N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

CALMAR 2 : UNE NOUVELLE VERSION DU PROGRAMME CALMAR DE REDRESSEMENT D'ÉCHANTILLON PAR CALAGE

Olivier Sautory[1]

RÉSUMÉ

Calmar2 est la nouvelle version du programme Calmar de redressement d'échantillon par calage. Elle contient deux principaux développements.

Lorsque la collecte de l'information est réalisée à différents niveaux (par exemple ménages et individus), des calages simultanés sur les différents échantillons permettent d'assurer une cohérence sur les statistiques issues des échantillons.

En présence de non-réponse totale, la méthode de calage généralisé permet une nouvelle écriture des équations de calage, qui met en jeu deux ensembles de variables, les variables de calage proprement dites et les variables expliquant la non-réponse. On réalise ainsi une correction de la non–réponse même lorsque les variables qui l'expliquent ne sont pas connues sur les non-répondants de l'échantillon.

MOTS CLÉS : Calage, calage généralisé, calages simultanés, non-réponse.

1. LES MACROS CALMAR

1.1 Un peu d'histoire

Calmar est un programme, écrit dans le langage macro de SAS, qui permet de mettre en œuvre les méthodes de calage élaborées par Deville et Sarndäl (1992). Ce programme permet de redresser un échantillon, par repondération des individus, en utilisant une information auxiliaire disponible sur un certain nombre de variables, appelées variables de calage. Les pondérations produites par la méthode assurent le calage de l'échantillon sur des totaux de variables quantitatives connus sur la population, et sur des effectifs de modalités de variables catégorielles connus sur la population.

Calmar est un acronyme pour CALage sur MARges : on désigne ainsi la technique de redressement qui permet d'ajuster les marges (estimées à partir d'un échantillon) d'un tableau de contingence, croisant deux (ou plus) variables catégorielles, aux marges connues dans la population. Mais le programme est plus général que le "calage sur marges" stricto sensu, puisqu'il permet de caler sur des totaux de variables quantitatives.

Le programme Calmar a été développé en 1990 à l'Institut National de la Statistique et des Études Économiques (Insee), où il est depuis régulièrement mis en œuvre pour le redressement des enquêtes. De nombreux instituts statistiques français ou étrangers l'utilisent également.

La nouvelle version de ce programme, Calmar2, développée en 2003, propose à l'utilisateur de nouvelles facilités pour réaliser un calage, et permet de mettre en œuvre la méthode de traitement de la non-réponse par calage généralisé, proposée par Deville (1998).

Calmar est téléchargeable sur le site Web de l'Insee www.insee.fr, et Calmar2 sera également mis à disposition sur ce site courant 2004.

[1] Olivier Sautory, Cepe-Insee, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France, sautory@ensae.fr

1.2 Les méthodes de calage de Calmar

Rappelons le principe des méthodes de calage que permet de mettre en œuvre Calmar (voir aussi Deville et al., 1993).

On considère une population U d'individus, dans laquelle on a sélectionné un échantillon probabiliste s. Soit ***Y*** une variable d'intérêt, dont on désire estimer le total sur la population $Y = \sum_{k \in U} y_k$.
L'estimateur habituel de Y est l'estimateur de Horvitz-Thompson :

$$\hat{Y}_{HT} = \sum_{k \in s} \frac{1}{\pi_k} y_k = \sum_{k \in s} d_k \, y_k$$

On suppose que l'on connaît les totaux sur la population de J variables auxiliaires[2] $X_1 \ldots X_j \ldots X_J$, disponibles sur l'échantillon :

$$X_j = \sum_{k \in U} x_{jk}$$

On va chercher de nouvelles pondérations, les "poids de calage" w_k , qui soient aussi proches que possible, au sens d'une certaine fonction de distance, des pondérations initiales d_k (qui sont en général les "poids de sondage", égaux aux inverses des probabilités d'inclusion π_k), et qui assurent le calage sur les totaux des variables X_j , i.e. qui vérifient les **équations de calage** :

$$\forall \; j = 1 \cdots J \quad \sum_{k \in s} w_k \, x_{jk} = X_j \tag{1}$$

La solution de ce problème est donnée par $w_k = d_k F(x_k' \, \lambda)$, où $x_k' = (x_{1k} \ldots x_{Jk})$, λ est un vecteur de J multiplicateurs de Lagrange associés aux contraintes (1), et F une fonction dont l'expression dépend du choix de la fonction de distance : elle est appelée fonction de calage.

