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

Article

Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse

par Carl-Erik Särndal et Sixten Lundström

Décembre 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2010
Vol. 36, N° 2, pp. 141-156
Statistique Canada, N° 12-001-X au catalogue

Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse

Carl-Erik Särndal et Sixten Lundström [1]

Résumé

Le présent article décrit l'élaboration d'outils de calcul, appelés indicateurs, qui permettent de juger de l'efficacité de l'information auxiliaire utilisée pour contrôler le biais de non-réponse dans les estimations par sondage, obtenues ici par calage. L'étude est motivée par le contexte dans lequel sont réalisés les sondages dans plusieurs pays, surtout en Europe du Nord, où de nombreuses variables auxiliaires possibles concernant les ménages et les particuliers sont tirées de registres administratifs fiables. Un grand nombre de vecteurs auxiliaires pouvant donc être composés, il est nécessaire de les comparer afin de déterminer dans quelle mesure ils peuvent réduire le biais. Les indicateurs décrits dans le présent article sont conçus pour répondre à ce besoin. Ils sont utilisés dans les enquêtes réalisées par Statistics Sweden. Nous considérons des conditions générales d'enquête où un échantillon probabiliste est tiré de la population finie selon un plan d'échantillonnage arbitraire et où des cas de non réponse se produisent. La probabilité d'inclusion dans l'échantillon est connue pour chaque unité de la population ; la probabilité de réponse est inconnue, ce qui cause un biais. La variable étudiée (variable y) n'est observée que pour l'ensemble de répondants. Quel que soit le vecteur auxiliaire utilisé dans un estimateur par calage (ou dans toute autre méthode d'estimation), un biais résiduel persiste systématiquement. Le choix du vecteur auxiliaire (le meilleur possible) est guidé par les indicateurs proposés dans le présent article. Dans les premières sections, nous décrivons le contexte de leur élaboration et leurs caractéristiques de calcul, puis nous exposons leur contexte théorique. Les dernières sections sont consacrées aux études empiriques. L'une de ces études illustre la sélection des variables auxiliaires dans une enquête réalisée par Statistics Sweden. Une deuxième illustration empirique consiste en une simulation à partir d'une population finie synthétique ; un certain nombre de vecteurs auxiliaires possibles sont classés par ordre de préférence à l'aide des indicateurs.

Mots clés : Pondération par calage ; correction de la non-réponse ; biais de non-réponse ; variables auxiliaires ; indicateur de biais.

1. Introduction

De nos jours, on peut s'attendre à un taux de non-réponse élevé dans de nombreuses enquêtes, de sorte qu'il faut élaborer des méthodes qui permettent de réduire autant que possible le biais de non-réponse dans les estimations. Il faut donc disposer d'information auxiliaire puissante. Les fichiers de données administratives sont l'une des sources de ce genre d'information. À cet égard, les pays scandinaves et certains autres pays européens, notamment les Pays Bas, sont dans une situation avantageuse. De nombreuses variables auxiliaires possibles (appelées variables x) peuvent être tirées de registres administratifs de haute qualité dans lesquels les valeurs des variables auxiliaires sont spécifiées pour l'entièreté de la population. Les variables mesurant divers aspects de la collecte des données représentent une autre catégorie utile de données auxiliaires. Des mesures efficaces peuvent être prises pour contrôler le biais de non réponse. Au delà du plan d'échantillonnage, le *plan d'estimation* devient, dans ces pays, une composante importante du plan d'enquête global. Statistics Sweden a consacré des ressources considérables à l'élaboration de méthodes en vue de sélectionner les meilleures variables auxiliaires possibles.

De nombreux articles traitent de la pondération dans les sondages présentant une non réponse et de la sélection des « meilleures variables auxiliaires ». Eltinge et Yansaneh (1997), Kalton et Flores-Cervantes (2003), ainsi que Thomsen, Kleven, Wang et Zhang (2006) en sont des exemples. Une attention particulière est accordée à la pondération dans le cas des enquêtes par panel avec érosion du panel dans, par exemple, Rizzo, Kalton et Brick (1996), selon lesquels « le choix des variables auxiliaires est important, probablement plus important que celui de la méthode de pondération ». La revue effectuée par Kalton et Flores-Cervantes (2003) fournit de nombreuses références à des travaux antérieurs. Comme dans le présent article, Deville (2002) et Kott (2006) privilégient une approche de pondération par calage pour corriger la non-réponse.

Certaines méthodes antérieures sont des cas particuliers de la perspective adoptée dans le présent article, laquelle est fondée sur l'utilisation systématique d'information auxiliaire en effectuant un calage à deux niveaux. Récemment, la recherche d'une méthode de pondération efficace a pris deux directions, à savoir i) fournir des conditions plus générales que les méthodes populaires, mais limitées, de pondération par cellule et ii) quantifier la recherche de variables auxiliaires à l'aide d'indicateurs calculables.

1. Carl-Erik Särndal, professeur et Sixten Lundström, conseiller principal en méthodologie, Statistics Sweden. Courriel : carl.sarndal@scb.se.

Särndal et Lundström (2005, 2008) proposent ce genre d'indicateurs, tandis que Schouten (2007) adopte une perspective différente pour justifier un indicateur. Un article à consulter à ce sujet est celui de Schouten, Cobben et Bethlehem (2009).

