CALMAR 2 : UNE NOUVELLE VERSION DE LA MACRO CALMAR DE REDRESSEMENT D'ÉCHANTILLON PAR CALAGE

Josiane LE GUENNEC [(*)], Olivier SAUTORY[(**)]

[(*)] Cepe - Ensai
[(**)] Cepe - Insee

La macro SAS CALMAR (CALage sur MARges) 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. Ces nouvelles pondérations, les "poids de calage" w_k, sont aussi proches que possible, au sens d'une certaine distance, des pondérations initiales d_k (qui sont en général les "poids de sondage", égaux aux inverses des probabilités d'inclusion p_k).
On trouvera dans [2] et [5] une présentation détaillée de la théorie du calage.

Ce programme est utilisé à l'Insee depuis 1990 pour redresser les échantillons des enquêtes-ménages, ainsi que par de nombreux organismes statistiques français et étrangers dans différents domaines de la statistique d'enquête.

Une nouvelle version de la macro, *Calmar 2*, offre à l'utilisateur de nouvelles fonctionnalités qui sont présentées dans ce papier.

1. Calages simultanés

1.1 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 opérer des calages indépendants sur les différents niveaux d'observation, ou bien opérer des calages simultanés : dans ce dernier cas, on est assuré par exemple d'avoir les mêmes poids pour tous les individus d'un même ménage, et plus généralement d'avoir une cohérence entre les statistiques obtenues à partir des différents fichiers de l'enquête.

Ainsi, dans le cas d'une enquête ménages et individus, un calage simultané sur ces deux niveaux assure d'obtenir la même statistique pour :

- estimer le nombre de ménages dont le chef est ouvrier, à partir du fichier-ménages
- estimer le nombre de chefs de ménage ouvriers, à partir du fichier-individus.

1.2 La méthode

Exemple 1

Examinons le cas d'une enquête ménage où tous les individus du ménage sont interrogés. Le calage est réalisé au niveau d'observation le plus élevé, ici le ménage.

Soient : 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 de l'échantillon

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

On calcule d'abord les totaux par ménage : $z_m = \sum_{i \in m} z_{m,i}$ (on alors $Z = \sum_{m \in U_m} z_m$).

Le vecteur de variables de calage pour le ménage m devient (x_m, z_m), et le vecteur des totaux (X, Z).

On réalise alors le calage sur l'échantillon des ménages s_m :

$$\sum_{m \in s_m} \frac{1}{p_m} F(x_m' \lambda + z_m' \mu)(x_m, z_m) = (X, Z)$$

où F désigne une fonction de calage.

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 dans l'échantillon s_m.

Exemple 2

Examinons le cas d'une enquête ménage où tous les individus du ménage sont interrogés, ainsi qu'un individu "Kish" k_m tiré parmi les e_m individus "éligibles" du ménage m. On note $U_i^e (\subset U_i)$ la

population des individus éligibles. Le calage est réalisé au niveau d'observation le plus élevé, le ménage.

La pondération initiale de l'individu-Kish du ménage m est : $d_{k_m} = d_m e_m$.

Soient : v_{k_m} = vecteur des variables auxiliaires connues pour tout individu-Kish k_m de l'échantillon d'individus-Kish s_K

$V = \sum_{i \in U_i^e} v_i$ = vecteur des totaux de ces variables sur U_i^e, connus.

On calcule, au niveau ménage, de nouvelles variables $e_m v_{k_m}$.

Le vecteur de variables de calage pour le ménage m devient $(x_m, z_m, e_m v_{k_m})$, et le vecteur des totaux (X, Z, V).

On réalise alors le calage sur l'échantillon des ménages s_m :

$$\sum_{m \in s_m} \frac{1}{p_m} F(x_m' \lambda + z_m' \mu + e_m v_{k_m}' \gamma)(x_m, z_m, e_m v_{k_m}) = (X, Z, V)$$

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 dans l'échantillon s_m, et la pondération de l'individu-Kish k_m du ménage m dans l'échantillon d'individus-Kish vaut $w_{k_m} = w_m e_m$.

On a bien : $\sum_{k_m \in s_K} w_{k_m} v_{k_m} = V$.

La méthode est présentée de façon plus détaillée dans [1].

1.3 Mise en œuvre dans Calmar 2

Le programme Calmar 2 permet de réaliser simultanément le calage pour les différentes unités observées, y compris dans le cas d'enquêtes comprenant des individus-Kish.

L'utilisateur doit fournir autant de tables en entrée qu'il y de niveaux d'observation, contenant pour chacune les variables de calage relatives au niveau considéré et la variable de pondération initiale (poids de sondage). Il doit également constituer les tables des marges correspondantes.