Le vecteur λ est déterminé par la résolution du système non linéaire de J équations à J inconnues résultant des équations de calage :

$$\sum_{k \in s} d_k \, F(x_k' \;\; \lambda) \, x_k = X$$

L'estimateur du total d'une variable d'intérêt sera alors l'estimateur "calé" $\hat{Y}_w = \sum_{k \in s} w_k \, y_k$.

4 méthodes de calage, correspondant à 4 fonctions de distance, étaient proposées dans la première version de Calmar ; elles sont caractérisées par la forme de la fonction F :

- la méthode *linéaire* : l'estimateur calé est alors l'estimateur par régression généralisée :

$$\hat{Y}_{reg} = \hat{Y}_{HT} + \left(X - \hat{X}_{HT}\right)\hat{B}_s \qquad où \quad \hat{B}_s = \left(\sum_{k \in s} d_k \; x_k \; x_k' \right)^{-1} \left(\sum_{k \in s} d_k \; x_k \; y_k \right)$$

- la méthode *exponentielle* : dans le cas où les variables de calage sont toutes catégorielles, cette méthode est la méthode du raking ratio (Deming et Stephan, 1940)

[2] Il s'agit de variables quantitatives ou d'indicatrices associées aux modalités de variables catégorielles.

- la méthode *logit* : cette méthode permet de donner des bornes inférieure L et supérieure U aux rapports de poids w_k / d_k

- la méthode *linéaire tronquée*, très similaire à la méthode logit.

Ces deux dernières méthodes permettent de contrôler l'étendue de la distribution des rapports de poids. C'est la méthode logit qui est la plus souvent utilisée, car elle permet d'éviter les poids trop élevés, qui entraînent des risques de manque de robustesse des estimations, et les poids trop faibles, voire négatifs, auxquels peut conduire la méthode linéaire.

Précision

Les estimateurs calés Y_w ont tous la même précision (asymptotique), quelle que soit la méthode utilisée : la variance approchée de Y_w est donc égale à celle de l'estimateur par régression $\hat{Y}_{reg}$: $AV\left(\hat{Y}_w\right) = \sum\sum_U \Delta_{k\ell}\left(d_k\ E_k\right)\left(d_\ell\ E_\ell\right)$, où

$$E_k = y_k - x_k'\ B,\ avec\ B = \left(\sum_{k\in U} x_k\ x_k'\right)^{-1}\left(\sum_{k\in U} x_k\ y_k\right).$$

E_k est le résidu de la régression de $\boldsymbol{Y}$ sur les X_j dans la population U.

Cette variance est d'autant plus faible que la corrélation entre la variable d'intérêt $\boldsymbol{Y}$ et les variables de calage $X_1 \ldots X_j \ldots X_J$ est élevée.

Un estimateur de variance est donné par $\hat{V}\left(\hat{Y}_w\right) = \sum\sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}}\left(d_k\ e_k\right)\left(d_\ell\ e_\ell\right)$, où $e_k = y_k - x_k'\ B_s$, avec

$$B_s = \left(\sum_{k\in s} d_k\ x_k\ x_k'\right)^{-1}\left(\sum_{k\in s} d_k\ x_k\ y_k\right).$$

e_k est le résidu de la régression (pondérée par les d_k) de $\boldsymbol{Y}$ sur les X_j dans l'échantillon s.

1.3 Les nouveautés de Calmar2

Calmar2 (Le Guennec et Sautory, 2003) propose les 4 méthodes de calage qui viennent d'être présentées, mais également :

- le traitement de calages simultanés pour différents niveaux d'une même enquête

- la correction de la non-réponse totale par la méthode de calage généralisé.

Ces deux points seront abordés en détail dans les paragraphes suivants.

Calmar2 propose une solution au problème posé par les colinéarités entre les variables de calage : l'utilisation de matrices inverses généralisées permet de calculer les poids de calage, alors que Calmar générait dans ce cas un message d'erreur

D'autre part, une nouvelle fonction de distance est proposée dans Calmar2, la fonction *sinus hyperbolique généralisée*, dépendant d'un paramètre α. Cette méthode donne des poids toujours positifs, comme la méthode exponentielle, mais conduit à des distributions de poids moins étendues que cette dernière du côté des poids élevés.

D'autre part, le coefficient α permet de réduire l'étendue de la distribution des poids, comme le font les méthodes logit et linéaire tronquée, mais à l'aide d'un seul paramètre (Roy *et al.*, 2001).

