Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 141–150
Statistique Canada

Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire

F. DUPONT[1]

RÉSUMÉ

L'estimation par régression et sa généralisation, l'estimation par calage introduite par Deville et Särndal en 1993 permet de réduire a posteriori la variance des estimateur, grâce à l'utilisation de l'information auxiliaire. Dans les enquêtes réalisées par sondage, on dispose souvent d'information additionnelle exploitable répartie selon un schéma complexe notamment lorsque le sondage est réalisé en plusieurs phases. Une adaptation de l'estimation par régression a été proposée avec ses variantes dans le cadre du sondage à deux phases par Särndal et Swensson en 1987. Cet article propose d'étudier les stratégies d'estimation alternatives selon deux configurations alternatives pour l'information auxiliaire, en reliant les deux approches possibles pour aborder le problème: modèle de régression et estimation par calage.

MOTS CLÉS: Information auxiliaire; estimateur par régression; estimateur par calage; sondage en deux phases.

1. INTRODUCTION

L'estimateur par régression étudié par Fuller (1975), Cassel, Särndal et Wretman (1976), Särndal (1980), Gourieroux (1981), Isaki et Fuller (1982), et Wright (1983) permet d'améliorer, a posteriori, c'est-à-dire une fois le sondage réalisé, l'estimation d'un total d'une variable d'intérêt sur la base de variables auxiliaires $x_1, \ldots, x_k$ pour lesquelles on dispose d'information additionnelle. On réduit en effet la variance par rapport à l'estimateur d'Horwitz-Thompson en utilisant l'estimateur par régression à condition de connaître la valeur sans aléa, des totaux sur la population cible des variables auxiliaires, qui constitueront l'information additionnelle appelée information auxiliaire. Deville et Särndal ont proposé en 1992 une classe d'estimateurs dérivée d'une approche par repondération qui répond à la même problématique de réduction de variance: les estimateurs par calage. Le calage des poids de sondage réalisé permet en effet d'intégrer a posteriori une information auxiliaire du type totaux $X_1, \ldots, X_k$ de k variables $x_1, \ldots, x_k$ dans l'estimation réalisée à partir des nouvelles pondérations et donc de l'améliorer. Cette approche généralise l'estimation par régression qui constitue l'un des éléments de la classe.

Toutefois, dans les enquêtes réalisées par sondage, on dispose souvent d'information additionnelle exploitable répartie selon un schéma plus complexe que ce qui vient d'être décrit, notamment lorsque le sondage est réalisé en plusieurs phases. Cet article étudie différentes stratégies d'utilisation de cette information auxiliaire complexe dans le cadre d'un sondage en deux phases, la généralisation à plus de deux phases étant possible.

Lorsque le plan de sondage comporte deux phases, l'information auxiliaire est constituée de l'information connue pour la population toute entière, mais aussi de l'information connue pour l'échantillon issu de la première phase de sondage. Or, ces deux informations peuvent porter sur des variables différentes.

Särndal et Swensson proposent dans leur article de 1987, un estimateur utilisant toute l'information auxiliaire disponible pour un tirage en deux phases, avec une information auxiliaire différente pour la population toute entière et la population issu du tirage de la première phase. Il s'agit d'un estimateur adaptant le principe de l'estimateur par régression lorsque l'information connue pour les individus issus du tirage première phase est considérée comme substituable à l'information agrégée et de meilleure qualité que l'information disponible pour la population cible toute entière, vis-à-vis de l'estimation de la variable d'intérêt. Or, Il arrive en pratique que ces deux informations soient complémentaires plutôt que substituables. On a donc cherché dans cette étude à développer l'estimation par régression, dans un contexte où les informations auxiliaires sont complémentaires.

Par ailleurs, dans la mesure où l'estimation par calage généralise l'estimation par régression lorsque l'information auxiliaire n'est constituée que d'un seul niveau, on a cherché à adapter l'estimation par calage dans ce contexte. On passe en revue les stratégies de calages afin de proposer les plus adaptées, en cherchant à les relier aux généralisations de l'estimation par régression possibles dans ce contexte.

On montre (section 2) que l'utilisation conjointe de deux informations auxiliaires différentes conduit à deux modèles de régression et trois décompositions de la variable d'intérêt associées. L'approche assistée par modèle de régression (AAMR) permet donc de dériver 3 estimateurs alternatifs.

L'approche calage (AC) (section 3) permet quant à elle de dériver 4 estimateurs. Chacun de ces estimateurs peut être relié (associé) aux trois estimateurs dérivés de l'approche modèle de régression.

[1] F. Dupont, Unité Méthodes Statistiques, Institut National de la Statistique et des Études Économiques, (INSEE), 18 Blvd. Adolphe Pinard, 75675 Paris Cedex.

Ainsi (section 4), les deux approches peuvent être reliées et débouchent sur trois classes d'estimateurs associées chacune à une décomposition de la variable d'intérêt. Les estimateurs d'une même classe ont la même variance asymptotique.

L'évaluation des stratégies fondée sur le seul plan de sondage oriente le choix vers la troisième classe d'estimateurs qui domine les deux autres du point de vue de la variance.

L'évaluation des stratégies, lorsqu'elle s'appuie sur une modélisation de la variable d'intérêt conduit à privilégier la classe d'estimateurs associée à la modélisation adoptée.

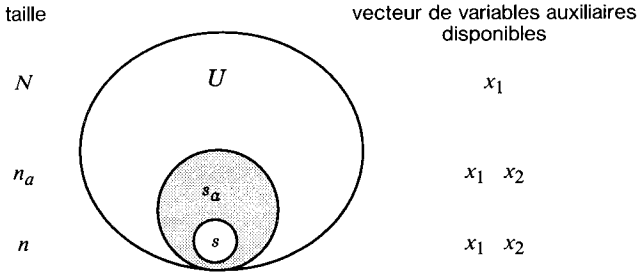
Dans un contexte où l'on souhaite effectuer le redressement d'une enquête, et que l'on souhaite simultanément corriger les biais qu'induiraient l'utilisation des pondérations brutes et réduire la variance, les conclusions doivent être adaptées: les modifications introduites dans les pondérations pour corriger les biais sont plus importantes que les corrections pour réduction de variance. Dès lors, on intégrera dans le calage les variables, dès que l'on pense qu'elles interviennent dans la probabilité de sélection et donc participent à la création du biais.

