Master MIASH C2ES, 1ère année: Économétrie 2

Endogénéité et estimation par variables instrumentales

Michal Urdanivia, Université de Grenoble Alpes, Faculté d'Économie, GAEL
Courriel: michal.wong-urdanivia@univ-grenoble-alpes.fr

Année universitaire 2016-2017

1. Introduction

Dans ce notes nous commençons le traitement de la méthode des variables instrumentales(VIs par la suite) pour corriger le problème de l'endogénéité des régresseurs dans le modèle de régression.

2. Endogénéité

Considérons le modèle de régression partitionné suivant,

$$\begin{aligned} Y_i &= X_i^\top \beta + U_i \\ &= X_{i1}^\top \beta_1 + X_{i2}^\top \beta_2 + U_i \end{aligned} \tag{1}$$

où X_{i1} est un vecteur $(K_1 \times 1)$ et X_{i2} est un vecteur $(K_2 \times 1)$, de régresseurs, β_1 est un vecteur $(K_1 \times 1)$ et β_2 est un vecteur $(K_2 \times 1)$ de paramètres inconnus, et $K_1 + K_2 = K$. Supposons que X_{i1} est endogène,

$$\mathbb{E}(X_{i1}U_i) \neq 0$$

par opposition à X_{i2} qui est (faiblement)exogène,

$$\mathbb{E}(X_{i2}U_i) = 0$$

(L'hypothèse $\mathbb{E}(U_i|X_{i2}) = 0$ est appelée exogénéité forte).

2.1. Sources d'endogénéité.

Variables omises. Considérons l'équation de salaire suivante,

$$\begin{aligned} \log Sal_i &= \alpha + \beta_1 Etudes + \gamma Genre + \delta Abilit + V_i \\ &= \alpha + \beta_1 Etudes + \gamma Genre + U_i \end{aligned}$$

Étant donné que *Abilit* est inobservable elle se retrouve dans le terme d'erreur du modèle $U_i = \delta Abilit + V_i$. Nous pouvons considérer que la variable *Genre* est exogène, mais *Abilit* est vraisemblablement corrélée avec le niveau d'études , et par conséquent *Etudes* est endogène.

Érreurs de mesure. Supposons que le vrai modèle soit,

$$Y_i = \tilde{X}_{i1}^\top \beta + X_{i2}^\top \beta_2 + V_i$$

où cependant $\tilde{X}_{i1}$ est inobservable. On observe à la place, $X_{i1} = \tilde{X}_{i1} + \epsilon_i$ où ϵ est un vecteur de bruits indépendant de $\tilde{X}_{i1}$, et X_{i2}. Substituons $\tilde{X}_{i1}$ dans l'équation précédente,

$$Y_i = X_{i1}^\top \beta + X_{i2}^\top \beta_2 - \epsilon_i^\top \beta + V_i$$

Posons $U_i = -\epsilon_i^\top \beta + V_i$. Alors que X_{i2} est exogène, X_{i1} est endogène car corrélé avec U_i par le biais de ϵ_i.

Simultanéité. Considérons l'équation suivante,

$$Heures_i = \beta_1 Enfants_i + X_{i2}^\top \beta_2 + U_i$$

où $Heures_i$ est le nombre d'heures travaillées par semaine, $Enfants_i$ est le nombre d'enfants dans une famille, et X_{i2} est un vecteur de variables exogènes. Alors que le nombre d'enfant affecte l'offre de travail, il est raisonnable de penser que les décision de carrière affectent la taille de la famille, i.e., on doit considérer une autre équation qui détermine le nombre d'enfants dans la famille,

$$Enfants_i = \gamma Heures_i + Z_{i1}^\top \gamma_2 + + V_i$$

où Z_{i1} est un autre vecteur de variables exogènes. En substituant l'expression pour les heures dans l'équation pour le nombre d'enfants, nous obtenons(en supposant que $1 - \beta_1\gamma_1 \neq 0$),

$$Enfants_i = X_{i2}^\top \left(\frac{\beta_2\gamma_1}{1-\beta_1\gamma_1}\right) + Z_{i1}^\top \left(\frac{\gamma_2}{1-\beta_1\gamma_1}\right) + \left(\frac{\gamma_1}{1-\beta_1\gamma_1}\right) U_i + \left(\frac{1}{1-\beta_1\gamma_1}\right) V_i$$