Enfin, l'ergonomie du programme a été améliorée, sur deux points en particulier :

- la macro accepte en entrée des variables de calage catégorielles sans que l'utilisateur ait à opérer un recodage préalable pour obtenir des modalités séquentielles ;
- l'utilisateur peut, s'il le souhaite, entrer les paramètres de façon interactive à l'aide d'écrans de saisie, qui le guident dans ses choix.

2. LES CALAGES SIMULTANÉS

2.1 Le problème

Dans certaines enquêtes, la collecte d'informations s'opère à différents niveaux d'observation :

- l'enquête PCV de l'Insee sur les conditions de vie des ménages comporte des questions sur les ménages (type de logement, nombre de personnes, profession du chef de ménage...), sur chacun des individus du ménage (sexe, âge, profession…), et en général un questionnaire spécifique sur un individu tiré au hasard parmi les personnes "éligibles" du ménage (souvent les 15 ans ou plus), appelé "individu–Kish" ;
- l'enquête annuelle d'entreprises (EAE) réalisée par le Ministère de l'Industrie comprend, en plus du questionnaire sur l'activité globale de l'entreprise, un volet concernant chacun de ses établissements.

Lors du redressement de l'enquête, on peut soit opérer des calages indépendants sur les différents niveaux d'observation, soit opérer des calages simultanés, ou "intégrés". La réalisation de calages simultanés permet d'obtenir *in fine* les mêmes poids pour tous les individus d'un même ménage, si tous les individus du ménage sont sollicités pour l'enquête ; cela permet aussi d'assurer une cohérence entre les statistiques obtenues à partir des différents fichiers de l'enquête. Par exemple, avec des calages indépendants sur un échantillon de ménages et sur l'échantillon des individus correspondants, le nombre de ménages d'une personne estimé à partir du premier échantillon n'a aucune raison de coïncider avec le nombre de personnes appartenant aux ménages d'une personne estimé à partir du deuxième échantillon.

2.2 La méthode

De façon plus générale, les situations décrites ci-dessus correspondent au cas où l'on a réalisé un sondage en grappes ou à plusieurs degrés, et où on dispose d'une information auxiliaire sur les grappes (ou les unités primaires) et les unités secondaires, et où les variables d'intérêt de l'enquête portent aussi bien sur les grappes (ou les UP) que sur les unités secondaires.

La méthode proposée pour réaliser des calages simultanés a été présentée par Sautory (1996). Elle est plus générale que la méthode proposée par Lemaître et Dufour (1987). Elle consiste à réaliser un seul calage, au niveau des UP : pour cela, on calcule pour chaque UP les estimations des totaux des variables de calage définies au niveau des US, et ces estimations sont utilisées dans le calage au niveau des UP, qui fait intervenir simultanément les variables de niveau UP et de niveau US.

Ainsi, si $\boldsymbol{X}$ est une variable de calage pour les US, on calcule pour chaque UP m l'estimation $\hat{X}_m = \sum_{k \in m} \pi_{k/m}\, x_k$, où $\pi_{k/m}$ désigne la probabilité d'inclusion de l'US k sachant que l'UP m a été tirée. L'équation de calage pour la variable $\boldsymbol{X}$ s'écrit alors $\sum_{m \in s_M} w_m \hat{X}_m = X$, où s_M désigne l'échantillon des UP.

2.3 Un exemple

Examinons l'exemple d'une enquête où on a tiré un échantillon de ménages s_M, pour lesquels on a recueilli certaines informations. On a interrogé tous les individus des ménages sélectionnés, qui constituent un échantillon s_I. De plus, on a interrogé, avec un questionnaire spécifique, un individu k_m (appelé individu-Kish) dans chaque ménage m sélectionné, tiré selon un sondage aléatoire simple sans remise parmi les e_m individus éligibles du ménage, par exemple les individus de 15 ans ou plus.

On note :

x_m = vecteur des variables auxiliaires connues pour tout ménage m de l'échantillon s_M de ménages

$X = \sum_{m \in U_M} x_m$ = vecteur des totaux de ces variables sur la population U_M de ménages, connus

$z_{m,i}$ = vecteur de variables auxiliaires connues pour tout individu i du ménage m

$Z = \sum_{i \in U_I} z_i$ = vecteur des totaux de ces variables sur la population U_I d'individus, connus.

v_{k_m} = vecteur de variables auxiliaires connues pour l'individu-Kish k_m du ménage m

$V = \sum_{i \in U_I^e} v_i$ = vecteur des totaux de ces variables sur la population U_I^e des individus éligibles, connus.

