LA CORRECTION DE LA NON-RÉPONSE PAR CALAGE GÉNÉRALISÉ

Jean-Claude DEVILLE

Ecole Nationale de la Statistique et de l'Analyse de l'Information
Laboratoire de Statistique d'Enquête, Campus de Ker-Lan

1. Quelques idées générales sur la correction pour non-réponse

La compensation de la non-réponse dans les enquêtes s'appuie sur deux techniques différentes : la repondération destinée essentiellement à corriger la non-réponse globale, et l'imputation destinée à corriger la non-réponse partielle.

Nous ne nous intéresseront ici qu'aux procédures de repondération.

Elles sont basées sur une modélisation du mécanisme de réponse. Celui-ci consiste en la sélection d'un échantillon r de répondants selon un "pseudo" plan de sondage $q(r|s)$, inconnu. En paramètrant ce pseudo plan à l'aide d'un paramètre β de dimension p on cherchera à estimer β et tout se passera comme si le paramètre estimé était le bon paramètre. C'est pour cela qu'on parle de méthode de pseudo-randomisation.

Parallèlement, la théorie de l'estimation en sondage a connu depuis dix ans des développements intéressants qui synthétisent et généralisent les types d'estimateurs qu'on utilisait traditionnellement (*Deville, Särndal (1992), Deville, Särndal, Sautory (1993)*). Il existe une analogie formelle (en fait assez profonde) entre ces estimateurs et les modèles linéaires généralisés qui permettent une description intéressante du mécanisme de réponse. Ceci conduit à une synthèse entre ces approches déjà ébauchée dans (*Deville, Dupont (1993)*) et qui trouvera ici une extension nouvelle basée sur la théorie généralisée du calage. On arrive par cette voie à de nouvelles procédures de repondération qui s'avèrent à la fois commodes et efficaces. En particulier on met en évidence des gains de variance dus à la procédure même d'estimation. On montre aussi qu'on peut utiliser dans les modèles de réponse des variables qui ne sont pas observées chez les non-répondants (le statut social, par exemple).

2. Théorie généralisée du calage : quelques idées utiles

2.1 Que savons-nous ?

Rappelons les lignes générales de la théorie classique du calage. A partir d'un estimateur sans biais, ou, du moins, "convergent", en général celui de Horvitz-Thompson, $\hat{Y} = \sum_s d_k y_k$ du total Y des y_k on cherche de nouveaux poids w_k proches des $d_k = 1/p_k$ (inverses des probabilités d'inclusion) au sens d'une certaine distance et vérifiant les équations de calage : $X = \sum_s w_k x_k$ où X est le p-vecteur de totaux de p variables réelles rangées dans le vecteur x_k. On trouve alors que les poids sont de la forme $w_k = d_k F\left(q_k x_k' l\right)$ où les q_k sont une famille de nombres positifs destinés à prendre en compte une certaine hétéroscédasticité, l un p-vecteur et F une fonction régulière de $\mathbf{R}$ dans $\mathbf{R}$ vérifiant $F(0) = 1$ et $F'(0) = 1$. C'est donc la solution des équations :

$$X = \sum_s d_k \, x_k \, F\left(q_k x_k' \, l\right) \qquad \text{(2-1)}$$

L'intérêt de cette procédure d'estimation est d'éliminer, dans la variance de l'estimateur, l'influence des variables x_k. De façon précise, on a le résultat suivant :

$$Var\left(\sum_s w_k y_k\right) = Var_{HT}\left(\sum_s w_k e_k\right) \qquad \text{(2-2)}$$

où les e_k sont les résidus de la régression $y_k = x_k' b + e_k$ obtenus par les moindres carrés pondérés par les nombre q_k.

2.2 Ebauche d'une théorie généralisée.

Les équations (2-1) sont obtenues, habituellement, à partir de considérations de distance entre les poids d'origine et les poids calés. Cependant, la pratique a montré que seule la fonction F et les constantes q_k ont une réelle importance sur les résultats tant théoriques que pratiques. On observe alors que les variables x_k interviennent à deux endroits dans (2-1) d'une façon qui n'est légitimée que par un argument technique. L'idée de base d'une généralisation du calage consiste à partir de la fonction F et de dissymétriser les deux apparitions de x_k pour des raisons qui vont apparaître bientôt.

On désire toujours caler nos poids de façon à retrouver à retrouver le p vecteur de totaux connus X. Pour y parvenir il est clair que nous aurons besoin (au minimum) de p paramètres d'ajustement que nous rangerons dans le vecteur l.

