Économétrie II

L3 Économétrie – L3 MASS

Ch. 5. $\exists\, i : E(\varepsilon_i x_i) \neq 0$: Endogénéité

Prof. Philippe Polomé, U. Lyon 2

Année 2015-2016

Rappel

1. $E(\varepsilon_i) = 0 \; \forall i$: **Espérance nulle**
2. ✓ $var(\varepsilon_i) = \sigma^2 \; \forall i$: **Homoscédasticité**
3. ✓ $cov(\varepsilon_t, \varepsilon_s) = 0 \; \forall t \neq s$: **Pas d'autocorrélation**
4. $E(\varepsilon_i x_i) = 0 \; \forall i$: **Exogénéité**
5. ✓ La matrice X est de plein rang : **Pas de multicolinéarité**
6. Le modèle est **correctement spécifié**
7. La variable dépendante Y est **continue**

Table des matières

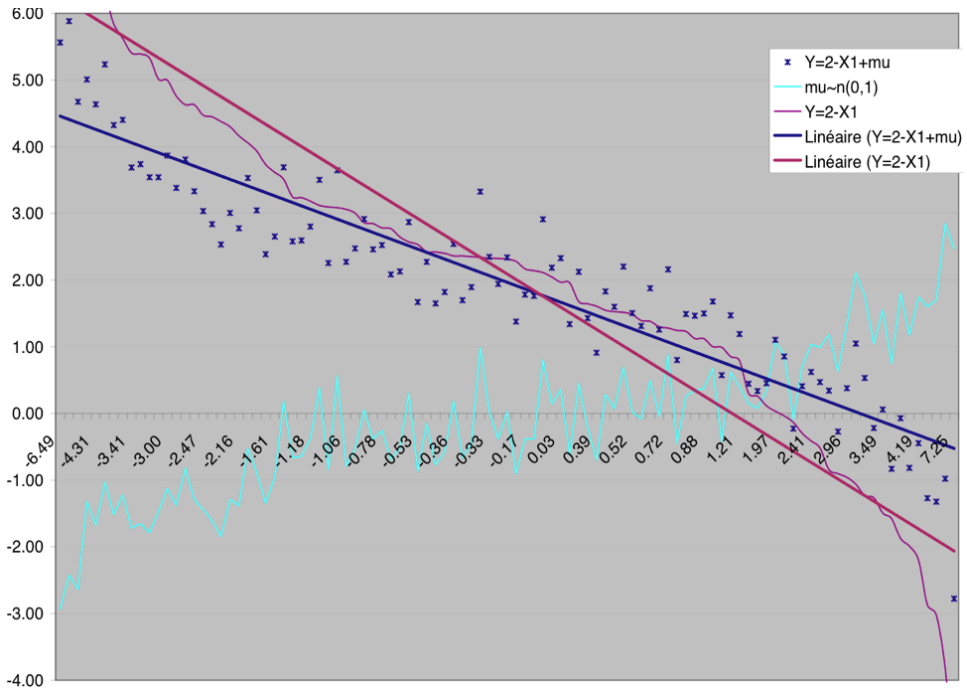
Table des matières

Définition

- Exogénéité :
  - Coupe transv. $E(\varepsilon_i X_i) = 0$
    - Pas de corrélation entre l'erreur et chaque régresseur pour un même i
    - On écrit aussi $E(\varepsilon_i | X_i) = 0$: **espérance conditionnelle nulle**
  - Série temp : $E(\varepsilon_t x_t) = 0 \, \forall t \, \forall x$ **pas de corrélation contemporaine**
    - Lorsque $E(\varepsilon_t | x_s) = 0 \, s = 1, \ldots, T$, x est **strictement exogène**
- Rupture de cette hypothèse = **endogénéité**
- Interprétation : Un choc aléatoire ε induit un choc sur Y et sur X pour un même i
  - Donc difficile de séparer les effets "confondants"

Conséquence : **Inconsistance** de l'estimateur MCO

- Corrélation positive : à des valeurs élevées (basses) de ε correspondent des valeurs élevées (basses) de X
  - ε grand : $Y > X\beta$ et ε petit : $Y < X\beta$
  - Donc : droite estimée par MCO pente plus forte que la réalité



- Monte Carlo : fichier tableur en ligne **Endogeneite.ods**

Pourquoi l'endogénéité est-elle un problème ?

