CALAGE SUR INFORMATION AUXILIAIRE INCERTAINE : PROPOSITION D'ALGORITHME DE REDRESSEMENT RIDGE

Flavien Alleaume[1] & Lorie Dudoignon[2]

[1] *Médiamétrie, 70 rue Rivay, 92532 Levallois-Perret Cedex, falleaume@mediametrie.fr*
[2] *Médiamétrie, 70 rue Rivay, 92532 Levallois-Perret Cedex, ldudoignon@mediametrie.fr*

Résumé. Le redressement d'échantillons s'opère généralement par une méthode de calage sur marges et plus précisément à l'aide d'algorithmes qui proposent le plus souvent un calage « exact » de l'échantillon sur les marges avec lesquelles on souhaite être en adéquation. Si, les informations auxiliaires proviennent d'une enquête exhaustive, cette approche se justifie pleinement. Or dans la pratique, on est parfois amené à estimer les marges à partir d'une enquête par sondage. Dans ce contexte, il peut être intéressant d'autoriser un relâchement des contraintes, c'est-à-dire de permettre une tolérance dans le respect des marges associées aux variables de calage. L'adaptation de la logique de la régression ridge aux redressements d'échantillons permet de répondre à cette problématique. Nous proposons ici un algorithme qui fait la synthèse entre les approches de Beaumont et Bocci (2008) et celle de Singh et Mohl (1996).

Mots-clés. Calage sur marges, algorithme itératif, régression ridge

Abstract. Survey sample are usually adjusted using a method of calibration on marginal counts, and more specifically by using algorithms which provide "exact" calibration of the sample on the population totals we want to match up. If the auxiliary information is from an exhaustive study, this approach is fully justified. In practice, we sometimes have to estimate the marginal counts from a survey. In this context, it may be interesting to authorise a relaxing of the constraints, i.e. be more tolerant when respecting the marginal counts relating to the calibration variables. By adapting the logic of ridge regression to the sample adjustments it is possible to tackle this problem. We propose here an algorithm which makes the synthesis between the Beaumont and Bocci approach and the Singh and Mohl one.

Keywords. Calibration, iterative algorithm, ridge regression

1 Introduction

Médiamétrie a lancé, au printemps 2012, le recrutement du Panel Multi-Écrans, nouveau panel de mesure d'audience de la télévision et d'Internet sur tous les écrans, i.e. sur téléviseur, ordinateur, téléphone mobile ou tablette. L'objectif de ce nouveau dispositif est d'analyser les comportements croisés entre les différents médias mesurés. Le redressement du panel intègre classiquement des variables de calage socio-démographiques, mais également des variables d'audience TV. L'intégration de ces dernières se justifie par un besoin de cohérence avec Médiamat, l'étude de référence sur le marché français pour la mesure d'audience de la télévision. Dans ce cadre, il peut être intéressant de caler parfaitement sur les marges des variables socio-démographiques et de laisser une tolérance sur les variables d'audience TV, en tenant ainsi compte des marges d'erreur associées aux résultats du panel Médiamat.

C'est la raison pour laquelle nous nous sommes intéressés aux approches basées sur l'utilisation de la régression ridge. Nous avons testé les méthodes proposées par Beaumont et Bocci (2008) et Singh et Mohl (1996), en comparaison avec celles implémentées dans CALMAR par Sautory (1993), notamment la méthode sinus hyperbolique de Roy et Vanheuverzwyn (2001). Les

difficultés opérationnelles (complexité du choix des paramètres et temps de calcul importants) auxquelles nous avons été confrontés nous ont conduits à développer une méthode mixte.

Les travaux présentés dans cet article reposent sur les recherches effectuées par Marie Pitré lors de son stage de fin d'études au sein de la Direction Analyses et Méthodes Scientifiques de Médiamétrie en 2012 et Loïc Faure, son maître de stage.

2 Rappels sur le calage sur marges

On considère un échantillon S de taille n issu d'une population de taille N. On note d_k le poids initial de l'individu k (inverse du poids de sondage).

L'objectif du calage sur marges est de déterminer les poids de redressement ω_k qui vérifient les équations de calage :

$$\sum_{k \in S} \omega_k \, x_k = T$$

où x_k est le vecteur de variables auxiliaires pour l'individu k et T le vecteur des marges (totaux des variables auxiliaires sur l'ensemble de la population).

Les poids obtenus doivent s'éloigner le moins possible des poids initiaux. Pour cela, on considère une fonction de distance G à minimiser. Le problème général de calage sur marges peut alors s'écrire comme :

$$\text{Min} \sum_{k \in S} G(\omega_k, d_k) \quad s.c. \quad \sum_{k \in S} \omega_k \mathbf{x}_k = T$$

Cette généralisation de la problématique de calage est proposée par Deville et Särndal (1992).