Maintenant, à chaque k de la population, on associe donc une *fonction de calage* $F_k : \mathrm{R}^p \to \mathbf{R}$ vérifiant $F_k(0) = 1$, régulière. On note $z_k = grad F_k(0) \in \mathbf{R}^p$ le vecteur des dérivées partielles de F_k en 0. On s'autorise des modifications de poids de la forme :

$$w_k = d_k \, F_k(l)$$

(2-3)

où l sera un p-paramètre unique indépendant de k .

De façon générale, on aura donc $w_k = d_k\,(1+z_k^{'}\,l+O(\|l\|^2))$; le cas le plus simple (dit linéaire) est obtenu quand $F_k(u)=1+z_k^{'}\,l$, z_k étant une variable à *p* composantes connue sur tout l'échantillon. Le cas habituel dans les applications (et qui figure dans CALMAR II) est le cas linéaire généralisé où on se donne une unique fonction F , monotone, régulière (en fait dérivable suffit) et vérifiant *F(0)= 1*. Les poids calés valent alors $w_k = d_k\;F(z_k^{'}\,l)$.

Il reste à poser et à résoudre les équations de calage :

$$X=\sum_s d_k\,x_k\,F_k(l).$$

Dans le cas linéaire, cela donne :

$$X=\sum_s d_k\,x_k(1+z_k^{'}\,l\,),$$

et dans le cas linéaire-généralisé :

$$X=\sum_s d_k\,x_k\;F(\,z_k^{'}\,l\,).$$

Comme dans le cas standard étudié dans *Deville-Särndal(1992)*, on obtient :

$$l=\left(T_{szx}^{'}\right)^{-1}\left(X-\hat{X}\right)+O\left(\left\|X-\hat{X}\right\|^2\right)$$

avec $T_{szx}=\sum_s d_k z_k\,x_k^{'}$, matrice supposée de plein rang. Géométriquement, ceci signifie la chose suivante.

Notons L_X le sous espace de R^N engendré par les *p* variables coordonnées des x_k et L_z le sous espace de R^N engendré par les *p* variables coordonnées des z_k . La condition s'écrit géométriquement $L_X \cap L_z^{\perp} = 0$.

Sous les mêmes hypothèses techniques que dans *Deville-Särndal (1992)* on obtient des résultats analogues, à savoir :

- l'estimateur calé généralisé est convergent et de biais négligeable
- La valeur des z_k étant fixée, tous les estimateurs de la famille ainsi obtenue ont asymptotiquement la même variance.

Celle-ci peut donc être étudiée à partir du cas linéaire auquel nous allons prêter une attention particulière (bien que, pratiquement, comme nous allons le voir, ce ne soit que rarement le plus naturel).

2.3 Le cas linéaire et la variance de l'estimateur calé généralisé.

Le cas particulier le plus simple est le cas linéaire où on prend simplement $F_k(l)=1+z_k^{'}\,l$ avec un vecteur de variables "instrumentales" (la raison du choix de ce terme apparaîtra dans la suite) z_k .

On a, dans ce cas :

$$\begin{aligned}\hat{Y}_C &= \hat{Y}+\left(X-\hat{X}\right)^{'}T_{szx}^{'-1}\sum_s d_k z_k\;y_k\\ &= \hat{Y}+\left(X-\hat{X}\right)^{'}\tilde{b},\end{aligned}$$

où $\tilde{b}$ est solution de : $\sum_s d_k z_k \quad \left(y_k - x_k^{'}b\right)=0$

On constate les faits suivants :

- $\tilde{b}$ est le vecteur des coefficients de la régression instrumentale (*Fuller (1987)* par exemple)) utilisant les z_k comme instruments.
- Géométriquement, celle-ci s'interprète comme la projection, dans R^N, du vecteur des y_k sur L_X le long de $L_Z^{\perp}$.
- Les poids de régression peuvent s'obtenir par minimisation d'une distance aux anciens poids. En effet, il suffit d'utiliser une distance déduite d'un produit scalaire pour lequel L_X et $L_Z^{\perp}$ sont orthogonaux. Si on appelle P la matrice donnant la projection sur L_X le long de $L_Z^{\perp}$, il suffit de prendre pour matrice métrique *P'P + (1-P)'(1-P).*
- La variance de l'estimateur se calcule en utilisant la technique des résidus. La différence avec le cas standard est qu'il faut utiliser les résidus de la régression instrumentale.
- L'estimateur de variance utilise le même principe (en utilisant les résidus empiriques dans un logiciel du genre POULPE).
- Les "instruments" z_k n'ont besoin d'être connus que sur l'échantillon : *ils ne constituent pas une information auxiliaire.*

2.4 Exemples

2.4.1 Estimateur par ratio.

X et x_k sont unidimensionnels. La variable instrumentale est la variable "gratuite" z_k=1. L'équation de calage s'écrit : $X = \sum_s d_k x_k (1 + z_k l)$ d'où $\tilde{b} = \frac{\hat{Y}}{\hat{X}} = \hat{R}$ et les résidus valent $y_k - \hat{R}x_k$.

