Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimation de la variance par linéarisation pour les estimateurs par calage généralisé en présence de non-réponse

par Julia D'Arrigo et Chris Skinner

Décembre 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2010
Vol. 36, N° 2, pp. 197-209
Statistique Canada, N° 12-001-X au catalogue

Estimation de la variance par linéarisation pour les estimateurs par calage généralisé en présence de non-réponse

Julia D'Arrigo et Chris Skinner [1]

Résumé

Diverses formes d'estimateurs de variance par linéarisation pour les estimateurs par calage généralisé sont définies en choisissant différents poids à appliquer a) aux résidus et b) aux coefficients de régression estimés utilisés dans le calcul des résidus. Des éléments de théorie sont présentés pour trois formes de l'estimateur par calage généralisé, à savoir l'estimateur par ratissage croisé classique, l'estimateur par calage basé sur le « maximum de vraisemblance » et l'estimateur par la régression généralisée, ainsi que pour les estimateurs de variance par linéarisation connexes. Une étude par simulation est effectuée en se servant des données d'une enquête sur la population active et d'une enquête sur les revenus et dépenses. Les propriétés des estimateurs sont évaluées en fonction de l'échantillonnage ainsi que de la non-réponse. L'étude révèle peu de différences entre les propriétés des divers estimateurs par calage pour un plan d'échantillonnage et un modèle de non-réponse donnés. En ce qui concerne les estimateurs de variance, l'approche consistant à pondérer les résidus par les poids de sondage peut être fortement biaisée en présence de non-réponse. L'approche de pondération des résidus par les poids calés a tendance à produire un biais nettement plus faible. Le choix de différents types de poids pour produire les coefficients de régression a peu d'incidence.

Mots clés : Calage ; non-réponse ; ratissage ; estimation de la variance ; poids.

1. Introduction

Dans les sondages, le recours à la pondération pour corriger le biais de non-réponse est une approche très répandue. L'estimation par calage généralisé (Deville, Särndal et Sautory 1993) fournit une classe de méthodes de pondération qui peuvent être utilisées quand les totaux de population des variables auxiliaires sont disponibles. Ces méthodes peuvent, en principe, éliminer le biais de non-réponse (en grand échantillon) quand la probabilité de non-réponse est reliée aux valeurs des variables auxiliaires par un modèle linéaire généralisé.

Dans le présent article, nous présentons certains éléments de théorie concernant l'estimation de la variance par linéarisation pour ce genre de méthodes en présence de non-réponse. Nous décrivons également une étude par simulation des propriétés de divers estimateurs par calage et des estimateurs de variance connexes dans des conditions choisies pour imiter deux enquêtes européennes réalisées par des instituts nationaux de statistiques. Nous considérons trois formes d'estimateur par calage, à savoir l'estimateur par ratissage croisé (raking ratio) classique, l'estimateur par calage du « maximum de vraisemblance » (Brackstone et Rao 1979 ; Fuller 2002) et l'estimateur par la régression généralisée (GREG). Le premier estimateur a été utilisé en pratique dans l'Enquête sur la population active (EPA) du Royaume-Uni, qui est la première enquête sur laquelle est fondée notre étude par simulation. Une version du deuxième estimateur a été utilisée en pratique dans l'Enquête sur les revenus et les dépenses (ERD) de l'Allemagne, qui est la deuxième enquête sur laquelle s'appuie notre étude par simulation. L'estimateur GREG est d'usage très répandu dans de nombreuses enquêtes, en particulier dans le contexte de la non-réponse (Särndal et Lundström 2005).

Un certain nombre de méthodes de pondération, qui n'entrent pas dans la catégorie des méthodes de calage généralisé considérées ici, ont été proposées. Voir Särndal et Lundström (2005) pour un compte rendu historique et Kott (2006), ainsi que Chang et Kott (2008) pour certains développements récents, où les variables auxiliaires pour lesquelles l'information au niveau de la population est disponible peuvent différer des variables utilisées comme covariables dans le modèle linéaire généralisé de la probabilité de non-réponse.

Le présent article porte avant tout sur l'estimation de la variance et, en particulier, sur les méthodes de linéarisation, pour lesquelles un certain nombre de formes légèrement différentes d'estimateurs de variance sont décrites dans la littérature. Dans notre étude par simulation, nous comparerons les propriétés de divers estimateurs par calage et des estimateurs de variance connexes en ce qui concerne les effets de l'échantillonnage ainsi que de la non-réponse. Une étude par simulation antérieure effectuée par Stukel, Hidiroglou et Särndal (1996) n'a révélé que peu de différences entre deux formes d'estimateur par linéarisation en ce qui concerne l'échantillonnage. Cependant, il existe des raisons pour lesquelles la non-réponse pourrait entraîner des écarts plus importants. Les conditions pour l'absence de biais dans les méthodes d'estimation par calage sous des modèles de non-réponse varient selon la méthode

1. Julia D'Arrigo et Chris Skinner, Université de Southampton. Courriel : C.J.Skinner@soton.ac.uk.

d'estimation (par exemple, Kalton et Maligalig 1991 ; Kalton et Flores-Cervantes 2003), et le choix de l'estimateur de variance pourrait importer davantage en présence de non-réponse (par exemple, Fuller 2002, section 8).

La plan de l'article est le suivant : les estimateurs par calage généralisé sont définis à la section 2 et, après la présentation d'un cadre asymptotique, le biais de ces estimateurs est examiné à la section 3. Les estimateurs de variance par linéarisation sont définis à la section 4. L'étude par simulation est présentée à la section 5, les résultats sont discutés à la section 6 et certaines conclusions sont énoncées à la section 7.

2. Estimation par calage généralisé

Nous considérons la classe des estimateurs pondérés d'un total de population $T_y = \sum_U y_i$, qui peut être exprimé sous la forme $\hat{T}_y = \sum_s w_i y_i$, où y_i est la valeur d'une variable étudiée pour une unité i dans un échantillon s tiré d'une population U *et* w_i est le *poids de sondage* qui peut dépendre de l'échantillon, mais non du choix de la variable étudiée. Nous supposons ici que l'échantillon s est constitué de l'ensemble restant de répondants après l'échantillonnage et l'éventuelle non-réponse totale. Le calage généralisé est une forme d'estimation pondérée qui peut être employé quand l'information auxiliaire au niveau de la population est disponible sous la forme d'un vecteur $T_x = \sum_U x_i$ des totaux de population des valeurs x_i d'un vecteur de variables auxiliaires, où la valeur x_i est connue pour toutes les unités présentes dans s. À l'instar de Deville et Särndal (1992), nous disons que les poids w_i sont *calés* s'ils satisfont aux *équations de calage* $\sum_s w_i x_i = T_x$. Le vecteur T_x est appelé vecteur des *totaux de calage*. La classe des poids de calage généralisé w_i est obtenue en minimisant la fonction objectif :

$$\sum_s d_i G\ (w_i / d_i), \tag{2.1}$$

sous la contrainte que les poids w_i soient calés, où $G(.)$ est une fonction objectif spécifiée qui satisfait à certains critères (voir Deville et coll. 1993) et d_i est un poids initial. Nous le prendrons ici égal au poids de sondage, c'est-à-dire $d_i = \pi_i^{-1}$, où π_i est la probabilité que l'unité i soit échantillonnée. Deville et Särndal (1992) montrent que (sous la contrainte que $G(.)$ obéisse à certaines conditions), la solution du problème d'optimisation contrainte susmentionné peut s'exprimer sous la forme :

