Calage serré des poids d'enquête

Monique Graf [1]

[1] *Institut de Statistique, Faculté des Sciences économiques, Université de Neuchâtel*
Pierre-à-Mazel 7, CH-2000 Neuchâtel
monique.p.n.graf@bluewin.ch

Résumé. Le calage des poids de sondage se réfère à la recherche de corrections multiplicatives des poids, de telle sorte que les totaux extrapolés des variables de calage coïncident avec les totaux de population correspondants, supposés connus. Il est souvent souhaitable d'imposer des limites sur la variabilité des corrections de poids, spécialement si l'on prévoit de faire des estimations, non seulement pour la population entière ou les catégories utilisées pour le calage, mais aussi pour des domaines coupant ces catégories. Les propriétés d'optimalité des poids calés ne fournissant aucune garantie dans ce cas, il est intéressant de limiter au maximum la variabilité des corrections de poids. Dans la pratique, le problème du calage est résolu en minimisant d'une fonction de perte convexe dépendant de limites définies a priori sur la correction des poids. On décrit ici une méthode pour trouver les limites les plus serrées possible pour le calage des poids de sondage, telles que le problème soit toujours réalisable. Malgré la taille du problème, la mise en oeuvre dans R à l'aide de matrices creuses s'est avérée facile à gérer pour les enquêtes en taille réelle, d'au moins plusieurs milliers d'unités. On donne un exemple réel et un exemple de simulation qui prouvent la faisabilité de la méthode.

Mots-clés. calage, poids d'enquête, algorithme du simplexe, matrice creuse ...

Abstract. Calibration of survey weights refers to finding weight corrections, so that extrapolated totals coincide with known population totals. It is of interest to restrict the range of the weight corrections, especially if it is intended to make estimations in smaller domains. Indeed there is no guarantee that the calibrated weights give any better results than the original weights in domains running across the categories defined by the calibration variables. The restricted range makes the resulting weights more robust. It is often desirable to impose bounds on the weight corrections, but it may then happen that calibration becomes infeasible. In practice, the calibration problem is solved by minimizing a convex loss function, using a priori defined bounds on weight corrections. This paper describes a method for finding the tightest bounds for calibration of survey weights, so that the problem is still feasible. Despite the size of the problem, implementation in R using a sparse matrix formulation has proven manageable for real size surveys, at least several thousands of units. A real example and a simulation example are given, proving the feasibility of the method.

Keywords. calibration, survey weights, simplex algorithm, sparse matrix ...

1 Le problème du calage

Le problème du calage est bien connu (Deville et al. (1993)). La macro CALMAR (Sautory, 1993) écrite en SAS est largement utilisée en statistique publique. Un logiciel SPSS pour le calage, g-CALIB, existe (Vanderhoeft, 2001). Des packages R ont été écrits dans le contexte des sondages (Lumley, 2012; Tillé et Matei, 2012; Zardetto, 2013) proposant des fonctions de calage. A cause du caractère explosif du problème combinatoire sous-jacent, la recherche des bornes de calage les plus serrées possible n'a pas été abordé, même si la solution de principe par la programmation linéaire a été reconnue (Deville et al., 1993; Vanderhoeft, 2001). Si la question est abordée, ces bornes sont obtenues par tâtonnement (Chauvet et al. 2005) ou par une règle empirique (Zardetto, 2013). Vanderhoeft (2001) donne une condition suffisante sur les bornes pour que le calage converge, si toutes les variables auxiliaires sont positives. La méthode présentée ici permet de trouver des bornes de calage à écart minimum, sans la contrainte de positivité sur les variables auxiliaires.

Notations Supposons qu'il y ait p marges de calage, et soit $\mathbf{a}_i$, $i = 1, \cdots, n$ les $(p \times 1)$ vecteurs correspondant aux variables auxiliaires pour les n unités échantillonnées. Notons par $\mathbf{d} = (d_1, d_2, ..., d_n)^t$ le vecteur des poids d'échantillonnage préliminaires. Définissons la matrice $p \times n$ suivante:

$$\mathbf{B} = (\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_n)\text{diag}(\mathbf{d}). \tag{1}$$

Si $\mathbf{A}$ is the vecteur des totaux de population des p variables auxiliaires et $\mathbf{g}$ le n-vecteur des corrections des poids à déterminer, les equations de calage sont données par

$$\mathbf{Bg} = \mathbf{A}. \tag{2}$$

Le calage est un problème d'optimisation convexe. Etant donné une fonction de perte G et des bornes L and U sur les g-poids, le but est de trouver

$$\begin{aligned} &\arg\min_{g_1, g_2, ..., g_n} \left\{ \sum_{k=1}^{n} d_k G(g_k) \right\} \\ &\text{sous les conditions (2) et } L \leq g_i \leq U, \quad i = 1, ..., n, \quad U \geq 1,\ L \leq 1. \end{aligned} \tag{3}$$

1.1 Bornes L et U les plus serrées

En général dans la pratique, la matrice $\mathbf{B}$ n'a pas d'éléments négatifs, mais nous ne ferons pas cette hypothèse. Nous supposons que $L \leq 1$ et $U \geq 1$. Cette condition n'est pas très restrictive et peut facilement être levée par des ajustements a priori des poids.

