}_{\ell,i}^e / 1, \tilde{\mathbf{z}}_i^e]; y_i] \text{ pour } \ell = 1, \dots, K - M.$$

On a également :

$$b_{\ell,0} = \text{Cov}[EL[\tilde{e}_{\ell,i}^e / 1, \tilde{\mathbf{z}}_i^e]; \tilde{e}_{\ell,i}^e]^{-1} \text{Cov}[EL[\tilde{e}_{\ell,i}^e / 1, \tilde{\mathbf{z}}_i^e]; y_i] \text{ pour } \ell = 1, \dots, K - M.$$

# Outline

1. Modèles à variables explicatives endogènes
2. La notion de variable instrumentale et l'estimateur des VI
3. L'estimateur des 2MC
4. Choix des VI et identification
5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée
6. Les variables de contrôle de l'hétérogénéité

## **5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée**

## **5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée**

Dans les sections précédentes, on a utilisé les techniques de VI, *i.e.* l'estimateur des 2MC, pour contourner le problème des variables explicatives.

On a déterminé des VI pour les variables endogènes, défini des conditions d'orthogonalité estimantes et on en a déduit l'estimateur des 2MC (dont l'estimateur des VI n'est qu'un cas particulier).

Cette approche peut être utilisée, pour peu qu'on dispose de VI, pour tout type de problème d'endogénéité :

- simultanéité,
- omission de variables explicatives pertinentes

ou :

- erreurs de mesure sur les variables explicatives.

**Les objectifs sont ici de :**

1. Présenter l'*approche dite du « contrôle des problèmes d'endogénéité »*
2. Présenter une autre utilisation des VI : l'approche par les *fonctions de contrôle*
3. Présenter un test *d'exogénéité des variables explicatives*

et :

4. De fournir des bases pour présenter une autre approche que celle des VI pour gérer les problèmes d'endogénéité dus à l'omission de variables explicatives pertinentes : les variables de contrôle de l'hétérogénéité présentées dans la section suivante.

## 5.1. Approche par les fonctions de contrôle

### Intérêts :

1. Utilisable dans des contextes plus « complexes » que ceux traités ici (ratio de Mills, panel, ...)
2. Simple dans le cas de la régression
3. Introduction au test d'exogénéité dit test de la régression augmentée.

### Principe de fonctionnement des fonctions de contrôle de l'endogénéité

*Décomposer le terme d'erreur* en deux parties :

- (i) Une partie qui « *capte* » ou « *contrôle* » le *problème d'endogénéité* : la *fonction de contrôle*  
et
- (ii) Une *partie résiduelle* qui ne pose pas de problème d'endogénéité  
pour *transformer le modèle en un modèle de régression*.

On considère ici le modèle linéaire général à une variable explicative endogène :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + u_i \quad \text{avec} \quad E[u_i / \mathbf{x}_i^x] = E[u_i] \equiv 0.$$

On suppose qu'on dispose d'un vecteur de VI « externes » pour  $x_i^e$ ,  $\tilde{\mathbf{z}}_i^e$  :

$$E[u_i / \mathbf{x}_i^x, \tilde{\mathbf{z}}_i^e] = E[u_i] \equiv 0.$$

Le modèle considéré est donc un modèle à VI et comme d'habitude, on notera :

$$\mathbf{z}_i \equiv \begin{bmatrix} \mathbf{x}_i^x \\ \tilde{\mathbf{z}}_i^e \end{bmatrix} \quad \text{et} \quad \mathbf{a}_0 \equiv \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}$$

Le problème posé par l'estimation de  $\mathbf{a}_0$  est dû à :

$$E[u_i / \mathbf{z}_i, \tilde{x}_i^e] = E[u_i / \mathbf{x}_i^x, \tilde{\mathbf{z}}_i^e, \tilde{x}_i^e] \neq 0.$$



### Principe de la fonction de contrôle (de l'endogénéité d'une variable)

Si on connaît la forme de  $E[u_i/\mathbf{z}_i, \tilde{x}_i^e]$ , e.g. :

$$E[u_i/\mathbf{z}_i, \tilde{x}_i^e] = \rho \times \lambda(\mathbf{z}_i, \tilde{x}_i^e),$$

on peut écrire le modèle de  $y_i$  sous la forme :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + \rho \lambda(\mathbf{z}_i, \tilde{x}_i^e) + \varepsilon_i \text{ avec } E[\varepsilon_i/\mathbf{z}_i, \tilde{x}_i^e] = E[\varepsilon_i] \equiv 0,$$

i.e. sous la forme d'un **modèle de régression** (éventuellement non linéaire).

On utilise ici la **décomposition d'une variable** en la somme de son **espérance conditionnelle** et du **résidu associé** à cette espérance :

$$E[u_i/\mathbf{z}_i, \tilde{x}_i^e] = \rho \lambda(\mathbf{z}_i, \tilde{x}_i^e) \Leftrightarrow u_i = \rho \lambda(\mathbf{z}_i, \tilde{x}_i^e) + \varepsilon_i \text{ avec } E[\varepsilon_i/\mathbf{z}_i, \tilde{x}_i^e] = 0$$

où :

$$\varepsilon_i \equiv u_i - \rho \lambda(\mathbf{z}_i, \tilde{x}_i^e).$$

On utilise ici *deux projections linéaires* pour définir la fonction de contrôle :

(i) celle de  $\tilde{x}_i^e$  sur  $\mathbf{z}_i$  qui donne  $\tilde{x}_i^e = EL[\tilde{x}_i^e / \mathbf{z}_i] + e_i$  et  $E[\mathbf{z}_i e_i] = \mathbf{0}$

(ii) celle de  $u_i$  sur  $e_i$  qui donne  $u_i = EL[u_i / 1, e_i] + \varepsilon_i$  et  $E[e_i \varepsilon_i] = E[\varepsilon_i] = 0$ .

**Rmq.** Cette technique ne requiert aucune *hypothèse particulière* (en dehors de celles contenues dans la définition du modèle à VI) puisque *les projections sont des outils mathématiques* et sont donc « en libre accès ».

**Rmq.** Dans le cas de problèmes d'endogénéité plus « complexes » (que celui de l'endogénéité d'une variable explicative dans un modèle de forme linéaire), la définition de la forme de la fonction de contrôle est plus compliquée et repose généralement sur une hypothèse (plus ou moins) *ad hoc* concernant la forme de  $E[u_i / \mathbf{z}_i, x_i^e]$ .

En notant :  $EL[\tilde{x}_i^e/\mathbf{z}_i] = \gamma' \mathbf{z}_i$ , la **première projection** permet de décomposer  $\tilde{x}_i^e$  en une somme de deux termes :

$$\tilde{x}_i^e = \gamma' \mathbf{z}_i + e_i,$$

(i) la projection  $\gamma' \mathbf{z}_i$  qui *est exogène* par rapport à  $u_i$  car  $E[\mathbf{z}_i e_i] = \mathbf{0}$ .

et :

(ii) le résidu de projection  $e_i$  *qui contient la source de l'endogénéité de  $\tilde{x}_i^e$*  par rapport à  $u_i$  car  $\boxed{Cov[\tilde{x}_i^e; u_i] = E[e_i; u_i]}$  puisque  $E[\mathbf{z}_i e_i] = \mathbf{0}$

En notant :  $EL[u_i/1, e_i] = \rho e_i$ , La **seconde projection** permet de décomposer  $u_i$  en une somme de deux termes :

$$u_i = \rho e_i + \varepsilon_i$$

(i) la projection  $\rho e_i$  qui « *contrôle* » le *problème d'endogénéité de  $\tilde{x}_i^e$* ,  $e_i$ , par rapport à  $u_i$

et :

(ii) le résidu de projection  $\varepsilon_i$  dont on sait que  $E[e_i \varepsilon_i] = 0$  et  $E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}$ .

**Rmq.** On a :

(a)  $E[\mathbf{z}_i e_i] = \mathbf{0}$  par construction, *i.e.*  $e_i$  est un résidu de projection sur  $\mathbf{z}_i$ .