Lorsque les variables auxiliaires sont qualitatives, l'arbitrage entre utilisation a priori et a posteriori de l'information auxiliaire, c'est-à-dire entre, utilisation au niveau de l'échantillonnage ou au niveau du redressement, repose encore sur la distinction entre les deux modélisations de la variable d'intérêt.

Il est possible d'étendre ces résultats au cas de sondages en plus de deux phases.

2. NOTATIONS

Le cadre est celui d'un sondage en deux phases. On suppose qu'on dispose d'information auxiliaire à deux niveaux différents: population cible et population issue du tirage 1ière phase. La situation peut être schématisée de la manière suivante:



Où U représente la population cible pour laquelle les valeurs du vecteur de variable x_1 sont connues ou à défaut le total $X_1 = \sum_{i\in U} x_{i1}$. s_a représente un niveau intermédiaire de sondage pour lequel les valeurs des vecteurs de k_1 variables x_1 et k_2 variables x_2 sont connues pour tous les individus. On note π_{ia}, la probabilité de sélection du tirage associée à la première phase du sondage. s représente l'échantillon final pour lequel les valeurs de la variable y dont on cherche à estimer le total sont disponibles, ainsi que les valeurs des vecteurs de variables auxiliaires x_1 et x_2. On note $\pi_i = P(i \mid s_a)$.

On souhaite utiliser au mieux toute cette information auxiliaire pour améliorer les estimations qui seront effectuées à partir des données recueillies sur l'échantillon issu de la deuxième phase de sondage s.

Une première idée naturelle consiste à chercher la généralisation de l'estimateur par régression dans ce contexte.

3. APPROCHE ESTIMATION PAR RÉGRESSION

3.1 L'information contenue dans x_1 est jugée substituable à l'information contenue dans x_2 pour estimer y et de moins bonne qualité que x_2

Särndal, Swensson et Wretman proposent dans leur ouvrage l'estimateur par régression suivant pour l'estimation du total de y:

$$\hat{Y}_1 = \sum_{i\in s} \frac{y_i}{\pi_i \pi_{ai}} + \left(\sum_{i\in s_a} \frac{x'_{i2}\hat{b}_2}{\pi_i} - \sum_{i\in s} \frac{x'_{i2}\hat{b}_2}{\pi_i \pi_{ai}} \right) + \left(\sum_{i\in U} x'_{i1}\hat{b}_1 - \sum_{i\in s_a} \frac{x'_{i1}\hat{b}_1}{\pi_{ai}} \right)$$

où le deuxième terme est la correction pour mauvaise estimation sur s_a et le troisième est la correction pour mauvaise estimation sur s.

L'estimateur peut aussi s'écrire:

$$\hat{Y}_1 = \sum_{i\in U} x'_{i1}\hat{b}_1 + \sum_{i\in s_a} \frac{(x'_{i1}\hat{b}_1 - x'_{i2}\hat{b}_2)}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x'_{i2}\hat{b}_2)}{\pi_i \pi_{ai}}$$

où le deuxième terme est la correction pour mauvaise approximation de y_i par $x'_{i1}\hat{b}_1$ et le troisième est la correction pour mauvaise approximation de y_i par $x'_{i2}\hat{b}_2$;

avec $$\hat{b}_1 = \left(\sum_{i\in s_a} \frac{x_{i1}x'_{i1}}{\pi_{ai}} \right)^{-1} \left(\sum_{i\in s} \frac{x_{i1}y_i}{\pi_i \pi_{ai}} \right)$$

et $$\hat{b}_2 = \left(\sum_{i\in s} \frac{x_{i2}x'_{i2}}{\pi_i \pi_{ai}} \right)^{-1} \left(\sum_{i\in s} \frac{x_{i2}y_i}{\pi_i \pi_{ai}} \right).$$

L'idée sous-jacente est que l'on a deux modèles concurrents pour y qui sont:

(1) $y_i = x'_{i1}b_1 + u_{i1}$ avec $E(u_{i1}) = 0$ et $V(u_{i1}) = \sigma_1^2$

et

(2) $y_i = x'_{i2}b_2 + u_{i2}$ avec $E(u_{i2}) = 0$ et $V(u_{i2}) = \sigma_2^2$

dont on pense que le second est a priori meilleur pour prédire la valeur de y_i. x_1 fonctionne donc dans cette optique modèle, comme une "*proxy*" de x_2. Une situation de ce genre correspond par exemple au cas où x_2 représente la mise à jour c'est-à-dire à la date de l'enquête de la variable extraite de la base de sondage x_1. Autrement dit si x_2 était disponible au niveau de la population toute entière, l'estimateur utilisé serait

$$\sum_{i\in U} x'_{i2}\,\hat{b}_1 + \sum_{i\in s} \frac{(y_i - x'_{i2}\,\hat{b}_2)}{\pi_i \pi_{ai}}.$$

Imaginons maintenant le cas d'une enquête par sondage en deux phases effectuée auprès des ménages. Supposons que la base de sondage soit constituée de logements pour lesquels on dispose d'une information constituée de la taille du logement notée x_1, connue donc pour tous les individus de la population cible . Si tous les individus issus de la première phase de sondage sont interrogés sur la composition du ménage notée x_2, notamment sur le nombre d'enfants du ménage, les deux informations apparaissent plutôt complémentaires que substituables, lorsqu'il s'agit d'étudier le budget du ménage. Ceci est encore renforcé si au lieu de la composition du ménage on recueille l'âge du chef de ménage ou bien sa profession.

Dans une optique modèle, la situation alternative où l'information contenue dans x_1 est jugée complémentaire de celle contenue dans x_2 pour estimer y, apparaît donc naturellement.