- Ne vaut-il pas mieux prédire Y le mieux possible ?
- Trois cas
  - **Prédiction** : on veut prédire Y conditionnellement à X
    - Si on connait X "ε inclus", $x(\varepsilon)$, ce qui n'est pas évident,
    - Dans ce cas, l'effet de l'erreur sur X est inclus, donc prédiction MCO $\hat{Y}$ correcte
  - **Contrôle** : on choisit X, quel sera Y ? [p.e. effet d'une politique]
    - Si on choisit X, l'erreur n'y est pas, donc prédiction MCO incorrecte
  - Si l'on souhaite **comprendre** la relation entre Yet X il faut traiter l'endogénéité
- Dans les 2 derniers cas : ce n'est pas une bonne idée "d'ajuster une droite au mieux" dans le nuage de points

5 sources de l'endogénéité

1. Hétérogénéité inobservée
2. Erreur de mesure
3. Simultanéité
4. Échantillonnage endogène
5. Autocorrélation en séries temporelles

Table des matières

2 cas d'hétérogénéité inobservée

- Variable omise
- Coefficients aléatoires

Variable omise

- Le modèle correctement spécifié est $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  - Mais le modèle estimé est $Y = \beta_0 + \beta_1 x_1 + \nu$
- L'effet du régresseur manquant se retrouve dans l'erreur du modèle estimée : $\nu = \beta_2 x_2 + \varepsilon$
  - = hétérogénéité inoberservée : Des facteurs inobservés affectent à la fois la variable expliquée et un régresseur
- Si le régresseur manquant est corrélé à un régresseur présent
  - Alors le terme d'erreur du modèle estimé est corrélé avec au moins un régresseur présent
  - De plus vraisemblablement :
    - Hétéroscédasticité si $var(x_{2t}) \neq var(x_{2s}), t \neq s$
    - Autocorrélation si $corr(x_{2t}, x_{2s}) \neq 0, t \neq s$
    - $E(\nu) \neq 0$ l'intercept du modèle est biaisé

Que faire en cas de variable omise ?

1. **Ignorer** le problème : inconsistance des estimateurs
2. Essayer de trouver un **proxy** acceptable pour la variable inobservée
   - Proxy = mesure approximative de la variable inoberservée (ci-dessous)
3. Si données de panel et si la variable inobservée ne change pas dans le temps (mais seulement entre les agents)
   - **Modèle "à effets fixes"** (programme de M2)
4. Laisser la variable inobservée dans le terme d'erreur mais utiliser un estimateur qui reconnait sa présence
   - Estimateur **Variable Instrumentale** ci-dessous

Proxy

- Variable inobservée : on sait que le modèle devrait inclure un régresseur, mais on n'a pas de donnée
- Modèle $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  - x_2 pas observée
- **Proxy** z pour x_2 :
  - z observée mais pas explicative dans le modèle
  - z corrélée à x_2 : $x_2 = \delta_0 + \delta_1 z + \mu$
    - On ne peut tester cette corrélation puisque x_2 pas observée
  - P.e. (salaire) éducation x_2 et nombre d'années d'étude z
- La proxy n'est pas une erreur de mesure
  - Ni un instrument (plus loin)

Utilisation d'une proxy

- La variable proxy est **substituée** à la variable inobservée dans $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  - On peut estimer $Y = \pi_0 + \pi_1 x_1 + \pi_2 z + \xi$
  - Que dit ce modèle sur le précédent ?

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 (\delta_0 + \delta_1 z + \mu) + \varepsilon \\ &= \beta_0 + \beta_2 \delta_0 + \beta_1 x_1 + \beta_2 \delta_1 z + \beta_2 \mu + \varepsilon \\ &= \pi_0 + \pi_1 x_1 + \pi_2 z + \xi \end{aligned}$$

- Donc si μ n'est pas corrélé avec x_1, estimer $Y = \pi_0 + \pi_1 x_1 + \pi_2 z + \xi$ par MCO
  - Sans bias et consistant pour $\beta_1 = \pi_1$
  - Les autres coef. π_0 et π_2 n'ont pas d'interprétation directe

Coefficients aléatoires

- Autre forme d'hétérogénéité inobservée
- Modèle vrai $Y_i = \beta_0 + \xi_{1i} x_{1i} + \eta_i$ avec ξ_{1i} aléatoire t.q.
  - $\xi_{1i} = \gamma_1 + \mu_{1i}$
  - γ_1 pas aléatoire (pourrait dépendre de régresseurs)
  - μ_{1i} est un bruit blanc
  - P.e. rendement éducation
- On estime $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$ donc
  - $\varepsilon_i = \mu_{1i} x_{1i} + \eta_i$
  - $\beta_1 = \gamma_1$
- Solution
  - Variable instrumentale (ci-dessous)
  - Modélisation explicite par Maximum de Vraisemblance (on ne voit pas)