(b)  $EL[u_i/1, e_i] = \rho e_i$ , *i.e.* sans « constante », car  $E[u_i] = 0$  et  $E[e_i] = 0$ .

(c)  $E[e_i \varepsilon_i] = 0$  par construction, *i.e.*  $\varepsilon_i$  est un résidu de projection sur  $e_i$

(d)  $E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}$  car :

(i)  $E[\mathbf{z}_i u_i] = \mathbf{0}$  puisque  $\mathbf{z}_i$  est un vecteur de VI et

(ii)  $E[\mathbf{z}_i u_i] = E[\mathbf{z}_i (\rho e_i + \varepsilon_i)] = E[\mathbf{z}_i \varepsilon_i]$  puisque  $E[\mathbf{z}_i e_i] = \mathbf{0}$ .

On peut alors montrer que :

$$\text{Cov}[\tilde{x}_i^e; \varepsilon_i] = E[\tilde{x}_i^e \varepsilon_i] = 0$$

car :

$$\text{Cov}[\tilde{x}_i^e; \varepsilon_i] = E[\tilde{x}_i^e \varepsilon_i] = E[(\gamma' \mathbf{z}_i + e_i) \varepsilon_i] = \gamma' E[\mathbf{z}_i \varepsilon_i] + E[e_i \varepsilon_i] = 0.$$

On obtient alors le modèle suivant :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + \rho \times (\tilde{x}_i^e - \gamma' \mathbf{z}_i) + \varepsilon_i,$$

avec

$$E[\varepsilon_i / \mathbf{z}_i] = E[\varepsilon_i] \equiv 0 \quad \text{et} \quad E[\tilde{x}_i^e \varepsilon_i] = 0$$

qui est un **modèle de régression** (non linéaire) grâce à la décomposition :

$$u_i = \rho e_i + \varepsilon_i = \rho(\tilde{x}_i^e - \gamma' \mathbf{z}_i) + \varepsilon_i,$$

*i.e.* avec  $\lambda(\mathbf{z}_i, \tilde{x}_i^e) \equiv \tilde{x}_i^e - \gamma' \mathbf{z}_i$ .


Ce modèle est dit « **de la régression augmentée** » (par  $\rho e_i = \rho(\tilde{x}_i^e - \gamma' \mathbf{z}_i)$ ).

**Rmq.** Dans un modèle linéaire la condition  $E[\tilde{x}_i^e \varepsilon_i] = 0$  est suffisante pour l'exogénéité de  $\tilde{x}_i^e$  (c'est pour ça que les projections suffisent ici).


**Rmq.** Le terme  $\rho e_i = \rho(\tilde{x}_i^e - \gamma' \mathbf{z}_i)$  est **extrait de**  $u_i$ , il n'augmente pas vraiment le modèle.

## En résumé : *Principe de la fonction de contrôle*

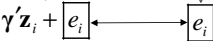
Les termes  $\tilde{x}_i^e$  et  $u_i$  sont liés dans le modèle d'intérêt

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \boxed{\tilde{x}_i^e} + \boxed{u_i}$$


On « *isole* » **cette liaison** par les décompositions  $\tilde{x}_i^e = \gamma' \mathbf{z}_i + e_i$  et  $u_i = \rho e_i + \varepsilon_i$

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 (\gamma' \mathbf{z}_i + \boxed{e_i}) + (\rho \boxed{e_i} + \varepsilon_i)$$


Ceci garantit que  $\tilde{x}_i^e$  **est exogène vis-à-vis de**  $\varepsilon_i$  dans le **modèle « augmenté »**

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \underbrace{\tilde{x}_i^e}_{\gamma' \mathbf{z}_i + \boxed{e_i}} + \overbrace{\rho \times (\tilde{x}_i^e - \gamma' \mathbf{z}_i)}^{\text{Fonction de contrôle}} + \varepsilon_i$$


Les paramètres de la *régression augmentée* :

$$y_i = \alpha_0' \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + \rho e_i + \varepsilon_i \text{ avec } e_i \equiv \tilde{x}_i^e - \gamma' \mathbf{z}_i$$

peuvent être estimés en *deux étapes (de régression)*.

**Procédure. Estimation des paramètres de la régression augmentée**

1. On estime les paramètres de la projection,  $\gamma$ , par les MCO ce qui permet de calculer les  $\hat{e}_{i,N} \equiv \tilde{x}_i^e - \mathbf{z}_i' \hat{\gamma}_N^{MCO}$  pour  $i = 1, \dots, N$ .
2. On estime les paramètres  $\alpha_0$ ,  $\beta_0$  et  $\rho$  à partir du modèle approché :

$$y_i = \alpha_0' \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + \rho \hat{e}_{i,N} + \hat{\varepsilon}_{i,N}$$

où

$$\hat{\varepsilon}_{i,N} \equiv \varepsilon_i + \rho(e_{i,N} - \hat{e}_{i,N}) = \varepsilon_i + \rho \mathbf{z}_i' (\gamma_N^{MCO} - \gamma)$$

par les MCO.

- On utilise le fait que l'erreur de mesure sur  $e_i = \tilde{x}_i^e - \gamma'z_i$  introduite par le remplacement de  $\gamma$  par  $\gamma_N^{MCO}$  dans l'**étape 2** est négligeable lorsque  $N$  est suffisamment grand.

En effet,  $p \lim_{N \rightarrow +\infty} (\gamma_N^{MCO} - \gamma) = \mathbf{0}$  implique que :

$$p \lim_{N \rightarrow +\infty} (\hat{e}_{i,N} - e_{i,N}) = p \lim_{N \rightarrow +\infty} (\hat{e}_{i,N} - \varepsilon_i) = 0.$$

- L'**erreur de mesure** sur  $e_i = \tilde{x}_i^e - \gamma'z_i$  due au remplacement de  $\gamma$  par  $\gamma_N^{MCO}$  :

- (i) n'affecte pas les **propriétés de convergence** de l'estimateur des MCO de l'**étape 2** (elle converge en probabilité vers  $\mathbf{0}$ )

**mais :**

- (ii) introduit de l'aléa dans la régression augmentée qui **modifie la distribution as. de l'estimateur** de l'**étape 2**. Ce provient de ce que le terme  $\sqrt{N}(\gamma_N^{MCO} - \gamma)$  ne converge pas vers  $\mathbf{0}$ .

En fait, les  $\hat{e}_{i,N}$  sont des « **régresseurs estimés** ».



- La dernière remarque implique que la technique par les fonctions de contrôle n'a pas beaucoup d'intérêt pour l'estimation des paramètres de modèles linéaires à VI. Elle est même un peu « dangereuse ».

*Néanmoins :*

- (i) On a montré ici qu'on pouvait « gérer » autrement les problèmes d'endogénéité que par la construction de conditions de moment estimantes avec des VI.
- (ii) Cette technique est employée pour « gérer » des problèmes d'endogénéité moins « standard » que celui considéré ici
- (iii) L'**étape 1** de la procédure d'estimation permet l'examen de la qualité de  $\tilde{\mathbf{z}}_i^e$  pour « instrumenter »  $x_i^e$  :

Les éléments de  $\gamma_N^{MCO}$  correspondant à  $\tilde{\mathbf{z}}_i^e$  doivent être statistiquement (très)  $\neq 0$ .

et :

- (iv) On va voir que le travail présenté ici est utile pour **tester l'exogénéité de  $\tilde{x}_i^e$** .