Les probabilités d'inclusion des ménages m sont notées π_m, et on pose $d_m = 1/\pi_m$. Les probabilités d'inclusion des individus (m,i) sachant que le ménage m a été tiré valent 1. Enfin, la probabilité d'inclusion de l'individu-Kish k_m sachant que le ménage m a été tiré vaut $1/e_m$.

La méthode consiste à réaliser un seul calage, au niveau ménage, en calculant pour chaque ménage m, les totaux des variables de calage des individus $Z_m = \sum_{(m,i) \in men_m} z_{m,i}$, et les totaux estimés des variables de calage des individus-Kish $\hat{V}_m = e_m v_{k_m}$.

Le vecteur de variables de calage pour le ménage m devient $(x_m, Z_m, \hat{V}_m)$, et le vecteur des totaux (X, Z, V). Les équations de calage s'écrivent :

$$\sum_{m \in s_M} d_m F(x_m' \lambda + Z_m' \mu + \hat{V}_m' \gamma)(x_m, Z_m, \hat{V}_m) = (X, Z, V)$$

λ, μ, γ désignent les composantes du vecteur des multiplicateurs de Lagrange.

Les solutions $w_m = d_m F(x_m' \lambda + Z_m' \mu + \hat{V}_m' \gamma)$ de ces équations sont les nouvelles pondérations attribuées aux ménages. La pondération $w_{m,i}$ attribuée à l'individu i du ménage m dans l'échantillon d'individus est alors égale à la pondération w_m du ménage m. La pondération w_{k_m} attribuée à l'individu-Kish du ménage m est alors égale à $e_m w_m$. On vérifie qu'avec ces pondérations les différents échantillons sont bien calés sur les totaux X, Z et V :

$$\sum_{i \in s_I} w_{m,i} z_{m,i} = \sum_{m \in s_M} w_m \left(\sum_{(m,i) \in men_m} z_{m,i} \right) = \sum_{m \in s_M} w_m Z_m = Z$$

$$\sum_{k_m \in s_K} w_{k_m} v_{k_m} = \sum_{k_m \in s_K} w_m e_m v_{k_m} = \sum_{k_m \in s_K} w_m \hat{V}_m = V$$

Cette méthode pouvait être mise en œuvre avec Calmar (voir Caron et Sautory, 2004), mais cela nécessitait un peu de programmation en SAS. Le programme Calmar2 réalise toutes les opérations nécessaires pour se ramener à un calage unique. L'utilisateur doit fournir les différentes tables en entrée correspondant aux différents niveaux d'observation, ainsi que les totaux des variables de calage.

Estevao et Sarndäl (2003) comparent plusieurs méthodes de calage dans le cas de plans de sondage à deux degrés, dont la méthode présentée ci-dessus

3. LE CALAGE GÉNÉRALISÉ

3.1 Le principe

Contrairement à la présentation habituelle du calage à l'aide de fonctions de distance entre poids, Deville (par exemple 2002) pose directement les équations de calage, avec des fonctions de calage définies sous une forme très générale : $F_k : \lambda \in \mathbb{R}^J \rightarrow F_k(\lambda) \in \mathbb{R}$, avec $F_k(0) = 1$, où λ est un vecteur de J paramètres d'ajustement.

Les équations de calage généralisé s'écrivent : $\sum_{k \in s} d_k \, F_k(\lambda) x_k = X$, où x_k désigne comme précédemment le vecteur des J variables de calage. La solution en λ de ce système conduit aux nouveaux poids $w_k = d_k \, F_k(\lambda)$.

Résultat fondamental

On pose $grad \, F_k(0) = z_k$, vecteurs que l'on appellera "instruments" (voir ci-dessous). On montre que les estimateurs calés fondés sur les mêmes instruments et les mêmes variables de calage sont tous équivalents asymptotiquement.

En effet, on peut réécrire les équations de calage $X = \sum_{k \in s} d_k \left(1 + z_k' \, \lambda + O\|\lambda\|^2\right) x_k$.

D'où : $X - \hat{X}_{HT} = \left[\sum_{k \in s} d_k \, x_k \, z_k'\right] \lambda + \sum_{k \in s} d_k \, x_k \, O\|\lambda\|^2$, soit $\lambda_s = (T_{szx}')^{-1}\left(X - \hat{X}_{HT}\right) + O\left\|X - \hat{X}_{HT}\right\|^2$ en posant $T_{szx} = \sum_{k \in s} d_k \, z_k \, x_k'$, que l'on suppose de plein rang.