Table des matières

Définition & traitement

- Modèle $y = \beta_0 + \beta_1 x_1^* + \varepsilon$
  - On n'observe pas x_1^* mais bien $x_1 = x_1^* + \nu$
    - ν est une erreur de mesure
  - "Classical Error-in-Variables" (CEV)
- Équation estimée : avec x_1
  - $y = \beta_0 + \beta_1 (x_1^* + \nu) + (\varepsilon - \beta_1 \nu) = \beta_0 + \beta_1 x_1 + \mu$
  - Donc, $cov(x_1, \mu) = cov(x_1^* + \nu, \varepsilon - \beta_1 \nu) = -\beta_1 \sigma_\nu^2 \neq 0$
    - Pour autant que erreur de mesure ν pas corrélée avec x_1^*
- Les erreurs de mesure sont la norme
  - Endogénéité pas toujours préoccupante
- Solution : variable instrumentale ci-dessous
- Une erreur de mesure sur y accroît la variance des erreurs
  - mais ne cause pas d'endogénéité

Table des matières

Définition & exemples

- 2 variables sont causales l'une de l'autre
- $y = \beta_0 + \beta_1 x + \varepsilon$ (x cause y) **et** $x = \gamma_0 + \gamma_1 y + \mu$ (y cause x)
  - Donc : $x(y)$ mais $y(\varepsilon)$ donc $x(\varepsilon)$
    - d'où corrélation entre x et l'erreur dans $y = \beta_0 + \beta_1 x + \varepsilon$
- Exemples
  - Publicité et vente :
    - La publicité accroît/soutient les ventes
    - Le budget publicité est calculé en proportion aux ventes
  - Fonction de coût $C(Q, W)$
    - Q = production, W vecteur des prix d'intrants
    - Lemme de Sheppard $\partial C / \partial W = d(Q, W)$: demande d'intrants est fonction de la production mais production est fonction des intrants utilisés

Exemple de simultanéité : modèle keynésien

- Deux équations : **forme structurelle** = forme économique
  - Consommation $C_t = \beta_0 + \beta_1 Y_t + \varepsilon_t$ avec Y le PIB
  - Identité comptable $Y_t = C_t + I_t$ avec I l'investissement, ici exogène
    - Économie fermée sans état
- La consommation et le revenu sont donc déterminés **simultanément**
  - C et Y sont deux endogènes

Forme réduite

- **Forme réduite** = toutes les endogènes à gauche
  - $C_t = \frac{1}{1-\beta_1}[\beta_0 + \beta_1 I_t + \varepsilon_t] = \delta_0 + \delta_1 I_t + \mu_t$
  - $Y_t = \frac{1}{1-\beta_1}[\beta_0 + I_t + \varepsilon_t] = \gamma_0 + \gamma_1 I_t + \nu_t$
- Clairement, Y est corrélé à ε
  - DONC : **Endogénéité dans le modèle structurel** en estimant l'équation de consommation, même seule

Moindres Carrés Indirects

- MCO **toujours** consistant pour forme réduite
- **Identification** : coef. forme structurelle peuvent-ils être récupérés de la forme réduite ?
  - Ici, en estimant chaque équation de la FR : $\hat{\delta}_0, \hat{\delta}_1, \hat{\gamma}_0, \hat{\gamma}_1$
  - On calcule les coef. structurels par $\beta_1 = \dfrac{\delta_1}{1+\delta_1}$ ect
- Si **une seule** manière de récupérer **tous** les coef. structurels : système **exactement identifié**
  - **Moindres Carrés Indirects MCI** = appliquer MCO à la forme réduite & résoudre pour obtenir les coef. structurels
- Si certains coef. ne peuvent être ainsi retrouvés : **sous-identifié**
- Si certains coef. retrouvés de **plus d'une** manière : **sur-identifié**
  - "Bonne" manière ? Variable Instrumentale

Table des matières

Méthode des Moments MM

- Interprétation d'inversion de MCO dite "méthode des moments"
- Soit A un estimateur de β dans $Y = X\beta + \varepsilon$
  - alors on peut écrire $Y - XA = \hat{\varepsilon}$
- Hypothèse exogénéité $E(X\varepsilon) = 0$
- Stratégie "**méthode des moments**"
  - Cette condition sur les moments de la **population** est **imposée** aux moments de **l'échantillon**
  - C'est-à-dire on veut A tel que $X'\hat{\varepsilon} = 0$
  - Donc : $X'(Y - XA) = 0$: ce sont les CPO de moindres carrés
  - Donc :$A = (X'X)^{-1}X'Y = \hat{\beta}$

Variables instrumentales

- Hypothèse exogénéité $E(X\varepsilon) = 0$ ne tient plus
- Supposons qu'on puisse trouver un ensemble de variables Z ou "**instruments**" telles que
  - Z et X soient de mêmes dimensions
  - $E(\varepsilon|Z) = 0$
  - $Corr(Z, X)$ soit élevée
- Donc Z permet d'inverser la relation $Y = X\beta + \varepsilon$
  - Via $Z'Y = Z'XA + Z'\varepsilon$, on a $(Z'X)^{-1}Z'Y = A + (Z'X)^{-1}Z'\varepsilon$
  - Si on a $Plim\, Z'\varepsilon = 0$ (à la limite Z et ε ne sont pas corrélés)
  - Alors : **Estimateur Méthodes des Moments** :
    - A t.q. $Z'(Y - XA) = 0$
    - $A = (Z'X)^{-1}Z'Y = \hat{\beta}_{VI}$
    - $\equiv$ **Estimateur Variables Instrumentales** VI