Pour que le problème ait une solution, l'ensemble admissible (i.e. l'ensemble des points $(g_1, ..., g_n)$ vérifiant toutes les contraintes) doit être non vide. D'une part, le système (2)

définit un espace affine de dimension $n-p$ de $\mathbb{R}^n$; d'autre part, les bornes L et U sur les g-poids définissent un hypercube de $\mathbb{R}^n$. L'intersection des ces deux domaines (équation 4), donne l'ensemble des points admissibles:

$$\mathcal{D} = \{\mathbf{g} \in \mathbb{R}^n \,|\, \{\mathbf{Bg} = \mathbf{A}\} \cap \{L \leq g_i \leq U\,, \forall i = 1, \cdots, n\}\} \neq \emptyset. \tag{4}$$

1.2 La recherche des bornes serrées de calage comme un problème de programmation linéaire

Le calage, décrit dans Deville et al. (1993), est un problème de programmation convexe. Une solution existe chaque fois que l'ensemble admissible $\mathcal{D}$ est non vide. Ici, nous voulons trouver les bornes les plus serrées possible U et L, telles que le problème ait toujours une solution, i.e telles que $\mathcal{D} \neq \emptyset$. Deville et al. (1993); Vanderhoeft (2001) ont vu que la recherche d'un domaine admissible $\mathcal{D}$ peut être considéré comme un problème de programmation linéaire (lp), mais cette idée n'a, à notre connaissance, jamais été appliquée dans la pratique des enquêtes.

La solution proposée consiste à ajouter L and U à l'ensemble des variables $g_1, ..., g_n$ et à trouver l'écart minimum $U - L$, tel que l'ensemble (4) soit non vide. C'est un problème de programmation linéaire, formalisé à la Table 1.

Table 1: Programme linéaire pour trouver les bornes de calage les plus serrées possible

Minimiser $U - L$	
sous les conditions:	
$\mathbf{Bg} = \mathbf{A}$,	contraintes de marges (voir équation 2),
$L \leq g_i \leq U$,	contraintes d'inégalité
$U \geq 1$, $L \leq 1$,	$i = 1, ..., n,$.

Comme ce problème ne dépend que des contraintes, le résultat est indépendant de la méthode de calage. Il y a $n + 2$ variables à déterminer (L, U et les g-poids $g_1, ..., g_n$). Le programme lp cherche toujours des solutions à coordonnées positives; il n'est donc pas nécessaire de spécifier $L \geq 0, g_i \geq 0, i = 1, ..., n$. Une méthode classique est donnée par l'algorithme du simplexe de Dantzig et al.(1955); voir aussi Dantzig (1998). Cet algorithme est basé sur le principe suivant: la fonction linéaire à optimiser (ici $U - L$) a ses extrêmes aux sommets d'un polyhèdre convexe défini par l'ensemble des inégalités et des contraintes affines $\mathbf{Bg} = \mathbf{A}$. Il suffit donc de parcourir les sommets de ce polyhèdre pour trouver l'optimum.

La solution est en général non unique dans les g-poids, mais nous savons, grâce à la solution trouvée, que l'ensemble admissible $\mathcal{D}$ est non vide pour les valeurs L et U obtenues. Cette solution n'optimise cependant pas la fonction convexe choisie pour le calage. La stratégie est donc 1. de trouver L and U par cette méthode; 2. d'introduire ces bornes (amplifiées par une petite tolérance pour éviter des problèmes numériques) dans un programme de calage comme la macro CALMAR de Sautory (1993).

2 Mise en oeuvre avec R

Il existe plusieurs possibilités pour l'implémentation de la recherche des bornes les plus serrées dans R. Dans la plupart des applications du calage, le nombre d'observations est de l'ordre de plusieurs milliers, ou même plus, et la matrice $\mathbf{B}$ contient plusieurs indicatrices de catégories, donc beaucoup de zéros (elle est creuse). La fonction choisie doit pouvoir accepter le format creux, qui est utile lorsque la matrice considérée contient beaucoup de zéros. Il est défini comme suit:

Definition Codification d'une matrice creuse

Soit $\mathbf{U} = (U_{ij})$ une $p \times n$ matrice avec m cellules non nulles. Elle est représentée comme une matrice creuse lorsque seules les cellules non nulles sont spécifiées, sous la forme:

$$i, j, U_{ij}.$$

Le format creux est donc utile chaque fois que $pn > 3m$.

