Calage sur marges

Enquêtes et Sondages - CNAM - UE STA 108

Sylvie Rousseau

Sommaire

1. Objectif et intérêts du calage
2. La méthode
3. Un exemple

Objectif du calage

- Redresser un échantillon pour que les résultats soient cohérents avec des informations synthétiques connues par ailleurs.

  Ainsi, après calage, l'échantillon peut restituer

  - les totaux de variables quantitatives connus sur la population,
  - les effectifs de modalités de variables catégorielles connus sur la population.

Intérêts du calage

- Assurer la cohérence entre les résultats de plusieurs enquêtes ;
- Améliorer la précision des estimateurs des paramètres d'intérêt d'une enquête

  Pourvu que les critères de calage soient liés aux variables d'intérêt

Principe de la méthode

- Re-pondérer les individus échantillonnés en utilisant une information auxiliaire disponible sur un certain nombre de variables, appelées variables de calage.
- Exemple : les estimateurs par le ratio et par régression sont des estimateurs re-pondérés, resp. : $w_i = d_i X / \sum_{i \in s} d_i x_i$

$$w_i = d_i \left(1 + \frac{x_i}{\sum_{i \in s} d_i x_i^2} \left(X - \sum_{i \in s} d_i x_i \right) \right)$$

Notations

- Population : U
- Echantillon : s
- Variable d'intérêt : y
- De total sur la population : $Y = \sum_{k \in U} y_k$
- Estimateur d'Horvitz-Thompson : $\hat{Y}_\pi = \sum_{k \in s} \frac{1}{\pi_k} y_k = \sum_{k \in s} d_k y_k$
- On suppose connus les totaux sur U de J variables auxiliaires $x_1 \ldots x_j \ldots x_J$: $X_j = \sum_{k \in U} x_{jk}$

  Ces variables sont aussi mesurées sur l'échantillon.

Formalisation mathématique

- On cherche un estimateur "calé" de Y de la forme $\hat{Y}_w = \sum_{k \in s} w_k y_k$

où les poids w_k sont

- proches des poids de sondage d_k
- et vérifient les équations de calage (I)

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

Résolution

- Utilisation d'une fonction de distance,notée G, entre les w_k et les d_k

  avec G positive, convexe et G(1)=G'(1)=0
- Recherche des poids w_k (k $\in$ s) solutions de

$$\underset{w_k}{\text{Min}} \sum_{k \in s} d_k \, G(w_k / d_k)$$

sous les contraintes des équations de calage (I)

Solution

$$w_k = d_k F(x_k^{'} \lambda)$$

Avec :

- F : fonction réciproque de la dérivée de G ;
- $x_k^{'} = (x_{1k} \ldots x_{Jk})$ décrit le k^{ème} individu ;
- λ : vecteur des J multiplicateurs de Lagrange associés aux contraintes (I).

Résolution du système non linéaire de J équations à J inconnues issu des équations de calage avec l'algorithme de Newton : $\sum_{k \in s} d_k F(x_k' \lambda) x_k = X$

Les fonctions de distance disponibles

G	$F = G'^{-1}$	Type de distance
$\frac{1}{2}(x-1)^2$	$1+u$	*Khi-deux* Méthode linéaire (1) i.e. estimateur par la régression
$x\log x - x + 1$	$\exp u$	Entropie Méthode du raking -ratio (2)
$\frac{1}{A}\left[(x-L)\log\left(\frac{x-L}{1-L}\right)+(U-x)\log\left(\frac{U-x}{U-1}\right)\right]$ $= \frac{U-L}{(1-L)(U-1)}$; $x \in [L,U]$, $(\infty \text{ sinon})$	$\frac{L(U-1)+U(1-L)\exp u}{(U-1)+(1-L)\exp u}$ $\in]L,U[$	Logistique Méthode du raking ratio tronquée (3)
$\frac{1}{2}(x-1)^2$ si $x \in [L,U]$ ∞ sinon	$1+q_i u$ $\in [L,U]$	*Khi-deux tronquée* Méthode linéaire tronquée (3)

Choix des fonctions de distance

- Méthode linéaire
  - converge toujours en 2 étapes
  - redonne l'estimateur par régression
  - peut donner des poids négatifs
  - rapports de poids non bornés supérieurement
- Méthode exponentielle
  - poids positifs
  - redonne l'estimateur du raking ratio («règle de 3»)
  - rapports de poids non bornés supérieurement, en général supérieurs à la méthode linéaire
- Méthodes logit, linéaire tronquée
  - poids positifs
  - contrôle des rapports de poids

