Les méthodes de calage

Olivier Sautory

Les méthodes de calage, élaborées par Deville et Särndal [3] et [4] permettent 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 ces méthodes 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. Elles permettent également d'améliorer la précision des estimations des totaux des variables d'intérêt bien corrélées aux variables de calage.

Ces méthodes sont mises en œuvre à l'Insee depuis 1990 à l'aide de la macro SAS Calmar[1] (voir Sautory [10]). 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 package R Icarus, proposé par A. Rebecq [8], permet également de mettre en œuvre ces méthodes. Il est disponible sur le CRAN.

1. Aspects théoriques du calage

1.1 Le problème, et la solution

On considère une population U d'individus, dans laquelle on a sélectionné un échantillon probabiliste s. Pour tout individu k de U, on note π_k sa probabilité d'inclusion dans 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 de Y à partir des données de l'enquête est dans la quasi-totalité des cas de la forme $\hat{Y} = \sum_{k \in s} d_k \, y_k$, où les d_k sont des poids d'estimation associés aux observations de l'échantillon. Ces poids sont souvent les « poids de sondage », égaux aux inverses des probabilités d'inclusion π_k : l'estimateur obtenu est alors l'estimateur d'Horvitz-Thompson : $\hat{Y}_{HT} = \sum_{k \in s} \frac{1}{\pi_k} y_k$.

On suppose que l'on connaît les totaux sur la population de J variables auxiliaires[2] ***X₁ … X_j … X_J***, disponibles pour toutes les observations de 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 » G, des pondérations initiales d_k, et qui assurent le calage sur les totaux des variables $\boldsymbol{X}_j$, i.e. qui vérifient les **équations de calage** :

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

La fonction de distance G, d'argument $r = w_k / d_k$, utilisée pour mesurer les distances entre les w_k et les d_k, est positive et convexe, et vérifie $G(1) = 0$. Les poids cherchés w_k minimisent la quantité $D = \sum_{k \in s} d_k G(w_k / d_k)$ sous les contraintes de calage (1).

[1] téléchargeable sur le site web de l'Insee www.insee.fr

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

La solution de ce problème est donnée par $w_k = d_k\, F(x_k' \lambda)$, où $x_k' = (x_{1k} \dots x_{Jk})$, λ est un vecteur de J multiplicateurs de Lagrange associés aux contraintes (1). F, appelée fonction de calage, est la fonction réciproque de la dérivée de la fonction G.

Le vecteur λ est déterminé par la résolution du système non linéaire de J équations à J inconnues résultant des équations de calage : $\sum_{k\in s} d_k\, F(x_k'\ \lambda)\, x_k = X$, où X désigne le vecteur des totaux X_j.

On peut résoudre numériquement ce système par la méthode itérative de Newton ; on calcule une suite de vecteurs $\lambda^{(i)}$ définis par une relation de récurrence, en initialisant l'algorithme avec le vecteur $\lambda^{(0)} = 0$. La convergence est obtenue lorsque les rapports de poids w_k/d_k obtenus lors de deux itérations successives « ne bougent presque plus » :

$$\underset{k\in s}{\text{Max}} \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \varepsilon \ ^{3}$$

Une fois les poids de calage w_k calculés, l'estimateur du total de toute variable d'intérêt ***Y*** sera alors l'estimateur dit « calé », de la forme $\hat{Y}_w = \sum_{k\in s} w_k\, y_k$.

1.2 Les fonctions de calage

4 méthodes de calage, correspondant à 4 fonctions de distance, sont proposées dans la macro SAS Calmar et dans le package R Icarus. Elles sont définies par la forme de la fonction F. On indique ci-dessous pour chacune des méthodes la fonction G(r) (où $r = w_k/d_k$ désigne le « rapport de poids »), et la fonction F(u) (où $u = x_k' \lambda$).

a) <u>méthode « linéaire »</u>

$$G(r) = \frac{1}{2}(r-1)^2,\ r \in \mathbb{R} \qquad F(u) = 1 + u\ (\in \mathbb{R})$$

D est alors une distance de type khi-deux entre les poids d_k et w_k. La forme linéaire de F donne son nom à cette méthode, et l'estimateur calé est alors l'estimateur par régression généralisée :

