# Calage non linéaire

Éric Lesage
eric.lesage@ensai.fr

Ensai-Crest, Laboratoire de Statistique d'Enquête

Journées de Méthodologie Statistique 2009

# Plan de la présentation

# Introduction

L'objectif d'une enquête est d'estimer un certain nombre de paramètres de la population totale (nombre de chômeurs, salaire moyen,... ), avec la meilleure précision possible.

Après la collecte, on peut disposer d'**information auxiliaire** comme la population totale ventilée par sexe et âge, provenant d'un recensement.

Cette information peut mettre en évidence un **déséquilibre de l'échantillon** qui se manifeste par une erreur dans l'estimation des totaux des variables auxiliaires.

Pour corriger cette erreur d'estimation, on utilise l'information auxiliaire pour construire un nouvel estimateur qui permet de **redresser l'erreur d'échantillonnage**.

# Estimateurs par calage

Une des méthodes de redressement repose sur **l'estimation par calage** (Deville et Särndal, 1992).

Ces estimateurs par calage ont la propriété d'estimer exactement les **totaux** des variables auxiliaires.
Le problème est justement qu'ils ne prennent en compte qu'une information auxiliaire linéaire.

On entend par **information auxiliaire linéaire** le fait d'utiliser les totaux des variables auxiliaires.

# Estimateurs par calage

## Problématique

Comment peut-on construire de nouveaux estimateurs par calage en utilisant une **information auxiliaire non linéaire**, comme un ratio, une médiane ou un indice de Gini ?

# Calage non linéaire

## Le principe

Procéder à un calage non linéaire en ayant recours à une équation de calage linéaire.

## Méthodes

1. équations estimantes
2. linéarisation (calage approché)

# Sondage en population finie

- U, une population finie de taille $N$
- $t_y = \sum_{k \in U} y_k$ paramètre d'intérêt
- $s$ échantillon de taille $n$ tiré suivant un plan de sondage $p(s)$
- $\pi_k$ probabilité d'inclusion de $k$ dans l'échantillon $s$
- $d_k = \frac{1}{\pi_k}$ poids d'échantillonnage
- $\hat{t}_{y\pi} = \sum_{k \in s} d_k y_k$, estimateur de Horvitz-Thompson
- $x_1, ..., x_p$, $p$ variables auxiliaires, connues sur $s$
- $t_{x_1}, ..., t_{x_p}$ totaux de ces variables auxiliaires également connus

# Calage (linéaire)

L'estimateur par calage de $t_y$ est : $\hat{t}_y = \sum_{k \in s} w_k y_k$, où $\{w_k\}_{(k \in s)}$ est une série de poids que l'on veut proche des poids $\{d_k\}_{(k \in s)}$. Les $\{w_k\}_{(k \in s)}$ sont obtenus par résolution du programme d'optimisation suivant :

$$\min_{\{w_k\}_{(k \in s)}} \sum_{k \in s} d(w_k, d_k)$$

sous contraintes :

$$\begin{cases} \hat{t}_{x_1} = t_{x_1} \\ \dots \\ \hat{t}_{x_p} = t_{x_p} \end{cases}$$

# Exemple : distance du $\chi^2$

Si $d(w_k, d_k) = \frac{1}{2}\frac{(w_k - d_k)^2}{d_k}$

$$\min_{\{w_k\}_{(k\in s)}} \sum_{k\in s} \frac{1}{2}\frac{(w_k - d_k)^2}{d_k}$$

Solution : $w_k = d_k(1 + x_k'\lambda)$ ( $\lambda$ le vecteur de taille P des multiplicateurs de Lagrange).

# Calage non linéaire : problématique

$\theta(\{x_{1,k}, ..., x_{P,k}\}_{k \in U})$ un paramètre **complexe** connu sur $U$, de dimension 1 (par ex. : $\sigma_x^2$). $\theta$ n'est pas linéaire (i.e. ne se ramène pas à un total).

$\hat{\theta}(\{x_{1,k}, ..., x_{p,k}\}_{k \in s})$ son estimateur par substitution (en utilisant les poids de calage)

## Calage non linéaire

On veut construire un estimateur par calage qui a la propriété de calage non linéaire suivante :

$$\hat{\theta}(\{x_{1,k}, ..., x_{p,k}\}_{k \in s}) = \theta(\{x_{1,k}, ..., x_{p,k}\}_{k \in U})$$

# Exemple de la variance

La variance de la variable auxiliaire $x$ se définit par la formule :

$$\sigma_x^2 = \frac{1}{N}\sum_{k\in U} x_k^2 - \left(\frac{\sum_{k\in U} x_k}{N}\right)^2 = \theta(\{1, x_k, x_k^2\}_{k\in U})$$