$$w_i = d_i\ F\ (x_i' \hat{\lambda}), \tag{2.2}$$

où $F\ (u) = g^{-1}(u)$ désigne la fonction réciproque de $g\ (u) = dG(u)/du$ et $\hat{\lambda}$ est le multiplicateur de Lagrange qui résout les équations de calage :

$$\sum_s d_i\ F\ (x_i' \hat{\lambda})\ x_i = T_x. \tag{2.3}$$

Deville et Särndal (1992) discutent des divers choix de la fonction $G(.)$ et de la fonction $F(.)$ connexe. Nous examinons les trois choix suivants :

linéaire :

$$G_L(u) = (1/2)(u-1)^2,\ F_L(u) = 1 + u\ ;$$

multiplicative (*ratissage croisé*) :

$$G_M(u) = u \log(u) - u + 1,\ F_M(u) = \exp(u)\ ;$$

calage basé sur le maximum de vraisemblance :

$$G_{ML}(u) = u - 1 - \log(u),\ F_{ML}(u) = (1-u)^{-1}.$$

Voir également Deville et coll. (1993) et Fuller (2009, section 2.9) en ce qui concerne la terminologie susmentionnée pour ces fonctions. Dans le cas de la forme linéaire de $G(.)$, le problème d'optimisation possède une solution analytique et l'estimateur par calage généralisé devient $\hat{T}_y = \hat{T}_{yd} + (T_x - \hat{T}_{xd})' \hat{B}_s$, c'est-à-dire l'*estimateur par la régression généralisée* (GREG), où $\hat{T}_{yd} = \sum_s d_i y_i$, $\hat{T}_{xd} = \sum_s d_i x_i$ et

$$\hat{B}_s = \left(\sum_s d_i x_i x_i' \right)^{-1} \sum_s d_i x_i y_i. \tag{2.4}$$

Dans le cas de la forme multiplicative de $G(.)$, l'estimateur calé de T_y est l'estimateur par le ratissage croisé classique (Brackstone et Rao 1979) quand T_x contient les dénombrements de population dans les catégories d'au moins deux variables auxiliaires catégoriques. Par exemple, dans le contexte de l'Enquête sur la population active du Royaume-Uni, x_i désigne le vecteur de variables indicatrices de trois variables auxiliaires catégoriques : $x_i = (\delta_{1..i}, ..., \delta_{A..i}, \delta_{.1.i}, ..., \delta_{.B.i}, \delta_{..1i}, ..., \delta_{..Ci})'$, où $\delta_{a..i} = 1$ si l'unité i se trouve dans la catégorie a de la première variable auxiliaire et 0 autrement, $\delta_{.b.i} = 1$ si l'unité i est dans la catégorie b de la deuxième variable auxiliaire et 0 autrement, et ainsi de suite. Le total de population T_x de ce vecteur contient donc les dénombrements de population dans chacune des catégories (marginales) de chacune des trois variables auxiliaires. La construction des poids pour l'estimation par le ratissage croisé classique s'appuie habituellement sur l'ajustement proportionnel itératif (Brackstone et Rao 1979). Ireland et Kullback (1968) démontrent que cette méthode converge vers une solution du problème d'optimisation susmentionné.

La fonction $G_{ML}(u)$ mène à une version distincte basée sur le « maximum de vraisemblance » du redressement par calage, quand x_i prend la même forme désignant des variables indicatrices pour les variables auxiliaires catégoriques. Dans ce cas, la fonction objectif (2.1) peut être

interprétée comme une quantité proportionnelle à moins une log-vraisemblance dans le cas de l'échantillonnage aléatoire simple avec remise (Brackstone et Rao 1979 ; Fuller 2002).

3. Cadre asymptotique et biais de non-réponse

Nous examinons maintenant les propriétés asymptotiques de $\hat{T}_y$ en regard du plan d'échantillonnage ainsi que du mécanisme de non-réponse. Nous supposons que ce dernier est tel que chaque unité de la population répond, si elle est échantillonnée, avec la probabilité q_i, où cette probabilité est indépendante du choix de l'échantillon et où les diverses unités répondent de manière indépendante. Nous considérons un cadre asymptotique défini en fonction de suites de populations finies et de plans d'échantillonnage probabilistes et de mécanismes de réponse connexes (Fuller 2009, section 1.3), avec les termes d'ordre de grandeur exprimés en fonction de $n = \sum_U \pi_i q_i$, le nombre prévu d'unités répondantes, et de N, la taille de la population. Nous supposons qu'il existe des constantes positives K_1, K_2 et K_3 telles que $K_1 < nN^{-1}d_i < K_2$ et $K_3 < q_i$ pout tout i.

Nous supposerons que les estimateurs d'Horvitz-Thompson des moyennes convergent vers les moyennes de population finie correspondantes et que le théorème de la limite centrale est vérifié (tel qu'exprimé formellement dans les conditions du théorème 1.3.9 de Fuller 2009). En particulier, nous supposons que les suites et la fonction $F(.)$ sont telles qu'il existe une solution unique λ de

$$\sum_U q_i F(x_i'\lambda)\, x_i = T_x, \tag{3.1}$$

avec

$$\hat{\lambda} = \lambda + O_p(n^{-0,5}), \tag{3.2}$$

et que

$$\hat{T}_y = \sum_U q_i F(x_i'\lambda) y_i + O_p(Nn^{-0,5}). \tag{3.3}$$

Deville et Särndal (1992) montrent que $\lambda = 0$ sous certaines hypothèses (leur résultat 2). Toutefois, leurs hypothèses ne s'appliquent qu'à la distribution induite par le plan d'échantillonnage et comprennent la contrainte que $N^{-1}(\hat{T}_{xd} - T_x) \to 0$ en probabilité. Toutefois, dans le cas de la non-réponse, cette dernière est souvent peu plausible (voir Fuller 2002, page 15) et nous n'exigeons pas que λ soit le vecteur nul.

L'une de nos hypothèses clés sera :

Condition C : il existe un vecteur α tel que $F(x_i'\alpha) = q_i^{-1}$.

Si la condition C est vérifiée, α est une solution de (3.1) et donc $\lambda = \alpha$. Il découle de (3.3) que $\hat{T}_y$ converge vers T_y pour tout choix de la variable y si cette condition est vérifiée. Donc, nous pouvons considérer la condition C comme une condition suffisante de l'absence de biais de non-réponse (asymptotique). Cette propriété de la condition C a été discutée par Fuller, Loughlin et Baker (1994), Fuller (2009, page 284) ainsi que Särndal et Lundström (2005, proposition 9.2) pour le cas où F est linéaire. Fuller (2002, page 15), Kott (2006), ainsi que Chang et Kott (2008) considèrent aussi l'estimation des probabilité de réponse en utilisant des modèles généraux de la forme $q_i^{-1} = F(x_i'\alpha)$.

Afin d'illustrer ce qui pourrait arriver si la condition C n'était pas vérifiée, supposons que x_i est simplement une grandeur scalaire avec $x_i \equiv 1$. Alors, la solution unique de (3.1) est $\lambda = g(N/\sum_U q_i)$ et $p\lim(\hat{T}_y) = N(\sum_U q_i y_i)/(\sum_U q_i)$. D'où, le biais asymptotique de non-réponse qui ne disparaîtra que pour les variables étudiées qui sont « non corrélées » aux probabilités de réponse q_i.