Calmar 2 réalise alors toutes les opérations nécessaires pour se ramener, comme dans l'exemple ci-dessus, à un calage unique :

- "remontée" de l'information au niveau le plus élevé,
- constitution d'une nouvelle table de marges contenant toute l'information auxiliaire utilisée par les différents calages,
- réalisation du calage au niveau supérieur, calcul des poids de calage à ce niveau,
- "redescente" des poids de calage aux niveaux d'observation inférieurs.

2. Le traitement de la non-réponse totale par calage généralisé

La ***méthode de calage généralisé appliquée au traitement de la non–réponse*** consiste à écrire les équations de calage sous la forme suivante :

$$\sum_{k \in r} d_k F(z_k' \ l) x_k = X$$

où z_k désigne le vecteur des variables explicatives de la non-réponse, connues sur l'échantillon r des répondants, et x_k le vecteur des variables de calage, bien corrélées aux variables d'intérêt, connues sur l'échantillon, et dont on connaît les totaux (vecteur X) dans la population. La fonction F est une des fonctions de calage disponibles dans le programme.

La méthodologie est présentée en détail dans [3].

Cette méthode, programmée dans Calmar 2, permet donc une correction de la non–réponse même lorsque les variables qui l'expliquent ne sont observées que sur l'échantillon des répondants, en particulier lorsque ces variables sont des variables d'intérêt. Elle produit une réduction du biais dû à la non-réponse grâce aux variables Z, et une diminution de la variance grâce aux variables X.

Dans le cas où la fonction F est linéaire, cette généralisation du calage peut se voir comme une méthode d'estimation par régression avec variables instrumentales (les variables z_k).

La mise en œuvre de cette méthode par Calmar 2 ne pose pas de difficulté particulière : tables de données et tables des marges se présentent sous la forme habituelle, l'utilisateur devant indiquer dans la table des marges le statut des variables de calage : variables "Z" ou variables "X".

3. Autres nouveautés de Calmar 2

3.1 Une nouvelle fonction de distance

Suite à une suggestion d'utilisateurs de Calmar (voir [4]), une nouvelle fonction de distance est proposée dans Calmar 2, la fonction *sinus hyperbolique généralisée*, définie par :

$$G_a(x) = \frac{1}{2a} \int_1^x sh[\, a\,(t - \frac{1}{t})\,]\, dt \qquad \text{où } a \text{ est un coefficient positif}$$

Cette méthode donne des poids toujours positifs, comme la méthode exponentielle (ou raking ratio), mais conduit à des distributions de poids moins étendues que cette dernière du côté des poids élevés.

D'autre part, le coefficient a permet de réduire l'étendue de la distribution des poids, comme le font les méthodes logit et linéaire tronquée. Un avantage de cette distance est que cette diminution de l'étendue des poids s'obtient en agissant sur un seul paramètre, au lieu de deux (les limites L et U définissant l'étendue des rapports de poids dans les méthodes logit et linéaire tronquée).

3.2 Prise en compte des colinéarités

La résolution des équations de calage nécessite l'inversion d'une matrice de la forme :

$$F = \sum_{k \in s} x_k x_k' \,.$$

Il est donc nécessaire que les variables de calage ne soient pas colinéaires.

Le programme Calmar éliminait automatiquement les colinéarités structurelles qui apparaissent lorsque plusieurs variables catégorielles figurent dans les variables de calage (la somme des indicatrices relatives aux modalités d'une même variable catégorielle vaut toujours 1).

Mais d'autres colinéarités entre variables de calage peuvent apparaître, parfois de façon insidieuse.

Par exemple, si l'on veut caler l'échantillon d'une enquête ménages sur deux répartitions "croisées", *taille du ménage ´ région* et *CS du chef de ménage ´ région*, on obtient une équation redondante par région.

Calmar détectait ces colinéarités, mais l'utilisateur devait redéfinir les variables de calage, de façon à supprimer ces colinéarités.

Calmar 2 propose une autre solution à ce problème, qui est d'utiliser la technique des **matrices inverses généralisées**.

3.3 Points divers

La contrainte imposée à l'utilisateur que les variables de calage *catégorielles* soient obligatoirement codées 1, 2, 3... est supprimée : désormais Calmar 2 accepte n'importe quel codage pour les variables catégorielles.

L'utilisateur peut indiquer un coefficient multiplicateur à appliquer aux poids initiaux avant calage : ceci permet par exemple de réaliser une correction uniforme de la non-réponse.

Bibliographie

[1] Caron, N. et Sautory, O. (1998), "Calages simultanés pour différentes unités d'une même enquête", *note interne*, INSEE.

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

[3] Deville, J.-C. (2002), "La correction de la non-réponse par calage généralisé", *Papier présenté aux Journées de méthodologie statistique de l'INSEE, décembre 2002.*

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

[5] Sautory, O. (1991). "Redressement d'échantillons d'enquêtes auprès des ménages par calage sur marges", *Document de travail de la Direction des Statistiques Démographiques et Sociales* n° F 9103, INSEE.