Modèle de sélection en présence de non-réponse non ignorable et population hétérogène

Éric Gautier [1]

[1] *Ensae, Crest, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France,*
Eric.Gautier@ensae-paristech.fr

Supposons que nous disposions de données d'enquête en présence de non-réponse non-ignorable et que nous souhaitions estimer la fonction de répartition d'une variable Y. La fonction de répartition permet de retrouver de nombreux paramètres d'intérêt, par exemple des indicateurs d'inégalité.

Une approche usuelle consiste à modéliser conjointement la variable d'intérêt et le mécanisme de non-réponse. Le modèle de non-réponse que nous considérons est très souple, il s'écrit

$$D = \mathbb{I}\left\{Z'\Gamma + \epsilon > 0\right\}.$$

Le cas $D = 1$ correspond au cas où la variable Y est renseignée et le cas $D = 0$ correspond à la non-réponse. Z est un vecteur de variables observées chez les répondants et les non-répondants et Γ est un vecteur de cœfficients aléatoires dont on ne spécifie pas la distribution, on ne spécifie pas non plus la loi du terme d'erreur. Les variables entrant dans le vecteur Z doivent être convenablement choisies car leur loi jointe (sachant X, des variables de contrôle, si nécessaire pour justifier l'indépendance) doit être indépendante de la loi jointe de (Γ', ϵ, Y) (sachant X si nécessaire). On les appelle des variables instrumentales. L'introduction de cœfficients aléatoires autorise des comportements de non-réponse différents pour chaque individu. La non-réponse est ignorable du fait de la dépendance entre (Γ', ϵ) et Y. La méthode proposée permet de conditionner par (Γ', ϵ) et de retrouver toute la distribution de Y. Ce modèle faisant apparaître plusieurs sources d'hétérogénéité ne restreint pas le modèle de non-réponse comme le ferait un modèle de la forme

$$D = \mathbb{I}\left\{\pi(Z) + \epsilon > 0\right\}$$

où π est la fonction score qui implique l'hypothèse classique et critiquable de monotonie. En présence d'hétérogénéité inobservée dans l'équation de sélection il est connu que les méthodes à variables instrumentales n'utilisant pas explicitement de modèle de non-réponse ne permettent même pas de retrouver la moyenne de Y.

Outre l'hypothèse de variables instrumentales nous aurons besoin de quelques hypothèses techniques pour montrer qu'il est possible d'estimer la fonction de répartition de Y sachant (X, Γ, ϵ) et donc de Y car la densité de (Γ, ϵ) peut être estimée. Même si le problème d'estimation fait intervenir un problème inverse mal posé, l'estimateur est une simple somme sur l'échantillon.

Bibliographie

Gautier, E., et Hoderlein, S. (2012). Estimating treatment effects with random coefficients in the selection equation, Preprint : hal-00618469.