4. Estimation de la variance par linéarisation

Nous nous penchons maintenant sur la variance asymptotique de $\hat{T}_y$ et sur son estimation. Comme à la section précédente, nous définissons la variance par rapport à la loi conjointe induite par l'échantillonnage ainsi que la non-réponse.

Commençons par noter qu'en général (et en particulier pour $G_M(.)$ et $G_{ML}(.)$), une itération est nécessaire pour résoudre les équations de calage. On trouve dans la littérature (voir Deville et coll. 1993) des travaux visant à estimer la variance de $\hat{T}_y$ après un nombre fini d'itérations. Nous suivons plutôt l'approche de Deville et coll. (1993) et, par exemple, de Binder et Théberge (1988) en approximant la variance de $\hat{T}_y$ par la variance de l'estimateur « convergé », c'est-à-dire l'estimateur hypothétique issu d'un nombre infini d'itérations, représenté par $\mathrm{var}(\sum_s w_i y_i)$, où les w_i sont les poids « convergés » qui sont les solutions du problème d'optimisation sous contrainte de la section 2.

Un estimateur de variance par linéarisation s'obtient en approximant $\mathrm{var}(\sum_s w_i y_i)$ par $\mathrm{var}(\sum_s d_i z_i)$ pour une « variable linéarisée » z_i (Deville 1999). Nous cherchons maintenant à construire cette variable en utilisant un argument sous grand échantillon. Nous obtenons d'abord une expression de $\hat{\lambda}$. Un développement en série de Taylor du premier membre des équations de calage (2.3) donne

$$\sum_s d_i F(x_i'\hat{\lambda}) x_i = \sum_s d_i F_i x_i + \sum_s d_i f(x_i'\lambda^*) x_i x_i'(\hat{\lambda} - \lambda),$$

où $F_i = F(x_i'\lambda)$, λ^* est compris entre $\hat{\lambda}$ et λ, et il est supposé que $f(u) = dF(u)/du$ existe. En supposant que $f(.)$ est continue et que $\lim_{N\to\infty} N^{-1}\sum_U q_i f_i x_i x_i'$ existe, et en utilisant (3.2), nous avons

$$N^{-1}\sum_s d_i\, F\,(x_i'\hat{\lambda})x_i =$$

$$N^{-1}\sum_s d_i\, F_i\, x_i + N^{-1}\sum_s d_i\, f_i\, x_i x_i'(\hat{\lambda}-\lambda) + o_p(n^{-0,5}), \quad (4.1)$$

où $f_i = f(x_i'\lambda)$. Alors, en supposant que $\lim_{N\to\infty} N^{-1}$ $\sum_U q_i f_i\, x_i\, x_i'$ est non singulière et en utilisant (2.3), nous obtenons

$$\hat{\lambda}-\lambda = \left[\sum_s d_i f_i x_i x_i'\right]^{-1}\left[T_x - \sum_s d_i F_i x_i\right] + o_p(n^{-0,5}). \quad (4.2)$$

Voir Fuller (2009, preuve du théorème 1.3.9) pour une description formelle de la façon dont (4.1) et (4.2) peuvent être dérivées et des conditions de régularité sous-jacentes. Notons que, pour s'assurer que $\lim_{N\to\infty} N^{-1}\sum_U q_i\ f_i\, x_i\, x_i'$ est non singulière, il pourrait être nécessaire d'éliminer de x_i les variables redondantes et peut-être (comme dans Deville et Särndal 1992) de modifier l'estimateur pour les échantillons dont la probabilité est faible qui rendent cette matrice singulière.

Un argument similaire comportant le développement en série de Taylor de w_i dans (2.2) autour de λ donne :

$$w_i = d_i[F_i + f_i\, x_i'(\hat{\lambda}-\lambda)] + o_p(Nn^{-1,5}). \quad (4.3)$$

Alors, en supposant que les moments de population nécessaires existent afin que le terme résiduel dans (4.3) soit vérifié uniformément sur les i (Fuller 2009, Corollaire 2.7.1.1.), nous avons

$$\hat{T}_y \equiv \sum_s w_i y_i$$

$$= \sum_s d_i\left[F_i + f_i x_i'(\hat{\lambda}-\lambda)\right] y_i + o_p(Nn^{-0,5}) \quad (4.4)$$

et, donc, découlant de (4.2) et (4.4) :

$$\hat{T}_y = \sum_s d_i\, F_i\, y_i + B\left[T_x - \sum_s d_i\, F_i\, x_i\right] + o_p(Nn^{-0,5}), \quad (4.5)$$

où

$$B = \left[\sum_s d_i\, f_i\, y_i\, x_i'\right]\left[\sum_s d_i\, f_i\, x_i\, x_i'\right]^{-1}. \quad (4.6)$$

Notons que $F_i = f_i = 1$ sous les hypothèses de Deville et Särndal (1992) (puisque, dans ce cas, $\lambda = 0$ et il découle des hypothèses au sujet de $G(.)$ que $F(0) = f(0) = 1$). Donc, sous ces hypothèses, l'expression (4.5) correspond au résultat 5 de Deville et Särndal (1992), c'est-à-dire que l'estimateur par calage généralisé est asymptotiquement équivalent à l'estimateur GREG. Par conséquent, la variance asymptotique de $\hat{T}_y$ est la même que celle de $\sum_s d_i z_i$, où z_i est la variable linéarisée :

$$z_i = F_i(y_i - \beta x_i), \quad (4.7)$$

et nous supposons que B converge vers une matrice limite finie β. Une autre dérivation de cette expression est donnée par Demnati et Rao (2004, section 3.4).

Pour l'estimation de la variance par linéarisation, $\hat{T}_y$ est traité comme étant l'estimateur linéaire $\sum_s d_i \hat{z}_i$, où

$$\hat{z}_i = \hat{F}_i(y_i - \hat{B}\, x_i) \quad (4.8)$$

est traité comme une variable fixe.

Un certain nombre de choix de $\hat{F}_i$ et $\hat{B}$ ont été discutés dans la littérature. En ce qui concerne $\hat{F}_i$, le choix naturel impliqué par l'argument qui précède est $\hat{F}_i = F(x_i'\hat{\lambda})$. Toutefois, un choix plus simple consisterait à prendre $\hat{F}_i = 1$. Deville et Särndal (1992) soulignent que, dans leur théorie classique avec $\lambda = 0$, ces choix sont asymptotiquement équivalents, mais expriment une préférence pour $\hat{F}_i = F(x_i'\hat{\lambda})$. Sous nos conditions selon lesquelles il existe une non-réponse et où l'égalité $\lambda = 0$ n'est pas nécessairement vérifiée, le deuxième choix semble préférable et c'est ce qui est souligné par Fuller (2002, page 15). Notons que ces deux choix impliquent que $\sum_s d_i \hat{z}_i$ prend la forme $\sum w_i(y_i - \hat{B}\, x_i)$ quand $\hat{F}_i = F(x_i'\hat{\lambda})$ ou $\sum d_i(y_i - \hat{B}\, x_i)$ quand $\hat{F}_i = 1$. Nous désignerons donc ces choix comme étant les *résidus pondérés par* w_i ou les *résidus pondérés par* d_i.