2.4.2 Estimateur par régression pondérée.

Les instruments sont : $z_k = q_k x_k$.

2.4.3 Estimateur par régression optimal (Montanari (1987)).

Il s'agit de trouver le meilleur estimateur sans biais de la forme $\hat{Y} + B'(X - \hat{X})$ où X est le total d'un vecteur de variables auxiliaires connu, ces variables étant mesurées sur l'échantillon.

Un calcul immédiat donne $B = Var(\hat{X})^{-1} Cov(\hat{X}, \hat{Y})$ où $Var(\hat{X}) = \sum_U \sum_U \Delta_{kl} x_k x_l'$ et $Cov(\hat{X}, \hat{Y}) = \sum_U \sum_U \Delta_{kl} x_k y_l$ avec $\Delta_{kl} = \frac{p_{kl}}{p_k p_l} - 1$ (en utilisant les probabilités d'inclusion d'ordre deux). Pour utiliser cet estimateur, il faut, malheureusement, estimer les variances, ce qui conduit à $\hat{Y} = \sum_s d_k y_k + (X - \hat{X})' (\sum_s d_k z_k x_k')^{-1} (\sum_s d_k z_k y_k)$. On reconnaît donc l'estimateur par calage généralisé où on a donc utilisé les instruments : $z_k = \sum_{l \in U} \Delta_{kl} x_l$. Si les x_k sont connus sur la base de

sondage, les z_k sont utilisables sur s par une simple recodification. Sinon, il doivent eux même être estimés, ce qui est hasardeux.
Dans le cas d'un plan stratifié avec sondage aléatoire simple dans chaque strate, c'est malgré tout assez simple car on a : $z_k = (x_k - \bar{X}_h)\frac{N_h^2}{n_h}\left(1-\frac{n_h-1}{N_h-1}\right)$, avec des notations habituelles. Si les x_k ne sont disponibles que sur l'échantillon, seules les moyennes de strate doivent être estimées, ce qui limite la casse.

2.4.4 Un exemple non-linéaire.

Soit u_k une variable positive connue dans s, $l' = (a,b,c)$ et $F_k(l) = a + exp(bu_k)u_k^c$.
On a bien $F_k(O) = 1$ et on trouve :

$$\frac{\partial F_k}{\partial a} = 1$$
$$\frac{\partial F_k}{\partial b} = u_k \exp(bu_k)u_k^c \Rightarrow \frac{\partial F_k}{\partial b}(0) = u_k$$
$$\frac{\partial F_k}{\partial c} = Logu_k \exp(bu_k)u_k^c \Rightarrow \frac{\partial F_k}{\partial c}(0) = Logu_k$$

Cet exemple montre bien la nature des variables z qui peuvent être déduites d'une même variable à partir de transformations non linéaires (ici le logarithme).

3. Non-réponse pour une enquête exhaustive : repondération

3.1 Modèle de réponse et estimation.

Le mécanisme de réponse est modélisé par un plan de sondage $q(r\,;b)$ où b est un paramètre inconnu de $\mathbb{R}^P$. Ce modèle nous fournit des poids d'extrapolation "à la Horvitz-Thompson", $p_k^{-1} = F_k(b)$, ainsi que des probabilités d'inclusion à l'ordre deux, si nous en avons besoin pour exprimer la variance et l'estimer.

Le modèle le plus simple et, à bien des égards, le plus naturel est le modèle de Poisson :

$$q(r\,;b) = \prod_{k\in r} F_k^{-1}(b) \prod_{k\in U-r=o} (1-F_k^{-1}(b))$$

On peut aussi introduire des plans plus compliqués, avec par exemple des effets de grappe si on soupçonne qu'il y a des effets dus aux enquêteurs, par exemple.

La question maintenant est de savoir comment estimer b.

La réponse, assez étonnante, est : peu importe ! Maximum de vraisemblance, méthodes des moments, Chi 2 - minimum, ce sera comme on voudra (ou presque !). Tout ce qui compte, c'est le fait qu'on arrive à un ensemble de p équations estimantes, qu'on résoudra *quel que soit l'échantillon r possible*

; de ce fait, ces équations vérifiées pour tout r constituent des statistiques dont la variance est nulle .