Choix des fonctions de distance

- Les cinq mesures de distance sont équivalentes du point de vue biais et variance :
  - Elles produisent des estimateurs de même erreur quadratique moyenne asymptotique.
  - Ils sont équivalents asymptotiquement à l'estimateur par régression généralisée :

$$\hat{Y}_w = \sum_{i \in s} w_i(s) Y_i = \hat{Y}_{HT} + \hat{\beta}_s^{\tau}(X - \hat{X}_{HT}) + O_p(n^{-1})$$

$$\hat{\beta}_s = T_s^{-1} \sum_{i \in s} d_i x_i y_i \qquad T_s = \sum_{i \in s} d_i x_i x_i^{\tau}$$

Macro CALMAR

- Insee, 1993
- Macro SAS
- Disponible sur www.insee.fr
- Syntaxe *(paramètres obligatoires)*

  %CALMAR
  (data=, poids=,ident=,
  datamar=,
  M=, *LO=, UP=,*
  datapoi=, poidsfin);

Exemple – le programme

- /* 1. les données individuelles */

```
DATA ech;
INPUT nom $ x $ y $ z pond;
CARDS;
A  1  f   1  10
B  1  h  2    0
C  1  h  3    .
D  5  f   1  11
E  5  f   3  13
F  5  h  2    7
H  1  h  2    8
G  5  h  2    8
I   5   f   2    9
J  .    h  2  10
K   5 h  2   14
;
```

- /*2. la table des marges */

```
DATA marges;
INPUT var $ n mar1 mar2;
CARDS;
X 2    20    60
Y  2    30   50
Z  0  140     .
;
run ;
```

- /* 3. lancement de Calmar */

```
%CALMAR(DATA=ech, POIDS=pond,IDENT=nom,
     DATAMAR=marges,M=2,OBSELI=oui,
    DATAPOI=sortie,POIDSFIN=pondfin,
     LABELPOI=pondération raking ratio) ;
```

Exemple - Résultats et sorties

- Avant calage

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
X	1	18	20	22.50	25.00
	5	62	60	77.50	75.00
Y	f	43	30	53.75	37.50
	h	37	50	46.25	62.50
Z		152	140	.	.

- Après calage

Variable	Modalité	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
X	1	20.000	20	25.00	25.00
	5	60.000	60	75.00	75.00
Y	f	30.000	30	37.50	37.50
	h	50.000	50	62.50	62.50
Z		140.000	140	.	.

Méthode : raking ratio
Premier tableau récapitulatif de l'algorithme :
la valeur du critère d'arrêt et le nombre de poids négatifs après chaque itération

Itération	Critère d'arrêt	Poids négatifs
1	0.56651	0
2	0.17766	0
3	0.04198	0
4	0.00322	0
5	0.00002	0

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Deuxième tableau récapitulatif de l'algorithme :
les coefficients du vecteur lambda de multiplicateurs de Lagrange après chaque itération

Variable	Modalité	LAMBDA1	LAMBDA2	LAMBDA3	LAMBDA4	LAMBDA5
X	1	1.20511	1.70361	1.87331	1.88687	1.88695
X	5	1.32247	1.81959	1.99270	2.00648	2.00656
Y	f	-0.73974	-0.94297	-1.02331	-1.02984	-1.02987
Y	h	.	.	.	.	.
Z		-0.47287	-0.74661	-0.83348	-0.84035	-0.84039

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
et sur les pondérations finales

Univariate Procedure

Variable=_F_ Rapport de poids

Moments

N	8	Sum Wgts	8
Mean	1.031891	Sum	8.255131
Std Dev	0.444812	Variance	0.197858
Skewness	-1.21649	Kurtosis	0.18399
USS	9.903406	CSS	1.385006
CV	43.10651	Std Mean	0.157265
T:Mean=0	6.561485	Pr>|T|	0.0003
Num ¬ =0	8	Num > 0	8
M(Sign)	4	Pr>=|M|	0.0078
Sgn Rank	18	Pr>=|S|	0.0078
W:Normal	0.811594	Pr<W	0.0394

Quantiles(Def=5)

100% Max	1.385113	99%	1.385113
75% Q3	1.385113	95%	1.385113
50% Med	1.187493	90%	1.385113
25% Q1	0.755692	10%	0.213423
0% Min	0.213423	5%	0.213423
		1%	0.213423
Range	1.17169		
Q3-Q1	0.629421		
Mode	1.385113		

