Économétrie — TD 7

Exogénéité, instrumentation et sur-identification

Pierre Beaucoral

```
library(knitr)
knit_hooks$set(optipng = hook_optipng)
```

Plan de la séance

- Rappel : endogénéité et VI
- Identification : exact vs sur-identification
- Idée du test de sur-identification
- Statistique de Sargan (cas homoscédastique)
- Interprétation et limites

Rappel : endogénéité et variables instrumentales

- Modèle structurel simple :

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

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

- Problème : $\text{Cov}(x_i, u_i) \neq 0$
  estimateur MCO biaisé et non convergent
- Idée VI : trouver des instruments z_i tels que :

- **Pertinence** : $\text{Cov}(z_i, x_i) \neq 0$
- **Validité / exogénéité** : $\text{Cov}(z_i, u_i) = 0$

Notation VI (forme matricielle)

- On empile les données :
  - y : vecteur $(n \times 1)$
  - X : matrice $(n \times K)$ des régresseurs (dont certains endogènes)
  - W : régresseurs exogènes (optionnels)
  - Z : matrice $(n \times L)$ des instruments (et exogènes)
- Conditions clés :

$$E[Z'u] = 0 \quad \text{(validité des instruments)}$$

$$\text{rang}(E[Z'X]) = K \quad \text{(pertinence + identification)}$$

Identification : exactement vs sur-identifié

- K : nombre de variables **endogènes** à instrumenter
- L : nombre d'instruments (exogènes distincts)

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
  on dispose **d'information supplémentaire** sur les conditions d'orthogonalité $E[Z'u] = 0$

Idée de la sur-identification

- Quand $L > K$, plusieurs "combinaisons" d'instruments pourraient identifier β.
- Si **tous les instruments sont valides**, toutes ces manières d'identifier β devraient donner **la même vraie valeur**.
- Intuition :

  > *Les conditions d'exogénéité imposées par les instruments supplémentaires sont des **restrictions** supplémentaires sur le modèle.*

- On peut alors **tester** si ces restrictions supplémentaires sont compatibles avec les données.

C'est l'objet du **test de sur-identification de Sargan**.

Idée informelle du test de Sargan

- On estime le modèle par VI (2SLS) et on obtient les résidus :

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

- Si les instruments sont vraiment exogènes, on doit avoir :

$$E[z_{ji}\widehat{u}_i] = 0 \quad \text{pour tous les instruments } j$$

- Le test de Sargan vérifie donc **dans les données** si les résidus $\widehat{u}_i$ sont "orthogonaux" aux instruments Z.
- Idée pratique : si on peut **expliquer** les résidus par les instruments, alors ces derniers sont probablement **corrélés aux erreurs**, donc **invalides**.

Construction de la statistique de Sargan (cas homoscédastique)

Supposons que l'on a estimé le modèle par 2SLS :

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

Étapes :

1. **Étape 1** : estimer le modèle VI (2SLS) et récupérer les résidus $\hat{u}$.
2. **Étape 2** : régresser $\hat{u}$ sur **tous les instruments** Z (et en pratique aussi les exogènes inclus dans X) :

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

3. **Étape 3** : récupérer le R^2 de cette régression, noter $R^2_{\hat{u}\sim Z}$.
4. **Statistique de Sargan** :

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

où n est la taille de l'échantillon.

Loi asymptotique de la statistique

Sous les hypothèses suivantes :

- instruments valides : $E[Z'u] = 0$
- homoscédasticité des erreurs
- spécification correcte

alors, sous H_0 (tous les instruments sont valides) :

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

- $L - K$: nombre de **restrictions sur-identifiantes**
  - L : nb d'instruments
  - K : nb de variables endogènes instrumentées
- On peut alors calculer une **p-value** à partir de la loi χ^2_{L-K}.

Hypothèses du test de Sargan

- **Hypothèse nulle** H_0 : tous les instruments sont **exogènes**
  $\Rightarrow E[Z'u] = 0$
- **Hypothèse alternative** H_1 : au moins un instrument est **invalidé**
  (corrélé aux erreurs, mauvaise spécification, etc.)
- Attention : le test repose sur plusieurs hypothèses :
  - homoscédasticité des erreurs u_i
  - forme fonctionnelle correcte du modèle
  - aucune erreur de mesure "catastrophique" dans les variables, etc.