3.2 L'information contenue dans x_1 est jugée complémentaire de celle contenue dans x_2 pour estimer y

3.2.1 Décomposition $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$

Le modèle sous-jacent est alors:

$$y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i \quad \text{avec} \quad E(u_i) = 0 \quad \text{et} \quad V(u_i) = \sigma_1^2.$$

L'estimateur à utiliser est alors:

$$\hat{Y}_2 = \sum_{i\in U} x'_{i1}\hat{a}_1 + \sum_{i\in s_a} \frac{x'_{i2}\hat{a}_2}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2)}{\pi_i \pi_{ai}}$$

avec:

$$\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix} = \begin{pmatrix} \sum_{i\in s} \frac{x_{i1}x'_{i1}}{\pi_{ai}\pi_i} & \sum_{i\in s} \frac{x_{i1}x'_{i2}}{\pi_{ai}\pi_i} \\ \sum_{i\in s} \frac{x_{i2}x'_{i1}}{\pi_{ai}\pi_i} & \sum_{i\in s} \frac{x_{i2}x'_{i2}}{\pi_{ai}\pi_i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i\in s} \frac{x_{i1}y_i}{\pi_{ai}\pi_i} \\ \sum_{i\in s} \frac{x_{i2}y_i}{\pi_{ai}\pi_i} \end{pmatrix}.$$

En effet, la variable y se décompose en trois composantes $y_i = x'_{i1}\hat{a}_1 + x'_{i2}\hat{a}_2 + \hat{u}_i$. Le total de y se décompose donc en trois composantes qui sont chacune estimée au niveau le plus haut, c'est-à-dire avec le maximum de précision possible:

- U pour $x'_{i1}\hat{a}_1$,
- s_a pour $x'_{i2}\hat{a}_2$, et
- s pour $\hat{u}_i$.

3.2.2 Décomposition $y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

Si on souhaite utiliser au maximum l'information contenue dans x_1 disponible sur U, il est naturel d'introduire une autre écriture du même modèle $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$ qui isole tout ce qui dans y peut être pris en compte à travers x_1. Elle s'écrit:

$$y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i \quad \text{avec}$$
$$E(u_i) = 0 \quad \text{et} \quad V(u_i) = \sigma_1^2,$$

où M_{x_1} représente la projection orthogonale, dans la métrique associée aux poids $1/\pi_{ai}$, sur l'orthogonal de l'espace vectoriel engendré dans s_a (assimilé à $\Re^n$) par le groupe de variables x_1.

$M_{x_1}(x_{i2})$ est définie par:

$$M_{x_1}(x_{i2}) = x'_{i2} - \left(\sum_{i\in s_a} \frac{x_{i2}x'_{i1}}{\pi_{ai}}\right)\left(\sum_{i\in s_a} \frac{x_{i1}x'_{i1}}{\pi_{ai}}\right)^{-1} x_{i1}.$$

L'estimateur naturel associé est alors:

$$\hat{Y}_3 = \sum_{i\in U} x'_{i1}\hat{c}_1 + \sum_{i\in s_a} \frac{(M_{x_1}x'_{i2}\hat{c}_2)}{\pi_{ai}} + \sum_{i\in s} \frac{(y_i - x'_{i1}\hat{c}_1 - M_{x_1}x'_{i2}\hat{c}_2)}{\pi_i \pi_{ai}},$$

où $\hat{c} = \begin{pmatrix}\hat{c}_1\\ \hat{c}_2\end{pmatrix}$ est le coefficient de la régression $y = x'c_1 + (M_{x_1}x_2)'c_2 + u$ estimée sur s avec des poids $1/\pi_{ai}\pi_i$ (qui diffère légèrement de $\begin{pmatrix}\hat{b}_1\\ \hat{a}_2\end{pmatrix}$).

3.3 Les trois estimateurs dérivés de l'approche modèles

L'approche par la modélisation nous a permis de construire 3 estimateurs que l'on peut réécrire synthétiquement en introduisant de nouvelles notations. On écrira dans toute la suite pour un vecteur de variable z quelconque:

$$\hat{\hat{Z}} = \sum_{i\in s} \frac{1}{\pi_{ai}\pi_i} z$$

$$\hat{Z} = \sum_{i\in s_a} \frac{1}{\pi_{ai}} z.$$

Avec ces notations les trois estimateurs se réécrivent:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2' \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2' \hat{b}_2]$$

associé aux modèles:

et

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1}$$

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$$\hat{Y}_2 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$$

associé au modèle

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$$\hat{Y}_3 = [X_1' \hat{c}_1] + [M_{x_1} \hat{X}_2' \hat{c}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{\hat{X}}_2' \hat{c}_2]$$

associé au modèle

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i .$$

De la même façon que l'on généralise l'estimateur par la régression par les estimateurs par calage, on peut aborder le problème de l'utilisation de l'information auxiliaire à plusieurs niveaux à travers la théorie du calage, en essayant de construire des stratégies de calage adaptées à la configuration d'information auxiliaire étudiée dans cette article.

4. APPROCHE PAR CALAGE

4.1 Différentes stratégies possibles

Lorsque l'on cherche à généraliser l'estimation par calage proposée dans un contexte où l'information auxiliaire est présente à un seul niveau: celui de la population totale, plusieurs stratégies apparaissent naturellement:

Stratégie 1

a) caler la structure de l'échantillon 1ière phase s_a sur celle de la population totale U en termes de variable x_1, puis,

b) caler la structure de l'échantillon 2ième phase s sur celle de l'échantillon 1ière phase s_a en terme de variable x_2.

NB: pour cette dernière opération, mieux vaut tenir compte du calage précédent en x_1 pour déterminer la valeur de référence dans le calage en x_2 sur s_a. Si l'on ne tient pas compte du calage précédent, seules les estimations effectuées au niveau de s_a profiteront de l'amélioration apportée par l'étape a). Une bonne façon de s'en convaincre consiste à considérer le cas particulier extrême où $x_1 = x_2$.