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

Cette méthode est la plus rapide car l'algorithme de Newton converge toujours après deux itérations. Elle peut conduire à des poids w_k négatifs, et les poids ne sont pas bornés supérieurement.

b) <u>méthode « exponentielle », ou « raking ratio »</u>

$$G(r) = r \operatorname{Log} r - r + 1,\ r > 0 \qquad F(u) = \exp u\ (> 0)$$

D est alors une distance de type « entropie » entre les poids d_k et w_k. Lorsque les variables auxiliaires sont des variables catégorielles pour lesquelles on connaît les effectifs des modalités dans la population, le choix de cette fonction G conduit à une méthode classique de redressement, proposée par Deming et Stephan [2], sous le nom de raking ratio ; elle est aussi connue (dans SAS en particulier) sous le nom I.P.F. ("Iterative Proportional Fitting").

[3] ε est un seuil chois par l'utilisateur (10^{-4} par exemple)

Cette méthode conduit à des poids toujours positifs, mais non bornés supérieurement, d'ailleurs en général supérieurs (pour les poids les plus élevés) à ceux de la méthode linéaire.

c) méthode « logit »

On choisit deux réels L et U tels que L < 1 < U.

$$G(r) = \left[(r - L)\ \mathrm{Log}\frac{r-L}{1-L} + (U - r)\,\mathrm{Log}\frac{U-r}{U-1} \right] \frac{1}{A},\ \text{si } L < r < U\ (\text{ et } +\infty \text{ sinon})\ \text{ avec } A = \frac{U-L}{(1-L)\,(U-1)}$$

$$F(u) = \frac{L(U-1) + U(1-L)\exp(Au)}{U - 1 + (1-L)\exp(Au)} \in]L, U[$$

La forme logistique de la fonction F donne son nom à cette méthode, qui assure que les rapports de poids w_k / d_k sont compris dans l'intervalle]L, U[. Toutefois, on ne peut pas choisir *a priori* n'importe quelles valeurs pour L et U : il existe en général pour L une valeur maximale L_{max} (inférieure à 1), et pour U une valeur minimale U_{min} (supérieure à 1). Ces valeurs dépendent des données et des marges du calage : plus la structure de l'échantillon est différente de celle de la population concernant les variables de calage, plus ces valeurs sont éloignées de 1.

d) méthode « linéaire tronquée »

On choisit deux réels L et U tels que L < 1 < U.

$$G(r) = \frac{1}{2}(r-1)^2 \quad \text{si } L \le r \le U\ (+\infty \text{ sinon}) \qquad F(u) = 1 + u \in [L, U]$$

Cette méthode assure que les rapports w_k / d_k sont compris dans l'intervalle [L, U], et comme pour la méthode « logit » il existe en général des valeurs L_{max} et U_{min}.

C'est la méthode logit - ou linéaire tronquée - 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, inférieurs à 1 voire négatifs, auxquels peut conduire la méthode linéaire.

1.3 La précision

Les estimateurs calés $\hat{Y}_w$ ont tous la même précision (asymptotique), quelle que soit la méthode utilisée : la variance approchée de $\hat{Y}_w$ est donc égale à celle de l'estimateur par régression $\hat{Y}_{reg}$: elle est d'autant plus faible que la corrélation entre la variable d'intérêt ***Y*** et les variables de calage $\boldsymbol{X_1 \ldots X_j \ldots X_J}$ est élevée.

Si l'on dispose d'une formule - ou d'un logiciel - permettant d'estimer la variance de l'estimateur d'Horvitz-Thompson pour une variable d'intérêt ***Y***, la variance de $\hat{Y}_w$ est obtenue en en remplaçant dans la formule les valeurs y_k par les résidus de la régression (pondérée par les d_k ou par les poids de calage w_k) de ***Y*** sur les $\boldsymbol{X_j}$ dans l'échantillon s.