Propriétés de $\hat{\beta}_{VI}$

- Il s'agit d'un estimateur alternatif à MCO
  - En général **biaisé**
  - **Consistant** (si les instruments sont **valides**, voir ci-dessous)
- On peut démontrer que $\Sigma_{\hat{\beta}_{VI}} = \sigma_\varepsilon^2 \left(Z'X\right)^{-1} \left(Z'Z\right) \left(Z'X\right)^{-1}$
  - Cette variance est d'autant plus faible que la corrélation entre Z et X est forte
  - À la limite si Z et X non-corrélés, alors $Z'X \to 0$ et $\Sigma_{\hat{\beta}_{VI}} \to \infty$
- MCO peut être vu comme VI avec $Z = X$
  - $Corr(Z, X) = 1$
  - Donc : si pas endogénéité, MCO plus efficient que VI
- $\neq$ remplacer X par Z dans $Y = X\beta + \varepsilon$ [cfr. proxy]
  - Si on le faisait, le modèle serait $Y = Z\gamma + \mu$
  - Et l'estimateur MCO serait $\hat{\gamma}_{MCO} = \left(Z'Z\right)^{-1} Z'Y$

Exemple : Equation de salaire

- Eq mincérienne de salaire
  - $\ln w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \varepsilon$
  - w salaire ; $educ$ nbr années études ; $exper$ nbr années expérience
- Capacités Intellectuelles Intrinsèques (CII) de l'individu
  - Inobservées / inobservables
  - Corrélées $\oplus$ avec niveau d'éducation : $educ = \alpha_0 + \alpha_1 CII + \nu$
  - Corrélées $\oplus$ avec niveau de salaire : $\ln(w) = \delta_0 + \delta_1 CII + \mu$
- Rendement de l'éducation estimé par eq mincérienne
  - Sur- ou sous-estimé ?
- Données **card.gdt** de Wooldridge
  - Définir $\ln wage$, $exper^2$ puis GMM 1 étape
  - Instrument possible : proximité à un "college4"
  - On en verra d'autres plus loin

Instruments & Tests

- La difficulté fondamentale est de **trouver** des instruments
  - On verra quelques cas
  - En séries temporelles & panels : valeurs passées (retards)
  - En systèmes d'équations : régresseurs dans d'autres équations
- Avec VI, il faut exactement un instrument par variable de X (identification exacte)
  - Les variables non-endogènes sont leurs propres instruments
  - Plus d'un instrument pour une variable $\Longrightarrow$ il faut généraliser la méthode
- Les tests d'inférence n'ont plus de valeur qu'asymptotiquement
  - Le bootstrap reste valide
  - Le R^2 n'a plus de sens

Validité des instruments

- Difficile de tester l'exogénéité des instruments $cov(Z, \varepsilon) = 0$
  - Test OverId + loin
  - Si Z n'est pas exogène, VI sera inconsistant (par construction)
  - $\Rightarrow$ VI ne s'applique pas en cas d'échantillonnage endogène
- On peut mesurer la corrélation entre Z et X
- Soit $Y = \beta_0 + \beta_1 x + \varepsilon$
  - x est endogène, on a un instrument z
  - Si $cov(z, \varepsilon) \neq 0$ on peut montrer que $Plim\, \hat{\beta}_{1VI} = \beta_1 + \dfrac{cov(z, \varepsilon)}{cov(z, x)}$
    - Donc que si $cov(z, \varepsilon) \neq 0$ alors $\hat{\beta}_{VI}$ est inconsistant
  - De plus $Plim\, \hat{\beta}_{1VI} = \beta_1 + \dfrac{\sigma_\varepsilon corr(z, \varepsilon)}{\sigma_x corr(z, x)}$
    - Donc, si $corr(z, \varepsilon) \neq 0$ même faible, alors si $corr(z, x)$ est faible (mauvais instrument), $Plim\, \hat{\beta}_{1VI}$ ne sera pas proche de β_1

Illustration d'un mauvais instrument : Poids à la naissance

- Données **bwght.gdt** Wooldridge
  - Poids de l'enfant à la naissance (bwght) en log en fonction de
    - consommation de tabac (packs)
    - revenu familial (faminc) prix comme **proxy** d'autres facteurs (accès aux soins, ...)
    - On peut rajouter d'autres régresseurs
- La consommation de tabac pourrait être endogène
  - P.e. stress (= hétérogénéité inobservée)
- Instrument : *cigprice* prix des cigarettes
  - Équation d'instrumentation (ci-dessous)
    - ou corr(cigprice,packs)
    - On voit que c'est un mauvais instrument