Cette stratégie correspond aux équations de calage suivantes:

étape a:

$$\sum_{i \in s_a} \frac{F(x_{i1}' \beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1} = X_1$$

qui détermine β_1, puis

étape b:

$$\sum_{i \in s} \frac{F(x_{i1}' \beta_1)}{\pi_{ai} \pi_i} F(x_{i2}' \beta_2) x_{i2} = \sum_{i \in s_a} \frac{F(x_{i1}' \beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*$$

qui détermine β_2,

où F désigne tout au long de l'article la fonction utilisée dans le calage qui peut être linéaire, exponentielle, linéaire tronquée, logit (*cf.* Deville, Särndal 1993).

Stratégie 2

Caler la structure de l'échantillon 2ième phase s simultanément en terme de variables x_1 et x_2, c'est-à-dire:

- sur la structure de la population totale U en ce qui concerne x_1
- sur la structure de s_a pour x_2.

Cette deuxième stratégie nous conduit aux équations de calage suivantes:

$$\sum_{i \in s} \frac{F(x_{i1} \alpha_1 + x_{i2} \alpha_2)}{\pi_{ai} \pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \quad \text{et}$$

$$\sum_{i \in s} \frac{F(x_{i1}' \alpha_1 + x_{i2}' \alpha_2)}{\pi_{ai} \pi_i} x_{i2} = \sum_{i \in s_a} \frac{x_{i2}}{\pi_{ai}} = \hat{X}_2,$$

qui déterminent α_1 et α_2.

La première stratégie présente l'avantage de corriger les pondérations 1ière phase, c'est-à-dire d'intégrer l'information auxiliaire au niveau le plus élevé. La deuxième stratégie permet quant à elle, de corriger les pondérations qui seront effectivement utilisées dans l'estimation, et notamment d'obtenir une estimation parfaite du total de x_1.

Une troisième stratégie peut être proposée qui réunit les avantages des deux stratégies précédentes, et semble donc pouvoir les dominer:

Stratégie 3

a) caler la structure de l'échantillon 1ière phase s_a sur celle de la population totale U en termes de variables x_1, puis,

b) caler la structure de l'échantillon 2ième phase s simultanément en termes de variables x_1 et x_2, c'est-à-dire:
   - sur la structure de la population totale U en ce qui concerne x_1
   - sur la structure de s_a modifiée en tenant compte du calage précédent pour x_2.

Cette stratégie conduit aux équations de calage suivantes:

étape a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1}$$

qui détermine β_1, puis

étape b:

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{et}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

qui déterminent γ_1 et γ_2.

Enfin une quatrième stratégie peut être proposée qui peut être vue comme une variante de la stratégie précédente:

Stratégie 4

a) caler la structure de l'échantillon 1ière phase s_a sur celle de la population totale U en terme de variable x_1, puis,

b) caler la structure de l'échantillon 2ième phase s simultanément en terme de variables x_1 et x_2, en partant des poids modifiés par le calage précédent, soit:
   - sur la structure de la population totale U en ce qui concerne x_1
   - sur la structure de s_a modifiée en tenant compte du calage précédent pour x_2.

Cette stratégie conduit aux équations de calage suivantes:

étape a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1},$$

qui détermine β_1, puis

étape b:

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{et}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

qui déterminent δ_1 et δ_2.

Lorsque la fonction de calage est exponentielle, il est clair que les stratégies 3 et 4 coïncident.

Dans cette approche en termes de calage, le point de vue adopté est celui de la réduction de variance basée sur les caractéristiques du plan de sondage, sans considération de modèle. Deux questions se posent alors naturellement:

- Peut-on relier chacune de ces quatre stratégies à une approche en terme de modèle?
- Peut-on comparer ces quatre stratégies en termes de variance?

Nous étudierons d'abord le lien entre les trois stratégies définies par une approche calage et les stratégies définies par une approche modèle ou régression, puis dans un deuxième temps nous nous intéresserons au calcul des variances des estimateurs associés à chacune des stratégies.

4.2 Lien entre les différentes stratégies possibles et l'approche régression

Lorsque F est linéaire, chacun des estimateurs associés aux quatre stratégies peut se réécrire simplement.

Notations

Dans toute la suite on utilisera les notations suivantes pour un vecteur de variable z quelconque:

$$\hat{\hat{Z}}^* = \sum_{i\in s} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} z_i \qquad \hat{Z}^* = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} z_i.$$

On omettra également les indices i pour alléger la présentation lorsqu'il n'y a pas d'ambiguïté.

Stratégie 1

Les pondérations sont de la forme:

$$w_i^4 = \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} F(x_{i2}'\beta_2),$$

l'estimateur associé $\hat{Y}_4$ peut se réécrire en traduisant l'effet du deuxième calage en x_2:

$$\hat{Y}_4 = \hat{\hat{Y}}^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]'\hat{B}_2 \qquad \text{avec}$$

$$\hat{B}_2 = \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a\pi} x_2 x_2'\right)^{-1} \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a\pi} x_2 y\right),$$

puis en traduisant l'effet du premier calage en x_1:

$$\hat{Y}_4 = \hat{\hat{Y}} + [X_1 - \hat{X}_1]'\hat{B}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2^*]'\hat{B}_2,$$

soit encore:

$$\hat{Y}_4 = [X_1'\hat{B}_1] + [\hat{X}_2^{*\prime}\hat{B}_2 - \hat{X}_1'\hat{B}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime}\hat{B}_2],$$

$$\text{avec} \quad \hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a\pi}\right).$$

Or $\hat{Y}_1$ se réécrit:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2^{*\prime} \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime} \hat{b}_2].$$

On retrouve donc un estimateur semblable à l'estimateur $\hat{Y}_1$ issu de l'approche modèle dans le cas où l'information contenue dans x_1 est jugée substituable à l'information contenue dans x_2 pour estimer y et de moins bonne qualité. Les différences entre $\hat{Y}_1$ et $\hat{Y}_4$ portent sur les points suivants:

1. $\hat{B}_2$ est estimé en intégrant les modifications du calage en x_1 contrairement à $\hat{b}_2$.
2. L'estimation $\hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} \left(\sum_{s} \frac{x_1 y}{\pi_a \pi}\right)$ de B_1, est effectuée en partie sur s_a contrairement à $\hat{b}_1$.
3. Enfin on utilise les poids redressés $F(x_1\beta_1)/\pi_a\pi$ dans les sommes en x_2 sur s et sur S_a dans $\hat{Y}_4$ contrairement à ce que l'on fait pour $\hat{Y}_1$: l'estimation en x_2 est améliorée par la connaissance de x_1.