**Rmq.** On peut appliquer cette technique avec un vecteur de variables endogènes. Dans le modèle à VI :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta'_0 \tilde{\mathbf{x}}_i^e + u_i \text{ avec } E[u_i / \mathbf{z}_i] \equiv 0 \text{ et } \mathbf{z}_i \equiv (\mathbf{x}_i^x, \tilde{\mathbf{z}}_i^e),$$

il suffit d'utiliser :

$$\tilde{\mathbf{x}}_i^e = \Gamma \mathbf{z}_i + \mathbf{e}_i \text{ avec } E[\mathbf{e}_i \mathbf{z}_i'] = \mathbf{0}$$

où  $\Gamma \mathbf{z}_i \equiv EL[\tilde{\mathbf{x}}_i^e / \mathbf{z}_i]$  et  $\mathbf{e}_i \equiv \tilde{\mathbf{x}}_i^e - EL[\tilde{\mathbf{x}}_i^e / \mathbf{z}_i]$  et :

$$u_i = \rho' \mathbf{e}_i + \varepsilon_i \text{ avec } E[\mathbf{e}_i \varepsilon_i] = 0$$

où  $\rho' \mathbf{e}_i \equiv EL[u_i / \mathbf{e}_i]$  et  $\varepsilon_i \equiv u_i - EL[u_i / \mathbf{e}_i]$ .

Il suffit en fait de remarquer que  $\tilde{\mathbf{x}}_i^e = \Gamma \mathbf{z}_i + \mathbf{e}_i$  est un « empilement » d'*équations de régression linéaires*, une pour chaque élément de  $\tilde{\mathbf{x}}_i^e$ .

## 5.2. Tester l'exogénéité des variables explicatives

L'endogénéité des variables explicatives est un *problème fréquent et sérieux* en économétrie.

On pourrait être tenté de n'utiliser que des techniques à VI pour s'assurer contre tout problème d'endogénéité potentiel.

On doit cependant se garder de cet *excès de prudence*, pour trois raisons :

- (i) Le choix des VI est délicat et des VI ne sont pas toujours disponibles.
- (ii) Si la variable explicative « instrumentée » est en réalité exogène, on perd (souvent *beaucoup en pratique*) en efficacité d'estimation à utiliser les 2MC plutôt que les MCO

et :

- (iii) Il est possible de tester l'exogénéité de variables explicatives, tout au moins dans une certaine mesure.

On considère ici le cas d'un modèle linéaire de forme générale :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \text{ avec } E[u_i] \equiv 0$$

pour lequel on dispose d'un vecteur d'instruments  $\mathbf{z}_i$  valides.

**Propriété 40. Propriétés comparées de  $\hat{\mathbf{a}}_N^{MCO}$  et  $\hat{\mathbf{a}}_N^{2MC}$  selon l'exo/endogénéité de  $\mathbf{x}_i$**

$$\text{Si } E[u_i/\mathbf{x}_i] \neq 0 \text{ alors } \begin{cases} p \lim_{N \rightarrow +\infty} \hat{\mathbf{a}}_N^{MCO} \neq \mathbf{a}_0 \\ \sqrt{N}(\hat{\mathbf{a}}_N^{2MC} - \mathbf{a}) \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N}(\mathbf{0}; \Sigma_0^{2MC}) \end{cases}$$

$$\text{Si } E[u_i/\mathbf{x}_i] = 0 \text{ alors } \begin{cases} \sqrt{N}(\hat{\mathbf{a}}_N^{MCO} - \mathbf{a}) \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N}(\mathbf{0}; \Sigma_0^{MCO}) \\ \sqrt{N}(\hat{\mathbf{a}}_N^{2MC} - \mathbf{a}) \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N}(\mathbf{0}; \Sigma_0^{2MC}) \end{cases} \text{ et } \Sigma_0^{2MC} \gg \Sigma_0^{MCO}$$

**Conc :**  $\hat{\mathbf{a}}_N^{2MC}$  toujours convergent, mais moins efficace que  $\hat{\mathbf{a}}_N^{MCO}$  si  $\mathbf{x}_i$  exogène.  
 $\hat{\mathbf{a}}_N^{MCO}$  convergent que si  $\mathbf{x}_i$  exogène.

- La *comparaison des propriétés de convergence* est détaillée ci-avant !

- $\hat{\mathbf{a}}_N^{MCO}$  n'est convergent que si  $\mathbf{x}_i$  exogène

- Pour ce qui concerne l'*efficacité comparée de  $\hat{\mathbf{a}}_N^{MCO}$  et  $\hat{\mathbf{a}}_N^{2MC}$*  lorsque  $\mathbf{x}_i$  est exogène, *on considère ici le cas où les  $u_i$  sont homoscédastiques*.

- On sait que :

$$\Sigma_0^{MCO} = \sigma_0^2 E[\mathbf{x}_i \mathbf{x}_i']^{-1}$$

et

$$\Sigma_0^{2MC} = \sigma_0^2 \left\{ E[\mathbf{x}_i \mathbf{z}_i']^{-1} E[\mathbf{z}_i \mathbf{z}_i'] E[\mathbf{z}_i \mathbf{x}_i']^{-1} \right\}^{-1}.$$

- On utilise la projection linéaire de  $\mathbf{x}_i$  sur  $\mathbf{z}_i$  :

$$E[\mathbf{x}_i / \mathbf{z}_i] \equiv \Gamma \mathbf{z}_i \quad \text{avec} \quad \mathbf{e}_i \equiv \mathbf{x}_i - \Gamma \mathbf{z}_i$$

On a alors :

$$\Sigma_0^{2MC} = \sigma_0^2 \{ \Gamma E[\mathbf{z}_i \mathbf{z}_i'] \Gamma' \}^{-1} \quad \text{et} \quad \Sigma_0^{MCO} = \sigma_0^2 \{ \Gamma E[\mathbf{z}_i \mathbf{z}_i'] \Gamma' + E[\mathbf{e}_i \mathbf{e}_i'] \}^{-1},$$

car les propriétés des projections linéaires :

$$\mathbf{x}_i = \Gamma \mathbf{z}_i + \mathbf{e}_i \quad \text{avec} \quad EL[\mathbf{x}_i / \mathbf{z}_i] \equiv \Gamma' \mathbf{z}_i \quad \text{et} \quad E[\mathbf{z}_i \mathbf{e}_i'] = \mathbf{0},$$

impliquent que :

$$E[\mathbf{z}_i \mathbf{x}_i'] = E[\mathbf{z}_i \mathbf{z}_i'] \Gamma' \quad \text{et} \quad E[\mathbf{x}_i \mathbf{x}_i'] = \Gamma E[\mathbf{z}_i \mathbf{z}_i'] \Gamma' + E[\mathbf{e}_i \mathbf{e}_i'].$$

Ceci prouve que  $\Sigma_0^{2MC} \gg \Sigma_0^{MCO}$  puisque  $E[\mathbf{e}_i \mathbf{e}_i']$  est semi-définie positive.

### Interprétation

En utilisant des techniques de VI pour l'estimation d'un modèle de régression, on « perd » une partie des variations de  $\mathbf{x}_i$ , celle décrite par  $E[\mathbf{e}_i \mathbf{e}_i']$ .

Dans le calcul de l'estimateur des 2MC de  $\mathbf{a}_0$  on « remplace » implicitement les  $\mathbf{x}_i$  par les  $E[\mathbf{x}_i / \mathbf{z}_i] \equiv \Gamma \mathbf{z}_i$  dans le calcul de l'estimateur des MCO de  $\mathbf{a}_0$ .

Les propriétés comparées de  $\hat{\mathbf{a}}_N^{MCO}$  et  $\hat{\mathbf{a}}_N^{2MC}$  fondent l'idée des tests dits « **Tests d'Hausman** » de l'exogénéité des variables explicatives d'un modèle économétrique.

***Base statistique des Tests d'Hausman***

Sous Hyp0 :  $E[u_i/\mathbf{x}_i] = 0$  on a :

$$\sqrt{N}(\hat{\mathbf{a}}_N^{2MC} - \hat{\mathbf{a}}_N^{MCO}) \xrightarrow[N \rightarrow +\infty]{L} \mathcal{N}(\mathbf{0}; \Psi) \text{ et } \Psi \text{ est définie positive.}$$

Ce test est un exemple de **test de spécification**, et plus précisément de **compatibilité**. Ce n'est pas un simple test sur la valeur des paramètres du modèle, c'est un test sur des **éléments de sa structure**, *e.g.* régression ou pas.

