Économétrie — TD 7

Exogénéité, instrumentation et sur-identification

Pierre Beaucoral

1. Contexte : endogénéité et variables instrumentales

On part d'un modèle linéaire simple :

$$y_i = \beta x_i + \gamma' w_i + u_i$$

- y_i : variable expliquée
- x_i : régresseur **endogène** (corrélé au terme d'erreur)
- w_i : contrôles **exogènes**
- u_i : terme d'erreur

Problème : si

$$\mathrm{Cov}(x_i, u_i) \neq 0$$

alors l'estimateur MCO de β est **biaisé** et **inconsistant**.

Idée des variables instrumentales (VI) : introduire des variables z_i telles que :

1. (**Pertinence**) $\mathrm{Cov}(z_i, x_i) \neq 0$
2. (**Exogénéité / validité**) $\mathrm{Cov}(z_i, u_i) = 0$

Ces instruments permettent d'identifier β même si x_i est endogène.

---

2. Notation matricielle et identification

On empile les données :

- y : vecteur $(n \times 1)$
- X : matrice $(n \times K)$ des régresseurs (dont certains endogènes)
- W : variables exogènes incluses dans le modèle
- Z : matrice $(n \times L)$ des **instruments** (et exogènes)

Conditions clés pour les VI :

- **Validité** des instruments :
$$E[Z'u] = 0$$
- **Pertinence / rang** :
$$\text{rang}(E[Z'X]) = K$$

où K est le nombre de variables **endogènes à instrumenter**.

---

3. Exact-identification vs sur-identification

On note :

- K : nombre de variables endogènes instrumentées
- L : nombre d'instruments (hors exogènes inclus dans X)

Cas possibles :

- **Sous-identifié** : $L < K$
  pas assez d'instruments, le modèle n'est pas identifié.
- **Exactement identifié** : $L = K$
  autant d'instruments que de variables endogènes.
- **Sur-identifié** : $L > K$
  plus d'instruments que nécessaire.

Dans le cas **sur-identifié**, on a **plus de conditions d'exogénéité** $E[Z'u] = 0$ que nécessaire pour identifier les paramètres. Ces conditions supplémentaires peuvent être **testées** empiriquement : c'est le rôle du **test de sur-identification de Sargan**.

---

4. Intuition du test de sur-identification

Idée simple :

1. On estime le modèle par VI (typiquement 2SLS) et on obtient les **résidus** :

$$\hat{u}_i = y_i - \hat{y}_i$$

2. Si tous les instruments sont bien **exogènes**, alors ils doivent être **orthogonaux** aux erreurs vraies u_i, et donc approximativement à $\hat{u}_i$:

$$E[z_{ji}\hat{u}_i] \approx 0 \quad \forall j$$

3. Si l'on arrive à **expliquer** les résidus $\hat{u}_i$ par les instruments Z, cela suggère que ces instruments sont en fait **corrélés** au terme d'erreur, donc **invalides**.

Le test de Sargan mesure précisément la "force" de cette éventuelle corrélation entre $\hat{u}_i$ et Z.

---

5. Construction de la statistique de Sargan (cas homoscédastique)

5.1. Étape 1 : estimation VI

On estime le modèle :

$$y = X\beta + u$$

par 2SLS en utilisant Z comme instruments, et on obtient l'estimateur $\hat{\beta}_{2SLS}$ et les résidus $\hat{u}$:

$$\hat{u} = y - X\hat{\beta}_{2SLS}$$

5.2. Étape 2 : régression auxiliaire

On régresse ensuite les résidus $\hat{u}$ sur l'ensemble des instruments Z (et, en pratique, les exogènes inclus dans X) :

$$\hat{u}_i = \delta_0 + Z_i'\delta + v_i$$

On note $R^2_{\hat{u}\sim Z}$ le coefficient de détermination de cette régression.

5.3. Statistique de Sargan

La statistique de Sargan est définie par :

$$J = n \times R^2_{\hat{u} \sim Z}$$

où n est la taille de l'échantillon.

Intuition : plus R^2 est élevé, plus les instruments expliquent bien les résidus, donc plus il est suspect que les instruments soient corrélés à u.

---

6. Loi asymptotique et hypothèses

Sous les hypothèses suivantes :

- instruments valides : $E[Z'u] = 0$
- **homoscédasticité** des erreurs u_i
- spécification correcte du modèle (forme fonctionnelle, variables pertinentes, etc.)

alors, sous l'hypothèse nulle H_0 :

> Tous les instruments sont **exogènes**

la statistique de test J suit asymptotiquement une loi du chi-deux :

$$J \overset{a}{\sim} \chi^2_{L-K}$$

- les **degrés de liberté** sont : $L - K$ (nombre de **restrictions sur-identifiantes**)
  - L : nombre d'instruments
  - K : nombre de variables endogènes

On peut donc calculer une **p-value** à partir de cette loi χ^2_{L-K}.