En supposant que X_{i2}, Z_{i1}, V_i ne sont pas corrélés avec U_i, nous obtenons,

$$\begin{aligned}\mathbb{E}(U_i Enfants_i) &= \left(\frac{\gamma_1}{1-\beta_1\gamma_1}\right)\mathbb{E}(U_I^2) \\ &\neq 0\end{aligned}$$

2.2. **Propriétés de l'estimateur des moindres carrés en présence d'endogénéité.** Considérons l'estimateur des MCO de β_1 dans (1). Pour cela considérons l'écriture matricielle du modèle,

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{U}$$

où $\mathbf{Y}$ est le vecteur $(n \times 1)$ ayant pour élément i Y_i, $\mathbf{X}_1$ est la matrice $(n \times K_1)$ de régresseurs endogènes ayant pour ligne i $X_{i1}^\top$, $\mathbf{X}_2$ est la matrice $(n \times K_2)$ de régresseurs exogènes ayant pour ligne i $X_{i2}^\top$, et $\mathbf{U}$ est le vecteur $(n \times 1)$ ayant pour élément i U_i. L'estimateur des MCO de β_1 est,

$$\begin{aligned}\widehat{\beta}_{1n} &= (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{M}_2 \mathbf{Y} \\ &= \beta_1 + (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U}\end{aligned}$$

où $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top$. Nous avons,

$$n^{-1}\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1 = n^{-1}\sum_{i=1}^n X_{i1}X_{i1}^\top - n^{-1}\sum_{i=1}^n X_{i1}X_{i2}^\top \left(n^{-1}\sum_{i=1}^n X_{i2}X_{i2}^\top\right)^{-1} n^{-1}\sum_{i=1}^n X_{i2}X_{i1}^\top$$

$$n^{-1}\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U} = n^{-1}\sum_{i=1}^n X_{i1}U_i - n^{-1}\sum_{i=1}^n X_{i1}X_{i2}^\top \left(n^{-1}\sum_{i=1}^n X_{i2}X_{i2}^\top\right)^{-1} n^{-1}\sum_{i=1}^n X_{i2}U_i^\top$$

Supposons que,

- Les observations $\{(Y_i, X_i)\}_{i=1}^n$ sont i.i.d.
- $\mathbb{E}(X_{ik}^2) < \infty$ pour tout $k = 1, ..., K$.
- $\mathbb{E}(X_i X_i^\top)$ est définie positive.
- $\mathbb{E}(U_i^2) < \infty$.

Par la loi faible des grands nombre,

$$n^{-1}\sum_{i=1}^n X_{i1}X_{i1}^\top \xrightarrow{p} \mathbb{E}(X_{i1}X_{i1}^\top)$$

$$n^{-1}\sum_{i=1}^n X_{i1}X_{i2}^\top \xrightarrow{p} \mathbb{E}(X_{i1}X_{i2}^\top)$$

$$n^{-1}\sum_{i=1}^n X_{i2}X_{i2}^\top \xrightarrow{p} \mathbb{E}(X_{i2}X_{i2}^\top)$$

$$n^{-1}\sum_{i=1}^n X_{i2}U_i^\top \xrightarrow{p} 0$$

$$n^{-1}\sum_{i=1}^n X_{i1}U_i^\top \xrightarrow{p} \mathbb{E}(X_{i1}U_i)$$

Ainsi,

$$\begin{aligned} n^{-1}\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1 &\xrightarrow{p} \mathbb{E}(X_{i1}X_{i1}^\top) - \mathbb{E}(X_{i1}X_{i2}^\top)\left(\mathbb{E}(X_{i2}X_{i2}^\top)\right)^{-1}\mathbb{E}(X_{i2}X_{i1}^\top) \\ n^{-1}\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{U} &\xrightarrow{p} \mathbb{E}(X_{i1}U_i) - \mathbb{E}(X_{i1}X_{i2}^\top)\left(\mathbb{E}(X_{i2}X_{i2}^\top)\right)^{-1}\mathbb{E}(X_{i2}U_i) \\ &= \mathbb{E}(X_{i1}U_i) \\ &\neq 0 \end{aligned}$$

et nous concluons que $\widehat{\beta}_{1n}$ n'est pas convergent,

$$\begin{aligned} \widehat{\beta}_{1n} &\xrightarrow{p} \beta_1 + \left(\mathbb{E}(X_{i1}X_{i1}^\top) - \mathbb{E}(X_{i1}X_{i2}^\top)\left(\mathbb{E}(X_{i2}X_{i2}^\top)\right)^{-1}\mathbb{E}(X_{i2}X_{i1}^\top)\right)^{-1}\mathbb{E}(X_{i1}U_i) \\ &\neq \beta_1 \end{aligned}$$