Un estimateur calé $\hat{Y}_w = \sum_{k \in s} w_k \, y_k$ est donc équivalent asymptotiquement à $\sum_{k \in s} d_k \left[1 + z_k' \, (T_{szx}')^{-1}\left(X - \hat{X}_{HT}\right)\right] y_k = \hat{Y}_{HT} + \left(X - \hat{X}_{HT}\right)' T_{szx}^{-1} \sum_{k \in s} d_k \, z_k \, y_k = \hat{Y}_{HT} + \left(X - \hat{X}_{HT}\right)^{-1} \hat{B}_{szx} = \hat{Y}_{reg\,i}$

$\hat{B}_{szx}$ vérifie $\sum_{k \in s} d_k \, z_k \, y_k = \left(\sum_{k \in s} d_k \, z_k \, x_k'\right) \hat{B}_{szx}$: c'est le vecteur des coefficients de la régression (pondérée par les d_k) avec variables instrumentales de $\boldsymbol{Y}$ sur les variables $X_1 \ldots X_j \ldots X_J$, les variables composant les vecteurs z_k étant les "instruments" (voir par exemple Fuller, 1987). Par analogie avec l'estimateur par régression généralisée, l'estimateur $\hat{Y}_{reg\,i}$ est appelé estimateur par régression avec variables instrumentales.

3.2 Forme usuelle des fonctions de calage

Dans la pratique, les fonctions de calage F_k sont en général de la forme $F_k(\lambda) = F(z_k' \lambda)$, où z_k est un vecteur de J variables $\boldsymbol{Z}_j$ connues sur l'échantillon s, et F une fonction de $\mathbb{R}$ dans $\mathbb{R}$ telle que $F(0) = 1$ et $F'(0) = 1$ (d'où grad $F_k(0) = z_k$).

Les équations de calage s'écrivent alors : $\sum_{k \in s} d_k \, F(z_k' \, \lambda) x_k = X$

Dans le cas où la fonction F est linéaire, $F(z_k' \lambda) = 1 + z_k' \lambda$, l'estimateur calé est l'estimateur par régression avec variables instrumentales $\hat{Y}_{regi}$, car on a alors $\lambda_s = (T_{szx}')^{-1}(X - \hat{X}_{HT})$.

3.3 Précision

Par des démonstrations analogues à celles utilisées dans le cas du calage classique, on obtient les résultats suivants.

La variance approchée de l'estimateur calé a pour expression : $AV(\hat{Y}_w) = \sum_U \sum \Delta_{k\ell} (d_k E_k)(d_\ell E_\ell)$

où $E_k = y_k - x_k' B_{zx}$, *avec* $B_{zx} = \left(\sum_{k \in U} z_k x_k'\right)^{-1} \left(\sum_{k \in U} z_k y_k\right)$, est le résidu de la régression de $\boldsymbol{Y}$ sur les $\boldsymbol{X}_j$ dans U, avec les variables instrumentales $\boldsymbol{Z}_j$.

Un estimateur de variance est donné par $\hat{V}(\hat{Y}_w) = \sum_s \sum \frac{\Delta_{k\ell}}{\pi_{k\ell}} (d_k e_k)(d_\ell e_\ell)$,

où $e_k = y_k - x_k' B_{szx}$, *avec* $B_{szx} = \left(\sum_{k \in s} d_k z_k x_k'\right)^{-1} \left(\sum_{k \in s} d_k z_k y_k\right)$, est le résidu de la régression (pondérée par les d_k) de $\boldsymbol{Y}$ sur les $\boldsymbol{X}_j$ dans l'échantillon s, avec les variables instrumentales $\boldsymbol{Z}_j$.

4. LE CALAGE EN PRÉSENCE DE NON-RÉPONSE TOTALE

4.1 Les méthodes usuelles de correction de la non-réponse totale

La correction de la non-réponse totale est généralement réalisée par des techniques de repondération des unités répondantes. Ces techniques sont fondées sur une modélisation du mécanisme de réponse. Ce mécanisme est assimilé à un tirage aléatoire d'un échantillon r (de taille n_r) au sein de l'échantillon s. Ce tirage peut être vu comme une phase supplémentaire ajoutée au plan de sondage initial, définie par un "pseudo" plan de sondage noté q(r|s). À ce plan sont associées les probabilités de réponse individuelles $p_k = P(k \in r / k \in s)$.