Table des matières

Instrumentation

- Application particulière de VI
- Soit **l'équation structurelle** $Y = X\beta + \varepsilon$
  - Supposons que dans X, x_k soit endogène
    - et qu'on dispose d'un instrument z pour x_k
  - La matrice d'instruments serait Z,
    - identique à X sauf dernière colonne : remplacer x_k par z
- **Équation d'instrumentation**
  $x_k = \delta_0 + \delta_1 x_1 + \ldots + \delta_{k-1} x_{k-1} + \delta_k z + \mu = Z\delta + \mu$
  - Estimation MCO, valeurs ajustées de x_k
    - $\hat{x}_k = \hat{\delta}_0 + \hat{\delta}_1 x_1 + \ldots + \hat{\delta}_{k-1} x_{k-1} + \hat{\delta}_k z = Z\hat{\delta}$
  - On voit que $\hat{x}_k$ est un instrument valide pour x_k
    - si z est un instrument valide pour x_k
  - $\hat{X}$ la matrice X dans laquelle on a remplacé x_k par $\hat{x}_k$

MC en 2 Étapes

- Estimateur VI avec $\hat{X}$ est $\hat{\beta}_{VI} = \left(\hat{X}'X\right)^{-1}\hat{X}'Y$
  - Meilleur que $(Z'X)^{-1}Z'Y$ car la corrélation entre $\hat{x}_k$ et x_k au moins aussi élevée que entre z et x_k
- L'estimateur VI est équivalent à une estimation MCO en deux étapes (MC2E) :

1. Estimation de l'équation d'instrumentation $x_k = Z\delta + \mu$
2. Remplacer X par $\hat{X}$ dans l'équation structurelle
   - $Y = \pi\hat{X} + \nu$ (régression de 2nde étape)
   - et on estime par MCO
     - $\hat{\pi}_{MC2E} = \left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'Y$: c'est $\left(\hat{X}'\hat{X}\right)^{-1}$ et pas $\left(\hat{X}'X\right)^{-1}$ comme dans $\hat{\beta}_{VI}$
     - On peut montrer que $\hat{\pi}_{MC2E} = \hat{\beta}_{VI} = \left(\hat{X}'X\right)^{-1}\hat{X}'Y$

$\hat{\pi}_{MC2E} = \hat{\beta}_{VI} = \left(\hat{X}'X\right)^{-1}\hat{X}'Y$: Preuve

- On note qu'on peut écrire $\hat{X} = Z\left(Z'Z\right)^{-1}Z'X$
  - Pour la dernière colonne de $\hat{X}$, c'est $Z\hat{\delta}$
  - Pour les autres, ce sont les colonnes de X (exogènes)

$$\begin{aligned}
\hat{\pi}_{MC2E} &= \left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'Y \\
&= \left(X'Z\left(Z'Z\right)^{-1}Z'\left(Z\left(Z'Z\right)^{-1}Z'X\right)\right)^{-1}\hat{X}'Y \\
&= \left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}\hat{X}'Y \\
&= \left(\hat{X}'X\right)^{-1}\hat{X}'Y = \hat{\beta}_{VI}
\end{aligned}$$

Exemple : Equation de salaire

- Rendement de l'éducation
  - Estimé par $\ln w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \varepsilon$
  - Sur- ou sous-estimé ?
- Données **card.gdt** de Wooldridge
  - Instrument : proximité à un "college4"
  - Equation d'instrumentation $educ = \left(cst, exper, exper^2, college4\right)$
  - 2SLS : automatisé (Gretl "DMC") et "à la main" en mettant $\widehat{educ}$ comme régresseur dans l'équation de salaire
    - Mêmes coefficients, pas les mêmes t-tests
    - Mêmes résultats qu'avec VI (GMM 1 instrument) exemple antérieur

Plusieurs instruments

- Il faut au moins un instrument par variable explicative
  - Les exogènes sont leurs propres instruments
- Dans certains cas, on dispose de plus d'un instrument pour certains régresseurs
  - MC2E : on voit tout de suite comment intégrer ces instruments supplémentaires via la (ou les) équations d'instrumentation
- On peut démontrer que
  - parmi toutes les manières d'utiliser/combiner ces différents instruments
  - **MC2E est la plus efficiente**

Remarques

- Ne pas faire la régression en 2 étapes explicitement
  - utiliser la commande MC2E (2SLS)
  - sinon à la 2ème étape le logiciel va calculer une matrice de var cov selon la formule MCO et pas VI
- Des estimations robustes à l'hétéroscédasticité sont généralement disponibles pour MC2E et VI