La non convergence de l'estimateur des MCO de β_2 peut être montré de manière similaire. Nous avons,

$$\widehat{\beta}_{2n} = \beta_2 + (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{U}$$

où $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top$. Et nous avons,

$$\widehat{\beta}_{2n} \xrightarrow{p} \beta_2 + \left(\mathbb{E}(X_{i2}X_{i2}^\top) - \mathbb{E}(X_{i2}X_{i1}^\top)\left(\mathbb{E}(X_{i1}X_{i1}^\top)\right)^{-1}\mathbb{E}(X_{i1}X_{i2}^\top)\right)^{-1}\mathbb{E}(X_{i2}X_{i1}^\top)\mathbb{E}(X_{i1}X_{i1}^\top)^{-1}\mathbb{E}(X_{i1}U_i)$$

3. Estimation par variables instrumentales

Soit Z_{i1} un vecteur $(K_1 \times 1)$ de variables exogènes,

$$\mathbb{E}(Z_{i1}U_i) = 0$$

Il est important de noter que Z_{i1} est exclu du modèle (1), i.e., Z_{i1} ne contient aucun des éléments de X_{i2}. Définissons,

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix},$$
$$Z_i = \begin{pmatrix} Z_{i1} \\ X_{i2} \end{pmatrix}$$

Ici, X_i est le vecteur $(K \times 1)$ de régresseurs, et Z_i est le vecteur $(K \times 1)$ de *variables instrumentales*(VIs). Notons que les régresseurs exogènes apparaissent aussi dans les vecteur des VIs, et que pour chaque variables endogène nous avons une variables exogène(une VI) qui est exclue du modèle $Y_i = X_i^\top \beta + U_i$. Lorsque tous les régresseurs sont endogènes nous n'avons plus aucun élément commun à X_i et à Z_i.
Nous supposons que les VIs sont informatives par rapport aux régresseurs. Ceci est exprimé par la *condition de rang* suivante,

$$\text{Rang}\left(\mathbb{E}(Z_i X_i^\top)\right) = K \tag{2}$$

La condition dans (2) échouera si, par exemple, $\mathbb{E}(Z_{i1}X_i^\top) = 0$($Z_{i1}$ est exogène mais c'est un bruit aléatoire). La condition de rang échouera aussi si certains éléments de Z_{i1} sont des combinaisons linéaires des éléments dans les régresseurs exogènes inclus X_{i2}.

Exemple 1. *Reprenons le cas "Heures/Enfants". Angrist et Evans(1998) on suggéré d'utiliser la composition en termes de sexe des deux premier enfants comme instrument pour le nombre d'enfants dans une famille(l'échantillon utilisé est restreint aux femmes avec au moins deux enfants). Ceci est motivé par l'idée que si les deux premiers enfants sont du même sexe(fille-fille, ou garçon-garçon) la famille sera plus encline à avoir un troisième enfant que dans le cas où les deux premiers enfants sont de sexe différent. En conséquence, la variable indicatrice d'avoir deux premiers enfants du même sexe doit être positivement corrélée avec le nombre d'enfants. D'un autre côté , l'instrument est exogène car la composition en termes de sexe des deux premiers enfants est déterminée aléatoirement.*

