Groupe de lecture "Econométrie des données d'enquête"

Compte-rendu de la $5^{\text{ième}}$ réunion, 13 avril 2015
Extensions

Suivi par Marine Guillerm et Ronan Le Saout

Cette cinquième session s'articulait autour de deux présentations (par Thomas Balcone et Pham Trong-Hien) qui avaient pour but d'éclairer les modes de traitement de la non-réponse partielle dans les modèles économétriques.

Le premier article "What Do We Do with Missing Data? Some Options For Analysis of Incomplete Data." (Raghunathan, 2004) prenait pour exemples les données de santé. Des données manquantes y sont fréquentes. L'approche standard est rappelée. Elle consiste à supprimer les individus pour lesquelles des données sont manquantes (pour une des variables explicatives X ou la variable à expliquer Y). Sauf dans le cas où la non-réponse serait aléatoire (conditionnellement à (X, Y)), un biais de sélection est alors introduit. Trois mécanismes de correction sont ainsi présentés: la pondération, l'imputation multiple et la modélisation (par maximum de vraisemblance). Ces mécanismes permettent permettent de corriger la non-réponse partielle, si celle-ci est fontion de variables observées sur l'ensemble de la population. Si ce n'est pas le cas, seule la modélisation économétrique de la non-réponse (partielle ou totale) est possible. Les hypothèses ne peuvent alors être vérifiées que sur des jeux de données externes. Le deuxième article "Regression With Missing X's: A Review" (Little, 1992) adopte une approche plus théorique du problème. L'effet de la non-réponse partielle est étudié dans le cadre du modèle linéaire multiple. L'effet théorique des moindres carrés sur données imputées est étudié, mais l'imputation est restreinte à un cas atypique d'imputation par régression. Les approches bayésiennes sont également introduites.

Deux principales remarques ont été formulées. Premièrement, il faudrait étudier la validité des méthodes proposées en fonction du % de non-réponse partielle. Deuxièmement, la perspective des articles n'est pas la même que celle des sondeurs. L'imputation (et son effet sur les estimateurs) n'est traitée que très partiellement à travers un cas atypique d'imputation déterministe par régression. Il faudrait donc reprendre le cas des imputations usuelles en sondage et étudier leurs effets sur les modèles économétriques.

Au-delà des données d'enquête, la recherche en économie appliquée s'appuit sur des données individuelles administratives. Pour des raisons de confidentialité, ces données peuvent être bruitées (en tenant compte des corrélations entre les variables) et qualifiées de données synthétiques. Des recherches récentes (https://www2.vrdc.cornell.edu) étudient la validité des études économiques utilisant de telles données.