Exemple : Equation de salaire

- Rendement de l'éducation
  - Estimé par $\ln w = \beta_0 + \beta_1 educ + \beta_2 exper + \varepsilon$
  - Sur- ou sous-estimé ?
- Données **card.gdt** de Wooldridge
  - Instruments : Education de la mère et du père
  - Equation d'instrumentation
    $educ = (cst, exper, exper^2, Meduc, Feduc)$
  - Échantillon : drop missing values – 2SLS

Résumé

- Identification Exacte : 1 ! instrument par régresseur endogène
  - Utiliser Variable Instrumentale $\equiv$ MM
  - MC2E = + efficient des estimateurs VI
- Sous-identification : manque au moins un instrument
  - Estimation consistante impossible
- Sur-identification : plus d'un instrument pour au moins un régresseur endogène
  - MC2E avec eq d'instrumentation à pls instruments
  - = un cas de MM Généralisée : GMM

Table des matières

Test 1. Hausman : Endogénéité

	MCO	MC2E
Aucun régresseur endogène	consistant efficient	consistant inefficient
Au moins 1 régresseur endogène	inconsistant	consistant

- Donc : si endogénéité : MC2E – mais sinon : MCO
- H_0 : égalité des coefficients $\equiv \hat{\beta}_{MCO} - \hat{\beta}_{MC2E} = 0$
  - Si égaux, alors pas endogénéité : on préfère MCO
  - Sinon, on prend MC2E
- Test disponible sur **tous** les logiciels économétriques
  - Entre n'importe quelle paire d'estimateurs avec un consistant
    - VI contre MCGF par exemple
    - $var\left(\hat{\beta}_{MCO} - \hat{\beta}_{MC2E}\right)$ peut poser problème
- Aussi une bonne idée de comparer directement $\hat{\beta}_{MCO}$ et $\hat{\beta}_{MC2E}$

2 autres tests : Définitions

Equation structurelle : $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \mu_1$

- x_1 et x_2 sont exogènes
- On a aussi 2 autres exogènes x_3 et x_4
  - Qui ne sont pas dans l'équation structurelle
  - Qui sont corrélés à y_2
- On veut tester l'endogénéité de y_2

Forme réduite pour y_2 : $y_2 = \pi_0 + \pi_1 x_1 + \pi_2 x_2 + \pi_3 x_3 + \pi_4 x_4 + \nu_2$

Test 2. Test de régression / de corrélation des erreurs / Durbin–Wu–Hausman

- On veut tester l'endogénéité de y_2 dans eq structurelle
- Chaque x_j est non-corrélé avec μ_1
  - y_2 non-corrélé avec μ_1 **ssi** v_2 non-corrélé avec μ_1
- Estimer $y_2 = \pi_0 + \pi_1 x_1 + \pi_2 x_2 + \pi_3 x_3 + \pi_4 x_4 + v_2$ par MCO (consistant)
  - On obtient $\hat{v}_2$: une approximation à v_2
- Estimer $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \delta_1 \hat{v}_2 + \textit{erreur}$ par MCO
  - $\hat{v}_2$ significatif (t-stat) $\implies$ v_2 manquant dans eq struct.
    - v_2 est partie de μ_1, donc corrélés, donc y_2 endogène
  - $\hat{v}_2$ non-significatif n'implique rien

Test de régression : remarque

- On peut montrer que
  - $\hat{\beta}_{MCO}$ de $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \delta_1 \hat{v}_2 + erreur$ est identique à
  - $\hat{\beta}_{MC2E}$ de $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \mu_1$
- C'est une 2ème interprétation de $MC2E$
  - inclure $\hat{v}_2$ dans la régression MCO "nettoie" l'endogénéité de y_2
- Test sur plusieurs variables endogènes
  - tester conjointement (test F) la significativité des résidus de chaque équation d'instrumentation

Test 3. "OverID" Restrictions sur-identifiées : Exogénéité de l'instrument

- Si un seul instrument pour un régresseur endogène
  - **Impossible** de tester l'absence de corrélation entre l'instrument et le terme d'erreur : $corr(z, \varepsilon) = 0$
  - Modèle "juste / exactement" identifié
- Si on dispose de plusieurs instruments,
  - **Possible** de tester l'exogénéité d'un instrument
  - Le modèle est "sur-identifié"
- Dans notre exemple : x_3 et x_4 peuvent servir d'instruments pour y_2 dans l'équation structurelle $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \mu_1$

Étapes du test OverID

1. Estimer l'équation structurelle par VI en utilisant **seulement** x_3 comme instrument
   1.1 Calculer résidu $\hat{\mu}_{1MC2E} = y_1 - \hat{\beta}_0 + \hat{\beta}_1 y_2 + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_2$
2. Régresser résidu $\hat{\mu}_{1MC2E}$ sur **toutes** les variables **exogènes** du modèle (explicatives + instruments)
   2.1 $\hat{\mu}_{1MC2E} = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 + \xi$
   - Ce résidu est une approximation de μ_1 de l'éq. struct.