Il est aussi possible d'ajouter une contrainte supplémentaire afin de borner les poids de redressement : $Ld_k \leq \omega_k \leq Ud_k$, i.e. $g_k = \frac{\omega_k}{d_k} \in [L, U]$.

Il existe plusieurs algorithmes itératifs qui permettent de résoudre ce problème de minimisation sous contraintes. Singh et Mohl (1996) en fournissent un inventaire en distinguant deux grandes familles de méthodes :

- les algorithmes qui respectent, à chaque itération, les marges fixées et itèrent jusqu'à ce que les poids soient tous dans les bornes fixées,
- les algorithmes dont les poids sont compris dans les bornes à chaque itération et qui itèrent jusqu'à respecter les marges.

La macro CALMAR de l'INSEE développée par Sautory (1993) appartient plutôt à la seconde catégorie. Parmi la première famille, deux méthodes ont été adaptées en utilisant la régression ridge. Commençons par détailler ces deux algorithmes dans le cas standard.

2.1 Méthode Shrinkage Minimization

L'algorithme consiste en un enchaînement de calculs de pondération par la régression généralisée dont la solution est donnée par Särndal (1980). La pondération par la régression généralisée correspond au problème de calage sur marges avec la fonction de distance du χ^2.

Concrètement, l'algorithme procède de la manière suivante :

1- <u>Initialisation :</u>
   Le jeu de poids ω_k^0 est défini par la méthode donnant l'estimateur de régression généralisée.

$$\omega_k^0 = d_k + d_k \mathbf{x}'_k \left(\sum_{k \epsilon S} d_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \left(T - \sum_{k \epsilon S} d_k \mathbf{x}_k \right)$$

2- Itération ν :

i) Critère d'arrêt :

Si tous les poids se trouvent à l'intérieur de l'intervalle [L ; U], l'algorithme s'arrête.

ii) Rétrécissement :

Les poids sont « rétrécis » de sorte qu'ils soient tous à l'intérieur ou aux bornes de [L;U]

$$\omega_k^{\nu *} = \begin{cases} L' d_k & si\ \omega_k^\nu < L'' d_k \\ U' d_k & si\ \omega_k^\nu > U'' d_k \\ \omega_k^\nu & sinon \end{cases}$$

où pour $0 < \alpha \leq \eta \leq 1$ donnés, on a :

$$L' = \alpha L + (1 - \alpha), \quad U' = \alpha U + (1 - \alpha)$$
$$L'' = \eta L + (1 - \eta), \quad U'' = \eta U + (1 - \eta)$$

iii) Minimisation :

On recommence alors le calcul de pondération par la régression généralisée, non plus avec les poids initiaux mais les poids rétrécis à l'étape précédente.

On obtient :

$$\omega_k^{\nu+1} = \omega_k^{\nu *} + \omega_k^{\nu *} \mathbf{x}'_k \left(\sum_{k \epsilon S} \omega_k^{\nu *} \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \left(T - \sum_{k \epsilon S} \omega_k^{\nu *} \mathbf{x}_k \right)$$

A ce niveau, les marges sont bien respectées mais pas forcément les bornes. Dans ce cas, on passe à l'itération ν+1.

2.2 Méthode Scaled Modified Chi-Square

Là encore il s'agit d'une méthode qui respecte les marges à chaque itération et converge lorsque les bornes sont atteintes. La distance utilisée est toujours celle du χ^2. On utilise cette fois un coefficient réducteur (ou « facteur de cadrage » pour Singh et Mohl) des poids recalculé à chaque itération pour contrôler la distance entre les poids initiaux et les nouveaux poids ayant tendance à sortir de l'intervalle.

Concrètement, l'algorithme procède de la manière suivante :

1- Initialisation :

Le jeu de poids ω_k^0 est défini comme précédemment par la méthode donnant l'estimateur de régression généralisée.

2- Itération ν :

i) Critère d'arrêt :

Si tous les poids se trouvent à l'intérieur de l'intervalle [L ; U], l'algorithme s'arrête.

ii) Rétrécissement :

Calcul du facteur de cadrage qui permet de réduire l'étendue des poids.