L'application de l'idée du test d'Hausman dans le cas traité ici conduit au test de la régression augmentée.

## Test de la régression augmentée de l'exogénéité d'une variable explicative

On reprend ici le modèle :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + u_i \text{ avec } E[u_i / \mathbf{x}_i^x] = E[u_i] \equiv 0.$$

avec  $\tilde{\mathbf{z}}_i^e$  pour vecteur de VI externes pour  $\tilde{x}_i^e$  et  $\mathbf{z}_i \equiv (\mathbf{x}_i^x, \tilde{\mathbf{z}}_i^e)$ . On veut tester l'exogénéité de  $x_i^e$ , i.e.  $E[u_i / \mathbf{z}_i, \tilde{x}_i^e] = E[u_i / \mathbf{z}_i] = 0$ .

Le cadre d'analyse de la régression augmentée procure un test simple.

**Tout d'abord**, l'estimateur des MCO de  $(\alpha_0, \beta_0)$  est affecté d'un biais d'endogénéité de  $x_i^e$  si  $\text{Cov}[\tilde{x}_i^e; u_i] \neq 0$ . Or on a :

$$\text{Cov}[\tilde{x}_i^e; u_i] = \text{Cov}[\gamma' \mathbf{z}_i + e_i; \rho e_i + \varepsilon_i] = \rho V[e_i]$$

et donc :

$$\text{Cov}[\tilde{x}_i^e; u_i] \neq 0 \Leftrightarrow \rho \neq 0.$$



On a donc :

$$\text{Hyp0 : } x_i^e \text{ est exogène par rapport à } u_i \Leftrightarrow \rho = 0.$$

**Ensuite**, la régression augmentée approchée contient le paramètre  $\rho$  :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + \rho \hat{e}_{i,N} + \hat{\varepsilon}_{i,N}$$

**Enfin**, on a  $\hat{\varepsilon}_{i,N} \equiv \varepsilon_i + \rho(e_{i,N} - \hat{e}_{i,N}) = \varepsilon_i + \rho \mathbf{z}'_i (\boldsymbol{\gamma}_N^{MCO} - \boldsymbol{\gamma})$ . Aussi on a :

$$\text{Sous Hyp0 : } y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + \varepsilon_i \text{ est un modèle de régression.}$$

Ceci implique qu'on a pas besoin de corriger la distribution de l'estimateur de  $\rho$  de seconde étape,  $\hat{\rho}_N^{MCO}$ , pour le remplacement de  $\boldsymbol{\gamma}$  par  $\boldsymbol{\gamma}_N^{MCO}$  si on teste  $\rho = 0$  puisque les erreurs introduites par ce remplacement disparaissent du modèle sous Sous Hyp0 :  $\rho = 0$ .

**Procédure. Test de la régression augmentée de l'exogénéité de  $\tilde{x}_i^e$**

1. On estime les paramètres de la projection,  $\gamma$ , par les MCO ce qui permet de calculer les  $\hat{e}_{i,N} \equiv \tilde{x}_i^e - \mathbf{z}_i' \hat{\gamma}_N^{MCO}$  pour  $i = 1, \dots, N$ .
2. On estime les paramètres  $\alpha_0$ ,  $\beta_0$  et  $\rho$  à partir du modèle approché :

$$y_i = \alpha_0' \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + \rho \hat{e}_{i,N} + \hat{\varepsilon}_{i,N}$$

où

$$\hat{\varepsilon}_{i,N} \equiv \varepsilon_i + \rho(e_{i,N} - \hat{e}_{i,N}) = \varepsilon_i + \rho \mathbf{z}_i' (\gamma_N^{MCO} - \gamma)$$

par les MCO.

3. On test l'hypothèse  $\text{Hyp0} : \rho = 0$  à partir des tests usuels (*i.e.* sans tenir compte de ce que  $\gamma$  a été remplacé par  $\gamma_N^{MCO}$  dans l'étape 2 pour le calcul de la distribution de l'estimateur des MCO de l'étape 2).

**Utilisation du test du résultat du test de  $\rho = 0$** 

(i) « Hyp0 :  $\rho = 0$  » n'est pas rejetée par la procédure de test. On a :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + u_i \text{ avec } E[u_i / \mathbf{x}_i^x] = E[u_i] \equiv 0 \text{ et } E[\tilde{x}_i^e \varepsilon_i] = 0.$$

Ce modèle est un ***modèle de régression***. On estime  $(\alpha_0, \beta_0)$  en utilisant ***l'estimateur des MCO (gain d'efficacité)***.

(ii) « Hyp0 :  $\rho = 0$  » est rejetée par la procédure de test. On a :

$$y_i = \alpha'_0 \mathbf{x}_i^x + \beta_0 \tilde{x}_i^e + u_i \text{ avec } E[u_i / \mathbf{z}_i] = E[u_i] \equiv 0 \text{ et } E[\tilde{x}_i^e \varepsilon_i] \neq 0.$$

Ce modèle est ***un modèle à VI*** où  $\tilde{x}_i^e$  est une variable explicative ***endogène***. On estime  $(\alpha_0, \beta_0)$  en utilisant ***l'estimateur des 2MC*** avec  $\mathbf{z}_i$  comme vecteur de VI.

On obtient les mêmes estimations pour  $(\alpha_0, \beta_0)$  par les 2MC et par la régression augmentée, mais pas pour les écarts-types des estimateurs (c'est pour ça qu'***il faut ré-estimer  $(\alpha_0, \beta_0)$  par les 2MC***).

## Remarque sur les tests d'exogénéité de variables explicatives

Il est important de tester l'exogénéité des variables explicatives pour les raisons invoquées ci-dessus *mais* :

D'une part, *ça n'est pas toujours possible* puisqu'il faut des VI de la variable dont l'exogénéité est testée.

D'autre part, et cette remarque est liée à la précédente, les résultats de ces tests ne testent pas la validité « complète » de la spécification du modèle.

1. *Le test n'est valide que si les VI utilisées sont elles-mêmes valides*, ce qui ne peut être testé qu'avec d'autres VI,...
2. *Le test n'est valide que si les variables explicatives supposées exogènes le sont réellement*, ce qui ne peut être testé qu'avec des VI, ...

En tout état de cause, si les tests d'exogénéité sont utiles, ils ne peuvent remplacer le gros du travail de l'économètre :

*La spécification de l'ensemble d'information du modèle, i.e. ses variables explicatives exogènes et ses variables instrumentales.*

Ce travail est un travail d'*analyse du processus générateur des données* qui suppose une *bonne connaissance du phénomène modélisé*. C'est en cela que :

« L'économétrie, c'est d'abord de l'économie »

## Remarques sur les $R^2$

On considère ici un modèle à VI de forme générale :

$$y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i \text{ avec } E[u_i / \mathbf{z}_i] \equiv 0.$$

Les logiciels d'économétrie donnent généralement une mesure d'ajustement de type  $R^2$  pour ces modèles :

$$R^2_{2MC} = 1 - \frac{N^{-1} \sum_{i=1}^N (u_{i,N}^{2MC})^2}{N^{-1} \sum_{i=1}^N (y_i - N^{-1} \sum_{i=1}^N y_i)^2} \text{ avec } u_{i,N}^{2MC} \equiv y_i - \mathbf{x}'_i \hat{\mathbf{a}}_N^{2MC}.$$

Plusieurs remarques s'imposent quant à ce *critère d'ajustement*.