La modélisation sous-jacente est donc bien ici: (1) $y_i = x_{i1}' b_1 + u_{i1}$ et (2) $y_i = x_{i2}' b_2 + u_{i2}$ dont on pense que le second est a priori meilleur pour prédire la valeur de y_i.

Stratégie 2

On obtient des poids

$$w_i^5 = \frac{F(x_{i1}'\alpha_1 + x_{i2}'\alpha_2)}{\pi_{ai}\pi_i},$$

l'estimateur associé se réécrit:

$$\hat{Y}_5 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

On retrouve donc exactement l'estimateur $\hat{Y}_2$ proposé dans l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y. Le modèle sous-jacent est bien ici $y_i = x_{i1} a_1 + x_{i2} a_2 + u_i$.

Stratégie 3

On obtient des poids

$$w_i^6 = \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i},$$

l'estimateur associé se réécrit:

$$\hat{Y}_6 = \hat{\hat{Y}} + [X_1 - \hat{\hat{X}}_1]' \hat{a}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2]' \hat{a}_2$$

soit:

$$\hat{Y}_6 = [X_1' \hat{a}_1] + [\hat{X}_2^{*\prime} \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

Or,

$$\hat{X}_2^* = \sum_{s_a} \frac{x_2}{\pi_a} + \left(\sum_{s_a} \frac{x_2 x_1'}{\pi_a}\right)^{-1} \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right) \left[X - \sum_{s_a} \frac{x_1}{\pi_a}\right].$$

On en déduit en remplaçant dans $\hat{Y}_6$ que:

$$\hat{Y}_6 = [X_1' \hat{C}_1] + [M_{x_1} \hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2],$$

avec

$$\hat{C}_1 = \hat{a}_1 + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a}\right) \hat{a}_2.$$

On retrouve donc un estimateur proche de l'estimateur $\hat{Y}_3$ proposé dans l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y. Le modèle sous-jacent est ici $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$. Les différences entre $\hat{Y}_3$ et $\hat{Y}_6$ portent sur les coefficients estimés: $\begin{pmatrix} \hat{C}_1 \\ \hat{a}_2 \end{pmatrix}$ diffère légèrement de $\begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix}$ et $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$ diffère légèrement de $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{X}_2' \hat{c}_2]$. En revanche ces quantités sont asymptotiquement équivalentes.

Stratégie 4

On obtient des poids

$$w_i^7 = \frac{F(x_{i1}'\beta_1) F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i},$$

l'estimateur associé se réécrit:

$$\hat{Y}_7 = \hat{\hat{Y}}^* + [X_1 - \hat{\hat{X}}_1^*]' \hat{a}_1^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{a}_2^*.$$

En modifiant les poids initiaux en $d_i = F(x_{i1}\beta_1)/\pi_{ai}\pi_i$, on obtient de la même façon:

$$\hat{Y}_7 = [X_1' \hat{a}_1^*] + [\hat{X}_2^{*\prime} \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1^{*\prime} \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*].$$

En remplaçant $\hat{X}_2^*$ par son expression trouvée plus haut on obtient:

$$\hat{Y}_7 = [X_1' \hat{C}_1^*] + [M_{x_1} \hat{X}_2' \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1'^* \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*],$$

avec

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a}\right) \hat{a}_2^*.$$

$\hat{Y}_7$ et $\hat{Y}_6$ sont asymptotiquement équivalents.

En effet, posons $w = y - x_1' \hat{a}_1^* - x_2' \hat{a}_2^*$. Alors $\hat{Y}_7 = \hat{Y}_6 + [\hat{\hat{W}}^* - \hat{\hat{W}}]$. Or, asymptotiquement $[\hat{\hat{W}}^* - \hat{\hat{W}}]$ est un infiniment petit négligeable devant $\hat{Y}_6$:

$$[\hat{\hat{W}}^* - \hat{\hat{W}}] = \left(\sum_s \frac{wx_1'}{\pi_a \pi}\right)\left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1}[X_1 - \hat{\hat{X}}_1],$$

et

$$\left(\sum_s \frac{wx_1'}{\pi_a \pi}\right) \text{ tend vers zéro et } [X_1 - \hat{\hat{X}}_1] = O\left(\frac{1}{\sqrt{m}}\right).$$

On obtient finalement: $\hat{Y}_7 \cong \hat{Y}_6$.

En conclusion, lorsque la fonction de calage est exponentielle l'estimateur $\hat{Y}_7$ coïncide exactement avec le précédent. Lorsque F est linéaire, $\hat{Y}_7$ est proche du précédent et correspond donc encore à l'approche modèle de régression dans le cas où l'information contenue dans x_1 est jugée complémentaire à l'information contenue dans x_2 pour estimer y et où on utilise la décomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$.

Conclusion: les trois classes d'estimateurs

On vient de voir que les quatre stratégies dérivées d'une approche par calage pouvaient être associées à une modélisation par régression. On obtient ainsi trois classes d'estimateur qui sont:

$Y_4 \cong Y_1$ associés aux modèles

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1},$$

et

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$\hat{Y}_5 = \hat{Y}_2$ associé au modèle

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$\hat{Y}_6 \cong \hat{Y}_3$ et $\hat{Y}_7 \cong \hat{Y}_3$ associés au modèle

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_1.$$

L'approximation $\cong$ qui indique que les estimateurs sont rattachés au même modèle de régression prend tout son sens lorsqu'on s'intéresse au calcul de variance de ces différents estimateurs. Les estimateurs qui sont rattachés au même modèle de régression ont en effet la même variance asymptotique.

5. ESTIMATION DE VARIANCES

Étudions les variances des différents estimateurs $\hat{Y}_1$, ..., $\hat{Y}_7$ définis précédemment. VA désigne la variance asymptotique d'un estimateur qui est obtenue lorsque N, n et m tendent vers l'infini dans un rapport constant.