$q_k^{[\nu]} = q_k^0 \times \ldots \times q_k^\nu$

où $q_k^\nu = \begin{cases} 1 & si\ \xi_k^\nu < 1/2 \\ 1 - \beta(\xi_k^\nu - 1/2)^2 & si\ 1/2 \leq \xi_k^\nu < 1 \\ (1 - \beta/4)/\xi_k^\nu & sinon \end{cases}$ et $\xi_k^\nu = \begin{cases} \frac{\omega_k^\nu - d_k}{d_k(L'-1)} & si\ \omega_k^\nu \leq d_k \\ \frac{\omega_k^\nu - d_k}{d_k(U'-1)} & sinon \end{cases}$

avec $0 < \beta \leq 1$.

iii) Minimisation :

On recommence alors le calcul de pondération par la régression généralisée en introduisant le facteur de cadrage.

$$\omega_k^{\nu+1} = d_k + d_k q_k^{[\nu]} \mathbf{x'}_k \left(\sum_{k \in S} d_k q_k^{[\nu]} \mathbf{x}_k \mathbf{x'}_k \right)^{-1} \left(T - \sum_{k \in S} d_k \mathbf{x}_k \right)$$

A ce niveau, les marges sont bien respectées mais pas forcément les bornes. Dans ce cas, on passe à l'itération ν+1.

2.3 Méthode Scaled Modified Chi-Square pour d'autres distances

Beaumont et Bocci (2008) proposent une légère modification de la méthode précédente pour y introduire une autre fonction de distance que celle du χ^2.

L'algorithme se déroule de la même manière que le précédent, seul le facteur de cadrage change de manière à y introduire la fonction G souhaitée :

$$q_k^{[\nu]} = \frac{2(\omega_k^{\nu} - d_k)}{d_k \frac{dG(\omega_k^{\nu}, d_k)}{d\omega_k^{\nu}}}$$

3 Redressement ridge

Nous ne cherchons plus à caler exactement les marges mais à les respecter le plus possible. Le but est ici de relâcher les contraintes sur l'échantillon pour moins déformer les poids et donc rester plus facilement entre les bornes voulues.

3.1 Dans le cas du redressement par la régression généralisée

Le problème de minimisation change donc de forme. Bradley et Chambers (1984) suggèrent de chercher le jeu de poids qui s'éloigne le moins possible des poids initiaux selon la distance du χ^2 et pour lequel les totaux sont les plus proches possibles des marges spécifiées.

$$Min \left(\sum_{k \in S} \frac{(\omega_k - d_k)^2}{2 d_k} + \lambda \underbrace{\left(\sum_{k \in S} \omega_k \boldsymbol{x}_k - T \right)' C \left(\sum_{k \in S} \omega_k \boldsymbol{x}_k - T \right)}_{\hat{\mathrm{D}}} \right) \quad o\grave{u} \quad C = \begin{pmatrix} c^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c^j \end{pmatrix}$$

La seconde partie, notée D, correspond au biais créé vis-à-vis des marges spécifiées pour le calage, pondérée par une matrice de coût C. Pour tout i, c^i correspond au coût de la i^{ème} variable de redressement. Ainsi le respect de certains critères peut être accentué par rapport à d'autre, voir rendu obligatoire avec $c^i \to \infty$. A l'inverse, $c^i = 0$ correspond à la suppression du critère de redressement. λ sert de facteur d'échelle au biais.

Il existe une solution formelle à ce problème mais qui dépend de λ et de C^{-1} :

$$\omega_k(\lambda) = d_k + d_k \mathbf{x'}_k \left(\sum_{k \in S} d_k \mathbf{x}_k \mathbf{x'}_k + \lambda C^{-1} \right)^{-1} \left(T - \sum_{k \in S} d_k \mathbf{x}_k \right)$$

Le premier point à noter est qu'il est plus facile d'expliciter C^{-1}. En effet, si l'on souhaite que la marge i soit parfaitement respectée, il faut que $c^i = \infty$ et donc $(c^i)^{-1} = 0$. Dans le cas inverse, il suffit d'enlever la variable de la liste des critères de calage.

Une fois C choisie, on peut faire varier λ pour analyser l'impact du facteur d'échelle. λ est défini sur $\mathbb{R}+$. Pour simplifier le choix, Beaumont et Bocci proposent de s'intéresser à $\lambda^* = \lambda/(\lambda+1)$ défini entre 0 et 1. Nous utiliserons cette notation par la suite.

3.2 Méthode Scaled Modified Chi-Square pour d'autres distances version ridge

Il s'agit en fait du même problème de minimisation que la régression généralisée version ridge mais avec d'autres fonctions de distance possibles.

$$\text{Min}\left(\sum_{k\in S} G(\omega_k, d_k) + \frac{\lambda^*}{(1-\lambda^*)}\left(\sum_{k\in S}\omega_k \boldsymbol{x}_k - T\right)' C \left(\sum_{k\in S}\omega_k \boldsymbol{x}_k - T\right)\right)$$

En utilisant le résultat précédent 2.3 et par analogie au problème 3.1, nous obtenons :

$$\omega_k(\lambda^*) = d_k + d_k q_k^{\omega(\lambda^*)} \mathbf{x}'_k \left(\sum_{k\in S} d_k q_k^{\omega(\lambda^*)} \mathbf{x}_k \mathbf{x}'_k + \frac{\lambda^*}{(1-\lambda^*)} C^{-1}\right)^{-1} \left(T - \sum_{k\in S} d_k \mathbf{x}_k\right)$$

$$\text{où} \qquad q_k^{\omega(\lambda^*)} = \frac{2(\omega_k(\lambda^*) - d_k)}{d_k \dfrac{dG(\omega_k(\lambda^*), d_k)}{d\omega_k}}$$

Pour λ et C donnés, l'algorithme 2.3 s'adapte donc très facilement à la version ridge.