En ce qui concerne $\hat{B}$, il découle de notre argument concernant les choix de $\hat{F}_i$ que f_i dans (4.2) devrait être remplacée par $\hat{f}_i = f(x_i'\hat{\lambda})$, ce qui donne :

i) $\hat{B} = [\sum_s d_i\, \hat{f}_i\, y_i\, x_i']\,[\sum_s d_i\, \hat{f}_i\, x_i\, x_i']^{-1}$, comme l'ont également proposé Demnati et Rao (2004).

Les autres choix sont :

ii) $\hat{B} = \hat{B}_s$, comme dans (2.4), comme l'ont proposé Deville et coll. (1993).

iii) $\hat{B} = [\sum_s w_i\, y_i\, x_i']\,[\sum_s w_i\, x_i\, x_i']^{-1}$, comme l'ont proposé Deville et Särndal (1992, équation 3.4), ce qui pourrait être plus facile à calculer que $\hat{B}_s$ pour les utilisateurs des fichiers de données d'enquête qui contiennent les poids w_i, mais non les poids d_i.

La mesure dans laquelle ces choix diffèrent dépend du choix de la fonction $G(.)$. Dans le cas linéaire, $f(u) = 1$, de sorte que les estimateurs donnés en (i) et (ii) sont identiques. Dans le cas du redressement par calage classique, $f(u) = F(u) = \exp(u)$ de sorte que $\hat{f}_i = \hat{F}_i$ et $d_i\hat{f}_i = w_i$ et les estimateurs (i) et (iii) sont identiques. Pour l'estimateur par calage basé sur le « maximum de vraisemblance », nous avons $F(u) = (1-u)^{-1}$ et $f(u) = (1-u)^{-2}$, de sorte que $d_i\hat{f}_i = w_i^2/d_i$ et les trois estimateurs de variance sont tous distincts.

Après avoir déterminé la forme de $\hat{z}_i$ dans (4.8), nous obtenons l'estimateur de la variance par linéarisation pour

$\hat{T}_y$ par estimation de la variance de l'estimateur linéaire $\sum_s d_i \hat{z}_i$, en traitant d_i et $\hat{z}_i$ comme étant fixes. Dans le cas d'un plan d'échantillonnage à plusieurs degrés stratifié, en supposant que le tirage des unités primaires d'échantillonnage (UPE) dans les strates se fait « avec remise », un estimateur standard de la variance (par exemple, Stukel et coll. 1996) est donné par :

$$\hat{V}(\hat{T}_y) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} (z_{hj} - \bar{z}_h)^2 \qquad (4.9)$$

où $z_{hj} = \sum_k d_{hjk} \hat{z}_{hjk}$, $\bar{z}_h = \sum_j z_{hj}/n_h$ et $\hat{z}_{hjk}$ est la valeur de la variable définie en (4.8) pour le k^{e} individu dans la j^{e} UPE sélectionnée dans la strate h. Cet estimateur demeure approprié en présence de non-réponse si la réponse individuelle dans chaque UPE est indépendante de la réponse dans toutes les autres UPE et qu'au moins un individu est observé dans chaque UPE sélectionnée (Fuller et coll. 1994, page 78).

5. Études par simulation

Afin de comparer la performance des estimateurs pondérés à celle des estimateur de variance correspondants, nous avons effectué deux études par simulation en construisant des populations artificielles en nous servant des données de l'Enquête sur la population active (EPA) du Royaume-Uni et de l'Enquête sur les revenus et les dépenses (ERD) de l'Allemagne. Dans chaque cas, nous avons généré $R = 1\,000$ échantillons à partir de ces populations en procédant d'abord à l'échantillonnage de manière à calquer le plan de sondage réel moyennant certaines simplifications, puis en éliminant les cas de non-réponse conformément à deux modèles de non-réponse. Le premier modèle suppose une non-réponse de forme multiplicative qui, d'après la condition C de la section 3, pourrait donner lieu à un biais plus faible pour la méthode de calage par le raking ratio. Le deuxième modèle s'appuie sur l'hypothèse d'une non-réponse additive pour pourrait donner lieu à un biais plus faible pour l'estimateur GREG.

Pour chacun des R échantillons, nous avons calculé les estimations ponctuelles des paramètres en nous servant des diverses méthodes de calage généralisé présentées à la section 2 et estimé les variances en utilisant les diverses méthodes par linéarisation présentées à la section 4. Ensuite, nous avons résumé les propriétés des estimateurs.

5.1 Étude fondée sur l'Enquête sur la population active du Royaume-Uni

La première étude s'appuyait sur des données provenant du trimestre de mars à mai 1998 de l'EPA du Royaume-Uni, qui est une enquête auprès de la population à domicile du Royaume-Uni, conçue pour fournir des renseignements sur le marché britannique du travail et réalisée par l'Office for National Statistics (ONS). Nous avons traité l'échantillon d'environ 58 000 ménages comme une population artificielle. Nous avons tiré des échantillons répétés de cette population de façon à imiter le plan de sondage utilisé pour l'EPA (ONS 1998, section 3). Chaque échantillon comprenait 1 211 ménages sélectionnés par échantillonnage aléatoire stratifié avec répartition proportionnelle entre 19 strates définies selon la région de résidence. Ces régions ont été définies de façon qu'elles correspondent aux secteurs affectés aux intervieweurs, qui définissaient les strates dans l'EPA. Dans le cadre de cette enquête, toutes les personnes faisant partie d'un ménage échantillonné sont interviewées dans la mesure du possible. Dans notre étude par simulation, nous avons retenu tous les répondants compris dans un ménage échantillonné, sauf ceux de moins de 16 ans, qui sont sans pertinence pour les estimations d'intérêt.

Pour déterminer si les personnes échantillonnées avaient répondu, nous nous sommes servis des deux modèles de non-réponse qui suivent, basés sur les résultats d'une étude de Foster (1998).

Modèle de non-réponse multiplicatif :

$$q_i^{-1} = 1{,}15 \times 1{,}17 \text{ (si Londres)} \times 1{,}13 \text{ (si moins de 35 ans)} \times 1{,}1 \text{ (si sexe féminin)}$$

Modèle de non-réponse additif :

$$q_i^{-1} = 1{,}15 + 0{,}20 \text{ (si Londres)} + 0{,}15 \text{ (si moins de 35 ans)} + 0{,}10 \text{ (si sexe féminin)}$$

où q_i est la probabilité de réponse définie au début de la section 3 et la forme du modèle est choisie de manière à satisfaire la condition C.

Trois paramètres d'intérêt sont définis pour la population artificielle : les nombres totaux de personnes en chômage, occupées ou inactives. Nous avons conçu les poids pour les répondants individuels en nous servant de totaux de calage qui correspondaient aux dénombrements de population dans les catégories des trois variables auxiliaires catégoriques et des poids initiaux d'Horvitz-Thompson d_i, comme à la section 2. Ce choix des variables auxiliaires avait pour but de calquer celles utilisée dans l'EPA. Cependant, étant donné l'échelle réduite de notre population artificielle et les nombres conséquemment plus petits de personnes dans les strates, nous avons simplifié les variables de calage de l'EPA pour obtenir les trois facteurs catégoriques suivants, qui définissent 83 totaux de contrôle :

- région de résidence avec 23 catégories ;
- une classification croisée du sexe selon dix groupes d'âge (consistant en années uniques pour les personnes de 16 à 24 ans et en un groupe d'âge distinct pour les 25 ans et plus) avec 20 catégories ;

- une classification croisée de la région (Nord de l'Angleterre ; Londres et Sud-Est ; Midlands et East Anglia ; Écosse) selon le sexe et l'âge par tranche de 15 ans (16 à 29 ans, 30 à 44 ans, 45 à 59 ans, 60 à 75 ans et 75 ans et plus) avec 40 catégories.