Une fonction implémentant l'algorithme du simplexe avec le format creux existe dans le package `Rglpk` (Hornik and Theussl, 2012). Cette fonction est appelée par notre propre fonction LPCALIB qui calcule les bornes les plus serrées.

3 Application à l'enquête SILC

L'enquête SILC suisse est l'implémentation en Suisse de l'enquête de l'Union européenne sur les statistique des revenus et des conditions de vie (EU-SILC). En 2009, l'enquête suisse comprenait 17561 personnes répondantes, réparties dans 7372 ménages. L'Office fédéral de la statistique utilise 54 variables de calibration pour ajuster les poids d'échantillonnage. Un ajustement préliminaire pour la non réponse a été introduit, si bien que les poids initiaux ne sont pas trop éloignés des poids calés finaux. Les bornes les plus serrées trouvées par la fonction LPCALIB, lorsqu'on utilise l'enquête au niveau "personnes", sont $L = 0.747568$ and $U = 1.838922$. C'est la règle dans cette enquête d'imposer que les poids soient les mêmes pour toutes les personnes d'un même ménage. Pour respecter cette clause, on agrège les colonnes de la matrice $\mathbf{B}$ correspondant aux individus d'un même

ménage. Les bornes les plus serrées pour les poids des ménages sont alors $L = 0.3969588$ and $U = 1.8810312$.

Nous avons ensuite exécuté la macro SAS CALMAR2 de Sautory (1993) avec les bornes arrondies ($L = 0.39$ and $U = 1.89$). Le calage a été réalisé avec la méthode linéaire (7 itérations), ainsi qu'avec la méthode logit (9 itérations). Les poids résultants sont alors attribués à chaque personne d'un même ménage.

4 Discussion

L'intérêt de réduire la variabilité des corrections de poids vient de ce que les estimations prévues sur la base de l'enquête ne se cantonnent pas aux catégories définies par les variables de calage. En effet, le calage n'offre aucune garantie d'optimalité si on prévoit d'effectuer des estimations par plus petits domaines. Le calage serré rend alors les corrections de poids plus robustes. On peut démontrer que les bornes L et U sont presque uniques, c'est-à-dire qu'une petite variation seulement autour de la solution trouvée est possible dans certains cas. Une simulation (non présentée ici) comparant les performances de deux fonctions R de calage avec les bornes serrées, la fonction CALIB de Tillé et Matei (2012) et la fonction CALIBRATE de Lumley (2012), a montré que les bornes calculées par notre fonction LPCALIB a permis la convergence de CALIB et/ou CALIBRATE, dans 1000 cas sur 1000 pour la méthode "linéaire tronquée", et dans 999 cas sur 1000 pour la méthode "logit". Cela prouve la faisabilité de la méthode mise en oeuvre avec LPCALIB. Il existe un grand nombre d'implémentations du programme du simplexe. Il faut en choisir une qui accepte le format creux pour pouvoir effectuer les calculs pour une enquête de taille raisonnablement grande. La fonction LPCALIB invoquant le R package `Rglpk` fait facilement ce travail.

Remerciements

Ce travail a été financé par l'Office fédéral de la statistique. Yves Tillé est remercié pour son soutien et les nombreuses discussions.

Bibliographie

[1] Chauvet, G., Deville, J.-C., El Haj Tirari, M., and Le Guennec, J. (2005). Evaluation de trois logiciels de calage : g-calib 2.0, calmar 2 et bascula 4.0. Technical report, Statistics Belgium WorkingPaper.
[2] Dantzig, G. B. (1998). *Linear Programming and Extensions.* Princeton University Press, ISBN 0691059136.
[3] Dantzig, G. B., Orden, A., and Wolfe, P. (1955). The generalized simplex method for minimizing a linear form under inequality restraints. *Pacific J. Math.*, 5:183–195.

[4] Deville, J., Särndal, C., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1033–1020.
[5] Hornik, K. and Theussl, S. (2012). *Rglpk: R/GNU Linear Programming Kit Interface*. R package version 0.3-10.
[6] Lumley, T. (2012). *survey: analysis of complex survey samples*. R package version 3.28-2.
[7] Sautory, O. (1993). *La macro CALMAR: Redressement d'un Echantillon par Calage sur Marges*. Document de travail de la Direction des Statistiques Demographiques et Sociales, no. F9310.
[8] Tillé, Y. and Matei, A. (2012). *sampling: Survey Sampling*. R package version 2.5.
[9] Vanderhoeft, C. (2001). Generalized calibration at statistic belgium. Technical Report 3, Statistics Belgium WorkingPaper.
[10] Zardetto, D. (2013). *ReGenesees: R evolved Generalized software for sampling estimates and errors in surveys*. R package version 1.4.