5.1 Estimateur $\hat{Y}_1$ et $\hat{Y}_4$: modèle $y_i = x_{i1}' b_1 + u_{i1}$ et (2) $y_i = x_{i2}' b_2 + u_{i2}$.

• Estimateur $\hat{Y}_1$

La variance de cette estimateur et son estimation sont données dans l'ouvrage de Särndal, Swensson et Wretman (1991). La variance se décompose en deux termes qui mesurent la part de variance due respectivement à la première et à la deuxième phase du sondage.

$$\mathrm{VA}(\hat{Y}_1) = \left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{u_{1i} u_{1j}}{\pi_i \pi_j}\right) + \left(E_{s_a} \sum_{i,j \in s_a} \Delta_{ij}^2 \frac{u_{2i} u_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}}\right),$$

avec: $\Delta_{ij}^2 = \pi_{ij} - \pi_i \pi_j$,

$\Delta_{ij}^1 = \pi_{aij} - \pi_{ai} \pi_{aj}$,

$u_{1i} = y_i - x_{i1}' b_1$,

$u_{2i} = y_i - x_{i2}' b_2$,

$$b_1 = \left(\sum_{i \in U} x_{i1} x_{i1}'\right)^{-1}\left(\sum_{i \in U} x_{i1} y_i\right),$$

$$b_2 = \left(\sum_{i \in U} x_{i2} x_{i2}'\right)^{-1}\left(\sum_{i \in U} x_{i2} y_i\right).$$

Ainsi l'estimateur de variance se décompose aussi en deux termes qui estiment la part de variance relative à chacune des phases de tirage. On retrouve que par construction de $\hat{Y}_1$, x_1 permet de réduire la variance apportée par la première phase et x_2 permet de réduire la variance apportée par la deuxième phase

$$\hat{V}(\hat{Y}_1) = \left(\sum_{i \in s}\sum_{j \in s} \frac{\Delta_{ij}^1}{\pi_{ij} \pi_{aij}} \frac{\hat{u}_{1i} \hat{u}_{1j}}{\pi_i \pi_j}\right) + \left(\sum_{i,j \in s} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\hat{u}_{2i} \hat{u}_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}}\right),$$

1ière phase 2ième phase

avec: $\hat{u}_{1i} = y_i - x_{i1}' \hat{b}_1$,

$\hat{u}_{2i} = y_i - x_{i2}' \hat{b}_2$.

Une telle décomposition repose sur l'écriture $V(\hat{Y}_1) = V(E[\hat{Y}_1 \mid s_a]) + E(V[\hat{Y}_1 \mid s_a])$ qui interviendra pour tous les autres estimateurs.

• Estimateur $\hat{Y}_4$

Les termes du développement au premier ordre en $1/\sqrt{m}$ de $\hat{Y}_1$ et $\hat{Y}_4$ coïncident exactement. On peut donc donner un sens plus précis à l'écriture $\hat{Y}_4 \cong \hat{Y}_1$. On en déduit que $\mathrm{VA}(\hat{Y}_1) = \mathrm{VA}(\hat{Y}_4)$. Ainsi:

$$\hat{V}(\hat{Y}_4) =$$

$$\left(\sum_{i,j\in s}\sum_{j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{u}_{2i}\tilde{u}_{2j}}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

avec: $\tilde{u}_{1i} = y_i - x'_{i1}\tilde{B}_1,$

$\tilde{u}_{2i} = y_i - x'_{i2}\tilde{B}_2.$

5.2 Estimateurs $\hat{Y}_2 = \hat{Y}_5$: modèle $y_i = x'_{i1}a_1 + x'_{i2}a_2 + u_i$

On montre (voir Dupont 1994) facilement que:

$$\mathrm{VA}(\hat{Y}_2) = \mathrm{VA}(\hat{Y}_5) \cong$$

$$\left(\sum_{i,j\in U}\Delta^1_{ij}\frac{v_iv_j}{\pi_i\pi_j}\right)+\left(E_{s_a}\sum_{i,j\in s_a}\Delta^2_{ij}\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

avec: $v_i = y_i - x'_{i1}a_1,$

$u_i = y_i - x'_{i1}a_1 - x'_{i2}a_2$

$$a = \begin{pmatrix}a_1\\a_2\end{pmatrix} = \begin{pmatrix}\sum_{i\in U}x_{i1}x'_{i1} & \sum_{i\in U}x_{i1}x'_{i2}\\ \sum_{i\in U}x_{i2}x'_{i1} & \sum_{i\in U}x_{i2}x'_{i2}\end{pmatrix}\begin{pmatrix}\sum_{i\in U}x_{i1}y_i\\ \sum_{i\in U}x_{i2}y_i\end{pmatrix}.$$

On en déduit que:

$$\hat{V}(\hat{Y}_2) = \hat{V}(\hat{Y}_5) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\hat{v}_i\hat{v}_j}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

avec: $\hat{v}_i = y_i - x'_{i1}\hat{a}_1,$

$\hat{u}_i = y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2.$

On retrouve dans cette écriture que par construction de $\hat{Y}_2 - \hat{Y}_5$, x_1 réduit la variance apportée par la première phase et x_1 et x_2 sont utilisés simultanément pour réduire la variance apportée par la deuxième phase.