2. Le calage en présence de non-réponse totale

La correction de la non-réponse totale est généralement réalisée par des techniques de repondération des unités répondantes. Les deux principales stratégies de calage pouvant être mises en œuvre en présence de non-réponse totale sont les suivantes (voir la fiche méthodologique *La correction de la non-réponse par repondération*, en particulier le § *V. Le calage sur marges*) :

- le calage après correction de la non-réponse : on réalise dans un premier temps une correction de la non-réponse totale par repondération, puis on effectue un calage classique, en partant des poids corrigés pour non-réponse $d_k^* = d_k / \hat{p}_k$, où les $\hat{p}_k$ sont les probabilités de réponse estimées (par exemple par la méthode des groupes de réponse homogène) ;
- le calage « direct », en partant des poids de sondage des répondants. Ceci est justifié si les variables de calage contiennent les variables explicatives de la non-réponse, et si on suppose une forme particulière du modèle de non-réponse (modèle linéaire généralisé en lien avec la fonction de calage choisie) (voir Dupont [5]).

3. Le calage pénalisé

(ce paragraphe est fortement inspiré de la communication de Rebecq [8] ; voir aussi la présentation de Rebecq [9] au séminaire de méthodologie statistique de l'Insee du 15 mars 2016).

Le calage pénalisé consiste à accepter que le calage ne soit pas parfaitement réalisé sur certaines marges, de façon à faciliter la convergence de la procédure, permettant ainsi d'augmenter le nombre de variables pour lesquelles la valeur de l'estimation après calage est « contrôlée », tout en préservant une distribution des rapports de poids peu étendue. La méthode (voir Beaumont et Bocci [1], et Guggemos et Tillé [7]) consiste à relâcher les contraintes de calage, et à les intégrer dans le programme d'optimisation.

On note $\hat{X}_w = \sum_{k \in s} w_k x_k$ le vecteur des estimateurs des totaux des variables de calage utilisant les « poids de calage ». On se donne un vecteur de coûts C, de taille égale au nombre de variables de calage J, et on note diag(C) la matrice diagonale de dimensions J×J où les coefficients diagonaux sont les valeurs du vecteur C.

Le programme de calage pénalisé s'écrit :

$$\min_{w_k} \sum_{k \in s} d_k G(w_k / d_k) + \lambda (\hat{X}_w - X)' \mathrm{diag}(C) (\hat{X}_w - X)$$

Le paramètre λ est compris entre 0 et $+\infty$ et représente l'importance relative donnée à la distance entre poids finaux et poids initiaux (1ère partie de la fonction à minimiser), par rapport à l'écart aux marges X des estimations redressées $\hat{X}_w$ (2e partie de la fonction à minimiser, la partie « coûts »). Si $\lambda \to +\infty$, le terme de coût est prépondérant : les contraintes de marges sont satisfaites en priorité, ce qui éloigne les rapports de poids de 1. Si $\lambda \to 0$, le terme de distance est prépondérant : les rapports de poids se rapprochent de 1, mais les contraintes sur les marges sont beaucoup relâchées.

Le coût associé à une variable de calage $\boldsymbol{X_j}$ est d'autant plus élevé que l'on souhaite une « bonne » proximité entre l'estimation $\hat{X}_{j\,w}$ et le total X_j. On peut requérir un calage exact en fixant un coût infini.

Le calage pénalisé peut être mis en œuvre à l'aide du package R Icarus. L'utilisateur choisit le vecteur de coûts et la valeur d'un paramètre *gap* qui spécifie une valeur maximale pour l'étendue de la distribution des rapports de poids : le programme détermine alors la plus grande valeur de λ. Il n'est (à ce jour) pas possible d'imposer *a priori* une erreur relative d'estimation pour les variables de calage : l'obtention d'une solution satisfaisante pour le statisticien se fait de manière empirique en jouant à la fois sur les paramètres de coût et de gap.

On peut utiliser différentes fonctions de distance G. Mais utiliser une distance « bornée » n'est pas nécessaire en raison du paramètre *gap*. Les deux méthodes proposées par Icarus sont donc la méthode linéaire, pour laquelle il existe une solution analytique, et la méthode exponentielle (Icarus utilise l'algorithme ICRS décrit par Bocci et Beaumont).