Contrairement au cas du  $R^2$  des modèles de régression :

- si, on a  $R_{2MC}^2 < 1$ , on peut avoir :

$$R_{2MC}^2 < 0$$

et :

- on a généralement :

$$R_{2MC}^2 \neq \frac{N^{-1} \sum_{i=1}^N \left[ \hat{y}_{N,i}^{2MC} - N^{-1} \sum_{i=1}^N \hat{y}_{N,i}^{2MC} \right]^2}{N^{-1} \sum_{i=1}^N (y_i - N^{-1} \sum_{i=1}^N y_i)^2} \quad \text{où} \quad \hat{y}_{N,i}^{2MC} \equiv \mathbf{x}_i' \hat{\mathbf{a}}_N^{2MC}.$$

Ceci provient de ce que, en général, on a :

$$N^{-1} \sum_{i=1}^N \mathbf{x}_i' \hat{\mathbf{u}}_{i,N}^{2MC} \neq \mathbf{0}.$$

## Interprétation

- *Le critère du calcul de  $\hat{\mathbf{a}}_N^{2MC}$  n'est pas un critère d'ajustement*, cet estimateur est fondé sur la condition d'orthogonalité « empirique » :

$$N^{-1} \sum_{i=1}^N \hat{\Gamma}_N \mathbf{z}_i \hat{u}_{i,N}^{2MC} = \mathbf{0}.$$

- Si dans le modèle  $y_i = \mathbf{a}'_0 \mathbf{x}_i + u_i$ , le terme d'erreur est « gros », *i.e.* très variable, alors :

$$N^{-1} \sum_{i=1}^N (u_{i,N}^{2MC})^2$$

prendra de « grosses » valeurs. Le calcul de  $\hat{\mathbf{a}}_N^{2MC}$ , contrairement à celui de  $\hat{\mathbf{a}}_N^{MCO}$  ne vise pas à minimiser  $N^{-1} \sum_{i=1}^N (y_i - \mathbf{x}'_i \mathbf{a})^2$  sur  $\mathbf{a}$ .

- En micro-économétrie appliquée, les  $R^2$  obtenus dépassent rarement 0.4.



En outre, en définissant  $R_{MCO}^2$  par :

$$R_{MCO}^2 = 1 - \frac{N^{-1} \sum_{i=1}^N (u_{i,N}^{MCO})^2}{N^{-1} \sum_{i=1}^N (y_i - N^{-1} \sum_{i=1}^N y_i)^2} \text{ avec } u_{i,N}^{MCO} \equiv y_i - \mathbf{x}_i' \hat{\mathbf{a}}_N^{MCO},$$

- on aura toujours :

$$R_{MCO}^2 \geq R_{2MC}^2,$$

*i.e.*, l'**estimateur des MCO** de  $\mathbf{a}_0$  impliquera toujours un **meilleur ajustement** de  $\hat{y}_{N,i}^{MCO} \equiv \mathbf{x}_i' \hat{\mathbf{a}}_N^{MCO}$  à  $y_i$  que  $\hat{y}_{N,i}^{2MC} \equiv \mathbf{x}_i' \hat{\mathbf{a}}_N^{2MC}$ , au sens de l'erreur quadratique moyenne. L'estimateur  $\hat{\mathbf{a}}_N^{MCO}$  est calculé pour ça.

- En outre, l'**écart**  $R_{MCO}^2 - R_{2MC}^2$  **sera d'autant plus élevé** que les variables explicatives endogènes du modèle  $y_i = \mathbf{a}_0' \mathbf{x}_i + u_i$  seront « fortement » liées à  $u_i$ , *i.e.* **que**  $\hat{\mathbf{a}}_N^{MCO}$  **sera biaisé** (même asymptotiquement).

En résumé :

**La *logique d'ajustement* est différente de la *logique d'identification***

# Outline

1. Modèles à variables explicatives endogènes
2. La notion de variable instrumentale et l'estimateur des VI
3. L'estimateur des 2MC
4. Choix des VI et identification
5. Fonctions de contrôle et le test d'exogénéité de la régression augmentée
6. Les variables de contrôle de l'hétérogénéité

## **6. Les variables de contrôle de l'hétérogénéité**

## 6. Les variables de contrôle de l'hétérogénéité

L'objectif est ici de présenter *une autre approche que celle des VI* pour gérer des problèmes dus à l'endogénéité de variables explicatives.

Cette approche *ne s'applique que pour contrôler les problèmes dus l'omission de variables explicatives pertinentes* :

- (i) Les variables explicatives d'un modèle réduisent la variance du terme d'erreur du modèle, *i.e.* « contrôle » une partie l'hétérogénéité de la variable à expliquer.
- (ii) L'introduction de ces variables permet de capter les effets de phénomènes qui font que la relation d'intérêt est difficile à identifier dans une *population hétérogène*.
- (iii) Ce que les économètres nomment « *biais de variables explicatives pertinentes omises* » sont nommés « *biais dus à des effets confondants* » par les statisticiens.

- L'approche dite « *approche par les variables de contrôle de l'hétérogénéité* » est utilisée *en priorité par les économètres* (bien que présentée en fin de chapitre ici).  
En effet elle peut permettre d'éliminer complètement un problème d'endogénéité.
- Ceci-dit, les techniques de VI et l'approche par les variables de contrôle de l'hétérogénéité sont souvent utilisées conjointement : les unes pour certaines variables endogènes, l'autre pour d'autres variables endogènes.
- Les techniques de VI et l'approche par les variables de contrôle de l'hétérogénéité sont des approches très différentes :
  - (i) *Les VI permettent de « contourner » le problème.*
  - (ii) *Les variables de contrôle de l'hétérogénéité permettent de « contenir » le problème.* A l'instar de celle par les fonctions de contrôle, l'approche par les variables de contrôle « éliminent » le problème d'endogénéité.

## 6.1. Remarques générales sur les « variables » de contrôle

Généralement, un économètre est intéressé par l'effet d'une variable, la *variable explicative d'intérêt*, sur une variable à expliquer : *l'effet d'intérêt*.

Ceci dit, il utilise *plusieurs variables explicatives* « *supplémentaires* », dites *variables de contrôle (de l'hétérogénéité)*. Par exemple, dans le modèle :

$$y_i = \alpha_0 + \beta_0' \mathbf{c}_i + b_0 \tilde{x}_i + u_i \text{ avec } E[u_i] \equiv 0,$$

- $\tilde{x}_i$  est la variable d'intérêt,
- $b_0$  est l'effet d'intérêt

et :

- $\mathbf{c}_i$  est le vecteur des variables de contrôle.

**Rmq.** Les effets représentés par  $\beta_0$  peuvent également être intéressants.

$$y_i = \alpha_0 + \beta'_0 \mathbf{c}_i + b_0 \tilde{x}_i + u_i \text{ avec } E[u_i] \equiv 0 \text{ et } E[u_i / \mathbf{c}_i] = 0$$

- ***Dans tous les cas***, l'introduction de  $\mathbf{c}_i$  dans le modèle *améliore la précision de l'estimation de  $b_0$* .
- Dans le modèle « simplifié » (sans contrôle par  $\mathbf{c}_i$ ) :

$$y_i = \delta_0 + b_0 \tilde{x}_i + v_i \text{ avec } E[v_i] \equiv 0,$$

le *terme d'erreur*  $v_i$  *contient l'essentiel des effets des  $\mathbf{c}_i$* .

*Une large part de l'hétérogénéité de  $y_i$  reste alors dans  $v_i$ .*

- On a alors  $V[v_i] \geq V[u_i]$  ce qui fait que l'*estimation de  $b_0$  est moins précise dans le modèle simplifié* que dans le modèle complet
- Le terme  $\beta'_0 \mathbf{c}_i$  « *capte* » ou « *contrôle* » une part de l'*hétérogénéité de  $y_i$* .



**Modèle « complet »**

$$y_i = \alpha_0 + \beta'_0 \mathbf{c}_i + b_0 \tilde{x}_i + u_i \text{ avec } E[u_i] \equiv 0 \text{ et } E[u_i / \mathbf{c}_i] = 0$$

**Modèle « simplifié »**

$$y_i = \delta_0 + b_0 \tilde{x}_i + v_i \text{ avec } E[v_i] \equiv 0$$

- **Dans certains cas**, la variable explicative d'intérêt  $\tilde{x}_i$  peut être :
  - *exogène par rapport* à  $u_i$ , *i.e.* dans le modèle (complet) avec  $\mathbf{c}_i$ ,
- mais :*
  - être *endogène par rapport* à  $v_i$ , *i.e.* dans le modèle (simplifié) sans  $\mathbf{c}_i$ .