Extremes

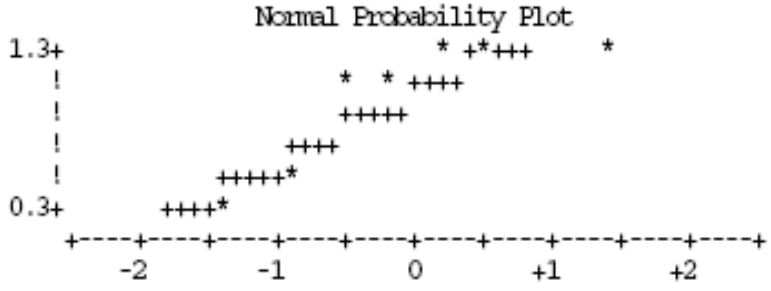
Lowest	ID	Highest	ID
0.213423	(E)	1.14602	(D)
0.494557	(I)	1.228966	(H)
1.016827	(A)	1.385113	(F)
1.14602	(D)	1.385113	(G)
1.228966	(H)	1.385113	(K)

Stem Leaf

```
  12 3999                  4
  10 25                    2
   8                       
   6                       
   4 9                     1
   2 1                     1
     ----+----+----+----+
Multiply Stem.Leaf by 10**-1
```

Boxplot

Normal Probability Plot



Méthode : raking ratio
Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
et sur les pondérations finales

Univariate Procedure

Variable=__WFIN Pondération finale

Moments

N	8	Sum Wgts	8
Mean	10	Sum	80
Std Dev	5.061209	Variance	25.61584
Skewness	0.439584	Kurtosis	1.109319
USS	979.3109	CSS	179.3109
CV	50.61209	Std Mean	1.789408
T:Mean=0	5.588441	Pr>|T|	0.0008
Num ¬ =0	8	Num > 0	8
M(Sign)	4	Pr>=|M|	0.0078
Sgn Rank	18	Pr>=|S|	0.0078
W:Normal	0.926636	Pr<W	0.4908

Quantiles (Def=5)

100% Max	19.39158	99%	19.39158
75% Q3	11.84356	95%	19.39158
50% Med	10	90%	19.39158
25% Q1	7.073401	10%	2.774494
0% Min	2.774494	5%	2.774494
		1%	2.774494
Range	16.61709		
Q3-Q1	4.770161		
Mode	2.774494		

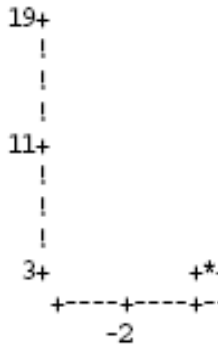
Extremes

Lowest	ID	Highest	ID
2.774494	(E)	9.831729	(H)
4.451013	(I)	10.16827	(A)
9.695789	(F)	11.0809	(G)
9.831729	(H)	12.60622	(D)
10.16827	(A)	19.39158	(K)

Stem Leaf

```
  18 4                    1
  16
  14
  12 6                    1
  10 21                   2
   8 78                   2
   6
   4 5                    1
   2 8                    1
     ----+----+----+----+
```

Boxplot



Normal Probability Plot

```
                    ********************
                    ***     BILAN     ***
                    ********************
*
*   DATE : 16 JUIN 2000                HEURE : 14:03
*
*   ***********************************
*   TABLE EN ENTRÉE : DON
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  : 11
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                : 3
*   NOMBRE D'OBSERVATIONS CONSERVÉES               : 8
*
*   VARIABLE DE PONDÉRATION : POND
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 2
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      X (2) y (2)
*
*   TAILLE DE L'ÉCHANTILLON (PONDÉRÉ) :              80
*   TAILLE DE LA POPULATION           :              80
*
*   NOMBRE DE VARIABLES NUMÉRIQUES : 1
*   LISTE DES VARIABLES NUMÉRIQUES :
      z
*
*
*   MÉTHODE UTILISÉE : RAKING RATIO
*   LE CALAGE A ÉTÉ RÉALISÉ EN 5 ITÉRATIONS
*   LES POIDS ONT ÉTÉ STOCKÉS DANS LA VARIABLE PONDFIN DE LA TABLE SORTIE
```

Un petit exemple commenté de calage sur marges
Liste des observations éliminées

Obs	nom	X	y	z	pond	__UN
1	B	1	h	2	0	1
2	C	1	h	3	.	1
3	J		h	2	10	1

Bibliographie

- Sautory O. (1993). « Redressement d'un échantillon par calage sur marges », Document de travail de la DSDS n°F9310,, www.insee.fr .
- Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). « Generalized raking procedures in survey sampling », Journal of the American Statistical Association, vol 88, n°423, pp. 1013-1020.
- Deville, J.-C. (1998). « La correction de la non-réponse par calage ou par échantillonnage équilibré ». Papier présenté au colloque de la Société Statistique du Canada, Sherbrooke.
- 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.
- Roy, G., et Vanheuverzwyn, A. (2001). « Redressement par la macro CALMAR : applications et pistes d'amélioration », Traitements des fichiers d'enquête, pp. 31-46. Presses Universitaires de Grenoble.