4. Aspects pratiques du calage

4.1 Les variables de calage

Une variable peut être utilisée dans un calage sur marges à condition qu'elle soit disponible pour chacune des observations de l'échantillon participant au calage d'une part, et que son total dans la population soit connu d'autre part. Il peut ainsi s'agir de variables de la base de sondage, ou de variables mesurées lors de la collecte de l'enquête et dont le total est connu par d'autres sources. Dans ce dernier cas, il est essentiel que la variable disponible sur l'échantillon corresponde exactement à la variable dont le total est connu : les variables doivent être mesurées au même moment, suivant les mêmes concepts, le total sur lequel on se cale doit correspondre à la population que l'échantillon cherche à décrire.

Ce total doit idéalement être connu exactement. Il peut également être estimé à partir d'une autre enquête à partir du moment où cette source permet d'obtenir des estimateurs beaucoup plus précis que l'enquête à laquelle est appliqué le calage. En pratique, on peut utiliser une enquête pour calculer des marges pour une autre enquête dès lors que l'échantillon de la première est dix fois plus gros que l'échantillon de la seconde. Ainsi, pour les enquêtes auprès des ménages de l'Insee, beaucoup de marges de calage sont calculées à partir de l'Enquête Emploi en Continu.

4.2 Les unités hors-champ et le calage

Du fait des imperfections des bases de sondage dans lesquelles sont sélectionnés les échantillons, des unités interrogées peuvent ne pas appartenir en réalité à la population visée par l'enquête (i.e. le « champ » de l'enquête) : entreprises qui ont cessé toute activité, logements vacants ou occupés à titre de résidence secondaire sont des exemples classiques d'unités hors-champ pour les enquêtes de l'Insee. Ces hors-champ sont le plus souvent détectés lors de la collecte.

Si les marges utilisées pour le calage sont issues de la base de sondage, alors les unités hors-champ détectées par la collecte, doivent participer au calage : en effet, les marges calculées dans la base de sondage sont relatives à une population qui contient à la fois des unités hors champ et des unités dans le champ de l'enquête. Si, à l'inverse, les marges ne sont relatives qu'au champ de l'enquête, alors les unités hors-champ ne doivent pas participer au calage. Si le calage sur marges utilise à la fois des marges relatives au champ de l'enquête et des marges calculées à partir de la base de sondage, alors les unités de l'échantillon hors-champ détectées par la collecte doivent participer au calage, mais, pour ces observations, les valeurs des variables de calage correspondant aux marges relatives au champ sont mises à 0.

5. Exemples

5.1 Enquêtes Sectorielles Annuelles

Les enquêtes sectorielles annuelles (ESA) servent à quantifier chaque année la décomposition des chiffres d'affaires des entreprises françaises suivant leurs différentes activités. L'échantillon comprend environ 160 000 entreprises, dont la moitié - les plus grandes - est interrogée exhaustivement et l'autre moitié sélectionnée aléatoirement parmi les petites et moyennes entreprises françaises.

Après la mise en œuvre de méthodes de correction de la non-réponse et de traitement des valeurs influentes (voir fiches méthodologiques *La correction de la non-réponse par repondération* et *Traitement des valeurs influentes dans les enquêtes*), un calage est réalisé : il consiste, autant que possible, à caler les unités des strates non exhaustives de l'échantillon sur le chiffre d'affaires fiscal total par groupe de la NAF 2008 et le nombre d'unités par division de la NAF 2008 dans la population privée des strates exhaustives. On utilise pour cela la macro Calmar avec comme paramètre par défaut la méthode linéaire tronquée, en contraignant les rapports de poids à rester dans l'intervalle [0,5 ; 2]. Lorsque le calage ne

converge pas, on élargit les bornes de rapport de poids, et si cela ne suffit pas, on regroupe des secteurs de calage.

5.2 Enquête auprès des ménages sur les Technologies de l'information et de la communication (TIC)

Cette enquête annuelle allie plusieurs modes de collecte : un échantillon de 3 500 ménages est enquêté par téléphone, un échantillon de 22 500 ménages est enquêté par internet/papier. Un calage est opéré après correction de la non-réponse totale, au niveau individuel, pour la métropole d'une part et pour les DOM d'autre part. Les données servant au calage sont celles de l'Enquête emploi en continu de l'année n-1.