Exemple : Supposons qu'on estime le b du modèle de Poisson ci-dessus par la méthode du maximum de vraisemblance. On obtient les équations estimantes suivantes :

$$\sum_r F_k(b)z_k^* = \sum_U z_k^* \quad \text{avec} \quad z_k^* = \frac{gradF_k(b)}{F_k(b)(F_k(b)-1)}$$

Si le modèle est spécifié plus précisément sous la forme d'un modèle linéaire généralisé $F_k(b) = F(z_k' b)$, alors $z_k^* = z_k \frac{\dot{F}}{F(F-1)}$. Si *F=1 - exp* (modèle log-linéaire), on a simplement $z_k^* = z_k$.

De façon générale, il vaudra sans doute mieux se fier à un principe de calage et utiliser les équations estimantes sans biais suivantes :

$$\sum_r F_k(\hat{b})x_k = \sum_U x_k$$

ou avec un *GLM* : (*)

$$\sum_r F(z_k' \hat{b})\, x_k = \sum_U x_k$$

Il est immédiat que la solution de ces équations a une variance en *1/n*, *n* étant la taille de *r*. Mais on peut aussi en donner l'interprétation suivante :

$$X = \sum_r x_k F_k(ß)\, G_k(?)$$

où b est la vraie valeur du paramètre et où on a posé :

$$G_k(?) = \frac{F_k(b+l)}{F_k(b)} \qquad (et\ on \quad a \quad bien\ G_k(0) = 1)$$

Ces équations ne sont autres que des équations de calage qui appellent les commentaires suivants :

- $\hat{b} = b + l$ est un estimateur de b .
- Si, par hasard, on dispose d'une autre estimateur $\hat{b}_0$ de b, on peut écrire $\hat{b} = b + (\hat{b}_0 - b) + l$, et l'interprétation est la même car l et l'erreur d'estimation $\hat{b}_0 - b$ sont de l'ordre de *1/n*.
- On n'a besoin de connaître les F_k (ou les z_k en pratique) que pour les répondants.
- L'effet sur la variance (et l'estimation de variance) est le même que celui obtenu dans le calage habituel.

3.2 Regardons plus en détail.

D'après ce qu'on sait de la théorie généralisée du calage, la variance est celle de l'"estimateur linéaire" de la même famille, à savoir :

$$\hat{Y} = \sum_r y_k F_k(b_0)(1 + z_k' l) \quad \text{avec} \quad l \quad \text{vérifiant}$$

$$X = \sum_r x_k F_k(b_0)(1 + z_k' l)$$

$$\text{et } z_k = grad G_k(0) = grad Log F_k(b_0).$$

Si on utilise un modèle linéaire généralisé, on a $F_k(b) = F(z_k^{*\prime} b)$ et

$$z_k = z_k^* \frac{\dot{F}(z_k^{*\prime} b_0)}{F(z_k^{*\prime} b_0)} = z_k^* \ (Log F(z_k^{*\prime} b_0)\dot{}) = z_k^* \ q_k .$$

On a alors :

$$Var(\hat{Y}) = Var(\sum_r \tilde{e}_k F_k(b_0))$$

avec : $\tilde{e}_k = y_k - \tilde{B}' x_k$, où $\tilde{B}$ est solution des équations normales de la régression instrumentale :

$$\sum_r z_k F_k(ß_0)\,(y_k - \tilde{B}' x_k) = 0.$$

Dans le cas d'un GLM, cette équation devient : $\sum_r q_k z_k^* \ F_k(ß_0)\,(y_k - \tilde{B}' x_k) = 0$.

Si $z_k^* = x_k$ alors on a : $\sum_r q_k x_k F_k(ß_0)\,(y_k - \tilde{B}' x_k) = 0$, et l'interprétation est alors soit celle des variables instrumentales, soit celle des moindres carrés pondérés.

3.3 Des exemples.

3.3.1 Redressement par ratio.

x_k est une variable positive, et $z_k = 1$. Autrement dit, les réponses manquent au hasard mais on cale sur le total X des x_k . Alors : $\hat{Y} = X \frac{Y_r}{X_r}$ où Y_r et X_r sont les totaux sur les répondants (il s'agit, en somme, de la théorie de la règle de trois). Avec le modèle de réponse de Poisson on obtient la variance estimée suivante :

$$\frac{X}{X_r}\left(\frac{X}{X_r} - 1\right) \sum_r (y_k - R x_k)^2 .$$

Si $x_k = 1$, on obtient :

$$\frac{N}{n}\left(\frac{N}{n} - 1\right) \sum_r (y_k - \bar{y})^2 = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) \frac{\sum_r (y_k - \bar{y})^2}{n} .$$

ce qui est assez naturel.

