Econométrie des données de panel: Modèle dynamique et variables instrumentales

Guillaume Horny*

*Banque de France

M2 SE

Chapitre 5 : Modèle dynamique et variables instrumentales

Plan

Plan

La persistance de la variable dépendante

- Les modèles vus jusqu'ici ne prennent pas en compte le fait que la variable dépendante puisse dépendre de ses valeurs passées.
- C'est toutefois courant. Par exemple, les variables de bilan sont souvent très persitantes.
- En effet, les comportements économiques sont par essence dynamiques : l'investissement d'une entreprise aujourd'hui va dépendre de ses investissements passés, idem pour la consommation d'un ménage.

Modèle dynamique

Le modèle dynamique le plus commun est :

$$y_{it} = \gamma y_{it-1} + x_{it}^{'}\beta + \alpha_i + \epsilon_{it},$$

où γ est un coefficient autoregressif. Pour $\gamma > 0$, les y_{it} sont positivement autocorrélés et pour $\gamma < 0$, les y_{it} sont négativement autocorrélés. Pour $\gamma = 1$, on a une racine unitaire impliquant que y_{it} suit une marche aléatoire avec une possible dérive. On fait généralement l'hypothèse que le modèle est stationaire, cad $|\gamma| < 1$. De plus, on considère le cas des panels courts.

On suppose que les α_i et les ϵ_{it} sont sans corrélation, et que les ϵ_{it} ne sont pas autocorrélés.

Quel effet sur la moyenne et la variance des y_{it} ?

Considérons un modèle sans variable explicative :

$$y_{it} = \gamma y_{it-1} + \alpha_i + \epsilon_{it}.$$

Par récursion, on peut montrer que :

$$\begin{aligned} y_{it} &= \gamma(\alpha_i + \epsilon_{it-1}) + \gamma^2(\alpha_i + \epsilon_{it-2}) + \gamma^3(\alpha_i + \epsilon_{it-3}) + \dots \\ &= (1-\gamma)^{-1}\alpha_i + \sum_j \gamma^j \epsilon_{i,t-j}. \end{aligned}$$

D'où $\mathrm{E}(y_{it}|\alpha_i) = (1-\gamma)^{-1}\alpha_i$ et $\mathrm{Var}(y_{it}|\alpha_i) = (1-\gamma^2)\sigma_\epsilon^2$. Pour un α_i donné, la présence d'autocorrélation va déplacer y_{it} par rapport à sa moyenne et atténuer sa variance.

Dépendance à l'état et hétérogénéité inobservée (1/2)

Les α_i sont inobservés, et ils interviennent aussi dans l'autocorrélation des y_{it} qui ne dépend donc pas que de γ :

$$\begin{aligned}\text{cor}(y_{it}, y_{it-1}) &= \text{cor}(\gamma y_{it-1} + \alpha_i + \epsilon_{it}, y_{it-1}) \\ &= \gamma + \text{cor}(\alpha_i, y_{it-1}) \\ &= \gamma + \frac{1-\gamma}{1 + (1-\gamma)\sigma_\epsilon^2/(1+\gamma)\sigma_\alpha^2}.\end{aligned}$$

La dernière relation provient d'un calcul similaire à celui effecté pour le calcul de la corrélation dans le modèle à effet aléatoire.

Dépendance à l'état et hétérogénéité inobservée (2/2)

cor(y_{it}, y_{it-1}) montre les différentes causes d'autocorrélation :

- la **dépendance à l'état**, qui provient des valeurs passées de y_{it}. A l'extrême, σ_α^2 tend vers 0,
- l'autocorrelation due **l'hétérogénéité inobservée**, lorsque $\gamma = 0$ et que cor(y_{it}, y_{it-1}) dépend du ratio $\sigma_\epsilon^2/\sigma_\alpha^2$.

On peut avoir cor(y_{it}, y_{it-1}) qui tend vers 1, que ce soit car γ tend vers 1 avec σ_α^2 tend vers 0, où car $\sigma_\epsilon^2/\sigma_\alpha^2$ tend vers 0.