Nous avons,

$$\mathbb{E}(Z_i U_i) = 0$$

L'application de la méthode des moments suggère un estimateur solution du système suivant de K équations,

$$n^{-1}\sum_{i=1}^{n} Z_I \left(Y_i - X_i^\top \widehat{\beta}_n^{VI}\right) = 0$$

d'où,

$$\begin{aligned}\widehat{\beta}_n^{VI} &= \left(\sum_{i=1}^{n} Z_i X_i^\top\right)^{-1} \sum_{i=1}^{n} Z_i Y_i \\ &= (\mathbf{Z}^\top \mathbf{X})^{-1} \mathbf{Z}^\top \mathbf{Y}\end{aligned}$$

où $\mathbf{X}$ est la matrice $(n \times K)$ ayant pour élément i $X_i^\top$, et $\mathbf{Z}$ est la matrice $(n \times K)$ ayant pour élément i $Z_i^\top$.
L'estimateur $\widehat{\beta}_n^{VI}$ est appelé *estimateur des variables instrumentales* de β. Nous étudions maintenant sa convergence, et sa normalité asymptotique. Pour cela nous supposons que,

- Les observations $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ sont i.i.d.
- $\mathbb{E}(Z_i U_i) = 0$ pour tout $k = 1, ..., K$.
- $\mathbb{E}(X_{ik}^2) < \infty$ pour tout $k = 1, ..., K$.
- $\mathbb{E}(Z_{i1k}^2) < \infty$ pour tout $k = 1, ..., K_1$.
- $\mathbb{E}(Z_i X_i^\top)$ est de rang K.
- $\mathbb{E}(U_i^2 Z_i Z_i^\top)$ est définie positive.

Écrivons,

$$\widehat{\beta}_n^{VI} = \beta + \left(n^{-1} \sum_{i=1}^n Z_i X_i^\top \right)^{-1} n^{-1} \sum_{i=1}^n Z_i U_i \tag{3}$$

Notons que sous les hypothèses faites plus haut, par l'inégalité de Cauchy-Schwartz,

$$\begin{aligned} \mathbb{E}(|Z_{ir} X_{is}|) &\leq \sqrt{\mathbb{E}(Z_{ir}^2)\mathbb{E}(X_{is}^2)} \\ &< \infty \text{ pour tout } r, s = 1, ..., K. \end{aligned}$$

Par conséquent, par le théorème de Slutsky,

$$\begin{aligned} \widehat{\beta}_n^{VI} &\xrightarrow{p} \beta + \mathbb{E}(Z_i X_i^\top)^{-1} \mathbb{E}(Z_i U_i) \\ &= \beta \end{aligned}$$

Afin de montrer la normalité asymptotique nous supposons en outre que,

- $\mathbb{E}(Z_{ik}^4) < \infty$, pour tout $k = 1, ..., K$.
- $\mathbb{E}(U_i^4) < \infty$.

Écrivons (3) comme suit,

$$n^{1/2}(\widehat{\beta}_n^{VI} - \beta) = \left(n^{-1} \sum_{i=1}^n Z_i X_i^\top \right)^{-1} n^{-1/2} \sum_{i=1}^n Z_i U_i$$

Notons que du fait des hypothèses précédentes,

$$\begin{aligned} \mathbb{E}(\left|U_i^2 Z_{ir} Z_{is}\right|) &\leq (\mathbb{E}(U_i^4))^{1/2} \mathbb{E}(Z_{ir}^4 Z_{is}^4)^{1/4} \\ &< \infty \end{aligned}$$

Par conséquent, par le théorème central-limite et le théorème de convergence de Cramer,

$$\begin{aligned} n^{1/2}(\widehat{\beta}_n^{VI} - \beta) &\xrightarrow{d} (\mathbb{E}(Z_i X_i^\top))^{-1} \mathcal{N}\left(0, \mathbb{E}(U_i^2 Z_i Z_i^\top)\right) \\ &= \mathcal{N}\left(0, (\mathbb{E}(Z_i X_i^\top))^{-1} \mathbb{E}(U_i^2 Z_i Z_i^\top)(\mathbb{E}(X_i Z_i^\top))^{-1}\right) \end{aligned}$$

La matrice de variances-covariances asymptotique prend une forme en sandwich et peut être estimé de manière convergente par,

$$\left(n^{-1} \sum_{i=1}^n Z_i X_i^\top \right)^{-1} n^{-1} \sum_{i=1}^n (\widehat{U}_i^2 Z_i Z_i^\top) \left(n^{-1} \sum_{i=1}^n X_i Z_i^\top \right)^{-1}$$

où $\widehat{U}_i = Y_i - X_i^\top \widehat{\beta}_n^{VI}$.