Pour la métropole, les variables de calage retenues sont les structures par : croisements sexe-âge (14 modalités), sexe-diplôme (9 modalités), âge-diplôme (7 modalités), CS (11 modalités), taille d'aire urbaine croisée avec la catégorie de commune dans le zonage en aire urbaine (8 modalités), taille d'unité urbaine (8 modalités), nouvelle région (12 modalités), nombre de personnes de 15 ans et plus dans le ménage croisé avec 3 grandes tranches d'âge (9 modalités), type de ménage (définition Eurostat - 4 modalités) et nationalité (2 modalités).

Pour les DOM, les variables de calage retenues sont la structure par sexe et âge pour l'ensemble des DOM (9 modalités), la structure par diplôme pour l'ensemble des DOM (4 modalités), la structure par CS (9 modalités), la répartition de la population par zones géographiques infra-départementales (15 modalités au total) et le nombre de ménages.

5.3 Enquête nationale sur les ressources des jeunes (ENRJ)

(ce paragraphe est extrait d'une note interne Insee de Gros [6]).

L'enquête ENRJ a été menée par la Drees et l'Insee du 1er octobre au 31 décembre 2014. Elle permet de décrire les ressources et les conditions de vie des jeunes adultes de 18 à 24 ans en France, résidant en logement ordinaire ou communauté.

Les seules données fiables et précises disponibles à l'Insee sur ce champ spécifique sont la pyramide des âges par sexe au 1er janvier 2015 issue du bilan démographique 2014. Les effectifs collectés via les sources administratives constituent plus des données de cadrage que des marges de calage à proprement parler : possibilité de doubles comptes dans les effectifs pour certaines variables, concepts légèrement différents entre source administrative et variable issue de l'enquête pour d'autres. Il a donc été décidé de mettre en œuvre un calage pénalisé, de la façon suivante :

- calage exact sur la pyramide des âges par sexe au 1er janvier 2015 ;
- calage approché sur un certain nombre de données administratives : le nombre de bacheliers par sexe, le nombre de bacheliers par filière, le nombre d'inscrits à l'université par sexe, le nombre d'inscrits en BTS par sexe, le nombre d'inscrits en écoles de commerce ou d'ingénieur, le nombre de boursiers sur critères sociaux par sexe, le nombre de jeunes non en couple bénéficiant des APL et enfin le nombre de jeunes femmes bénéficiant de l'allocation de base de la Paje.

Ce calage pénalisé a permis de contenir les rapports de poids dans l'intervalle [0.2 ; 1.8] et d'assurer un calage exact sur la pyramide des âges par sexe tout en corrigeant les principaux déséquilibres que l'on observait sur les données de cadrage lorsqu'on redressait uniquement par post-stratification sur la pyramide des âges par sexe.

RÉFÉRENCES

[1] Beaumont, J.-F and Bocci, C.(2008), "Another look at ridge calibration", *Metron*, vol 66.1, pp. 5-20.

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

[3] 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.

[4] 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.

[5] 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.*

[6] Gros, E. (2015), « Note de bilan relative aux calculs de pondérations dans l'enquête nationale sur les ressources des jeunes », note interne Insee.

[7] Guggemos, F. and Tillé (2010), Y., "Penalized calibration in survey sampling : Design based estimation assisted by mixed models", *Journal of statistical planning and inference*, 140.11 pp. 3199-3212.

[8] Rebecq, A. (2016), « Icarus : un package R pour le calage sur marges et ses variantes », *9e colloque francophone sur les sondages, Gatineau (Canada)* - http://sondages2016.sfds.asso.fr/.

[9] Rebecq, A. (2016), « Le calage pénalisé - Théorie et application », *Séminaire de méthodologie statistique de l'Insee « Miscellanées sur la calage »*, https://www.insee.fr/fr/information/2387498

[10] Sautory, O. (1993), « La macro Calmar. Redressement d'un échantillon par calage sur marges », *Document de travail F9310 de la DSDS*, Insee.

Département des méthodes statistiques
Version n°1, diffusée le 5 mars 2018