Mais les implications pour la prise de décision sont différentes : une persistance élevée des salaires n'a pas les mêmes enjeux de politique publique si elle relève de chocs persistants ou de caractéristiques inobservées.

Biais des estimateurs

Les estimateurs standards sont biaisés :

- l'estimateur OLS suppose que le terme d'erreur est $(\alpha_i + \epsilon_{it})$ et que y_{it-1} est une variable explicative. Or y_{it-1} dépend lui aussi de α_i, il y a donc corrélation entre le terme d'erreur et une variable explicative ;
- la transformation *within* n'élimine pas le terme autorégressif, qui dépend de ϵ_{it-1}. Or ϵ_{it-1} intervient dans le calcul de $\epsilon_{it} - \epsilon_{i.}$, il en résulte un biais.
  - On a toutefois que $\epsilon_{i.}$ tend vers 0 lorsque T est grand, l'estimateur à effet fixe est donc convergent pour $T \to \infty$. En pratique, on considère que le biais devient négligeable lorsque $T > 30$.
- les estimateurs GLS et FGLS sont des estimateurs de classe λ, c'est-à-dire des combinaisons linéaires des estimateurs *within* et *between*. Comme les estimateurs *within* et *between* sont biaisés, les estimateurs GLS et FGLS le sont aussi.

Modèles dynamique et modèle à erreurs corrélées

Les extensions des modèles à effet fixe ou à effet aléatoire nous conduisent à des problèmes d'endogénéité :

- la présence de corrélation entre les α_i et les x_{it} implique une corrélation entre l'erreur composée et les variables explicatives
- l'ajout d'un terme autorégressif fait de même

Sans oublier qu'une variable explicative peut aussi être endogène dans les modèles à effet fixe ou aléatoires

Tous ces problèmes peuvent toutefois être résolus avec des variables instrumentales.

Plan

Principe général des variables instrumentales

$$Y = X\beta + \epsilon, \mathrm{E}(X'\epsilon) \neq 0.$$

Les erreurs sont corrélées avec les explicatives et les estimateurs habituels sont biaisés. L'idée est de trouver des variables Z, que l'on va appeler "instruments" telles que :

- $\mathrm{E}(Z'\epsilon) = 0$, les instruments sont sans corrélation avec le terme d'erreur. On dira alors qu'ils sont **valides**,
- $\mathrm{E}(Z'X) \neq 0$, les instruments sont corrélés avec les variables endogènes. S'ils ne le sont pas, on parlera d'instruments **faibles**.

Chacune de ces deux propriétés peut être testée (test de Sargan, de Stock et Yogo...). Dans la pratique, trouver des variables Z satisfaisant ces deux propriétés est difficile et demande souvent de tester de nombreuses variables candidates...

Exemple de variables instrumentales

La littérature économique regorge d'instruments :

- pour les **prix** de marché : tout choc qui affecte l'offre sans réduire la demande est un bon instrument. Les exemples comprennent une météo favorable aux récoltes pour les prix des produits agricoles, ou défavorable aux pêcheurs pour le prix du poisson,
- pour les **salaires** : un bon instrument sera corrélé avec le niveau d'études et pas avec les caractéristiques inobservées des individus. La distance à une université a été utilisée, de même que le trimestre de naissance

Estimateur IV avec une seule variable explicative

On cherche à estimer :

$$\beta = \frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = \frac{dy/dz}{dx/dz}.$$

Un moyen est d'estimer le numérateur par la régression de Y sur Z, et le dénominateur par la régression de X sur Z. D'où :

$$\beta_{IV} = \frac{(Z'Z)^{-1}Z'Y}{(Z'Z)^{-1}Z'X} = (Z'X)^{-1}Z'Y.$$

On peut donc estimer β_{IV} directement par les covariances empiriques :

$$\beta_{IV} = \frac{\text{cov}(z,y)}{\text{cov}(z,x)}.$$

Estimateur IV avec plusieurs variables explicatives (1/2)

Lorsqu'il y a autant d'instrument que de variables endogènes, la généralisation est :

$$\beta_{IV} = (Z'X)^{-1}Z'Y.$$

Il s'ensuit :

$$\begin{aligned}\beta_{IV} &= (Z'X)^{-1}Z'(X\beta + \epsilon) \\ &= \beta + (Z'X)^{-1}Z'\epsilon\end{aligned}$$

L'estimateur est convergent lorsque $Z'\epsilon$ tend vers 0 et que $(Z'X)$ ne tend pas vers 0. On retrouve ici les conditions que les instruments soient valides et ne soient pas faibles.