3.3.2 Poststratification et raking-ratio formel.

Tout va de soi.

Dans le cas de la poststratification formelle, $x_k = z_k$ est le vecteur des indicatrices d'appartenance aux H modalités h d'une variable qualitative, désignant les 'poststrates', confondues avec les cellules d'un modèle de réponse homogène par cellules dans la terminologie de *Särndal, Swensson, Wretman (1992)*. Autrement dit, x_k est le vecteur de dimension H dont les cordonnées sont nulles sauf la $h^{\text{ième}}$ qui vaut *1* si l'unité *k* appartient à la 'poststrate' d'indice h.

Dans le cas du raking-ratio formel utilisé pour redresser la non-réponse, on a encore $x_k = z_k$, mais ici x_k désigne l'empilement des *d* variables qualitatives, 'vectorisées' comme ci-dessus, qu'on utilise pour le redressement. Le modèle de réponse postule que la probabilité de réponse est le produit de *d* facteurs liés à l'appartenance de *k* à une cellule d- dimensionnelle du produit cartésien des *d* variables. Autrement dit, la probabilité de réponse de l'unité *k* est postulée être de la forme *exp(-x_k' β)*, les composantes de β étant positives.

3.3.3 Un exemple pratique plus original.

Le vecteur x_k est encore une fois celui des indicatrices d'une variable qualitative à *I* modalités *i = 1 à I*. L'effectif N_i est supposé connu dans *U*. z_k est le vecteur des indicatrices d'une autre variable qualitative de même dimension indicé par *a*. On peut imaginer, par exemple, que x_k est une situation connue dans la base de sondage (la catégorie socioprofessionnelle de l'occupant d'un logement au moment du recensement), tandis que z_k est la situation mesurée lors de l'enquête (la catégorie socioprofessionnelle de l'occupant du même logement au moment de l'enquête – supposée exhaustive, rappelons-le, mais entachée de non réponse dépendant de la catégorie socioprofessionnelle de *l'occupant*). On note R_{ia} l'effectif des répondants classés à la modalité *i* de x_k et *a* de z_k (donc, des répondants de la CSP *a* occupant un logement 'de la CSP *i*' au moment du recensement). La probabilité de réponse est modélisée en prenant pour cellules homogènes de réponse les modalités *a* (CSP de ***l'occupant***, donc, ***inconnue*** pour les non-répondants). Pour se ramener au modèle formel linéaire, on posera donc *Pr(k répond)= P_a = 1/(1+b_a)* si *k* est classé dans la catégorie *a* (appartient à la CSP *a*).

Le vecteur $b = (...,b_a,...)$ des paramètres du modèle de réponse est estimé par les équations 'de type calage' suivantes :

$$N_i = \sum_a R_{ia}\left(1+\hat{b}_a\right) \ .$$

Sous l'hypothèse (3-1) d'un mécanisme de réponse Poissonnien, la variance de l'estimateur calé vaut, avec l'approximation d'usage :

$$\sum_a P_a(1-P_a)\sum_i \sum_{k\in U_{ai}} \left(y_k - \overline{Y}_i\right)^2$$

et elle est estimée par :

$$\sum_a \hat{b}_a\left(1+\hat{b}_a\right)\sum_i \sum_{k\in r_{ia}} \left(y_k - \hat{\overline{Y}}_i\right)^2$$

où $\hat{\overline{Y}}_i$ est l'estimation de la moyenne des y_k pour *k* classé dans la modalité *i* de x_k, soit :

$$\hat{\bar{Y}}_i = \frac{1}{N_i} \sum_a (1 + \hat{b}_a) \sum_{r_{ia}} y_k$$

4. Non-réponse après échantillonnage

4.1 Position du problème.

Un plan de sondage p sur la population U fournit un échantillon s auquel sont associés les poids d'extrapolation d_k. Conditionnellement à s, la non-réponse est régie par un plan de sondage $q(r|s;b)$ fournissant l'échantillon r. Les poids d'extrapolation de r vers s sont de la forme $F_k(b_0)$ conformément à ce qui se passait dans le paragraphe 3, de sorte que les poids d'extrapolation généraux, à partir des répondants sont $d_k\ F_k(b_0)$. Comme cela est bien connu *(Särndal, Wretman (1987)*), nous avons affaire à un sondage en deux phases dont la variance comporte deux termes qui s'estiment séparément. Dans la suite nous supposerons que nous disposons d'un logiciel (comme POULPE, *Caron, (1998*), *Petit (1998))* qui calcule de façon automatique ces deux formes quadratiques : $\hat{Var}(\hat{Y}) = Q_1(y_r) + Q_2(y_r)$, où $y_r = \{y_k ; k \in r\}$.

