Groupe de lecture "Econométrie des données d'enquête"

Compte-rendu de la 6ième réunion, 18 mai 2015
Le traitement de la non réponse totale

Suivi par Marine Guillerm et Ronan Le Saout

Cette sixième séance du groupe de lecture "Sondage et économétrie" avait pour thème: le traitement de la non réponse totale. Elle s'est articulée autour de deux présentations. Eric Lesage a d'abord présenté un travail mené en collaboration avec David Haziza (département de mathématique de l'Université de Montréal) et Xavier d'Haultfœuille (CREST-Ensae): "Risque d'amplification du biais de l'estimateur par calage généralisé en présence de non-réponse". Ensuite Benjamin Vignolles a présenté l'article de D'Haultfœuille (2009) "A new instrumental method for dealing with endogenous selection".

1 Présentation de l'article 'Risque d'amplification du biais de l'estimateur par calage généralisé en présence de non-réponse" (E.Lesage, 2015)

La présentation a pour objet le traitement de la non réponse totale par calage généralisé. La non réponse est classiquement corrigée en deux étapes. Une première étape consiste à estimer la probabilité de réponse de chaque unité répondante. Les nouveaux poids correspondent aux poids de sondage initiaux multipliés par l'inverse de la probabilité de réponse de l'unité i. Une deuxième étape ajuste ces poids par calage (sur marges) pour que les estimations de l'enquête coïncident avec certains totaux connus sur la population.

Le calage généralisé permet de traiter la non réponse en une étape. Les équations dans le cadre du calage généralisé sur données complètes ont été présentées, puis il a été montré comment ces équations s'adaptent en cas de non réponse totale pour traiter simultanément la non réponse et caler les données. Cette méthode s'appuie sur des variables z appelées instruments de calage dont les totaux ne sont connus que sur les seuls répondants (contrairement au calage "classique" où les totaux doivent être connus sur l'ensemble de la population) et sur des variables de calage "classiques" notées x (dont on connait le total sur la population). L'écriture du modèle de non réponse conduit à des équations estimantes qu'il n'est pas possible d'estimer si on ne connait pas les totaux de z sur l'ensemble de la population ou les valeurs pour l'ensemble des individus de l'échantillon (donc y compris sur les non répondants). On a alors recours à des variables x qui servent de proxy à z: variable auxiliaire (leur totaux sur la population sont connus), corrélée à z et qui n'intervient pas dans le modèle de non-réponse (relation d'exclusion sur x). On suppose de plus que z est corrélée à y (la variable d'intérêt) et à la non réponse, mais que conditionnellement à z, y est indépendante de la non réponse (relation d'exclusion sur y). On montre que le modèle de non réponse conduit alors à de nouvelles équations estimantes qui sont analogues à des équations de calage généralisé.

Sous ces hypothèses (et que le modèle de non-réponse soit bien spécifié), les estimateurs par calage généralisé sont convergents. La corrélation entre x et z est un point important : la variance est amplifiée en cas de faible corrélation. De plus, lorsque l'hypothèse d'exclusion sur x n'est pas vérifiée, l'estimateur n'est pas convergent et son biais est amplifié par la faible corrélation entre z et x. Un développement limité de l'erreur de non réponse permet de comprendre le rôle joué par cette corrélation.

2 Présentation de l'article "A new instrumental method for dealing with endogenous selection" (X. D'Haultfœuille, 2009)

On s'intéresse aux caractéristiques d'une variable aléatoire Y (moment, distribution, effet d'un traitement, etc.). La variable Y n'est observée que pour une partie des individus et cette sélection dépend de la variable d'intérêt (la sélection est dite non-ignorable). La sélection peut prendre plusieurs formes qui seront traitées indistinctement ici. Plusieurs exemples ont été donnés: la non-réponse dépend de la consommation de drogue

qui est la variable d'intérêt, le prix des biens et services n'est connu que pour ceux ayant fait l'objet d'une transaction, et pour évaluer une politique publique on souhaite disposer d'un contrefactuel.

Ignorer la sélection aboutirait à des estimateurs biaisés. Une solution classique par le modèle d'Heckman consiste à paramétriser le processus de sélection et à estimer un modèle à deux équations. Cette méthode pose deux problèmes: la variable z expliquant la sélection ne doit pas influencer la variable d'intérêt et l'estimation repose sur une relation paramétrique. L'article présenté ici propose une approche alternative par variables instrumentales.

On cherche un instrument z tel que: (1) $Y = \phi(z, \varepsilon)$, i.e. z doit expliquer en partie la variable d'intérêt, et (2) conditionnellement à Y et éventuellement à d'autres variables de contrôle x, la sélection ne dépend pas de z. Sous des hypothèses additionnelles, on peut estimer le modèle par une méthode similaire à du calage généralisé. Trois méthodes d'estimation (avec support fini, paramétrique et non paramétrique) sont présentées. D'après les simulations, l'estimateur paramétrique semble un bon compromis: il convergence et est le plus précis. L'article propose également une procédure de test de l'exogénéité de l'instrument z et de relâcher cette hypothèse le cas échéant.

3 Discussion

L'objectif du calage généralisé est de traiter la non-réponse non ignorable. Dans ce cadre, la seule définition de classes de non-réponse homogène ne suffit pas.

La méthode de calage généralisé se rapproche de la littérature économérique sur les variables instrumentales. Il ne faut donc pas que X et Z (Y dans l'article de Lesage *et al.* 2015) soient tous les deux corrélés au processus de sélection (la non-réponse).

Les méthodes de calage généralisé ont pour l'instant fait l'objet de peu d'applications pratiques (une à la DARES). Un des objectifs des papiers présentés est d'ouvrir la boîte noire, i.e. de comprendre les fondements probabilistes de la méthode. La mise en oeuvre demande encore une importante réflexion préalable.

Un lien peut être effectué avec le choix de pondérer ou non les modèles économétriques. Dans le cas d'un plan de sondage construit à partir de la variable dépendante (par exemple des strates construites à partir d'un a priori sur Y), il est indispensable de pondérer. Dans le cas où la non-réponse est non-ignorable (typiquement elle est fonction de Y), si le calage généralisé a été bien conduit, il suffit ensuite de pondérer le modèle avec le nouveaux poids pour traiter la sélection non-ignorable.

Même si c'est en théorie possible, il semble difficile d'avoir un unique jeu de poids pour traiter l'ensemble des cas d'études économiques et de non-réponse non-ignorable.

En pratique, des macros pour effectuer du calage généralisé existent.