Dans ce dernier cas, le terme  $\beta'_0 \mathbf{c}_i$  « *élimine* » le *problème d'endogénéité* de  $\tilde{x}_i$ , *i.e.* « *contrôle* » ce problème grâce au *contrôle de l'hétérogénéité* de  $y_i$ .

Cela se produit lorsque  $\tilde{x}_i$  et  $v_i$  sont liées par, et uniquement par,  $\mathbf{c}_i$ .

Quand on passe du modèle complet :

$$y_i = \alpha_0 + \beta'_0 \mathbf{c}_i + b_0 \tilde{x}_i + u_i$$

au modèle simplifié :

$$y_i = \delta_0 + b_0 \tilde{x}_i + v_i,$$

en fait on *décompose*  $\beta'_0 \mathbf{c}_i$  en :

$\beta'_0 \mathbf{c}_i = \underbrace{E[\beta'_0 \mathbf{c}_i]}_{\substack{\text{Effets} \\ \text{"en moyenne"}}} + \underbrace{(\beta'_0 \mathbf{c}_i - E[\beta'_0 \mathbf{c}_i])}_{\substack{\text{Effets} \\ \text{"hors-moyenne"}}}$
---

ou encore en :

$\beta'_0 \mathbf{c}_i = \beta'_0 E[\mathbf{c}_i] + \beta'_0 (\mathbf{c}_i - E[\mathbf{c}_i])$
--

avec :

$$E[\beta'_0 (\mathbf{c}_i - E[\mathbf{c}_i])] = \beta'_0 E[\mathbf{c}_i - E[\mathbf{c}_i]] = 0,$$

par construction.

On a alors :

$$y_i = \alpha_0 + \underbrace{\beta'_0 E[\mathbf{c}_i]}_{\text{Effets "moyens" de } \mathbf{c}_i} + \underbrace{\beta'_0 (\mathbf{c}_i - E[\mathbf{c}_i])}_{\text{Effets "hors-moyenne" de } \mathbf{c}_i} + b_0 \tilde{x}_i + u_i$$

ce qui donne :

$$y_i = \underbrace{\{\alpha_0 + \beta'_0 E[\mathbf{c}_i]\}}_{\delta_0} + b_0 \tilde{x}_i + \underbrace{\{u_i + \beta'_0 (\mathbf{c}_i - E[\mathbf{c}_i])\}}_{v_i}$$

et finalement :

$$y_i = \delta_0 + b_0 \tilde{x}_i + v_i \text{ avec } E[v_i] \equiv 0.$$

De fait, la notion de « *contrôle* » est intimement liée à celle de « *toutes choses égales par ailleurs* » : si  $\tilde{x}_i$  et  $\mathbf{c}_i$  sont corrélés (et  $\beta_0 \neq 0$ ) alors  $\tilde{x}_i$  et  $v_i$  **varient nécessairement conjointement** (endogénéité)

## 6.2. Le modèle considéré ici

On utilise à nouveau le modèle linéaire général à une variable explicative endogène :

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + u_i \text{ avec } E[u_i / \tilde{\mathbf{x}}_i^x] = E[u_i] \equiv 0.$$

En fait, on aurait aimé pouvoir travailler sur le modèle :

$$y_i = \theta_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + \delta_0 q_i + v_i \text{ avec } E[v_i / \tilde{\mathbf{x}}_i^x, \tilde{x}_i^e, q_i] = E[u_i] \equiv 0.$$

car c'est un modèle de régression.

Le problème est que *la variable  $q_i$  n'est pas disponible* dans les données.

- On suppose ici que  $Cov[\tilde{\mathbf{x}}_i^x; q_i] = \mathbf{0}$ , i.e. que  $\tilde{\mathbf{x}}_i^x$  *ne fournit aucune information « statistique » sur  $q_i$ .*
- *On pose ici que  $E[q_i] \equiv 0$  pour simplifier la suite. C'est possible car il s'agit seulement d'une normalisation sur une variable « non observée ».*

On ne dispose pas de VI pour  $\tilde{x}_i^e$  mais, en revanche, ***on sait pourquoi elle est endogène : c'est à cause de l'omission*** (forcée) de  $q_i$ .

***On utilise cette information pour*** (tenter de) ***résoudre le problème***  
d'identification de  $(\alpha_0, \beta_0)$  ***posé par l'endogénéité de  $\tilde{x}_i^e$ .***

### **Exemple typique : Equation de salaire**

$y_i$  : salaire de  $i$

$\tilde{x}_i^e$  : niveau d'éducation de  $i$  (ou origine ethnique)

$\tilde{\mathbf{x}}_i^x$  : caractéristiques de  $i$  (expérience, âge, sexe, ...)

$q_i$  : aptitudes (capacités d'adaptation, ...) *ou* origine (ethnique) de  $i$

$$\tilde{\mathbf{x}}_i \equiv (\tilde{\mathbf{x}}_i^x, \tilde{x}_i^e)$$

### 6.3. Liens entre le modèle latent et le modèle utilisé

#### *Modèle utilisé*

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + u_i$$

#### *Modèle latent*

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + \delta_0 q_i + v_i \quad \text{avec} \quad E[u_i / \tilde{\mathbf{x}}_i^x, \tilde{x}_i^e, q_i] = E[u_i] \equiv 0.$$

#### *Lien entre les modèles latent et utilisé*

$$u_i = v_i + \delta_0 q_i \quad \text{avec} \quad E[u_i] = 0 \quad \text{et} \quad \text{Cov}[\tilde{\mathbf{x}}_i^x; u_i] = \mathbf{0}$$

(i) La variable  $q_i$  est seulement « *latente* » dans le modèle utilisé : ses effets sont seulement « *implicites* » dans ce modèle.

(ii) On a  $E[u_i] = E[v_i] + \delta_0 E[q_i] = 0 + \delta_0 \times 0$

(iii) On a  $\text{Cov}[\tilde{\mathbf{x}}_i^x; u_i] = \text{Cov}[\tilde{\mathbf{x}}_i^x; v_i] + \delta_0 \text{Cov}[\tilde{\mathbf{x}}_i^x; q_i] = \mathbf{0} + \delta_0 \times \mathbf{0}$

### *Interprétation*

(i)  $u_i$  contient l'effet de  $q_i$  sur  $y_i$ ,  $\delta_0 q_i$ , qui est un effet de moyenne nulle

(ii) Le terme  $\delta_0 q_i$  accroît *l'hétérogénéité* « *non-contrôlée* » de  $y_i$

La variable  $\tilde{x}_i^e$  est endogène dans le modèle utilisé :

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + u_i \text{ avec } E[u_i / \tilde{\mathbf{x}}_i^x] = E[u_i] \equiv 0$$

si  $Cov[\tilde{x}_i^e; u_i] \neq 0$ , par définition de l'endogénéité.

Ici on a :

$$Cov[\tilde{x}_i^e; u_i] = Cov[\tilde{x}_i^e; v_i + \delta_0 q_i] = 0 + \delta_0 \times Cov[\tilde{x}_i^e; q_i].$$

### Biais de variable explicative pertinente omise

La variable  $\tilde{x}_i^e$  est *endogène* dans le modèle utilisé si :

$$\delta_0 \text{Cov}[\tilde{x}_i^e; q_i] \neq 0$$

i.e., si :

(i)  $q_i$  est une *variable explicative pertinente* de  $y_i$ , i.e.  $\delta_0 \neq 0$

et :

(ii) la variable latente  $q_i$  est *corrélée* à la variable  $\tilde{x}_i^e$ , i.e.  $\text{Cov}[\tilde{x}_i^e; q_i] \neq 0$ .

