Projet Migs 2: Introduction à la théorie des sondages. Une présentation du calage et réduction du nombre de variables auxiliaires par ACP

Camelia Goga

13 octobre 2014

Ce projet a pour but une introduction à la théorie des sondages et il est constitué de deux parties : une partie théorique et suivie de la mise en oeuvre pratique à l'aide des logiciels SAS et R.

1 Présentation des données d'électricité irlandaise

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).

Réalisez une étude descriptive de ces données et essayez de réduire l'information (une ACP).

2 Présentation du calage **(pas plus de 5 pages)**

En théorie des sondages, nous sommes souvent confrontés à l'estimation d'un total ou d'une moyenne d'une variable d'intérêt Y

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

à partir d'un échantillon $s \subset U$ sélectionné selon un plan de sondage $p(\cdot)$. Les estimateurs utilisés en théorie des sondages sont des estimateurs pondérés 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. Par exemple, on peut avoir $w_{ks} = \frac{1}{\pi_k}, k \in s$ *les poids de sondages* avec π_k =la probabilité d'un individu d'être dans un échantillon et on obtient dans ce cas l'estimateur de Horvitz-Thompson. Cet estimateur est sans biais pour le total t_y.

On suppose maintenant qu'on dispose de p variables auxiliaires $X_1, \ldots, X_p$. On note par X_{kj} la valeur de la $j$ème variable pour le $k$ème individu.

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 tel 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 $t_{\mathbf{X}_j} = \sum_{k\in U} X_{kj}$.

1. Déterminer l'expression des poids de calage et l'estimateur par calage. Cet estimateur est-il sans biais pour le total t_y ?
2. Ecrire une fonction en `R` qui calcule ces poids de sondage (attention : utiliser le moins de produits de matrices pour optimiser les calculs).
3. **Application aux données 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.

   (a) Quel est le nombre de variables auxiliaires ?

   (b) Un échantillon aléatoire simple sans remise de taille $n = 600$ est sélectionné. Donnez une estimation de la consommation totale d'électricité de la semaine 2 par calage.

   (c) **Simulations** On considère $I = 500$ échantillons aléatoires simples sans remise de taille $n = 600$. Pour chacun des échantillons, calculer l'estimateur de Horvitz-Thompson et l'estimateur par le calage pour le total t_y. Comparez ces deux estimateurs en utilisant le critère suivant :

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

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

3 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 est de 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 (3) (surtout les points (b) et (c)). Comparez également avec le calage sur l'ensemble de variables auxiliaires et commentez.

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.