La question est donc encore ici celle de l'estimation du vecteur de paramètres β du modèle de réponse. Comme ci-dessus, il ne s'agit pas de chercher le meilleur estimateur au sens de la statistique universitaire officielle classique (et de ses arbitraires fonctions de perte !), mais celui qui conduit aux poids d'extrapolation $d_k\ F_k(\hat{b})$ ayant les meilleures propriétés. De ce fait , le recours à des idées de calage prend toute son importance, si on songe à diminuer la variance ; symétriquement, si on veut limiter le biais de réponse, on se doit d'utiliser un bon modèle de réponse, utilisant en particulier les variables dont *ON SAIT* qu'elles influencent la probabilité de réponse, même si *on sait aussi* que ces variables NE SONT (en général) PAS MESURÉES PARMI LES NON-RÉPONDANTS.

Remarque non dénuée d'intérêt : si on utilise un modèle où les inverses de probabilités de réponse sont données par une forme linéaire généralisée $F_k(b) = F(z_k^{'}\ b)$, le total du vecteur z_k des variables explicatives de la réponse peut parfaitement être une variable d'intérêt, et même dans certains cas, la plus importante d'entre elles. Ainsi, le calage généralisé fournit une réponse satisfaisante à une large classe de problèmes où la non-réponse est, comme on dit, 'non ignorable', c'est-à-dire, en clair, dans des cas où la variable d'intérêt est elle-même un facteur explicatif important de la non-réponse.

4.2 Utilisation d'une information auxiliaire connue au niveau de la population.

Dans la suite, nous allons nous placer dans le cas où l'information auxiliaire 'de calage' est un vecteur X de totaux connus, les variables auxiliaires x_k étant également disponibles chez les répondants. On utilisera en général, comme modèle pour les inverses de probabilités de réponse, une forme linéaire généralisée $F_k(b) = F(z_k^{'}\ b)$, où z_k est un vecteur de variables explicatives de la réponse de même dimension que x_k , et F une fonction numérique régulière prenant ses valeurs dans l'intervalle $[1, \infty[$.

Dans la logique de ce qui a été dit au paragraphe 3, on est conduit à estimer ● par les équations estimantes de calage suivantes :

$$X = \sum_r x_k d_k F_k(\hat{b}) = \sum_r x_k d_k F(z_k' \hat{b})$$

(4-1)

L'astuce mathématique du §3 fonctionne encore, à savoir qu'on peut réécrire ces équations sous la forme d'authentiques équations de calage (généralisé), avec possibilité d'en exploiter toutes les conséquences :

$$X = \sum_r x_k d_k F_k(ß)\, G_k(?)$$

(4-2)

où b est la vraie valeur du paramètre et où on a posé :

$$G_k(?) = \frac{F_k(b + l)}{F_k(b)} \qquad (4\text{-}2 \text{ bis}).$$

On a bien $G_k(0)=0$ et les fonctions G_k sont bien des fonctions de calage au sens de la théorie du §2. Les conclusions sont donc les mêmes dans ce contexte qu'au §3 :

- $\hat{b} = b + l$ est un estimateur de b .
- Si, par hasard, on dispose d'une autre estimateur $\hat{b}_0$ de b, on peut écrire $\hat{b} = b + \left(\hat{b}_0 - b\right) + l$, et l'interprétation est la même car l et l'erreur d'estimation $\hat{b}_0 - b$ sont de l'ordre de $1/n$.
- On n'a besoin de connaître les F_k (ou les z_k en pratique) que pour les répondants.
- L'effet sur la variance (et l'estimation de variance) est le même que celui obtenu dans le calage habituel.

En particulier, donc, si notre logiciel de variance calcule pour un estimateur à deux phases la quantité $V\hat{a}r\left(\hat{Y}\right) = Q_1(y_r) + Q_2(y_r)$, la variance de l'estimateur obtenu a partir de ce qui vient d'être dit se calcule de la même façon à condition de remplacer les y par les résidus de la régression de y sur x utilisant les instruments z (*calculés comme indiqué au § 3-2*), et où on utilisera les probabilités d'inclusion estimées pour la partie correspondante de l'estimateur de variance.