Estimateur IV avec plusieurs variables explicatives (2/2)

Lorsqu'il y a plus d'instruments que de variables endogènes, la formule précédente retourne un vecteur plus d'éléments que le modèle ne comprend de paramètres. Les doubles moindres carrés résolvent ce problème :

$$\beta_{2SLS} = [X'Z(Z'Z)^{-1}ZX]^{-1}[X'Z(Z'Z)^{-1}ZY].$$

Cet estimateur peut se calculer directement, ou bien s'obtenir :

1. en régressant les X endogènes sur les Z,
2. en prédisant des $\widehat{X}$,
3. en régressant ensuite Y sur les $\widehat{X}$.

Mais attention aux écarts-type ! Les résidus de la deuxième étape doivent être transformés pour refléter la prévision, et son aléa, effectuée en première étape

Application aux panels : l'estimateur FEIV (1/2)

On considère un modèle à effet fixe, où il peut y avoir une corrélation entre X et α. De plus, une des explicatives est corrélée avec le terme d'erreur :

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it}.$$

Écrivons le modèle pour faire apparaitre les indicatrices d'individus :

$$Y = D\alpha + X'\beta + \epsilon.$$

On dispose d'une matrice Z d'instruments. Comme l'endogénéité provient d'une colonne de X, les colonnes de D sont des variables exogènes. L'estimateur 2SLS consiste en la régression de Y sur (X, D) avec les instruments (Z, D).

Application aux panels : l'estimateur FEIV (2/2)

On a :

$$\beta_{2SLS} = [X'WZ(Z'WZ)^{-1}Z'WX]^{-1}[X'WZ(Z'WZ)^{-1}Z'WY].$$

Comme W est un projecteur orthogonal, on peut écrire :

$$\beta_{2SLS} = [(WX)'WZ\left((WZ)'WZ\right)^{-1}(WZ)'WX]^{-1} \\ [(WX)'WZ\left((WZ)'WZ\right)^{-1}(WZ)'WY].$$

L'estimateur FEIV peut ainsi se calculer en appliquant les doubles moindre carrés au modèle linéaire après transformation *within*. Les intruments doivent donc varier dans le temps.

Identification de l'estimateur FEIV

Intuitivement, on remplace les valeurs problématiques par des valeurs prédites avec les instruments. Si les instruments sont valides, les prévisions seront sans corrélation avec ϵ. Si les instruments sont faibles, les résultats de deuxième étape seront peu précis (vraisemblance "plate" pour le modèle de la deuxième étape).

L'estimateur FEIV peut être calculé avec tout logiciel permettant d'effectuer des 2SLS. Attention toutefois à l'estimation de la matrice de variance : elle doit être "panel robust"

Alternatives aux 2SLS (1/2)

Des alternatives aux 2SLS existent : Hausman et Taylor (1981) pour autoriser des régresseurs constants dans le temps, en supposant que les α_i sont sans corrélation avec certains des x_{it} et des z_{it}.

Pour les modèles dynamiques : Anderson et Hsiao (1981), Arellano et Bond (1991), Arellano et Bover (1995), Blundell et Bond (1998)...

Mise en oeuvre (2/2)

Les subtilités sont nombreuses (correction de la matrice de variance lors de la deuxième étape, gestion des variables incluses/excluses, plusieurs tests possibles d'instruments faibles...).

Dans le cas des variables instrumentales, il est prudent de ne pas se lancer dans la programmation manuelle de ce type d'estimateurs et de privilégier des **routines préprogrammées** et déjà testées.

A ma connaissance, Stata est le logiciel qui propose des fonctions pour le plus grand nombre de modèles (routines `ivreg2` et `xtivreg2`, à installer manuellement).