   2.2 Calculer le R^2 de cette régression
   - Si exogénéité R^2 devrait être faible
3. Sous l'hypothèse nulle (exogénéité de x_4) : $nR^2 \overset{a}{\sim} \chi^2_q$
   - q : nombre d'instruments en excès
     - nombre d'instruments total dans le modèle moins nombre de régresseurs endogènes
     - "Over-identification", ici $q = 1$ car 2 instruments x_3 et x_4 et un régresseur endogène y_2

OverID : remarques

- Il faut faire l'hypothèse que x_3 est exogène : on ne peut la tester
  - Si $nR^2 > \chi^2_{q;0.95}$ on rejette que
    - x_4 est exogène
    - OU que x_3 est exogène
  - Hypothèse de un instrument valide par régresseur endogène
- Test implémenté directement dans beaucoup de logiciels

Exemple : Equation de salaire

- Rendement de l'éducation
  - Estimé par $\ln w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \varepsilon$
  - Sur- ou sous-estimé ?
- Données **card.gdt** de Wooldridge
  - Instruments : Education de la mère et du père
  - Equation d'instrumentation
    $educ = (cst, exper, exper^2, Meduc, Feduc)$
  - Échantillon : drop missing values – 2SLS
  - Test d'endogénéité :
    - Test de Hausman et OverId (Sargan) dans sortie "DMC"
    - Test de régression : Résidu de l'équation de d'instrumentation dans MCO de l'équation de salaire
  - Autre instrument possible proximité à un "college4"

Table des matières

4° source d'endogénéité : Échantillonnage

- Si on n'observe pas un échantillon "purement" aléatoire ("simple")
  - mais plutôt un échantillon **sélectionné** dans lequel seuls certains individus sont admis
  - ou bien avec des **données manquantes**
- 3 cas
  - Sélection – ou **attrition** – purement aléatoire, ou basée sur des variables aléatoires exogènes
    - Pas de problème
  - Sélection basée sur un régresseur x_j corrélé à la dépendante y
    - Généralement pas de problème
  - Sélection basée sur dépendante y
    - Un problème d'**échantillon sélectionné** ou **troncature** se pose

Sélection basée sur un régresseur x_j corrélé à y

- Exemple : On estime une équation de salaires, mais on observe plus d'attrition pour les niveaux d'éducation faibles
  - Sans que cette attrition soit corrélée au revenu (salaire) par classe d'éducation
  - On observe l'éducation, mais pas le salaire
- Les statistiques descriptives sont biaisées
  - par exemple, le salaire moyen sera plus élevé que dans la réalité
- Les MCO restent sans biais et consistants
  - Les estimations "contrôlent" les dimensions des variables explicatives
- Pas de problème tant qu'il y a assez de variabilité dans les variables explicatives pour identifier les effets mesurés

Sélection sur y – cas 1. Troncature (Truncation)

- Inclusion dans l'échantillon est $y_i \leq c_i$ (sélection de troncature)
- Alors $\varepsilon_i \leq c_i - X_i\beta$
  - Car $y_i = X_i\beta + \varepsilon_i$
  - Donc : la sélection de troncature introduit une **corrélation contemporaine** entre l'erreur et les régresseur(s)
- Notes
  - On n'observe aucun $y_i > c_i$ ni aucun X_i correspondant
  - La sélection peut aussi être $y_i \geq c_i$

Sélection sur y – cas 2. Troncature accidentelle

- Modèle bivarié de sélection de l'échantillon (MBSE)
  - **Équation de participation** $Y_1 = \begin{cases} 1 & si\ Y_1^* > 0 \\ 0 & sinon \end{cases}$
  - **Équation de résultat** $Y_2 = \begin{cases} Y_2^* & si\ Y_1^* > 0 \\ - & sinon \end{cases}$
  - Donc : On n'observe Y_2^* que si $Y_1^* > 0$, c'est-à-dire que si on observe $Y_1 = 1$
- On suppose que la réalité est $\begin{aligned} Y_1^* &= X_1\beta_1 + \varepsilon_1 \\ Y_2^* &= X_2\beta_2 + \varepsilon_2 \end{aligned}$

