Projet Migs 2
Amélioration des estimations par calage et réduction du nombre de variables auxiliaires par ACP

Camelia Goga

20 septembre 2016

1 Présentation du calage

Nous sommes confrontés tous les jours avec des sondages. Par exemple, Médiamétrie (l'entreprise française privée qui mesure l'audience dans les medias) donne des estimations journalières d'audience TV, de radio, d'internet, INSEE estime tous les trois mois le taux de chômage et les instituts de sondage privés (BVA, SOFRES, IPSOS, ...) donnent des estimations d'intentions de vote dans le cas des enquêtes électorales. Toutes ces estimations sont calculées à partir des enquêtes par sondages.

Dans la plupart des cas, la quantité à estimer est le total (ou la moyenne) d'une variable d'intérêt Y

$$t_y = \sum_{k\in U} y_k \quad (y_k \quad \text{est la valeur de Y pour le kème individu})$$

et un échantillon d'individus s est sélectionné dans la population d'intérêt U selon un plan de sondage $p(\cdot)$. Les estimateurs utilisés en théorie des sondages sont des sommes pondérées des valeurs de Y enregistrées sur l'échantillon :

$$\hat{t}_{yw} = \sum_{k\in s} w_{ks} y_k$$

avec des poids w_{ks} ne dépendant pas (en général) de Y. Si $w_{ks} = \frac{1}{\pi_k}, k \in s$ *les poids de sondages* avec π_k =la probabilité d'un individu d'être dans un échantillon, on obtient dans ce cas l'estimateur de Horvitz-Thompson. Cet estimateur est sans biais pour le total t_y. Néanmoins, l'estimateur d'Horvitz-Thompson peut être très inefficace (sa variance est très grande) pour certains plans de sondage (le sondage aléatoire simple sans remise, le systématique, ...). Dans ces conditions, des variables auxiliaires $X_1, \ldots, X_p$ (quantitatives ou qualitatives) sont utilisées pour construire un estimateur plus efficace que l'estimateur d'Horvité-Thompson.

La calage (Deville and Särndal, 1992) est une méthode très populaire et utilisée beaucoup dans les instituts de sondage. Cette méthode consiste à déterminer des *poids de calage* $w_{ks}, k \in s$ situés le plus proche possible des poids de sondages (dans le sens de la distance de chi-deux) et tels que le total des variables auxiliaires $X_1, \ldots, X_p$ soient estimés exactement :

$$\begin{aligned} \mathbf{w}_s = (w_{ks})_{k\in s} &= \text{argmin}_{\mathbf{w}} \sum_{k\in s} \frac{(w_k - \pi_k^{-1})^2}{\pi_k^{-1}}, \\ \hat{t}_{w,\mathbf{x}_j} &= t_{\mathbf{x}_j}, \quad j = 1, \ldots, p \end{aligned}$$

avec $\hat{t}_{w,\mathbf{x}_j} = \sum_{k\in s} w_{ks} X_{kj}$ et $t_{\mathbf{x}_j} = \sum_{k\in U} X_{kj}$. Par exemple, dans les enquêtes politiques, les instituts de sondage déterminent ces poids en sorte que la proportion de gens qui ont voté pour un candidat lors des élections précédentes (et même 10 ans avant) soit estimée parfaitement quel que soit l'échantillon d'individus.

Travail demandé : Déterminer l'expression des poids de calage et de l'estimateur par calage ; étudier les propriétés statistiques de cet estimateur. Que pouvez-vous dire si X_j est une variable qualitative ? Comparer à l'aide des simulations l'estimateur par calage avec l'estimateur d'Horvitz-Thompson et en utilisant le critère suivant :

$$R = \frac{\sum_{i=1}^{I}(\hat{t}_{yw}^{(i)} - t_y)^2}{\sum_{i=1}^{I}(\hat{t}_{HT}^{(i)} - t_y)^2}, \quad I = 500$$

où $\hat{t}_{HT}$ est l'estimateur de Horvitz-Thompson.

Application aux données électricité. Nous considérons les données d'électricité irlandaise Commission for Energy Regulation (Irelande) (`http://www.cer.ie/`) résumées dans le fichier `smart278co.Rdata`. Il s'agit de la consommation d'électricité enregistrée toutes les 30 minutes pendant deux semaines (du lundi 5 octobre 2009 à 0 :00 au dimanche 18 octobre 2009 a 23 :59) pour 6291 individus : résidentiels (code 1), petites et moyennes entreprises (code 2) et autres (code 3). Ce fichier contient deux objets : `contrat.smartco` (le type du contrat du client) et `smart.278co` (la consommation d'électricité).

Supposons le scénario suivant : la consommation totale d'électricité de la semaine deux doit être estimée :

$$t_y = \sum_{k \in U} y_k, \quad y_k \quad \text{la consommation de l'individu k pendant la semaine deux}$$

et nous avons la possibilité d'utiliser comme information auxiliaire la consommation de la semaine 1, enregistrée toutes les 30 minutes pour chaque individu. Donnez une estimation de la consommation totale d'électricité de la semaine 2 par calage et à partir d'échantillon aléatoire simple sans remise de taille $n = 600$ et comparer de nouveau avec l'estimateur d'Horvitz-Thompson (à partir de simulations). Que remarquez-vous ?

2 Calage sur composantes principales

Le calage en présence de beaucoup de variables auxiliaires peut s'avérer instable surtout s'il y a des collinéarités entre les variables auxiliaires. Une méthode pour palier cet inconvénient consiste à réduire l'information auxiliaire en réalisant une ACP. Considérer maintenant un calage sur les composantes principales de l'information auxiliaire. Vous pouvez varier le nombre de composantes principales et refaire l'application pratique sur les données électricité et par simulations.

Bibliographie

1. Ardilly, P. (2006), *Les techniques de sondages*, éditions Technip.
2. Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.
3. Tillé, Y. (2001), *Théorie des sondages*, chez Dunod.