5.2 Étude fondée sur l'Enquête sur les revenus et dépenses de l'Allemagne

Notre deuxième étude est basée sur l'édition de 1998 de l'Enquête sur les revenus et dépenses (ERD) de l'Allemagne, qui est une enquête-ménage nationale réalisée tous les cinq ans par le Bureau fédéral de la statistique pour fournir des renseignements sur la situation économique et sociale des ménages, surtout en ce qui concerne la distribution des revenus et des dépenses (Muennich et Schulrle 2003). Nous avons utilisé les données provenant d'une population synthétique de 64 326 ménages, créée pour représenter 20 % des ménages de la région de Brême à l'exclusion de ceux dont le revenu mensuel net du ménage était égal ou supérieur à 35 000 DM (DM désigne le mark allemand). Un plan d'échantillonnage par quota a été utilisé pour cette enquête et nous n'avons pas essayé de le calquer. À la place, nous avons utilisé dans notre simulation un échantillonnage aléatoire simple avec un modèle de non-réponse. Nous avons tiré des échantillons aléatoires simples répétés de 1 340 ménages à partir de la population artificielle, ce qui représente une fraction d'échantillonnage d'environ 1/48. Nous avons construit les modèles de non-réponse en utilisant les résultats d'études portant sur des enquêtes similaires réalisées en Grande-Bretagne, à savoir l'Enquête sur les dépenses des familles et l'Enquête nationale sur les aliments (Foster 1998). Pour chaque échantillon sélectionné, nous avons déterminé le sous-ensemble de ménages répondants à l'aide des modèles de non-réponse suivants :

Modèle multiplicatif :

$$\begin{aligned} q_i^{-1} = 1{,}44 \;&\times\; 1{,}09 \text{ (si travailleur autonome)} \\ &\times\; 1{,}03 \text{ (si chômeur)} \\ &\times\; 0{,}97 \text{ (si travailleur)} \\ &\times\; 1{,}16 \text{ (si aucun enfant dans le ménage).} \end{aligned}$$

Modèle additif :

$$\begin{aligned} q_i^{-1} = 1{,}44 \;&+\; 0{,}13 \text{ (si travailleur autonome)} \\ &+\; 0{,}04 \text{ (si chômeur)} \\ &-\; 0{,}04 \text{ (si travailleur)} \\ &+\; 0{,}23 \text{ (si aucune enfant dans le ménage).} \end{aligned}$$

Les paramètres d'intérêt sont le revenu net total du ménage par trimestre et les dépenses totales du ménage par trimestre, calculés d'après les données sur la population finie artificielle.

Comme pour l'étude fondée sur l'EPA, nous avons attribué un poids à chaque ménage échantillonné. Dans l'ERD réelle, les poids sont construits essentiellement selon la méthode de calage basée sur le maximum de vraisemblance en ajustant simultanément les données d'échantillon aux distributions marginales de plusieurs caractéristiques, telles que le type de ménage, le statut socioéconomique de la personne de référence, la catégorie de revenu net du ménage et la région (land). Nous essayons d'imiter ce redressement dans la mesure du possible dans notre étude. Toutefois, comme dans le cas de l'EPA, en raison du problème que posent les strates contenant un petit nombre de ménages, nous simplifions les variables de calage de l'ERD pour obtenir les trois facteurs catégoriques suivants :

- type de ménage avec sept catégories
  - mère ou père seul plus un enfant,
  - mère ou père seul plus deux enfants ou plus,
  - couple avec un enfant – conjoint travailleur,
  - couple avec un enfant – conjoint chômeur,
  - couple avec deux enfants ou plus – conjoint travailleur,
  - couple avec deux enfants ou plus – conjoint chômeur,
  - autre ;
- statut social de la personne de référence avec cinq catégories
  - travailleur autonome,
  - fonctionnaire ou militaire,
  - employé,
  - ouvrier,
  - chômeur, pensionné, étudiant ou autre ;
- revenu net du ménage par trimestre avec trois catégories
  - 0 à 5 000 DM,
  - 5 000 à 7 000 DM,
  - 7 000 à 35 000 DM.

6. Résultats

6.1 Propriétés des estimateurs ponctuels

Le tableau 6.1 donne les propriétés des estimateurs ponctuels du nombre total de chômeurs dans l'étude de l'EPA pour diverses méthodes de calage et diverses hypothèses au sujet de la non-réponse. Les propriétés sont évaluées selon les pratiques habituelles dans les études par simulation. Par exemple, au tableau 6.1, le biais est calculé d'après $\hat{B}(\hat{T}_y) = \hat{E}(\hat{T}_y) - T_y$, où $\hat{E}(\hat{T}_y) = 1/R \sum_{r=1}^{R} \hat{T}_{y_r}$, $\hat{T}_{y_r}$ est la valeur de $\hat{T}_y$ pour l'échantillon r et R est le nombre d'échantillons simulés. Nous constatons, en examinant ce tableau, que l'erreur-type demeure presque constante pour les diverses méthodes de calage pour un modèle de non-réponse donné. La non-réponse accroît l'erreur-type dans tous les estimateurs comme il fallait s'y attendre (puisque la taille d'échantillon

est réduite). Le tableau 2 donne des preuves d'un biais de non-réponse, qui est d'ordre similaire pour chacune des méthodes de calage. Nous ne constatons pas que ce biais est moindre quand l'estimateur concorde avec le modèle de non-réponse (c'est-à-dire l'estimateur GREG pour la réponse additive et l'estimateur par calage (ratissage croisé) pour la réponse multiplicative) comme nous aurions pu nous y attendre. Cela pourrait tenir au fait que les covariables utilisées dans les modèles de non-réponse (par exemple, la variable d'âge égal ou supérieur à 35 ans) ne sont pas toutes incluses dans les variables de calage. Néanmoins, le biais de non-réponse est faible en ce sens que la racine carrée de l'erreur quadratique moyenne est très semblable à l'erreur-type dans chaque cas. En présence de non-réponse, la méthode de calage GREG produit certains points négatifs, tandis que les deux méthodes de ratissage permettent d'éviter ce problème, comme prévu. Un plus grand nombre de poids très grands sont toutefois observés pour l'estimateur par calage basé sur le « maximum de vraisemblance ».

Les résultats correspondants pour les données de l'ERD sont présentés au tableau 6.2. La tendance des résultats est généralement similaire, quoi qu'il n'y ait pas d'évidence de présence d'un biais de non-réponse significatif (c'est-à-dire que le biais observé peut être expliqué par des variations de simulation). Les erreurs-types et les racines carrées des erreurs quadratiques moyennes demeurent également presque constantes d'une méthode de pondération à l'autre pour un modèle de non-réponse donné.

