Séance de révision - MCO, Frisch-Waugh-Lovell et variables instrumentales

Support étudiant
Pierre BEAUCORAL - Magistère 3

Question directrice
Quelle variation des données détermine le coefficient estimé, et sous quelle hypothèse ce coefficient s'interprète-t-il comme un effet ?

1 Objectifs et déroulement

- Lire un coefficient en niveau, en log et avec un terme quadratique ; distinguer taille, précision et portée causale.
- Prédire et chiffrer un biais de variable omise.
- Reproduire le théorème de Frisch-Waugh-Lovell et voir la variation qu'utilisent les MCO.
- Relier première étape, forme réduite et estimateur VI ; diagnostiquer un instrument faible.
- Discuter une restriction d'exclusion et la portée d'un test de suridentification.

Temps	Travail
0 à 55 min	MCO (Card 1995) : exercices 1 et 2
55 à 60 min	Pause
60 à 110 min	Variables instrumentales (AJR 2001) : exercices 3 et 4
110 à 120 min	Transition vers le panel, billet de sortie

Les questions marquées [Papier] se traitent au fil des diapositives ; les questions [Machine] se font en binôme, sous Stata ou sous R. Chaque étiquette correspond à une pastille sur les diapositives.

2 Repères avant l'ordinateur

Ce sujet s'appuie sur les sujets de TD des séances 1 à 3 conçus par Ibrahim Goni Abdoulkadiri (Abdoulkadiri 2026). La séance reprend les deux applications des séances 1 à 3. Card part d'une équation de salaire pour mesurer le rendement de l'éducation ; le théorème de Frisch-Waugh-Lovell montre quelle partie de la variation des études identifie ce rendement. Acemoglu, Johnson et Robinson instrumentent la qualité des institutions par la mortalité des colons ; l'estimateur VI se lit alors comme le rapport de deux régressions.

Règle de lecture
Un coefficient décrit d'abord une comparaison statistique. Une interprétation causale exige une hypothèse d'identification formulée séparément.

3 Études de cas de la séance

Card (1995). Quel est le rendement causal d'une année d'études ? Hommes de la cohorte NLS Young Men, salaires de 1976, 3 010 observations. La régression multiple de la séance reproduit le tableau 2, colonne 1, de l'article ; l'instrument de l'auteur (la proximité d'une université en 1966) donne un rendement plus élevé que les MCO.

Acemoglu, Johnson et Robinson (2001). Les institutions causent-elles le développement ? 64 anciennes colonies ; la mortalité historique des colons instrumente la protection contre l'expropriation.

Références de l'étude. (Card 1995 ; Acemoglu et al. 2001).

4 Données et variables

Variable	Définition et unité
`card.dta` : 3 010 hommes, 1976	
`lwage`	Log du salaire horaire (`wage` en cents)
`educ`	Années d'études
`exper`, `expersq`	Expérience potentielle ($age - educ - 6$) et son carré
`black`, `south`, `smsa`	Indicatrices : noir, résidence dans le Sud, en zone métropolitaine
`colonial_origins.dta` : anciennes colonies ; échantillon de base `baseco == 1`	
`logpgp95`	Log du PIB par tête en 1995
`avexpr`	Protection moyenne contre le risque d'expropriation, 1985–1995 (0–10)
`logem4`	Log de la mortalité des colons
`lat_abst`	Latitude absolue de la capitale, divisée par 90
`euro1900`	Présence européenne en 1900 (second instrument)

Logiciels : Stata ou R

Stata (version 17 ou plus ; `estat weakrobust` demande la version 18) : do-file `S0_etudiant.do`.
R (4.3 ou plus) : `install.packages(c("fixest", "haven", "here"))` ; script `S0_etudiant.R`.
Ouvrir Stata ou R dans le dossier `S0_revision` décompressé.

5 Questions de préparation

1. Quelle est l'unité d'observation de chaque fichier, et quelle population représente-t-il ?
2. Quelle comparaison économique le coefficient de `educ` doit-il résumer ? Celui de `avexpr` ?
3. Quelle hypothèse faut-il pour employer le mot « effet » dans chacun des deux cas ?
4. Quelle sortie permettrait de détecter une interprétation fragile ?

