Pseudo-maximum de vraisemblance : comment estimer simplement des modèles qualitatifs statiques avec variables endogènes sur données de panel

Stéfan LOLLIVIER[1]

De plus en plus de travaux appliqués ont recours à des estimations économétriques dans lesquelles la variable expliquée, voire les variables expliquées et explicatives, sont qualitatives. Parallèlement, l'utilisation des données longitudinales, qui permettent de mieux cerner l'hétérogénéité individuelle inobservable, se répand. Or, les méthodes classiques de maximisation de la vraisemblance (MV), usuelles en coupes transversales, sont plus difficiles à mettre en œuvre avec la dimension longitudinale. Leur usage se heurte en effet à deux types de difficulté.

La première est que la méthode du MV requiert des hypothèses paramétriques explicites afin de procéder aux estimations. Par exemple, un modèle probit dichotomique en données longitudinales renvoie à des hypothèses paramétriques sur la matrice de variance du terme d'erreur, avec T(T-1) variables à estimer. Si on restreint la spécification, par exemple avec un modèle à effets individuels ou de type AR(1), on réduit le nombre de paramètres, mais au prix d'une hypothèse a priori. Si celle-ci est justifiée, on gagne en précision (et en simplification) dans l'estimation du modèle ; mais si cette hypothèse est invalide, on perd en général les propriétés de convergence pour tous les estimateurs du modèle !

La seconde difficulté est que les estimations par le MV sont complexes à réaliser sur données longitudinales. Même dans le cas du modèle probit dichotomique, il faut souvent recourir à des techniques de simulation, qui n'assurent le plus souvent la convergence des paramètres que lorsque le nombre de simulations tend vers l'infini.

L'intérêt des méthodes de pseudo-maximum de vraisemblance est de permettre d'éviter les deux difficultés évoquées ci-dessus, recours quasi-obligé à des restrictions paramétriques et complexité calculatoire. Ces méthodes remplacent la « vraie » vraisemblance par une vraisemblance approchée avec un modèle en général normal, mais qui respecte les moments conditionnels d'ordre un, voire deux, de la variable expliquée (Gouriéroux, Monfort, Trognon, 1984). On peut ainsi facilement montrer que l'estimateur des paramètres des variables explicatives du modèle probit simple empilé est convergent, même lorsque la matrice de variance du terme d'erreur est non diagonale, et quelle que soit la forme de cette matrice. Le calcul de la matrice de variance de ces estimateurs est plus complexe, mais on peut facilement l'obtenir grâce à des techniques de bootstrap.

[1] DSDS, Insee

Mais c'est en présence de variables explicatives endogènes que le recours aux méthodes de pseudo-maximum de vraisemblance est le plus décisif, car l'estimation par la méthode du MV d'un système de variables latentes sur données longitudinales est quasiment impossible. Le problème est déjà complexe lorsque la variable de l'équation d'intérêt est continue, et différents auteurs ont proposé des méthodes alternatives à celles du MV. Ainsi, même en coupe transversale, Heckman (1974, 1978) a montré que le recours à une procédure en deux étapes était à la fois plus simple à mettre en œuvre, mais aussi plus économe en hypothèses sur les lois des termes d'erreur. Wooldridge (1995) a généralisé cette approche en deux étapes aux données de panel, mais la variable expliquée demeure quantitative. D'autres auteurs (Kyriazidou, 1997 ; Dustmann Rocchina-Barrachina, 2000) ont mis au point des techniques d'estimation plus économes en hypothèses paramétriques, notamment sur les lois des termes d'erreur, mais au prix de difficultés d'estimation accrues.

La situation dans laquelle toutes les équations font intervenir des variables qualitatives fait l'objet d'une littérature encore peu développée. On propose ici une méthode en plusieurs étapes, comme chez Heckman et Wooldridge. La première étape consiste à estimer de façon convergente l'équation instrumentale. Dans un deuxième temps, on tire pour chaque observation un résidu dans la loi conditionnelle du terme d'erreur. La troisième étape réalise l'estimation de l'équation d'intérêt au moyen d'une méthode de pseudo-maximum de vraisemblance simulé (Gouriéroux, Monfort, 1993). Les estimateurs ainsi obtenus sont alors convergents et asymptotiquement normaux, dès lors que le nombre d'individus et le nombre de simulations tendent vers l'infini. L'estimation est assez simple lorsque l'équation instrumentale est univariée, plus complexe, mais réalisable, lorsque l'équation instrumentale est multivariée.