Ces résidus seront donc calculés de la façon suivante : les équations qui définissent le résidus dans la population sont les équations normales de la régression dans U , soit:

$$\sum_U z_k (y_k - x_k' B) = 0.$$

L'estimation à partir de r est donnée, théoriquement, par la solution des équations normales estimées (et qui, de ce fait, deviennent des équations estimantes). On utilise donc les inverses des probabilités de réponse (vraie, en tout cas si on fait confiance au modèle de réponse) dans les pondérations, ce qui conduit aux équations :

$$\sum_r d_k F_k(b)\, z_k (\, y_k - x_k' \hat{B}_0) = 0 \ .$$

Ici, donc, $\hat{B}_0$ est l'estimateur de B qu'on aurait si les vraies probabilités de réponse étaient connues. Comme elle sont seulement estimées, il reste donc à calculer les résidus utiles pour l'estimation de variance (et accessoirement, bien que cela n'aie aucun intérêt en soi, l'estimation des coefficients de régression) par :

$$\sum_r d_k F_k(\hat{b})\ z_k(\underbrace{y_k - x_k'\hat{B}}_{résidus}) = 0$$

On remarquera que l'estimation se fait en utilisant les poids 'calés', ce qui n'est pas dans la coutume.

4.3 Cas particuliers : la poststratification formelle et raking-ratio formel

Ce sont des cas fréquents. Les totaux X connus de façon externe sont les effectifs N_h de poststrates (des tranches d'ages, par exemple) ; cependant, on ignore l'âge des non-répondants, ce qui empêche toute estimation traditionnelle des probabilités de réponse par tranches d'âge. La poststratification formelle utilise l'estimateur aux formes équivalentes suivantes :

$$\hat{Y}_{postform} = \sum_h N_h \frac{\sum_{rh} d_k y_k}{\sum_{rh} d_k} = \sum_h \frac{N_h}{\hat{N}_h} \frac{\hat{N}_h}{\sum_{rh} d_k} \sum_{rh} d_k y_k \ .$$

Là-dedans, les $\hat{N}_h$ désignent les estimations des N_h qu'on aurait obtenues si on avait eu un moyen quelconque de le faire de façon traditionnelle (et en utilisant, d'ailleurs, n'importe quelle méthode d'estimation !).

Pour le cas du raking-ratio formel, examinons d'abord le cas assez général où on utilise un modèle linéaire généralisé pour décrire la probabilité de réponse. Comme au § 3-2, on est amené a résoudre des équations de calage, puis à trouver les instruments de la régression qui permet de calculer les résidus intervenant dans l'estimation de variance. Si alors on a pris $F_k(b) = F(z_k^{*'} b)$, les instruments ne sont autres que $z_k = z_k^* \frac{\dot{F}(z_k^*)}{F(z_k^*)} = z_k^* q_k$. S'il se trouve que $F_k(b) = F(x_k' b)$, c'est-à-dire que les variables du modèle de réponse sont les mêmes que celles sur lesquelles on cale, on retrouve l'estimateur calé "standard" avec des poids de régression q_k. Si, de plus, $F=exp$, la fonction exponentielle, alors on a tout bonnement $q_k = 1$. C'est ce qui se passe pour la raking-ratio formel, avec les variables déjà décrites au 3-3-2. Autrement dit, y compris dans le cas où elle est utilisée dans le cadre de la correction de la non-réponse, la méthode du raking-ratio est associée à un calcul de variance qui fait intervenir une régression des moindres carrés ordinaires.

5. Conclusions et extensions

Le calage généralisé est une extension de la méthode bien connue de calage mettant en jeu deux ensembles de variables de même dimension x et z. Son utilité est de construire des estimateurs dont la variance est celle de l'estimateur de Horvitz-Thompson appliquée aux résidus de la régression de la variable d'intérêt sur les x en utilisant les z comme variables instrumentales. Les z apparaissent ainsi comme une généralisation des poids de régression utilisés dans la méthode d'origine. La méthode est donc efficace surtout quand les deux ensembles de variables sont bien corrélés, et, de ce fait, peut sembler être d'une utilité douteuse puisqu'on peut toujours prendre $z=x$. Il y a cependant des cas où, à cause d'autres éléments du problème statistique, elle s'impose de façon naturelle. Il y a d'abord le cas d'un calage classique où les x sont des mesures entachées d'erreurs (sans biais) des quantités dont les totaux sont connus sans erreur. C'est, en fait, une transposition de ce qu'on fait pour estimer des modèles linéaires avec erreurs sur les variables (*Fuller(1987)*). Elle s'avère, par ailleurs, indispensable dans les questions d'estimation par la méthode de partage des poids avec information auxiliaire (voir *Lavallée(2002)*). Elle est aussi très efficace quand on désire corriger les biais dus à la non-réponse en utilisant des méthodes de calage, ce qui était l'objet de cette étude.

