CALMAR 2 : une nouvelle version de la macro Calmar de redressement d'échantillon par calage

Josiane Le Guennec (Cepe-Ensai)

Olivier Sautory (Cepe-Insee)

Calages simultanés

Le problème

Collecte d'informations à différents niveaux d'observation :

- ménages, individus, individus-Kish
- entreprises, établissements.

+ information auxiliaire disponible pour chaque niveau.

Que faire ?

- calages indépendants sur les différents niveaux d'observation ;
- calages simultanés :
  - mêmes poids pour tous les individus (établissements) d'un(e) même ménage (entreprise)
  - cohérence entre les statistiques provenant des différents fichiers de l'enquête.

La méthode

Principe :

calage réalisé au niveau le plus élevé, en "remontant" les variables, et l'information, auxiliaires relatives aux niveaux inférieurs.

Exemple : enquête ménages (échantillon S_M) + tous les individus du ménage (échantillon S_I)

Pondération du ménage m : $d_m = 1/\pi_m$

Pondération de l'individu i du ménage m : $d_{m,i} = d_m \; \forall i \in men_m$

x_m = vecteur des variables auxiliaires connues pour tout ménage m de S_M

$X = \sum_{m \in U_m} x_m$ = vecteur des totaux sur la population U_M , connus

$z_{m,i}$ = vecteur de variables auxiliaires connues pour tout individu (m,i) de S_I

$Z = \sum_{i \in U_I} z_i$ = vecteur des totaux sur la population U_I , connus

On calcule les **totaux par ménage** : $z_m = \sum_{(m,i)\in men_m} z_{m,i}$

(on a : $Z = \sum_{m\in U_M} z_m$)

Vecteur de variables de calage pour le ménage m : (x_m, z_m)

Vecteur des totaux : (X , Z)

Équations de calage :

$$\sum_{m\in s_M} \frac{F(x'_m\lambda + z'_m\mu)}{\pi_m}(x_m, z_m) = (X, Z) \rightarrow \text{ pondérations } w_m$$

w_m = pondération du ménage m dans s_M

$w_{m,i} = w_m$ = pondération de l'individu (m,i) du ménage m dans s_I

On a bien : $\sum_{i\in s_I} w_{m,i} z_{m,i} = \sum_{m\in s_M} w_m \sum_{(m,i)\in men_m} z_{m,i} = \sum_{m\in s_M} w_m z_m = Z$

Calmar 2 permet de réaliser de tels calages simultanés.

L'utilisateur doit fournir les différentes tables en entrée et les tables des marges correspondantes : le programme réalise toutes les opérations nécessaires pour se ramener à un calage unique, et affecte les pondérations adéquates dans les différentes tables.

Calage et colinéarités

La résolution des équations de calage nécessite l'inversion d'une matrice de la forme : $\Phi = \sum_{k \in s} x_k x_k'$

→ **les variables de calage ne doivent pas être colinéaires.**

Le programme élimine automatiquement les colinéarités structurelles qui apparaissent lorsque plusieurs variables catégorielles figurent dans les variables de calage.

D’autres colinéarités entre variables de calage peuvent apparaître.

Exemple :

calage d'une enquête ménages sur deux répartitions "croisées" :

taille du ménage $\times$ *région* et *CS du chef de ménage* $\times$ *région*

$\rightarrow$ une équation redondante par région

2 solutions

- on redéfinit les variables de calage ;
- on utilise la technique des **matrices inverses généralisées**.

Cette dernière option est disponible dans Calmar 2.

Une nouvelle fonction de distance

Une nouvelle fonction de distance est proposée dans Calmar2, la fonction *sinus hyperbolique généralisée* :

$$G_\alpha(r) = \frac{1}{2\alpha}\int_1^r \mathrm{sh}[\alpha(t-\frac{1}{t})]\,dt$$ où α est un coefficient positif

Caractéristiques de la méthode :

- poids toujours positifs ;
- distributions de poids moins étendues qu'avec la méthode exponentielle du côté des poids élevés ;
- possibilité de réduire l'étendue de la distribution des poids grâce au coefficient α .

Le traitement de la non-réponse totale par calage généralisé

Équations de calage

$$\sum_{k \in r} d_k F(z'_k \lambda) x_k = X$$

où F est l'une des fonctions habituelles de calage.

Propriétés de la méthode

- permet une correction de la non–réponse même lorsque les variables qui l'expliquent ne sont connues que sur les répondants
- traite en particulier le cas où les variables facteurs de la non-réponse sont des variables d'intérêt (*mécanisme de réponse non ignorable*)
- produit une réduction du biais dû à la non–réponse grâce aux variables z_k, et une diminution de la variance grâce aux variables x_k
- si F linéaire : estimation par régression avec variables instrumentales (z_k)

$\hat{Y}_{calé} = \hat{Y}_{HT} + (X - \hat{X}_{HT})'\tilde{b}$, où $\tilde{b}$ solution des équations normales :

$$\sum_{k \in r} d_k z_k (y_k - x_k' b) = 0$$

Cette méthode est programmée dans Calmar 2.