Tableau 6.1
Propriétés de simulation des estimateurs ponctuels du total de chômeurs en utilisant les données de l'EPA avec R = 1 000

Modèle de non-réponse/estimateur ponctuel	Biais (erreur-type de simulation)	Erreur-type	Racine carrée de l'erreur quadratique moyenne	Nombre de poids négatifs[1]	Nombre de poids très grands[1, 2]
Réponse complète :					
Calage GREG	7,6 (14,3)	452,8	452,8	0	0
Calage par ratissage classique	8,3 (14,3)	452,8	452,9	0	0
Calage basé sur le « MV »	9,0 (14,3)	453,3	453,4	0	1
Non-réponse multiplicative :					
Calage GREG	-45,6 (15,8)	498,3	500,3	4	1
Calage par ratissage classique	-42,1 (15,8)	498,8	500,6	0	2
Calage basé sur le « MV »	-39,7 (15,8)	499,4	501,0	0	7
Non-réponse additive :					
Calage GREG	-37,3 (15,7)	497,4	498,8	5	1
Calage par ratissage classique	-34,7 (15,7)	497,5	498,7	0	3
Calage basé sur le « MV »	-32,4 (15,8)	498,1	499,1	0	7

[1]Nombre de ces poids sur l'ensemble des unités échantillonnés sur l'ensemble des 1 000 échantillons.
[2]Nombre de poids égaux à plus de dix fois le poids de sondage correspondant.

Tableau 6.2
Propriétés de simulation des estimateurs ponctuels du revenu total en utilisant les données de l'ERD avec R = 1 000

Modèle de non-réponse/estimateur ponctuel	Biais (erreur-type de simulation)	Erreur-type	Racine carrée de l'erreur quadratique moyenne	Nombre de poids négatifs	Nombre de poids très grands
Réponse complète :					
Calage GREG	-172,2 (331,3)	10 477,3	10 478,7	0	0
Calage par ratissage classique	-170,6 (331,5)	10 484,1	10 485,8	0	0
Calage basé sur le « MV »	-169,8 (331,8)	10 491,5	10 492,9	0	0
Non-réponse multiplicative :					
Calage GREG	-495,7 (429,7)	13 586,8	13 595,8	0	0
Calage par ratissage classique	-493,8 (429,6)	13 584,6	13 593,5	0	0
Calage basé sur le « MV »	-463,5 (429,5)	13 582,8	13 590,7	0	0
Non-réponse additive :					
Calage GREG	-473,2 (430,5)	13 614,8	13 623,0	0	0
Calage par ratissage classique	-469,4 (430,5)	13 612,9	13 621,0	0	0
Calage basé sur le « MV »	-439,5 (430,5)	13 613,5	13 620,6	0	0

6.2 Propriétés des estimateurs de variance

Les propriétés des divers estimateurs des variances des estimateurs ponctuels du nombre total de chômeurs d'après l'EPA sont présentées au tableau 6.3 (dans ce tableau, l'« estimation de l'erreur-type » désigne la racine carrée de l'estimation de la variance). Nous faisons plusieurs observations :

- la pondération des résidus par w_i plutôt que par d_i réduit le biais et la racine carrée de l'erreur quadratique moyenne de l'estimateur de l'erreur-type. Le biais dû à l'utilisation des résidus pondérés par d_i est particulièrement important en cas de non-réponse (comme l'a fait remarquer Fuller 2002), mais nous constatons des réductions non négligeables du biais même en cas de réponse complète ;
- le choix du poids utilisé dans $\hat{B}$ pour le calcul des résidus semble avoir peu d'effet ;
- pour un modèle de non-réponse et un choix de pondération des résidus donnés, les résultats produits par les divers choix de l'estimateur ponctuel diffèrent peu.

Au tableau 6.4, les résultats du tableau 6.3 sont étendus afin de prendre en considération le biais relatif des estimateurs de l'erreur-type, plutôt que leur biais absolu et pour considérer deux paramètres supplémentaires, à savoir les nombres totaux de chômeurs et de personnes inactives. Nous voyons de nouveau que le biais relatif dû à l'utilisation des résidus pondérés par d_i peut être important en présence de non-réponse, supérieur à 20 % dans plusieurs cas, et qu'il est réduit en utilisant les résidus pondérés par w_i. Encore une fois, nous observons peu de changements du biais relatif en pourcentage des estimateurs de l'erreur-type quand nous utilisons divers choix de pondération dans le calcul de $\hat{B}$ pour les résidus.

Les résultats correspondants pour les données de l'ERD lorsque l'on estime le revenu total sont présentés au tableau 6.5. De nouveaux, les résultats sont d'allure généralement similaire à ceux obtenus pour les données de l'EPA au tableau 6.3. En cas de réponse complète, l'utilisation des résidus pondérés par w_i plutôt que par d_i produit une légère amélioration du biais et de la REQM des estimateurs de l'erreur-type. Pour les cas de non-réponse, les améliorations sont considérables. Nous observons peu de changements dans les estimateurs de l'erreur-type quand nous modifions le choix de la pondération utilisée pour estimer les coefficients de régression. Au tableau 6.6, les résultats du tableau 6.5 sont étendus afin de prendre en considération le biais relatif des estimateurs de l'erreur-type plutôt que leur biais absolu et de considérer un paramètre supplémentaire, à savoir les dépenses totales par trimestre. De nouveau, nous voyons que le biais relatif résultant de l'utilisation des résidus pondérés par d_i peut être important en présence de non-réponse, supérieur à 35 % dans tous les cas, et qu'il est réduit si l'on utilise les résidus pondérés par w_i pour lesquels le biais relatif n'est jamais supérieur à environ 3 %.

7. Conclusion

L'étude par simulation a révélé peu de différences entre le biais ou les propriétés de variance des trois estimateurs par calage considérés, à savoir l'estimateur GREG, l'estimateur par calage classique (ratissage croisé) et l'estimateur par calage basé sur le maximum de vraisemblance. Nous avons observé certains petits écarts dans la distribution des poids extrêmes, l'estimateur par calage fondé sur le maximum de vraisemblance produisant le plus grand nombre de poids très grands et l'estimateur GREG étant le seul produisant quelques poids négatifs.

En ce qui concerne les estimateurs de variance, la principale observation est le contraste entre l'approche consistant à pondérer les résidus par les poids de sondage et celle consistant à les pondérer par les poids calés. Nous avons constaté que le second estimateur de variance possédait systématiquement un bais plus petit et que cet effet était très marqué en présence de non-réponse, situation dans laquelle le premier estimateur pouvait être gravement biaisé. Le biais du second estimateur était généralement petit et le niveau de couverture des intervalles de confiance associés était généralement proche de la couverture nominale.

Nous avons considéré d'autres moyens de pondérer les observations pour construire les coefficients de régression lorsque l'on calcule les résidus de l'estimateur de variance par linéarisation, mais les effets observés étaient faibles et nous n'avons recueilli aucune preuve que ce choix est important en pratique.

En général, les constatations concernant les variables catégoriques dans l'Enquête sur la population active du Royaume-Uni étaient remarquablement comparables à celles pour les variables continues de l'Enquête sur les revenus et dépenses de l'Allemagne.