Interprétation du test

- On calcule $J = nR^2$ et la p-value associée à χ^2_{L-K}.
- **p-value élevée (par ex. $> 5\%$)** :
  - On **ne rejette pas** H_0.
  - On ne trouve pas de preuve contre la validité globale des instruments.
  - "Les instruments sont **globalement compatibles** avec les hypothèses d'exogénéité."
- **p-value faible (par ex. $< 5\%$)** :
  - On **rejette** H_0.
  - Au moins un des instruments est probablement corrélé à u_i.
  - "Les instruments **ne sont pas tous valides**."

Le test ne dit pas **quel** instrument est invalide, uniquement s'il y a un problème global.

Sargan vs Hansen (J-test robuste)

- La statistique de **Sargan** est valable **uniquement** sous **homoscédasticité**.
- En présence d'hétéroscédasticité (très fréquente en données micro ou panel), on utilise la version **robuste** :
  - Test de **Hansen J** (ou Sargan-Hansen), issu du cadre **GMM**.
  - Même logique : test de sur-identification avec loi χ^2_{L-K}.
  - Mais construit avec une matrice de pondération robuste à l'hétéroscédasticité.

En pratique :

- **Sargan** : 2SLS classique + hypothèse d'homoscédasticité.
- **Hansen J** : IV-GMM (ou 2SLS "robuste") + hétéroscédasticité possible.

Exemple stylisé : 1 variable endogène, 2 instruments

Modèle :

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

- x_i endogène
- instruments : z_{1i}, z_{2i} (et w_i inclus dans les instruments)

On a :

- $K = 1$ (une seule variable endogène),
- $L = 2$ (deux instruments),
  Sur-identification : $L - K = 1$ restriction testable.

Statistique de Sargan :

- Estimer y_i sur x_i (instrumenté par z_{1i}, z_{2i}) et w_i par 2SLS.
- Récupérer les résidus $\hat{u}_i$.
- Régresser $\hat{u}_i$ sur z_{1i}, z_{2i} et w_i, récupérer R^2.
- Calculer $J = nR^2$ et comparer à $\chi^2(1)$.

Exemple de sortie (interprétation en mots)

Imaginons que l'on obtienne :

- $J = 3{,}2$
- ddl $= L - K = 1$
- Valeur de la table à $5\% = 3,841$

Interprétation :

- A 5% : $3,841 > J$ on **ne rejette pas** H_0.
- On ne trouve pas de preuve que les instruments sont globalement invalides.

Si au contraire :

- $J = 7{,}9$, ddl $= 1$, $J > 3,841$

alors :

- on **rejette** H_0.
- Probablement un problème d'exogénéité d'un (ou plusieurs) instrument(s).

Limites et mises en garde

- Le test de Sargan **ne teste pas** :
  - la **pertinence** des instruments (corrélation avec x)
  - la **spécification complète** du modèle structurel
- Il peut rejeter H_0 non pas parce qu'un instrument est "mauvais", mais parce que :
  - le modèle est mal spécifié (omission d'une variable importante, non-linéarité, etc.)
  - l'homoscédasticité est violée (dans ce cas, préférer Hansen J)
- Ne pas interpréter "non rejet de H_0" comme une **preuve** que les instruments sont parfaits : c'est seulement "on ne détecte pas d'invalidité".

Lien avec les conditions d'orthogonalité

Rappel :

- Conditions d'exogénéité VI : $E[Z'u] = 0$
- Dans un modèle sur-identifié, on a plus de conditions que nécessaire pour identifier β.

Le test de Sargan :

- teste si ces conditions supplémentaires sont **consistantes entre elles**,
- en utilisant les résidus $\hat{u}$ comme approximation de u.

Interprétation en termes de moments :

- On veut que les "moments résiduels" $\frac{1}{n}Z'\hat{u}$ soient "proches de 0".
- Sargan combine ces moments en une statistique quadratique (type GMM) qui suit approximativement une loi χ^2 sous H_0.

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.