Troncature accidentelle : origine du biais

- Il semble raisonnable
  - de supposer que le terme d'erreur ε_1 de l'équation de participation peut être corrélé au terme d'erreur ε_2 de l'équation de résultat : $\varepsilon_1(\varepsilon_2)$
  - que certains régresseurs au moins soient communs entre X_1 et X_2
    - $X_1 \cap X_2 = X_{21}$
    - Le reste se nomme X_{22} et X_{11}
- On peut écrire l'équation de participation comme $X_{11}\beta_{11} + X_{21}\beta_{12} > \varepsilon_1(\varepsilon_2)$
  - **Donc**, la troncature accidentelle provoque une corrélation entre X_{21} et ε_2
  - et donc entre X_2 et ε_2 : endogénéité dans l'équation de résultat

Autre interprétation : moyenne conditionnelle de Y_2

- La moyenne de Y_2 conditionnellement à X_2 dépend de Y_1^* car si $Y_1^* \leq 0$, on n'observe pas Y_2
  - On suppose pour simplifier que X_2 est non-endogène
    - $E(X_2|\varepsilon_1) = E(X_2|\varepsilon_2) = X_2$
  - Donc $$\begin{aligned} E(Y_2|X_2, Y_1^* > 0) &= E(X_2\beta + \varepsilon_2 | X_1\beta_1 + \varepsilon_1 > 0) \\ &= X_2\beta_2 + E(\varepsilon_2|\varepsilon_1 > -X_1\beta_1) \end{aligned}$$
- Donc
  - Si ε_2 et ε_1 sont indépendants, le dernier terme est nul
  - Sinon, il faut corriger la moyenne conditionnelle pour **le biais de sélection** (ou **troncature accidentelle**)
  - Et en particulier **MCO** de Y_2 sur X_2 sera **biaisé** et **inconsistant**

Exemple : Equation de salaire

- Le salaire dépend de caractéristiques comme le niveau d'étude, l'age, le sexe, le nombre d'enfants...
- On n'observe un salaire que pout ceux/celles qui participent au marché du travail
- La décision de participer au marché du travail dépend certainement de facteurs similaires à ceux expliquant le salaire
- Donc équation de sélection : participation
- Équation de résultat : salaire
- Corrélation entre les deux
- MCO équation de salaire : biaisé et inconsistant

Estimation

- Rappel : Sources 1 (hétérogénéité inobservée), 2 (erreurs de mesure) et 3 (simultanéité) peuvent être adressées par Variable Instrumentale / MC2E
- Source 4 Sélection d'échantillonnage
  - VI inutile car VI a le même problème d'échantillonnage
  - La solution passe par une modélisation du processus de sélection :
    - Plusieurs estimateurs alternatifs (Heckman) – on regardera en M1
- En résumé
  - Toujours se poser la question de l'échantillonnage
  - Pourquoi certaines données sont manquantes

Table des matières

5° source : Autocorrélation en séries temporelles

- Exogénéité en série temp : $E(\varepsilon_t x_t) = 0 \, \forall t \, \forall x$ **pas de corrélation contemporaine**
  - On écrit aussi $E(\varepsilon_t | x_t) = 0 \forall t \, \forall x$: **espérance conditionnelle nulle**
    - Ce qui est la même chose
  - Lorsque $E(\varepsilon_t | x_s) = 0 \, s = 1, \ldots, T$ on dit que x est **strictement exogène**
    - ε_t n'est corrélé à aucun régresseur à aucune période
- En série temp., on ne fait pas l'hypothèse d'absence d'autocorrélation
  - L'absence de corrélation contemporaine suffit à ce que MCO soit consistant
    - Si les séries sont $I(0)$
  - Pour que MCO soit non-biaisé il faut l'exogénéité stricte

Implication de l'exogénéité stricte

Soit le modèle statique du taux de meurtre en fonction du nombre de policiers / habitant

$$TxMeurtre_t = \beta_0 + \beta_1 Pol/h_t + \varepsilon_t$$

1. Pol/h ne peut avoir aucun effet retardé sur $TxMeurtre$
   - Sinon, Pol/h_{t-1} serait dans ε_t ce qui romperait l'exogénéité stricte
2. ε_t ne peut causer aucun changement futur de Pol/h
   - Supposons que la ville ajuste Pol/h sur base des valeurs passées de $TxMeurtre$, alors Pol/h_{t+1} est corrélée avec ε_t

- Facile que l'exogénéité stricte ne tienne pas

Devoir #6 : VI

- Reprenez de ma feuille Tableur l'exemple avec un régresseur endogène
- Générez un instrument
  - Valide (= non-corrélé avec le terme d'erreur)
  - Bon (= corrélation élevée avec le régresseur endogène)
  - Basez-vous directement sur la façon dont le régresseur a été généré
- Estimez les coefficients du modèle par VI
- Examinez comment les coefficients estimés varient en fonction de la corrélation de l'instrument avec le régresseur
- Illustrez qu'un instrument non-valide amène à l'inconsistance
  - Examinez le degré d'inconsistance en fonction de la corrélation de l'instrument avec le régresseur