L'estimateur des MCO de  $(\alpha_0, \beta_0)$  est *biaisé* si  $\delta_0 \neq 0$  et  $\text{Cov}[\tilde{x}_i^e; q_i] \neq 0$

*Biais positif* sur  $\beta_0$  si  $\delta_0 \text{Cov}[\tilde{x}_i^e; q_i] > 0$

*Biais négatif* sur  $\beta_0$  si  $\delta_0 \text{Cov}[\tilde{x}_i^e; q_i] < 0$

Les effets de  $q_i$  et  $\tilde{x}_i^e$  sur  $y_i$  se « *confondent* » dans l'estimateur des MC de  $\beta_0$



### Exemple typique : *Equation de salaire*

**Avec**  $q_i$  = aptitudes scolaires :

(i)  $\delta_0 > 0$  : aptitudes « causent » salaire

et :

(ii)  $Cov[\tilde{x}_i^e; q_i] > 0$  : aptitudes, scolaires en particulier, aident à faire des études.

**Biais**  $> 0$  **sur**  $\beta_0$  : les plus aptes font des études et ont de bons salaires

**Avec**  $q_i$  = origine (pas type « caucasien » / type « caucasien ») :

(i)  $\delta_0 \leq 0$  : effet potentiel de « discrimination » pour l'emploi

et :

(ii)  $Cov[\tilde{x}_i^e; q_i] < 0$  : corrélation (pas effet causal !) entre origine et difficultés qui limitent les études.

**Biais**  $\geq 0$  **sur**  $\beta_0$  : si discrimination moins bons salaires et moins d'études pour les « pas caucasiens ».

- (i) Dans le modèle  $\beta_0$  est spécifié comme un *effet causal* : les *études* donnent des *compétences qui se valorisent sous forme de salaire*.
- (ii) L'*estimation par les MCO de  $\beta_0$  est biaisée vers le haut*.

Dans le cas « *aptitude* »

- $\hat{\beta}_N^{MCO}$  mesure l'effet causal d'intérêt plus l'*effet de sélection des étudiants sur leurs capacités*.
- Les plus « aptes » progressent plus vite dans leur carrière professionnelle et font plus d'études (de « moins aptes » auraient pu suivre ces études).

Dans le cas « *origine* »

- $\hat{\beta}_N^{MCO}$  mesure l'effet causal d'intérêt plus une *autre forme d'effet de sélection* s'il y a discrimination sur le marché du travail.
- Ceux qui font le plus d'études sont aussi ceux qui souffrent le moins de l'effet (potentiel) de la discrimination sur le marché de l'emploi.

## 6.4. Les variables de contrôle de l'hétérogénéité

Il n'est pas nécessaire de mesurer  $q_i$  pour résoudre le problème

Il nous faut *une variable ou un vecteur de variables*, disons  $\mathbf{c}_i$ , *qui « capte » tout ce qui lie  $q_i$  à  $\tilde{x}_i^e$*  tout en étant *exogène par rapport* à  $v_i$ .

On sait que :

$$u_i = v_i + \delta_0 q_i \quad \text{et} \quad \text{Cov}[\tilde{x}_i^e; u_i] = \delta_0 \times \text{Cov}[\tilde{x}_i^e; q_i]$$

*L'idée est ici double :*

- (i) On ne peut rien faire pour  $\delta_0 \neq 0$ , si  $q_i$  est pertinente pour expliquer  $y_i$ , elle le restera quoiqu'on fasse.
- (ii) En revanche, on peut parfois *contrôler la « partie » de  $q_i$  liée à  $\tilde{x}_i^e$* .

## Principe de fonctionnement du contrôle de l'endogénéité (en général)

*Décomposer le terme d'erreur en deux parties :*

- (i) Une partie qui « *contrôle* » le *problème d'endogénéité*
  - (ii) Une *partie résiduelle* qui ne pose pas de problème d'endogénéité
- pour *transformer le modèle en un modèle de régression*.

$$u_i = v_i + \delta_0 q_i \text{ avec } Cov[\mathbf{x}_i^x; u_i] = 0 \text{ et } Cov[\tilde{x}_i^e; u_i] = \delta_0 \times Cov[\tilde{x}_i^e; q_i]$$

## Principe de fonctionnement des variables de contrôle de l'endogénéité

*Décomposer*  $q_i$  (la partie de  $u_i$  qui pose problème) en deux parties :

- (i) Une partie qui « *contrôle* » le *problème d'endogénéité* :  $Cov[\tilde{x}_i^e; q_i]$
  - (ii) Une *partie résiduelle* qui ne pose pas de problème d'endogénéité
- pour *transformer le modèle en un modèle de régression*

### Propriétés essentielles de variables de contrôle de l'endogénéité de $q_i$

$$q_i = \gamma + \boldsymbol{\rho}' \mathbf{c}_i + \varepsilon_i$$

avec :

$$\text{Cov}[\tilde{x}_i^e \varepsilon_i] = 0, \text{Cov}[\mathbf{c}_i \varepsilon_i] = \mathbf{0} \text{ et } \text{Cov}[\mathbf{c}_i v_i] = \mathbf{0}.$$

- (i) les variables  $\tilde{x}_i^e$  et  $\mathbf{c}_i$  vecteur sont *exogènes par rapport à  $\varepsilon_i$  et à  $v_i$*  (les variables  $\tilde{\mathbf{x}}_i^x$  le sont nécessairement). Si  $\varepsilon_i$  ajoute du « bruit » dans  $u_i$ ,  $\varepsilon_i$  n'est pas corrélé à  $\tilde{x}_i^e$  et n'est donc pas source d'endogénéité de cette variable.

et :

- (ii) Le terme  $\boldsymbol{\rho}' \mathbf{c}_i$  « *capte* » tout ce qui lie  $q_i$  à  $\tilde{x}_i^e$  :

$$\text{Cov}[\tilde{x}_i^e \varepsilon_i] = 0 \Leftrightarrow \text{Cov}[\tilde{x}_i^e; q_i] = \text{Cov}[\tilde{x}_i^e; \boldsymbol{\rho}' \mathbf{c}_i]$$

Si on « extrait »  $\boldsymbol{\rho}' \mathbf{c}_i$  de  $q_i$  alors plus rien ne lie  $\tilde{x}_i^e$  à  $q_i$ .

On a nécessairement  $\boldsymbol{\rho} \neq \mathbf{0}$  si  $\text{Cov}[\tilde{x}_i^e; q_i] \neq 0$ .

Les principales *conditions requises pour*  $\mathbf{c}_i$  sont :

$$\text{Cov}[\tilde{x}_i^e \varepsilon_i] = 0 \quad \text{et} \quad \text{Cov}[\mathbf{c}_i \varepsilon_i] = \text{Cov}[\mathbf{c}_i v_i] = \mathbf{0},$$

i.e.  $\mathbf{c}_i$  « *capte* » *tout ce lie*  $\tilde{x}_i^e$  à  $q_i$  et  $\mathbf{c}_i$  est exogène par rapport à  $\varepsilon_i$  et  $v_i$ .

En substituant  $q_i$  par  $\gamma + \mathbf{p}'\mathbf{c}_i + \varepsilon_i$  dans le modèle latent :

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + \underbrace{\delta_0 (\gamma + \mathbf{p}'\mathbf{c}_i + \varepsilon_i)}_{q_i} + v_i,$$

on obtient :

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + \boxed{\delta_0 \gamma} + \delta_0 \mathbf{p}'\mathbf{c}_i + \boxed{\delta_0 \varepsilon_i} + v_i$$

et finalement :

$$y_i = \underbrace{(\alpha_0 + \delta_0 \gamma)}_{\theta} + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + \delta_0 \mathbf{p}'\mathbf{c}_i + \underbrace{(v_i + \delta_0 \varepsilon_i)}_{\eta_i}.$$

Ceci permet de définir un nouveau modèle utilisable empiriquement :

$$y_i = \theta + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \tilde{x}_i^e + \boldsymbol{\pi}' \mathbf{c}_i + \eta_i$$

où :

$$\theta \equiv \alpha_0 + \delta_0 \gamma, \quad \boldsymbol{\pi} \equiv \delta_0 \boldsymbol{\rho} \quad \text{et} \quad \eta_i \equiv v_i + \delta_0 \varepsilon_i$$

avec :

$$\text{Cov}[\tilde{\mathbf{x}}_i^x; \eta_i] = \mathbf{0}, \quad \text{Cov}[\tilde{x}_i^e; \eta_i] = 0 \quad \text{et} \quad \text{Cov}[\mathbf{c}_i; \eta_i] = \mathbf{0}.$$

Ce modèle, qui intègre les *variables de contrôle*,  $\mathbf{c}_i$ , *des effets de la variable omise*  $q_i$  est un *modèle de régression* (linéaire).


