*S*éminaire de *M*éthodologie *S*tatistique

mardi 15 mars 2016

14h-17h, Insee - Malakoff 1 - salle Malinvaud (1245)

Miscellanées sur le calage

Les différentes méthodes de calage – Le choix des « bornes » de calage
Emmanuel Gros – *Division Sondages, Insee*

Le calage pénalisé : théorie et application
Antoine Rebecq – *Division Sondages, Insee*

Discussion sur l'utilisation du calage en présence de non-réponse
Éric Lesage – *Division Études territoriales, Insee*
David Haziza – *Université de Montréal*

Estimation régionale de taux de pauvreté par une méthode de calage
Pascal Ardilly – *Département des méthodes statistiques, Insee*

Résumés des interventions

Les différentes méthodes de calage – Le choix des « bornes » de calage

Emmanuel Gros – *Division Sondages, Insee*

Plusieurs méthodes de calage sont possibles, les méthodes linéaire, exponentielle, logit et linéaire tronquée étant les plus couramment utilisées. Toutes ces méthodes sont asymptotiquement sans biais, et équivalentes au sens où elles conduisent à des estimations semblables dès lors que la taille de l'échantillon est suffisamment grande.

Cependant en pratique, on préfère en général utiliser des méthodes « bornées » – logit ou linéaire tronquée – qui permettent de contrôler les déformations maximales de poids induites par le calage, et d'éviter ainsi des poids après calage négatifs ou très élevés, susceptibles de détériorer la qualité des estimations dans des domaines de taille réduite ou moyenne, ou pour des variables d'intérêt peu corrélées aux variables de calage.

Après avoir rappelé les principes théoriques du calage, on présentera diverses innovations et études méthodologiques récentes sur le sujet, en s'intéressant plus particulièrement à la question du choix des bornes :

- Méthode de calage sur bornes « minimales » : il s'agit d'une procédure permettant, dans le cadre d'un calage borné via les méthodes logit ou linéaire tronquée, de déterminer les bornes L et U les plus « serrées » possibles autour de 1, i.e. conduisant à l'étendue U −L des rapports de poids la plus faible possible.
- Impact du choix des bornes de calage : on s'intéresse ici à l'impact sur la qualité – biais, variance – des estimateurs calés, du choix des bornes L et U lorsqu'on utilise des méthodes bornées. En particulier, vaut-il mieux chercher les bornes L et U « minimales », quitte à avoir des accumulations de rapports de poids à ces bornes, ou bien des bornes moins proches de 1 mais déformant moins la distribution des rapports de poids par rapport à une distribution unimodale autour de 1 ?

Le calage pénalisé : théorie et application

Antoine Rebecq – *Division Sondages, Insee*

Lorsque l'on met en œuvre des méthodes de calage, il arrive que l'on veuille introduire un grand nombre de variables de calage : dans ce cas, la procédure peut ne pas converger ou aboutir à des facteurs de calage (poids après calage / poids avant calage) éloignés de 1 et très dispersés, ce qui peut être néfaste pour la qualité des estimations, en particulier dans des domaines de taille réduite ou moyenne ou pour des variables d'intérêt peu corrélées aux variables de calage.

La méthode de calage pénalisé consiste à relâcher la contrainte de calage exact pour certaines variables, et à accepter un calage « approché » pour ces variables : on accepte un « petit » écart entre les estimations des totaux ou des distributions de ces variables et les vrais totaux ou distributions connus sur la population. La pondération obtenue permet d'éviter les déformations de poids trop importantes tout en contrôlant l'écart aux marges pour l'ensemble des variables, et conserve par ailleurs toutes les bonnes propriétés statistiques du calage classique.

On présentera le cadre théorique de cette méthode, son application à l'Insee dans le cadre du redressement de l'enquête nationale sur les ressources des jeunes, ainsi que son utilisation dans un package R.

Discussion sur l'utilisation du calage en présence de non-réponse

Éric Lesage – *Division Études territoriales, Insee*
David Haziza – *Université de Montréal*

Dans les enquêtes par sondage, il est classique de procéder à une repondération des individus répondants pour pallier le risque de biais engendré par la présence de non-réponse totale. Il existe essentiellement deux approches de pondération permettant de corriger la non-réponse :

(i) la pondération en deux étapes qui est l'approche majoritaire dans les instituts nationaux de statistique,

(ii) la pondération en une étape dont l'utilisation a crû à l'étranger depuis une quinzaine d'années et qui a fait l'objet d'un ouvrage (Särndal et Lundström, 2005).

La pondération en deux étapes consiste à ajuster les poids de sondage en deux temps : dans un premier temps, le poids de sondage des répondants est multiplié par un facteur d'ajustement, défini comme l'inverse de la probabilité de réponse estimée. Dans un deuxième temps, les poids ajustés sont de nouveau modifiés au moyen d'un calage.

La pondération en une étape consiste à effectuer un calage avec trois objectifs simultanés en tête :

(i) réduire le biais de non-réponse,

(ii) garantir la cohérence entre les estimations et les totaux connus au niveau de la population,

(iii) et, si possible, réduire la variance des estimateurs.

Contrairement à l'approche en deux étapes, une estimation explicite des probabilités de réponse n'est pas requise. Toutefois, le choix de la fonction de calage est important, différentes fonctions pouvant potentiellement conduire à des estimateurs exhibant des propriétés très différentes en termes de biais et de variance. Autrement dit, bien que la pondération en une étape n'utilise pas explicitement les probabilités de réponse estimées dans la construction des estimateurs, un exercice de modélisation est généralement inévitable afin d'assurer un bon choix de la fonction de calage.

Estimation régionale de taux de pauvreté par une méthode de calage

Pascal Ardilly – *Département des méthodes statistiques, Insee*

Dans le cas de la France, la production d'estimations régionales de pauvreté à partir de l'enquête annuelle SILC (*Statistics on Income and Living Conditions*) doit s'appuyer sur des sources auxiliaires. Ces dernières regroupent essentiellement la source fiscale et le recensement de la population. L'enquête SILC permet d'estimer, au niveau national, les relations individuelles entre les différentes formes de pauvreté et certaines variables explicatives. Si ces variables sont également présentes dans les sources auxiliaires, on applique les relations estimées aux individus de la population non échantillonnés, produisant ainsi des estimateurs régionaux (dits « synthétiques »). Mais cette méthode a l'inconvénient de nécessiter un ajustement spécifique de modèle propre à chaque variable d'intérêt, ce qui peut être compliqué à mettre en œuvre pour un organisme qui ne dispose pas de moyens d'ingénierie suffisants.

Il est néanmoins possible de contourner cet obstacle en utilisant une technique de calage de l'échantillon national sur les structures régionales associées aux variables auxiliaires. En utilisant la macro Calmar, on a produit ainsi un jeu de poids par région pour chaque individu de l'échantillon national qui, appliqué à toute variable liée à la pauvreté, fournit des estimations régionales satisfaisantes et très simples à obtenir. Le risque de biais est certes inhérent à la technique de modélisation, mais en la circonstance l'exploitation de la source fiscale (fichiers RDL - Revenus Disponibles Localisés) est sécurisante et le rapprochement effectué sur le taux de pauvreté monétaire entre la source fiscale et l'estimation localisée est encourageant. L'exposé reprendra la méthode utilisée et présentera les résultats obtenus pour les années d'enquête 2009 et 2010.