Si ces probabilités étaient connues, le total Y d'une variable intérêt serait estimé sans biais par $\hat{Y}_{exp} = \sum_{k \in r} y_k / (\pi_k p_k)$, appelé estimateur par expansion. En réalité, le plan q(r|s), et donc les probabilités p_k, sont inconnues : il faut donc les estimer, en postulant un modèle pour le mécanisme de réponse, et en utilisant une méthode d'estimation (maximum de vraisemblance, moments, …).

Un modèle naturel est le modèle de Poisson : $q(r|s) = \prod_{k \in r} p_k \prod_{k \in s \setminus r} (1 - p_k)$. Pour spécifier complètement ce modèle, il faut donner la forme des probabilités p_k. Nous présentons ci-dessous trois types de modélisation classique du mécanisme de non-réponse.

Modèle de réponse uniforme

On suppose que chaque individu a la même probabilité de réponse : $p_k = p \ \forall k \in U$. La méthode du maximum de vraisemblance conduit à l'estimation $\hat{p} = n_r / n =$ taux de réponse observé.

Groupes homogènes de réponse

On partitionne la population U en H groupes supposés homogènes du point de vue de la non-réponse : tous les individus du groupe h ont la même probabilité de réponse, notée p_h. La méthode du maximum de vraisemblance conduit aux estimations $\hat{p}_h = n_{rh} / n_h$, où n_h (resp. n_{rh}) est le nombre d'individus du groupe h appartenant à l'échantillon s (resp. l'échantillon r). $\hat{p}_h$ est donc le taux de réponse observé dans le groupe h.

Modèle linéaire généralisé

La probabilité de réponse est une fonction d'un vecteur z_k de variables $\boldsymbol{Z}_j$ explicatives de la non-réponse, et d'un paramètre inconnu β : $p_k = 1/H(z_k' \beta)$, où H est une fonction définie sur $\mathbb{R}$ et à valeurs dans $[1,+\infty[$ (en principe…). Pour estimer β, donc les p_k, les variables $\boldsymbol{Z}_j$ doivent être connues sur les répondants et les non-répondants.

On peut adopter une modélisation encore plus générale, de la forme $p_k = 1/H_k(\beta)$, où β est un vecteur de J paramètres d'ajustement, et H_k une fonction dépendant de l'individu k.

Examinons maintenant différentes stratégies de calage en présence de non-réponse totale.

4.2 Le calage après correction de la non-réponse

On suppose que l'on a réalisé une correction de la non-réponse totale, par exemple avec l'une des méthodes qui viennent d'être présentées. On peut alors réaliser un calage classique, en partant des poids corrigés pour non-réponse $d_k^* = d_k / \hat{p}_k$. Les équations de calage s'écrivent $\sum_{k \in r} d_k^* F^*(x_k' \lambda) x_k = X$, F^* étant l'une des fonctions de calage habituelles.

4.3 Le calage classique direct

Une deuxième stratégie consiste à réaliser un calage directement, sans correction préalable de la non-réponse. Les équations de calage s'écrivent : $\sum_{k \in r} d_k F(x_k' \lambda) x_k = X$.

Si parmi les variables de calage figure la variable constante égale à 1, ou au moins une variable catégorielle, on peut multiplier les d_k par une constante sans que cela change les $w_k = d_k F(x_k' \lambda)$. On peut donc réécrire les équations de calage $\sum_{k \in r} d_k \frac{1}{n_r / n} F(x_k' \lambda) x_k = X$, ce qui montre que cette stratégie équivaut à la stratégie précédente avec une correction de la non-réponse avec un modèle de réponse uniforme.

Dupont (1996) a comparé les deux stratégies, à l'aide de considérations théoriques et de simulations. Son étude a conduit aux résultats suivants.

Si la correction de la non-réponse est réalisée par un modèle linéaire généralisé, où la fonction H est l'une des fonctions usuelles F de calage, et si les variables de calage $\boldsymbol{X}_j$ contiennent les variables explicatives de la non-réponse $\boldsymbol{Z}_j$, alors les deux stratégies donnent des résultats très proches.

De plus, si les variables de calage $\boldsymbol{X}_j$ sont exactement les variables $\boldsymbol{Z}_j$ explicatives de la non-réponse, les deux stratégies suivantes sont équivalentes :

- réaliser une correction de la non-réponse par un modèle linéaire généralisé, avec comme fonction H la fonction exponentielle, puis un calage à partir des poids corrigés avec comme fonction de calage F^* la fonction exponentielle
- réaliser un calage direct, à partir des poids initiaux, avec comme fonction de calage F la fonction exponentielle.