L'*estimateur des MCO* est convergent pour les paramètres de ce modèle, et pour  $\beta_0$  en particulier.

## Principe « d'action » variables de contrôle de l'hétérogénéité


L'équation  $q_i = \gamma + \mathbf{p}'\mathbf{c}_i + \varepsilon_i$  indique que  $q_i$  peut être *décomposée* en :

- un terme,  $\mathbf{p}'\mathbf{c}_i$ , *qui contrôle l'endogénéité de  $\tilde{x}_i^e$*  dans le modèle,
- un terme,  $\gamma + \varepsilon_i$ , dont l'effet se répartit dans la constante ( $\gamma$ ) et le terme d'erreur ( $\varepsilon_i$ ) du modèle.

**Modèle sans contrôle :**  $\tilde{x}_i^e$  corrélée à  $u_i$ , *endogène*

$$y_i = \alpha_0 + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \boxed{\tilde{x}_i^e} + \underbrace{\delta_0 q_i}_{u_i} + v_i$$


**Modèle avec contrôle :**  $\tilde{x}_i^e$  pas corrélée à  $\eta_i$ , *exogène*

$$y_i = \underbrace{(\alpha_0 + \delta_0 \gamma)}_{\theta} + \mathbf{b}'_0 \tilde{\mathbf{x}}_i^x + \beta_0 \boxed{\tilde{x}_i^e} + \underbrace{\delta_0 \mathbf{p}'\mathbf{c}_i}_{\pi} + \underbrace{(v_i + \delta_0 \varepsilon_i)}_{\eta_i}$$




### Exemple typique : *Equation de salaire*

Avec  $q_i$  = aptitudes scolaires :

Le choix usuel de  $c_i$  se limite à des scores à des tests tels que celui du **QI**.

Ce test ne mesure qu'une partie des aptitudes des individus, les aptitudes au raisonnement logique. Mais ces aptitudes sont les plus utiles pour les études et pour la carrière professionnelle.

Avec  $q_i$  = origine :

Dans le cas de l'origine, il n'y a pas de variable de contrôle sauf  $q_i$  elle-même.

La **seule solution** dans ce cas passe par l'utilisation des **techniques de VI**.

Ceci-dit, sans mesure de  $q_i$ , on ne dispose d'aucun moyen de mesure statistique des phénomènes de discrimination.

## Remarques finales sur les variables de contrôle

- La forme linéaire de l'effet de  $\mathbf{c}_i$  sur  $q_i$ ,  $\gamma + \boldsymbol{\rho}'\mathbf{c}_i$ , n'est pas vraiment restrictive puisque les éléments de  $\mathbf{c}_i$  peuvent être choisis en fonction du problème. Le terme  $\gamma + \boldsymbol{\rho}'\mathbf{c}_i$  peut être défini comme  $EL[q_i/1, \mathbf{c}_i]$ .
- La condition  $Cov[\tilde{\mathbf{x}}_i^x; q_i] = \mathbf{0}$  est « artificielle ». Elle visait ici à distinguer « clairement » les vecteurs  $\tilde{\mathbf{x}}_i^x$  et  $\mathbf{c}_i$ .
- Dans le modèle « final », les rôles des variables  $\tilde{\mathbf{x}}_i^x$  et  $\mathbf{c}_i$  sont à la fois similaires et différents.
  - Ces deux vecteurs de variables *améliorent la prédiction de  $y_i$  et la précision des estimateurs calculés.*
  - Ceci-dit, la priorité demeure ici l'identification de l'effet causal de  $\tilde{x}_i^e$ . Si  $\mathbf{c}_i$  contrôle l'endogénéité de  $\tilde{x}_i^e$  dû à l'omission (forcée) de  $q_i$ , ce n'est pas le cas de  $\tilde{\mathbf{x}}_i^x$ .

- Les *variables instrumentales* de  $\tilde{x}_i^e$ ,  $z_i$ , et les *variables de contrôle* de son endogénéité,  $c_i$ , sont de natures très différentes.

- (i) Si  $z_i$  et  $c_i$  *doivent être* tous deux *liés* à  $\tilde{x}_i^e$ ,
- (ii)  $z_i$  *doit être non corrélé* à  $q_i$  alors que  $c_i$  *doit l'être le plus possible*.

- *On ne peut contrôler l'endogénéité* de  $\tilde{x}_i^e$  si cette dernière est due à des problèmes de *simultanéité* ou d'*erreurs de mesure sur*  $\tilde{x}_i^e$ .
  - (i) Si  $\tilde{x}_i^e$  et  $y_i$  sont simultanées, alors  $c_i$  devrait contrôler entièrement  $u_i$ , ce qui est impossible. On aurait alors un modèle « parfait ».
  - (ii) Si  $\tilde{x}_i^e$  était endogène par rapport à  $u_i$  parce  $\tilde{x}_i^e$  est une mesure bruitée d'une variable d'intérêt, alors  $c_i$  devrait contrôler entièrement l'erreur de mesure concernée. Cette dernière serait alors « connue ».
- On a utilisé des formes spécifiques de l'effet de  $q_i$  sur  $y_i$  et de l'effet de  $c_i$  sur  $q_i$ . En pratique, le raisonnement est souvent beaucoup plus informel.

## « Mauvais contrôle » : deux exemples typiques

Dans l'équation de salaire simplifiée :

$$\text{salaire}_i = \alpha_0 + b_0 \text{educ}_i + \delta_0 \text{aptitudes}_i + v_i,$$

il faut contrôler l'effet de  $\text{aptitudes}_i$ , i.e. des qualités « innées » de  $i$ .

Une bonne variable de contrôle de  $\text{aptitudes}_i$  est un test de  $QI_i$  réalisé lorsque  $i$  est jeune.

Utilisation de  $CSP_i$ , catégorie socio-professionnelle de  $i$ . Lié à  $\text{aptitudes}_i$

**mais :**

**$CSP_i$  est une conséquence de  $\text{educ}_i$ .** En fait,  $CSP_i$  contrôle l'effet de  $\text{aptitudes}_i$  mais surtout l'essentiel de l'effet causal de  $\text{educ}_i$  sur  $\text{salaire}_i$ .

Utilisation de  $\text{Test}_i$ , le résultat d'un test d'embauche. Lié à  $\text{aptitudes}_i$

**mais :**

si  $\text{Test}_i$  dépend de  $\text{aptitudes}_i$ ,  $\text{Test}_i$  dépend également de  $\text{educ}_i$ . Une partie de l'effet de  $\text{educ}_i$  **transite par**  $\text{Test}_i$ .

Sinon on peut répéter ici ce qui a été dit à propos de la recherche des VI en le complétant :

La spécification de l'ensemble d'information du modèle, *i.e.* ses variables explicatives exogènes, ***et de ses variables de contrôle de l'hétérogénéité en particulier***, et ses variables instrumentales, est une phase essentielle du travail de l'économètre appliqué.

Ce travail est un travail d'analyse du processus générateur des données qui suppose une bonne connaissance du phénomène modélisé. C'est en cela que :

« L'économétrie, c'est d'abord de l'économie »
--