5.3 Estimateurs $\hat{Y}_3$, $\hat{Y}_6$ et $\hat{Y}_7$: modèle $y_i = x'_{i1}c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

On montre que $\mathrm{VA}(\hat{Y}_6) = \mathrm{VA}(\hat{Y}_7) = \mathrm{VA}(\hat{Y}_3)$. Ainsi:

$$\mathrm{VA}(\hat{Y}_3) = \mathrm{VA}(\hat{Y}_6) = \mathrm{VA}(\hat{Y}_7) \cong$$

$$\left(\sum_{i,j\in U}\Delta^1_{ij}\frac{u_{1i}u_{1j}}{\pi_i\pi_j}\right)+\left(E_{s_a}\sum_{i,j\in s_a}\Delta^2_{ij}\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$u_{1i} = y_i - x'_{i1}c_1 = y_i - x'_{i1}b_1,$

$u_i = y_i - x'_{i1}c_1 - M_{x_1}x'_{i2}c_2 = y_i - x'_{i1}a_1 - x'_{i2}a_2.$

On en déduit les trois estimateurs de variance qui diffèrent en raison de coefficients estimés différents:

$$\hat{V}(\hat{Y}_3) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\hat{u}_{1i}\hat{u}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$\hat{u}_{1i} = y_i - x'_{i1}\hat{c}_1,$

$\hat{u}_i = y_i - x'_{i1}\hat{c}_1 - M_{x_1}x'_{i2}\hat{c}_2,$

$$\hat{V}(\hat{Y}_6) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{u}_i\tilde{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$\tilde{u}_{1i} = y_i - x'_{i1}\hat{C}_1,$

$\tilde{u}_i = y_i - x'_{i1}\hat{a}_1 - x'_{i2}\hat{a}_2,$

$$\hat{V}(\hat{Y}_7) =$$

$$\left(\sum_{i,j\in s}\frac{\Delta^1_{ij}}{\pi_{ij}\pi_{aij}}\frac{\tilde{\tilde{u}}_{1i}\tilde{\tilde{u}}_{1j}}{\pi_i\pi_j}\right)+\left(\sum_{i,j\in s}\frac{\Delta^2_{ij}}{\pi_{aij}}\frac{\tilde{\tilde{u}}_i\tilde{\tilde{u}}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$\tilde{\tilde{u}}_{1i} = y_i - x'_{i1}\hat{C}_1^*,$

$\tilde{\tilde{u}}_i = y_i - x'_{i1}\hat{a}_1^* - x'_{i2}\hat{a}_2^*,$

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a}\frac{x_1x'_1}{\pi_a}\right)^{-1}\left(\sum_{s_a}\frac{x_1x'_2}{\pi_a}\right)\hat{a}_2^*.$$

On retrouve que par construction de $\hat{Y}_3$, $\hat{Y}_6$ et $\hat{Y}_7$, x_1 est utilisée pour réduire *au maximum* la variance apportée par la première phase et x_2 permet de réduire la variance apportée par la deuxième phase.

6. CHOIX D'ESTIMATEURS EN PRÉSENCE DE BIAIS DE SÉLECTION

En pratique, lorqu'on effectue le redressement d'une enquête, il est fréquent que l'on souhaite non seulement améliorer l'estimation, mais aussi, et surtout corriger les biais qu'introduisent les sélections non contrôlées d'individus, comme la non réponse.

Étudions le cas d'un sondage en deux phases, dont la deuxième phase correspond à la non réponse totale. Les poids π_i du sondage $2^{\text{ième}}$ phase sont alors inconnus. Le calage de s va permettre d'estimer ces probabilités, tout

en apportant une réduction de variance (*cf.* Deville et Dupont 1993). Cependant, asymptotiquement, les corrections de biais à apporter aux poids sont plus importantes que les modifications à apporter pour améliorer les estimateurs. C'est donc le modèle de réponse implicite qui va orienter le choix entre les différents estimateurs:

Le modèle de réponse implicite de la première classe d'estimateurs est $p_i = 1/F(x'_{i2}B_2)$

Le modèle de réponse implicite de la deuxième et la troisième classe d'estimateurs est $p_i = 1/F(x'_{i2}A_2 + x'_{i1}A_1)$.

- Quelque soit le modèle de réponse, l'évaluation des trois classes d'estimateurs sur la base du seul plan de sondage donne encore la préférence à la troisième puisqu'elle convient pour tous les modèles de réponse.
- Si on évalue les stratégies sur la base d'une modélisation par régression, on n'utilisera la première classe d'estimateurs que si le mécanisme de réponse est bien expliqué par x_2 c'est-à-dire: $p_i = 1/F(x'_{i2}B_2)$. Or, on a vu que la modélisation associée à la première classe d'estimateurs prend son sens lorsque les variables x_1 et x_2 sont très corrélées. Il est donc assez probable dans ce contexte, que la variable x_2 suffise à expliquer le mécanisme de réponse. Dans le cas contraire, il faudrait s'orienter vers la troisième classe d'estimateurs.

La comparaison entre les trois stratégies peut donc être adaptée dans un contexte où on souhaite corriger les biais introduits par les sélections non contrôlées. Les conclusions restent sensiblement les mêmes.

Il est bien sûr possible d'établir selon le même principe, des comparaisons entre stratégies de redressement alternatives dans un contexte de sondages comportant plus de deux phases, et une ou plusieurs sélections non contrôlées.

7. UTILISATION A PRIORI ET A POSTERIORI DE L'INFORMATION AUXILIAIRE

L'estimateur par calage permet d'améliorer a posteriori l'estimation, en réduisant la variance et en corrigeant le biais, comme cela vient d'être mentionné. Toutefois, on peut souhaiter intégrer l'information auxiliaire a priori, au stade du sondage plutôt qu'a posteriori au niveau de l'estimation. On retrouve alors, dans un contexte plus complexe, l'opposition classique entre stratification ou poststratification, bien connue dans le cas d'un sondage en une phase, lorsque toutes les variables auxiliaires sont qualitatives.

Il est possible de transposer les termes de l'arbitrage entre, utilisation de l'information a priori et a posteriori, dans la configuration de sondage et d'information auxiliaire étudiée, lorsque les variables auxiliaires sont qualitatives. En effet, lorsque les variables auxiliaires sont qualitatives, un calage correspond exactement à une poststratification.

Nous avons vu précédemment, que pour déterminer la procédure de redressement adéquate, il fallait distinguer deux modélisations de la variable d'intérêt possibles, selon que les informations contenues dans x_1 et x_2, sont jugées substituables ou complémentaires. Chacune de ces deux modélisations conduisait alors à une ou des procédures de redressement différentes. De la même façon, ces deux modélisations apparaissent lorsqu'on s'interroge sur la stratégie d'échantillonnage qui permet d'intégrer au mieux l'information auxiliaire:

- Lorsque l'information contenue dans x_1 et celle contenue dans x_2 sont substituables, la modélisation de la variable d'intérêt y est:

  (1) $y_i = x_{i1}b_1 + u_{i1}$ et

  (2) $y_i = x_{i2}b_2 + u_{i2}$ où le second modèle est de meilleure qualité pour prédire la valeur de y_i.