Il en est de même si on effectue une correction de la non-réponse par un modèle de groupes homogènes de réponse, puis une post-stratification, où groupes et post-strates coïncident : ceci est en effet équivalent à réaliser une post-stratification "formelle" directe sur l'échantillon des répondants.

L'avantage du calage direct est qu'il ne nécessite pas une modélisation explicite du mécanisme de réponse. Lundström et Särndal (1999) ont également étudié les propriétés du calage direct, et ont proposé en particulier des estimateurs de variance prenant en compte la variance d'échantillonnage et la variance due à la non-réponse.

4.4 Le calage généralisé direct

Partons d'un système d'équations de calage sur l'échantillon des répondants, de la forme : $\sum_{k\in r} d_k\, H_k(\beta) x_k = X$. Ces équations peuvent s'interpréter de la façon suivante.

On postule un modèle de réponse de la forme $p_k = \frac{1}{H_k(\beta_0)}$, β_0 désignant la vraie valeur du paramètre du modèle.

On peut réécrire les équations de calage de la façon suivante, $\hat{\beta}$ désignant la solution du système :

$$X = \sum_{k\in r} d_k\, H_k(\beta_0) \frac{H_k(\hat{\beta})}{H_k(\beta_0)} x_k = \sum_{k\in r} d_k\, H_k(\beta_0) \frac{H_k(\beta_0+\lambda)}{H_k(\beta_0)} x_k = \sum_{k\in r} d_k \frac{1}{p_k} F_k(\lambda)\, x_k$$

$$\text{avec } \hat{\beta} = \beta_0 + \lambda \quad \text{et} \quad F_k(\lambda) = \frac{H_k(\beta_0+\lambda)}{H_k(\beta_0)}$$

Ces équations apparaissent donc comme des équations de calage généralisé, où les pondérations initiales sont les d_k / p_k , i.e. les poids de sondage corrigés pour non-réponse, et les fonctions F_k , qui vérifient $F_k(0) = 1$, sont les fonctions de calage. Les instruments sont $z_k^* = \text{grad}\, F_k(0) = \frac{1}{H_k(\beta_0)} \text{grad}\, H_k(\beta_0)$. Résoudre ce système revient donc à faire simultanément une correction de la non-réponse et un calage généralisé.

On peut utiliser les résultats du § 3.3 pour calculer la précision des estimateurs calés selon cette méthode.

La variance approchée $AV(\hat{Y}_w)$ utilise les résidus de la régression avec variables instrumentales dans la population $E_k = y_k - x_k' B_{z^*x}$.

L'estimateur de variance $\hat{V}(\hat{Y}_w)$ utilise les résidus de la régression avec variables instrumentales dans l'échantillon des répondants, pondérée par les $d_k H_k(\beta_0)$: $e_k = y_k - x_k' B_{rz^*x0}$, *où* $\sum_{k\in r} d_k\, H_k(\beta_0)\, z_k^* \left(y_k - x_k'\, B_{rz^*x0}\right) = 0$: B_{rz^*x0} est l'estimateur de B_{z^*x} que l'on calculerait si les probabilités de réponse $H_k^{-1}(\beta_0)$ étaient connues. Ces probabilités sont inconnues, à cause de β_0 : on les estime en remplaçant β_0 par $\hat{\beta}$. Les résidus deviennent $e_k = y_k - x_k' \hat{B}_{rz^*x}$, *où* $\sum_{k\in r} d_k\, H_k(\hat{\beta})\, z_k^* \left(y_k - x_k'\, \hat{B}_{rz^*x}\right) = 0$: il s'agit d'une régression avec variables instrumentales dans l'échantillon r, pondérée par les poids de calage $w_k = d_k H_k(\hat{\beta})$.

Remarque : la variance estimée $\hat{V}\left(\hat{Y}_w\right)$ s'écrit sous la forme $Q_1(e_k)+Q_2(e_k)$, où la forme quadratique $Q_1(e_k)$ désigne la variance estimée 1ère phase (tirage de l'échantillon s), et $Q_2(e_k)$ la variance estimée 2ème phase ("tirage" de l'échantillon r).