6 Exercice 1 Lire un coefficient MCO

Modèle avec contrôles : $\hat{\beta}_{educ} = 0{,}0740$, $\hat{\beta}_{exper} = 0{,}0836$, $\hat{\beta}_{expersq} = -0{,}0022$.

1. [Papier 1] Interpréter $\hat{\beta}_{educ}$ en pourcentage approximatif puis exact ; calculer l'expérience qui maximise le salaire. Le R^2 triple par rapport à la régression simple : l'estimation est-elle plus crédible ?
2. [Papier 2] `exper` est omise dans la régression simple. Quel est le signe de $\hat{\delta}_{exper}$, coefficient de `educ` dans la régression de `exper` sur `educ` ? Dans quel sens la régression simple est-elle biaisée ?
3. [Papier 3] Avec l'interaction `educ` × `black`, que mesure le coefficient de `black` seul ($-0{,}4146$) ? Comment obtenir un écart interprétable ? Que mesure le coefficient de l'interaction ($0{,}0182$) ?

7 Exercice 2 Sur machine : Frisch-Waugh-Lovell et biais de variable omise

1. [Machine 1] Estimer le modèle complet. Reproduire le coefficient de `educ` en trois régressions : résidu de `educ` sur les contrôles, résidu de `lwage` sur les contrôles, régression du second sur le premier sans constante.
2. [Machine 2] Calculer $\hat{\gamma}_{exper} \times \hat{\delta}_{exper}$ et le comparer à l'écart entre la régression simple (0,052) et le modèle complet.

8 Exercice 3 Endogénéité et instrument

1. [Papier 4] Erreur de mesure classique sur le régresseur, puis sur la variable expliquée : effet sur $\hat{\beta}_{MCO}$?
2. [Papier 5] Avec `euro1900` comme second instrument, le test de suridentification ne rejette pas ($p = 0{,}80$). Peut-on conclure que les instruments sont exogènes ?
3. [Papier 6] Défendre l'instrument en quatre phrases : mécanisme de pertinence, restriction d'exclusion, menace principale, diagnostic disponible.

9 Exercice 4 Sur machine : l'estimateur VI

Sur l'échantillon de base (`baseco == 1`), avec des écarts types robustes.

1. [Machine 3] Estimer la première étape (`avexpr` sur `logem4` et `lat_abst`) et la forme réduite (`logpgp95` sur les mêmes variables). Calculer le rapport des deux coefficients de `logem4`, puis estimer le modèle par doubles moindres carrés. Comparer.
2. [Machine 4] Lire le F robuste de première étape et calculer le test d'Anderson-Rubin de $\beta = 0$. L'instrument est-il faible ? Que peut-on conclure, et que ne peut-on pas conclure ?

10 Production attendue

– Une vérification numérique de Frisch-Waugh-Lovell et du terme de biais lié à l'expérience.
– Le rapport forme réduite sur première étape, comparé à la 2SLS, avec le diagnostic de l'instrument.
– Une défense écrite de l'instrument en quatre phrases.

Billet de sortie [Billet]
En six lignes, pour le coefficient de `educ` (0,074) puis pour le coefficient VI de `avexpr` (0,996) : quelle variation identifie le coefficient, quelle hypothèse permet une interprétation causale, quelle menace subsiste ?

Espace pour la synthèse

Références

Abdoulkadiri, Ibrahim Goni (2026). *Économétrie, Magistère 3 : sujets de travaux dirigés, séances 1 à 5*. Supports étudiants.

Acemoglu, Daron, Simon Johnson et James A. Robinson (2001). "The Colonial Origins of Comparative Development : An Empirical Investigation". In : *American Economic Review* 91.5, p. 1369-1401. doi : 10.1257/aer.91.5.1369.

Card, David (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling". In : *Aspects of Labour Market Behaviour : Essays in Honour of John Vanderkamp*. Sous la dir. de Louis N. Christofides, E. Kenneth Grant et Robert Swidinsky. Toronto : University of Toronto Press, p. 201-222.