3.3 Méthode Shrinkage Minimization version ridge (Ridge-shrinkage)

Rao et Singh (2009) proposent une évolution de la méthode Shrinkage Minimization de Singh et Mohl en version ridge. La méthode Shrinkage Minimization est une succession de régressions généralisées avec la particularité de considérer les poids de l'itération précédente rétrécis au lieu des poids initiaux d_k.

Le passage à une succession de régressions généralisées version ridge semble donc tout approprié. L'algoritme 2.1 est donc pratiquement inchangé. Seule l'étape de minimisation est modifiée pour y introduire une matrice de coût qui dépend de l'itération ν :

$$\omega_k^{\nu+1} = \omega_k^{\nu*} + \omega_k^{\nu*} \mathbf{x}'_k \left(\sum_{k\in S} \omega_k^{\nu*} \mathbf{x}_k \mathbf{x}'_k + C_\nu^{-1}\right)^{-1} \left(T - \sum_{k\in S} \omega_k^{\nu*} \mathbf{x}_k\right)$$

Le détail du calcul de la matrice de coût à chaque itération est développé dans Rao et Singh (2009).

3.4 Méthode mixte

Lors de nos applications, la méthode Scaled Modified Chi-Square pour d'autres distances version ridge mettait trop de temps pour atteindre des étendues de poids intéressantes. Le fait de toujours calculer la distance G par rapport au poids initial semblait poser problème, alors que dans l'algorithme ridge-shinkrage, la convergence est plus rapide.

Nous avons donc légèrement modifié la solution de l'étape de minimisation en mixant les deux approches :

$$\omega_k^{(\nu+1)} = \omega_k^{\nu*} + \omega_k^{\nu*} q_k^{w_k^{\nu*}} \mathbf{x}'_k \left(\sum_{k\in S} \omega_k^{\nu*} q_k^{\omega_k^{\nu*}} \mathbf{x}_k \mathbf{x}'_k + \frac{\lambda^*}{(1-\lambda^*)} C^{-1}\right)^{-1} \left(T - \sum_{k\in S} \omega_k^{\nu*} \mathbf{x}_k\right)$$

où $\quad q_k^{\omega_k^{\nu*}} = \dfrac{2(\omega_k^{\nu*} - d_k)}{d_k \dfrac{dG(\omega_k^{\nu*}, d_k)}{d\omega_k}}$

Les solutions obtenues lors de nos différentes applications sont conformes aux attentes et confirment les bonnes performances de l'algorithme en termes de temps de calcul.

Bibliographie

[1] Bardsley, P. and Chambers, R.L. (1984), Multipurpose estimation from unbalanced samples, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 33, 290–299.

[2] Beaumont, J.-F. and Bocci, C. (2008), Another look at ridge calibration, *Metron-International Journal of Statistics*, LXVI, 5–20.

[3] Chen, J., Sitter, R.R. and Wu, C. (2002), Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika*, 89, 230–237.

[4] Deville, J.-C. and Särndal, C.-E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376–382.

[5] Goga, C. and Shehzad, M.A. (2012), Overview of ridge regression estimators in survey sampling.

[6] Rao, J.N.K and Singh, A.C. (1997), A ridge-shrinkage method for range-restricted weight calibration in survey sampling, *Proceedings of the Section on Survey Research Methods, American Statistics Association*, 57–65.

[7] Rao, J.N.K and Singh, A.C. (2009), Range-restricted weight calibration for survey data using ridge regression, *Pakistan Journal of Statistics*, 25, 371–384.

[8] Roy, G. et Vanheuverzwyn, A. (2001), Redressement par la macro CALMAR : applications et pistes d'amélioration, *Traitement des fichiers d'enquêtes*, Presses Universitaires de Grenoble.

[9] Sautory, O. (1993), La macro calmar : redressement d'un échantillon par calage sur marges. *Documents INSEE*.

[10] Singh, A.C. and Mohl, C. (1996), Understanding calibration estimators in survey sampling, *Survey Methodology*, 22, 107–115.