Tableau 6.3
Propriétés des estimateurs de variance pour l'estimation du nombre total de chômeurs d'après l'EPA (R = 1 000)

Méthode de pondération	Résidus pondérés par w ou d[1]	Poids utilisé pour $\hat{B}$ dans le résidu[1]	Moyenne de l'estimateur de l'erreur-type	Biais de l'estimateur de l'e.-t. (e.-t. de simulation)	REQM de l'estimateur de l'e.-t.	Couverture[2] de l'intervalle de confiance (%)
Réponse complète :						
Calage GREG	d	d	433,9	-18,8 (0,9)	33,4	93,5
	d	w	434,3	-18,5 (0,9)	33,3	93,5
	w	d	442,8	-10,0 (1,0)	31,9	93,8
	w	w	441,9	-10,8 (1,0)	32,0	93,7
Calage par ratissage classique	d	d	433,9	-18,8 (0,9)	33,4	93,5
	d	w	434,2	-18,5 (0,9)	33,3	93,5
	w	d	443,0	-9,8 (1,0)	32,0	93,8
	w	w	442,0	-10,7 (1,0)	32,0	93,8
Calage basé sur le « MV »	d	d	433,9	-19,4 (0,9)	33,7	93,5
	d	w	434,3	-19,1 (0,9)	33,6	93,5
	d	df	435,4	-17,9 (0,9)	33,0	93,5
	w	d	443,7	-9,6 (1,0)	32,5	93,7
	w	w	442,3	-11,1 (1,0)	32,4	93,7
	w	df	441,6	-11,8 (1,0)	32,3	93,7
Non-réponse multiplicative :						
Calage GREG	d	d	385,7	-112,6 (0,9)	116,0	85,8
	d	w	386,1	-112,1 (0,9)	115,5	85,8
	w	d	489,5	-8,8 (1,2)	39,2	94,2
	w	w	487,8	-10,4 (1,2)	39,2	94,2
Calage par ratissage classique	d	d	385,7	-113,1 (0,9)	116,5	85,7
	d	w	386,1	-112,7 (0,9)	116,1	85,7
	w	d	490,3	-8,5 (1,2)	39,6	94,3
	w	w	488,4	-10,4 (1,2)	39,5	94,1
Calage basé sur le « MV »	d	d	385,7	-113,7 (0,9)	117,1	85,4
	d	w	386,2	-113,2 (0,9)	116,6	85,6
	d	df	387,8	-111,6 (0,9)	115,0	85,8
	w	d	491,9	-7,5 (1,3)	40,4	94,2
	w	w	488,9	-10,5 (1,2)	39,9	94,0
	w	df	487,5	-11,9 (1,2)	39,8	94,0
Non-réponse additive :						
Calage GREG	d	d	386,5	-110,9 (0,9)	114,4	86,0
	d	w	387,0	-110,5 (0,9)	113,9	86,0
	w	d	489,3	-8,2 (1,2)	39,0	94,6
	w	w	487,6	-9,8 (1,2)	39,0	94,6
Calage par ratissage classique	d	d	386,5	-111,0 (0,9)	114,4	85,8
	d	w	387,0	-110,6 (0,9)	114,0	85,8
	w	d	490,1	-7,4 (1,2)	39,2	94,7
	w	w	488,1	-9,4 (1,2)	39,1	94,6
Calage basé sur le « MV »	d	d	386,5	-111,6 (0,9)	115,0	85,6
	d	w	387,0	-111,1 (0,9)	114,6	85,6
	d	df	388,6	-109,5 (0,9)	113,0	85,9
	w	d	491,6	-6,5 (1,3)	40,0	94,7
	w	w	488,6	-9,5 (1,2)	39,5	94,6
	w	df	487,3	-10,8 (1,2)	39,4	94,6

[1] Voir le texte qui suit l'équation (4.8), où les choix df, d et w correspondent à $\hat{B}$ dans (i), (ii) et (iii) respectivement.
[2] Pourcentage d'intervalles de confiance à 95 % de la théorie normale contenant la valeur réelle.

Tableau 6.4
Biais relatif (%) des estimateurs de l'erreur-type des totaux de personnes chômeuses, occupées et inactives d'après l'EPA (R = 1 000)

Méthode de pondération	Résidus pondérés par w ou d[1]	Poids utilisé pour $\hat{B}$ dans le résidu[1]	Biais relatif de l'estimateur de l'erreur-type		
			Chômeuses	Occupées	Inactives
Réponse complète :					
Calage GREG	d	d	-4,2	-3,4	0,5
	d	w	-4,1	-3,3	0,6
	w	d	-2,2	-2,2	1,9
	w	w	-2,4	-2,3	1,7
Calage par ratissage classique	d	d	-4,2	-3,3	0,7
	d	w	-4,1	-3,2	0,8
	w	d	-2,2	-2,1	2,1
	w	w	-2,4	-2,2	1,9
Calage basé sur le « MV »	d	d	-4,3	-3,3	0,7
	d	w	-4,2	-3,3	0,8
	d	df	-4,0	-3,1	1,1
	w	d	-2,1	-2,0	2,3
	w	w	-2,4	-2,2	1,9
	w	df	-2,6	-2,3	1,8
Non-réponse multiplicative :					
Calage GREG	d	d	-22,6	-22,3	-18,2
	d	w	-22,5	-22,2	-18,1
	w	d	-1,8	-3,3	1,8
	w	w	-2,1	-3,5	1,5
Calage par ratissage classique	d	d	-22,7	-30,6	-18,4
	d	w	-22,6	-30,5	-18,3
	w	d	-1,7	-13,5	1,7
	w	w	-2,1	-13,7	1,3
Calage basé sur le « MV »	d	d	-22,8	-22,0	-18,4
	d	w	-22,7	-21,9	-18,3
	d	df	-22,3	-21,7	-17,9
	w	d	-1,5	-2,7	1,9
	w	w	-2,1	-3,1	1,3
	w	df	-2,4	-3,3	1,1
Non-réponse additive :					
Calage GREG	d	d	-22,3	-21,8	-18,5
	d	w	-22,2	-21,7	-18,4
	w	d	-1,6	-2,9	1,1
	w	w	-2,0	-3,1	0,8
Calage par ratissage classique	d	d	-22,3	-30,2	-18,0
	d	w	-22,2	-30,1	-17,9
	w	d	-1,5	-13,3	1,8
	w	w	-1,9	-13,5	1,4
Calage basé sur le « MV »	d	d	-22,4	-21,6	-18,0
	d	w	-22,3	-21,5	-17,9
	d	df	-22,0	-21,3	-17,6
	w	d	-1,3	-2,4	2,0
	w	w	-1,9	-2,8	1,5
	w	df	-2,2	-3,0	1,3

[1] Voir le texte qui suit l'équation (4.8), où df, d et w correspondent à $\hat{B}$ dans (i), (ii) et (iii), respectivement.