Cas d'un modèle linéaire généralisé

Dans la pratique, les fonctions $H_k(\beta)$ sont de la forme $H(z_k'\beta)$, où z_k est un vecteur de variables $\boldsymbol{Z}_j$ explicatives de la non-réponse. Les équations de calage s'écrivent :

$$\sum_{k\in r} d_k\, H(z_k'\ \beta)\, x_k = X \qquad \text{(E)}$$

Les instruments sont $z_k^* = z_k \dfrac{H'(z_k'\beta_0)}{H(z_k'\beta_0)}$. Ils sont "estimés" par $z_k \dfrac{H'(z_k'\hat{\beta})}{H(z_k'\hat{\beta})}$. Ils sont égaux aux z_k lorsque H est la fonction exponentielle.

Propriétés de la méthode

La dissociation, dans le système d'équations de calage (E), entre les variables $\boldsymbol{Z}_j$ explicatives de la non-réponse et les variables de calage $\boldsymbol{X}_j$, produit une réduction du biais provoqué par la non-réponse, grâce aux $\boldsymbol{Z}_j$, et une diminution de la variance, grâce aux $\boldsymbol{X}_j$.

La méthode exige que le nombre de variables $\boldsymbol{Z}_j$ (variables quantitatives et indicatrices de modalités de variables catégorielles) soit égal au nombre de variables de calage $\boldsymbol{X}_j$. De plus, cette méthode n'est efficace que si les corrélations entre les $\boldsymbol{Z}_j$ et les $\boldsymbol{X}_j$ sont suffisamment élevées.

Contrairement aux méthodes usuelles de correction de la non-réponse, la méthode permet une correction même lorsque les variables qui causent la non-réponse ne sont connues que sur les répondants. Elle traite en particulier le cas où les facteurs de la non-réponse sont des variables d'intérêt (mécanisme de réponse "non ignorable").

Calmar2 permet de mettre en œuvre cette méthode, où les fonctions H sont les fonctions usuelles de calage. Le Guennec (2004) montre un exemple d'application de la méthode sur les données d'une enquête.

RÉFÉRENCES

Caron, N. et Sautory, O. (2004), "Calages simultanés pour différentes unités d'une même enquête", *Document de travail Méthodologie statistique n° 0403, INSEE.*

Deming, W.E. and Stephan, F.F. (1940), "On a least squares adjustment of a sampled frequency table when the exal totals are known", *Annals of Mathematical Statistics*, 11, pp. 427-444.

Deville, J.-C. and Särndal, C.-E (1992), "Calibration estimation in survey sampling", .*Journal of the American Statistical Association*, 87, n°418, pp. 375-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993), "Generalized raking procedures in survey sampling", *Journal of the American Statistical Association*, 88, n°423, pp. 1013-1020.

Deville, J.–C. (1998), "La correction de la non-réponse par calage ou par échantillonnage équilibré", Actes du colloque de la Société Statistique du Canada, Sherbrooke, Canada.

Deville, J.–C. (2004), "La correction de la non-réponse par calage généralisé", *Actes des journées de méthodologie statistique, 16 et 17 décembre 2002, INSEE-Méthodes à paraître.*

Dupont, F. (1996), "Calage et redressement de la non-réponse totale", *Actes des journées de méthodologie statistique, 15 et 16 décembre 1993, INSEE-Méthodes n°56-57-58.*

Estevao, V. and Särndal, C.-E. (2003), "Calibration estimation in sample surveys : an overview and recent developments", article présenté au Joint Statistical Meetings de l'ASA, San Fransisco.

Fuller, W. (1987), "Measurement Error Models", New York : Wiley.

Lemaître, G. and Dufour, J. (1987), "An integrated method for weighting persons and families", *Survey Methodology*, 13, pp. 199-207.

Le Guennec, J. et Sautory, O. (2003). "La macro Calmar2, manuel d'utilisation", document interne INSEE.

Le Guennec, J. (2004). "Correction de la non-réponse par calage généralisé : une expérimentation", *Actes des journées de méthodologie statistique, 16 et 17 décembre 2002, INSEE-Méthodes à paraître.*

Lundström, S. and Särndal, C.-E. (1999), "Calibration as a standard method for treatment of nonresponse", *Journal of Official Statistics,* 15, pp. 305-327.

Roy, G. et Vanheuverzwyn, A. (2001). "Redressement par la macro CALMAR : applications et pistes d'amélioration", in *Traitements des fichiers d'enquête*. Presses Universitaires de Grenoble, pp. 31–46.

Sautory, O (1996) : "Calage sur des échantillons de ménages, d'individus, d'individus-Kish, issus d'une même enquête", communication invitée aux Journées de Statistique de l'ASU, Québec, Canada.