Son estimateur "naturel" (ou estimateur par substitution) est :

$$\hat{\sigma}_x^2 = \frac{1}{\hat{N}}\sum_{k\in s} w_k x_k^2 - \left(\frac{\sum_{k\in s} w_k x_k}{\hat{N}}\right)^2$$

où $\hat{N} = \sum_{k\in s} w_k 1$

# Principe d'estimation par équation estimante

Certains paramètres se définissent, ou peuvent se définir, comme solution d'une fonction implicite appelée **équation estimante sur** $U$ (Godambe et Thompson, 1986), i.e. :

$$\sum_{k \in U} \Phi(\theta, x_{1,k}, ..., x_{p,k}) = 0.$$

Dans ce contexte, on construit un estimateur de $\theta$, noté $\hat{\theta}_\pi$, qui est la solution de **l'équation estimante sur** $s$ :

$$\sum_{k \in s} d_k \Phi(\hat{\theta}_\pi, x_{1,k}, ..., x_{p,k}) = 0.$$

# Exemples de paramètres définis par une équation estimante

| Paramètre | $\Phi(\theta, x_{1,k}, ., x_{p,k})$ | équation estimante |
|---|---|---|
| moyenne $\mu$ | $(x_k - \mu)$ | $\sum_{k \in U} (x_k - \mu) = 0$ |
| ratio $R = \frac{\mu_1}{\mu_2}$ | $(x_{1k} - R x_{2k})$ | $\sum_{k \in U} (x_{1k} - R x_{2k}) = 0$ |
| médiane $m$ | $(\mathbb{1}_{x_k \leq m} - \frac{1}{2})$ | $\sum_{k \in U} (\mathbb{1}_{x_k \leq m} - \frac{1}{2}) = 0$ |
| coefficients de régression | $\begin{cases} (x_{1k} - a - b x_{2k}) \\ (x_{1k} - a - b x_{2k}) x_{2k} \end{cases}$ | $\begin{cases} \sum_{k \in U} (x_{1k} - a - b x_{2k}) = 0 \\ \sum_{k \in U} (x_{1k} - a - b x_{2k}) x_{2k} = 0 \end{cases}$ |

# Calage dans le cas de paramètres définis par des équations estimantes

## Proposition

*Dans le cas où $\theta$ est défini par une équation estimante, caler sur $\theta$ revient à imposer que $\theta$ soit la solution de l'équation estimante sur $s$ :*

$$\sum_{k \in s} w_k \Phi(\theta, x_{1,k}, ..., x_{p,k}) = 0.$$

*D'un point de vue pratique, on construit une nouvelle variable auxiliaire $z_k = \Phi(\theta, x_{1,k}, ..., x_{p,k})$ sur le total de laquelle on cale. On a donc comme équation de calage : $\hat{t}_z = t_z = 0$.*

# Linéarisation d'un paramètre $\theta$

Le principe des techniques de linéarisation est d'approcher des formulations d'estimateurs non linéaires par des expressions linéaires auxquelles on les assimile.

Donc, pour un estimateur $\hat{\theta}_\pi$ de $\theta$ linéarisable, on aura l'expression (sous certaines conditions, dont n grand) :

$$\hat{\theta}_\pi \approx \theta - t_z + \hat{t}_{z\pi}$$

où $z_k$ est la variable dite linéarisée de $\hat{\theta}_\pi$ et $\hat{t}_{z\pi} = \sum_{k \in s} d_k z_k$.
Remarque : on peut écrire $z_k = \Phi(\theta, \alpha_1, \ldots, \alpha_Q, x_{1,k}, ..., x_{p,k})$, où les $\alpha_q$ sont des paramètres (de nuisance) connus sur U.

# Calage dans le cas de paramètres linéarisables

## Proposition

*Soit $\theta$, un paramètre non linéaire, estimé par $\hat{\theta}_\pi$, son estimateur par substitution.*
*Si $\hat{\theta}_\pi$ est linéarisable, alors caler sur $t_z$ (où $z_k$ est la variable linéarisée de $\theta$) est équivalent à un calage approché (ou asymptotiquement exact) sur $\theta$.*

$$\hat{t}_z = t_z \Leftrightarrow \hat{\theta} \approx \theta$$

# Exemple du calage (approché) sur la variance

$$\sigma_x^2 = \frac{1}{N}\sum_{k\in U} x_k^2 - \left(\frac{\sum_{k\in U} x_k}{N}\right)^2 = f(t_x, t_{x^2}, N)$$

$$\hat{\sigma}_x^2 = \frac{1}{\hat{N}}\sum_{k\in s} w_k x_k^2 - \left(\frac{\sum_{k\in s} w_k x_k}{\hat{N}}\right)^2 = f(\hat{t}_x, \hat{t}_{x^2}, \hat{N})$$

La variable linéarisée de $\sigma_x^2$ est :

$$z_k = -2\frac{t_x}{N^2}x_k + \frac{1}{N}x_k^2 - \frac{1}{N}\left\{\sigma_x^2 - \left(\frac{t_x}{N}\right)^2\right\}$$

# Exemple du calage (approché) sur la variance

En faisant le calage sur la variable linéarisée, on n'obtient pas exactement $\hat{\sigma}_x^2 = \sigma_x^2$, mais :

$$\hat{\sigma}_x^2 = \sigma_x^2 - \left( \frac{t_x}{N} - \frac{\hat{t}_x}{\hat{N}} \right)^2$$

# Conclusion et piste de travail

Calcul de précision des estimateurs par calage non linéaire.

# Bibliographie


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

Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population : Their relationship and estimation. *Revue International de Statistique*, 54, 127-138.

Merci de votre attention.