La façon ‘traditionnelle’ de mettre en œuvre le calage (classique) en situation de non-réponse consiste à utiliser comme variables de calage à la fois des variables bien corrélées avec la variable d’intérêt et des variables bien corrélées avec la propension à répondre. Cette méthode procure bien les effets désirés mais présente néanmoins deux inconvénients. Le premier est un certain manque de parcimonie dans le nombre de paramètres d’ajustement (c’est-à-dire la dimension du problème de calage), ce qui peut devenir gênant quand on possède beaucoup d’information auxiliaire. Le second est plus définitif : on doit connaître les totaux de toutes les variables auxiliaires, même ceux des plus discriminantes en matière de non-réponse. Or, très généralement, les totaux de ces variables discriminantes de la non-réponse sont, presque par nature, inconnus.

Le recours au calage généralisé évite en grande partie ces deux écueils. D’abord, la taille du problème est réduite à celle des vecteurs de calage x. Ces variables ont pour utilité, comme dans la méthode classique, de réduire la variance de l’estimation. Les variables z, quant à elles, ont pour fonction de réduire les biais induits par la non-réponse. Or ces dernières ne constituent pas une information auxiliaire externe, et on a besoin de les collecter uniquement sur l’échantillon de répondants. Elle peuvent donc être elles-mêmes des variables d’intérêt, ce qui est particulièrement précieux. En effet, l’objet de l’enquête -la collecte de certaines variables- peut être une cause de non-réponse ; on devra donc les introduire parmi les variables z du calage.

Une remarque importante s’impose ici. Quand on utilise le calage (généralisé ou pas) en situation de non-réponse, les facteurs de calage (ou g-poids) s’interprètent comme des estimations des (inverses de) probabilités de réponse. On s’en rend compte facilement si on réalise que la correction pour biais de réponse est d’ordre fini (garde le même ordre de grandeur fini pour toute taille d’échantillon), tandis que le facteur de calage associé à un estimateur sans biais est un ‘infiniment petit’ (il est de l’ordre de grandeur de l’inverse de la taille des échantillons).

Cette étude a été volontairement limitée aux aspects essentiels de la théorie du calage généralisé. D’assez nombreuses extensions sont possibles. Il est, par exemple, presque immédiat de l’étendre au cas où les x et les z ne comportent pas le même nombre de variables. Dans un cas, on sera amené à caler sur une transformation linéaire des x choisie pour assurer une bonne corrélation avec les z , dans l’autre, on ne saura estimer les paramètres du modèle de réponse qu’à une transformation linéaire près, sans que cela aie d’influence sur l’estimation des probabilités de réponse. On peut aussi s’intéresser au cas où de l’information est disponible aux trois niveaux possibles : la population totale U, l’échantillon initial s ou l’échantillon de répondants r. Cette analyse peut se trouver, avec quelques autres babioles, dans le texte particulièrement ‘zippé’ de *Deville (1998)*.

Bibliographie

[1] CARON N., "Le logiciel POULPE : aspects méthodologiques", INSEE-Méthodes N°84-85-86 : Actes des Journées de Méthodologie Statistique de 1998, (1999).

[2] DEVILLE, J.C., "La correction de la non-réponse par calage généralisé ou par échantillonnage équilibré", Actes de la Société de Statistique du Canada, Université de Sherbrooke,(1998).

[3] DEVILLE, J.C, et SÄRNDAL, C.E, "Calibration estimators in survey sampling", Journal of the American Statistical Association, Vol 87, p 376-382 (1992).

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

[5] DEVILLE, J.C, et DUPONT, F. "Non-réponse : Principes et Méthodes", INSEE-Méthodes N°56-57-58 : Actes des Journées de Méthodologie Statistique de 1993, (1996).

[6] DUPONT, F. "Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire" Techniques d'enquête, vol. 21, n° 2, pp.141-150 (1995).

[7] FULLER,W.A., Measurement Error Models, Wiley (1987).

[8] LAVALLE, P. "Le Sondage Indirect ou la Méthode Généralisée de Partage des Poids", Ellipses (2002).

[9] MONTANARI, G.E. "Post sampling efficient prediction in large scale surveys", International statistical review, Vol 55, pp191-202 (1987).

[10] OH,H,L et SCHEUREN, F,J "Weighting Adjustments for Unit Nonresponse", Incomplete Data in Sample Surveys, Vol 2, pp143-184 (1983).

[11] PETIT, J.N, : Le logiciel POULPE : modélisation informatique, INSEE-Méthodes N°84-85-86 : Actes des Journées de Méthodologie Statistique de 1998, (1999).

[12] SÄRNDAL,C.E. et SWENSSON,B. "A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse", International statistical review, Vol 55, pp 279-294 (1987).

[13] SÄRNDAL,C-E, SWENSSON,B et WRETMAN,J "Model Assisted Survey Sampling", Springer (1992).