  Nous avons vu que l'utilisation de l'information auxiliaire a posteriori conduit à la stratégie de calage n° 1, soit à la première classe d'estimateurs. Si l'on souhaite tenir compte de l'information auxiliaire au niveau du sondage, il est naturel de proposer, un tirage stratifié en x_1 pour la première phase et un tirage stratifié en x_2 pour la deuxième phase.

  Toutefois, le parallèle entre procédure de redressement et procédure d'échantillonnage n'est pas complet: dans un calage, on peut n'utiliser que l'information marginale en x_1. On réalise alors une poststratification incomplète (Särndal et Deville 1992). En revanche, dans la procédure d'échantillonnage proposée comme alternative a priori, on est amené à utiliser tous les croisements des variables x_1. L'équivalent a priori d'un calage serait alors un sondage équilibré sur les marges du vecteur de variables x_1.

- Lorsque l'information contenue dans x_1 et celle contenue dans x_2 sont complémentaires, la modélisation de la variable d'intérêt y est $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$. Nous avons vu que dans ce cas l'utilisation de l'information auxiliaire a posteriori conduisait aux stratégies de calage 2, 3 et 4 aux classes d'estimateurs 2 et 3. Si l'on souhaite tenir compte de l'information auxiliaire au niveau du sondage, il est naturel de proposer, un tirage stratifié en $x1$ pour la première phase et un tirage stratifié en x_1 et x_2 pour la deuxième phase.

  De la même façon que précédemment, il n'y a pas un parallèle exact entre les procédures a priori et a posteriori puisque l'utilisation de l'information a priori mobilise tous les croisements entre les variables x_1 et x_2.

Ainsi, il est possible d'effectuer un arbitrage entre intégrer l'information a priori ou a posteriori, voire d'optimiser le plan de sondage, lorsque les variables auxiliaires sont qualitatives. Les termes de l'arbitrage sont les mêmes que dans un sondage en une phase avec un seul niveau d'information. Il s'y ajoute la multiplicité des strates crées par les croisements de x_1 et x_2 dans le cas où la modélisation utilisée est $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$ qui renforce les avantages de l'utilisation a posteriori de l'information.

Lorsque les variables auxiliaires sont quantitatives, l'arbitrage passe par leur transformation en variables qualitatives, l'extension correcte ne pouvant être réalisée qu'en utilisant le parallèle entre calage et sondage équilibré (*cf.* Deville 1992).

8. CONCLUSION

Dans un sondage en deux phases, lorsque deux informations différentes sont disponibles pour l'ensemble de la population d'une part et l'échantillon issu de la première phase d'autre part, plusieurs stratégies sont possibles lorsqu'on souhaite utiliser l'information auxiliaire pour améliorer l'estimation de totaux.

Deux approches naturelles différentes ont été utilisées pour dériver les estimateurs: une approche assistée par modèle de régression qui cherche à adapter l'idée de l'estimateur par régression, et une approche calage qui tente d'adapter l'idée du calage. Les estimateurs obtenus par les deux approches peuvent être reliés. On a fait apparaître 3 modélisations sous jacentes alternatives auxquelles peuvent se rattacher l'ensemble des estimateurs obtenus. On obtient ainsi trois classes d'estimateurs. Quelques stratégies de calage auxquelles on pouvait penser ont été éliminées d'emblée comme non pertinentes.

On a montré que les estimateurs d'une même classe, c'est-à-dire rattachés à un même modèle sont asymptotiquement équivalents et on donne la forme des variances dérivées dans le cas d'une fonction de calage linéaire, ces résultats restent valables compte tenu des équivalences asymptotiques pour une fonction de calage quelconque.

La forme des variances indique, conformément à l'intuition qu'une des classes d'estimateurs (estimateurs 3, 6 et 7 (stratégies de calage 3 et 4)) domine l'autre du point de vue de la variance, lorsque l'on se fonde uniquement sur le plan de sondage. Lorsqu'elle s'appuie sur une modélisation de la variable d'intérêt, l'évaluation des stratégies, conduit à privilégier la classe d'estimateurs associée à la modélisation adoptée.

Dans un contexte où l'on souhaite effectuer le redressement d'une enquête, et que l'on souhaite simultanément corriger les biais qu'induiraient l'utilisation des pondérations brutes et réduire la variance, les conclusions doivent être adaptées. Les modifications introduites dans les pondérations pour corriger les biais sont plus importantes que les corrections pour réduction de variance. Dès lors, on intégrera dans le calage les variables, dès que l'on pense qu'elles interviennent dans la probabilité de sélection et donc participent à la création du biais.

Lorsque les variables auxiliaires sont qualitatives, l'arbitrage entre utilisation a priori et a posteriori de l'information auxiliaire, c'est-à-dire entre, utilisation au niveau de l'échantillonnage ou au niveau du redressement, repose encore sur la distinction entre les deux modélisations de la variable d'intérêt.

Ces résultats se généralisent sans difficulté au cas d'un sondage à plus de deux phases.

REMERCIEMENTS

Je remercie chaleureusement Jean-Claude Deville, Louis Meuric et Carl-Erik Särndal pour toutes les suggestions dont ce texte a pu bénéficier.

BIBLIOGRAPHIE

CASSEL, C.M., SÄRNDAL, C.-E., et WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrica*, 63, 615-620.

DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, October 1992, Örebro.

DEVILLE, J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DEVILLE, J.-C., et DUPONT, F. (1993). Calage et redressement de la non-réponse totale. Journées de Méthodologie.

DUPONT, F. (1994). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. Document de travail de la Direction des Statistiques Démographiques et Sociales, F9409.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā* C, 37, 117-132.

GOURIEROUX, C. (1981). *Théorie des sondages*. Edition Economica Paris.

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

SÄRNDAL, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phases sampling and nonresponse. *International Statistical Review*, 55, 279-294.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.