---

7. Formulation du test

- **Hypothèse nulle** H_0 : *Tous les instruments sont valides (exogènes)*
  $E[Z'u] = 0$
- **Hypothèse alternative** H_1 : *Au moins un instrument est invalide*
  instruments corrélés au terme d'erreur, ou mauvaise spécification globale.

---

8. Interprétation pratique

On calcule la p-value associée à J sous la loi χ^2_{L-K} :

- **Si la p-value est élevée** (par ex. > 5 %) :
  - on **ne rejette pas** H_0 ;
  - on ne trouve pas de preuve statistique contre la validité **globale** des instruments ;
  - attention : cela **ne prouve pas** que les instruments sont parfaits, seulement qu'on ne détecte pas de violation.
- **Si la p-value est faible** (par ex. < 5 %) :
  - on **rejette** H_0 ;
  - il est probable qu'au moins un (ou plusieurs) instrument(s) soit(ent) corrélé(s) au terme d'erreur ;
  - cela peut aussi refléter une **mauvaise spécification** du modèle (variables omises, non-linéarités, etc.).

Important : le test ne dit pas **quel** instrument est problématique. Il teste seulement la **validité conjointe** de tous les instruments.

---

9. Sargan vs Hansen (J-test robuste)

Le test de Sargan repose sur l'hypothèse d'**homoscédasticité**.
En présence d'**hétéroscédasticité**, la loi de $J = nR^2$ n'est plus une bonne approximation.

Dans le cadre plus général du **GMM (Generalized Method of Moments)**, on peut construire un test de sur-identification **robuste à l'hétéroscédasticité** :

- Test de **Hansen J** (ou Sargan–Hansen) :

- même hypothèse nulle : validité globale des instruments ;
- même loi asymptotique : χ^2_{L-K} ;
- mais la statistique est calculée à partir d'une matrice de pondération robuste (type "sandwich").

En pratique :

- **Sargan** : adapté si l'on suppose l'homoscédasticité.
- **Hansen J** : à privilégier lorsque l'on utilise des estimateurs GMM ou des variances robustes.

---

10. Exemple stylisé : 1 endogène, 2 instruments

Considérons le modèle :

$$y_i = \beta x_i + \gamma' w_i + u_i$$

- x_i est endogène
- w_i est exogène et inclus dans le modèle
- on dispose de deux instruments z_{1i}, z_{2i} pour x_i

On a :

- $K = 1$ (une variable endogène à instrumenter)
- $L = 2$ (deux instruments)

Le modèle est **sur-identifié** avec $L - K = 1$ restriction sur-identifiante.

Étapes :

1. Estimer le modèle par 2SLS en utilisant z_{1i}, z_{2i} (et w_i) comme instruments.
2. Récupérer les résidus $\widehat{u}_i$.
3. Régresser $\widehat{u}_i$ sur z_{1i}, z_{2i} (et w_i), récupérer le R^2.
4. Calculer $J = nR^2$ et comparer à la loi χ^2 à 1 ddl.

Questions TDs

1- Appliquez le test d'exogénéité de ***Nakamura et Nakamura*** en utilisant trois ensembles d'instruments :

i)- ***motheduc*** (éducation de la mère)

ii)- ***motheduc*** et ***fatheduc*** (éducation du père)

iii)- ***motheduc***, ***fatheduc*** et ***huseduc*** (éducation du mari)

Afficher la réponse

Comme le TD, d'abord appliquer le 2SLS dans Eviews, puis réaliser le test nakamura nakamura dans les IV tests. Dans quels cas le test est intéressant?

2- Si cela vous semble pertinent, appliquez les ***doubles moindres carrés*** (***DMC***) en utilisant les trois ensembles d'instruments.

Afficher la réponse

Se baser sur le test de nakamura!!!

3- Appliquez, lorsque cela est possible, le test de ***sur-identification de Sargan***.

Afficher la réponse

Si le test de Nakamura a révélé un DMC dans lequel il y avait plus de un instrument, appliquer le test de suridentification.

4- Réalisez un test de ***White*** sur les estimations ***DMC*** et appliquez, si besoin est, la procédure de correction de ***White***. Qu'en concluez- vous ?

Afficher la réponse

Vous savez faire les tests, pour la correction, allez fouiller dans les options des estimations AVANT de lancer l'estimation.