Tableau 6.5
Propriétés des estimateurs de variance de l'estimateur du revenu total d'après l'EDR (R = 1 000)

Méthode de pondération	Résidus pondérés par w ou d[1]	Poids utilisé pour $\hat{B}$ dans le résidu[1]	Moyenne de l'estimateur de l'erreur-type	Biais de l'estimateur de l'e.-t. (e.-t. de simulation)	REQM de l'estimateur de l'e.-t.	Couverture[2] de l'intervalle de confiance (%)
Réponse complète :						
Calage GREG	d	d	10 338,8	-138,5 (6,9)	259,0	93,8
	d	w	10 339,2	-138,2 (6,9)	258,8	93,8
	w	d	10 377,9	-99,5 (6,9)	240,0	94,1
	w	w	10 376,8	-100,5 (6,9)	240,3	94,1
Calage par ratissage classique	d	d	10 338,8	-145,3 (6,9)	262,7	93,8
	d	w	10 339,2	-144,9 (6,9)	262,5	93,8
	w	d	10 370,0	-106,1 (6,9)	243,1	94,0
	w	w	10 376,9	-107,2 (6,9)	243,5	94,0
Calage basé sur le « MV »	d	d	10 338,8	-152,7 (6,9)	266,9	93,9
	d	w	10 339,2	-152,4 (6,9)	266,7	93,9
	d	df	10 340,3	-151,3 (6,9)	266,1	94,0
	w	d	10 378,3	-113,2 (6,9)	246,5	94,0
	w	w	10 377,1	-114,4 (6,9)	247,0	94,0
	w	df	10 376,7	-114,8 (6,9)	247,2	94,0
Non-réponse multiplicative :						
Calage GREG	d	d	8 104,7	-5 482,1 (7,4)	5 487,1	75,8
	d	w	8 105,5	-5, 81,3 (7,4)	5 486,3	75,8
	w	d	13 214,5	-372,3 (12,8)	549,7	94,5
	w	w	13 210,9	-375,9 (12,8)	551,7	94,5
Calage par ratissage classique	d	d	8 104,7	-5 479,8 (7,4)	5 484,9	75,8
	d	w	8 105,5	-5 479,1 (7,4)	5 484,1	75,8
	w	d	13 214,1	-370,4 (12,8)	549,4	94,5
	w	w	13 210,4	-374,2 (12,8)	551,5	94,5
Calage basé sur le « MV »	d	d	8 104,7	-5 478,1 (7,4)	5 483,1	75,8
	d	w	8 105,5	-5 477,3 (7,4)	5 482,3	75,8
	d	df	8 108,1	-5 474,7 (7,4)	5 479,7	75,9
	w	d	13 215,2	-367,6 (12,9)	549,4	94,5
	w	w	13 210,6	-372,2 (12,9)	551,6	94,5
	w	df	13 208,9	-373,9 (12,9)	552,3	94,5
Non-réponse additive :						
Calage GREG	d	d	8 106,3	-5 508,5 (7,4)	5 513,5	75,6
	d	w	8 107,1	-5 507,7 (7,4)	5 512,7	75,6
	w	d	13 207,9	-407,0 (12,8)	573,8	94,3
	w	w	13 204,3	-410,5 (12,8)	575,9	94,3
Calage par ratissage classique	d	d	8 106,3	-5 506,6 (7,4)	5 511,6	75,7
	d	w	8 107,1	-5 505,9 (7,4)	5 510,9	75,7
	w	d	13 207,7	-405,3 (12,8)	573,6	94,1
	w	w	13 203,9	-409,0 (12,8)	575,8	94,1
Calage basé sur le « MV »	d	d	8 106,3	-5 507,2 (7,4)	5 512,2	75,9
	d	w	8 107,1	-5 506,4 (7,4)	5 511,4	75,9
	d	df	8 109,7	-5 503,8 (7,4)	5 508,8	75,9
	w	d	13 208,9	-404,6 (12,9)	574,8	94,1
	w	w	13 204,2	-409,2 (12,9)	577,3	94,1
	w	df	13 202,5	-411,0 (12,9)	578,1	94,1

[1] Voir le texte qui suit l'équation (4.8), où les choix df, d et w correspondent à $\hat{B}$ dans (i), (ii) et (iii) respectivement.
[2] Pourcentage d'intervalles de confiance à 95 % de la théorie normale contenant la valeur réelle.

Tableau 6.6
Biais relatif (%) des estimateurs de variance des totaux des dépenses et des revenus d'après l'ERD (R = 1 000)

Méthode de pondération	Résidus pondérés par w ou d[1]	Poids utilisé pour $\hat{B}$ dans le résidu[1]	Biais relatif de l'estimateur de l'erreur-type	
			Dépenses	Revenus
Réponse complète :				
Calage GREG	d	d	0,7	-1,3
	d	w	0,7	-1,3
	w	d	1,3	-1,0
	w	w	1,3	-1,0
Calage par ratissage classique	d	d	0,7	-1,4
	d	w	0,7	-1,4
	w	d	1,2	-1,0
	w	w	1,2	-1,0
Calage basé sur le « MV »	d	d	0,6	-1,5
	d	w	0,6	-1,5
	d	df	0,6	-1,4
	w	d	1,2	-1,1
	w	w	1,2	-1,1
	w	df	1,2	-1,1
Non-réponse multiplicative :				
Calage GREG	d	d	-38,2	-40,4
	d	w	-38,2	-40,3
	w	d	-0,3	-2,7
	w	w	-0,3	-2,8
Calage par ratissage classique	d	d	-38,2	-40,3
	d	w	-38,2	-40,3
	w	d	-0,3	-2,7
	w	w	-0,3	-2,8
Calage basé sur le « MV »	d	d	-38,2	-40,3
	d	w	-38,2	-40,3
	d	df	-38,2	-40,3
	w	d	-0,3	-2,7
	w	w	-0,3	-2,7
	w	df	-0,4	-2,8
Non-réponse additive :				
Calage GREG	d	d	-38,1	-40,5
	d	w	-38,1	-40,5
	w	d	-0,2	-3,0
	w	w	-0,2	-3,0
Calage par ratissage classique	d	d	-38,1	-40,5
	d	w	-38,1	-40,5
	w	d	-0,2	-3,0
	w	w	-0,2	-3,0
Calage basé sur le « MV »	d	d	-38,2	-40,5
	d	w	-38,2	-40,5
	d	df	-38,1	-40,4
	w	d	-0,2	-3,0
	w	w	-0,3	-3,0
	w	df	-0,3	-3,0

[1] Voir le texte qui suit l'équation (4.8), où df, d et w correspondent à $\hat{B}$ dans (i), (ii) et (iii), respectivement.

Remerciements

Les commentaires de deux examinateurs nous ont aidé à améliorer considérablement le présent article. Nous remercions l'Office for National Statistics d'avoir mis à notre disposition les données de l'Enquête sur la population active, ainsi que Ralf Münnich et ses collègues du projet DACSEIS (http://www.dacseis.de/) de nous avoir fourni la population synthétique basée sur l'Enquête sur les revenus et dépenses de l'Allemagne. La présente étude a été financée par l'Economic and Social Research Council.

Bibliographie

Binder, D.A., et Théberge, A. (1988). Estimating the variance of raking ratio estimators. *Canadian Journal of Statistics*, 16, Supp. 47-55.

Brackstone, G.J., et Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyã*, Séries C, 41, 97-114.

Chang, T., et Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête (avec discussion). *Techniques d'enquête*, 30, 17-37.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-82.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-20.

Foster, K. (1998). Evaluating nonresponse on household surveys. *GSS Methodology Series*, 8, Office for National Statistics. Londres.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken : Wiley.

Fuller, W.A., Loughlin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

Ireland, C.T., et Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.

Kalton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kalton, G., et Maligalig, D.S. (1991). A comparison of methods for weighting adjustment for nonresponse. *Proceedings of the US Bureau of the Census 1991 Annual Research Conference*, 409-428.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.

Muennich, R., et Schulrle, J. (2003). Monte Carlo simulation study of European surveys, Workpackage 3, Deliverables 3.1 and 3.2. DACSEIS project. Disponible au http://www.uni-trier.de/index.php?id=29730.

Office for National Statistics (1998). *Labour Force Survey User Guide*, *Volume* 1: *Background and Methodology*. Londres.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, Angleterre.

Stukel, D.M., Hidiroglou, M.A. et Särndal, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.