Le présent article comporte quatre volets. Aux sections 2 à 4, nous exposons le contexte général de l'estimation en présence de non réponse. Aux sections 5 et 6, nous présentons les indicateurs pour le classement préférentiel des vecteurs **x** et nous discutons de leur calcul. Aux sections 7 et 8, nous présentons l'algèbre linéaire qui sous tend les indicateurs. Enfin, aux sections 9 et 10, nous présentons deux exemples empiriques. Le premier (section 9) s'appuie sur des données réelles provenant d'une grande enquête réalisée par Statistics Sweden. Le deuxième (section 10) décrit une simulation exécutée sur une population finie synthétique.

2. Estimateurs par calage pour une enquête avec non-réponse

Un échantillon probabiliste s est tiré de la population $U = \{1, 2, ..., k, ..., N\}$. Le plan d'échantillonnage donne à l'unité k la probabilité d'inclusion connue $\pi_k = \Pr(k \in s) > 0$ et le poids de sondage connu $d_k = 1/\pi_k$. Des cas de non-réponse ont lieu. L'ensemble de réponses r est un sous-ensemble de s ; la façon dont il a été généré est inconnue. Nous supposons que $r \subset s \subset U$, et que r est un ensemble non vide. Le taux de réponse (pondéré par les poids de sondage) est

$$P = \frac{\sum_r d_k}{\sum_s d_k} \tag{2.1}$$

(si A est un ensemble d'unités, $A \subseteq U$, la somme $\sum_{k \in A}$ s'écrira $\sum_A$). Habituellement, de nombreuses variables sont étudiées dans le cadre d'une enquête. Une variable étudiée typique, qu'elle soit continue ou catégorique, est désignée par y. Sa valeur pour l'unité k est y_k, enregistrée pour $k \in r$, et non disponible pour $k \in U - r$. Nous cherchons à estimer le total y de population, $Y = \sum_U y_k$. De nombreux paramètres d'intérêt dans la population finie sont des fonctions de plusieurs totaux, mais nous pouvons nous concentrer sur un seul d'entre eux.

L'information auxiliaire est de deux types, auxquels correspondent deux types de vecteurs, $\mathbf{x}_k^*$ et $\mathbf{x}_k^\circ$. L'*information auxiliaire de population* est transmise par $\mathbf{x}_k^*$, une valeur vectorielle connue pour chaque unité $k \in U$. Donc, $\sum_U \mathbf{x}_k^*$ est un total de population connu. Alternativement, nous permettons que $\sum_U \mathbf{x}_k^*$ soit importé d'une source extérieure et que $\mathbf{x}_k^*$ soit une valeur vectorielle connue (observée) pour chaque unité $k \in s$. L'*information auxiliaire d'échantillon* est transmise par $\mathbf{x}_k^\circ$, une valeur vectorielle connue (observée) pour chaque unité $k \in s$; le total $\sum_U \mathbf{x}_k^\circ$ est inconnu, mais est estimé sans biais par $\sum_s d_k \mathbf{x}_k^\circ$. La valeur vectorielle auxiliaire combinant les deux types de vecteurs est désignée $\mathbf{x}_k$. Ce vecteur et l'information qui y est associée sont

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}. \tag{2.2}$$

Le vecteur $(y_k, \mathbf{x}_k, \pi_k)$ est lié à la k^{e} unité. Ici, π_k est connue pour tout $k \in U$, y_k pour tout $k \in r$, la composante $\mathbf{x}_k^*$ de $\mathbf{x}_k$ fournit l'information de population et la composante $\mathbf{x}_k^\circ$ de $\mathbf{x}_k$ fournit l'information d'échantillon.

De nombreux vecteurs **x** peuvent être formés à l'aide des variables extraites des registres administratifs, des données sur les processus d'enquête ou d'autres sources. Parmi tous les vecteurs à notre disposition, nous voulons identifier celui qui est le plus susceptible de réduire le biais de non réponse, si ce n'est pas à une valeur nulle, au moins à une valeur quasi nulle.

Nous considérons les vecteurs ayant la propriété qu'il existe un vecteur non nul constant $\boldsymbol{\mu}$ tel que

$$\boldsymbol{\mu}' \mathbf{x}_k = 1 \text{ pour tout } k \in U \tag{2.3}$$

« Constant » signifie que $\boldsymbol{\mu} \neq \mathbf{0}$ ne dépend pas de *k*, ni de *s* ou de *r*. La condition (2.3) simplifie les démonstrations mathématiques et ne contraint pas sévèrement $\mathbf{x}_k$. En fait, la plupart des vecteurs **x** utiles en pratique sont couverts. À titre d'exemple, mentionnons : 1) $\mathbf{x}_k = (1, x_k)'$, où x_k est la valeur, pour l'unité *k*, d'une variable auxiliaire continue *x* ; 2) le vecteur représentant une variable *x* catégorique avec *J* classes mutuellement exclusives et exhaustives, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, ..., \gamma_{jk}, ..., \gamma_{Jk})'$, où $\gamma_{jk} = 1$ si *k* appartient au groupe *j*, et $\gamma_{jk} = 0$ autrement, $j = 1, 2, \ldots, J$; 3) le vecteur $\mathbf{x}_k$ utilisé pour codifier deux variables catégoriques, la dimension de $\mathbf{x}_k$ étant $J_1 + J_2 - 1$, où J_1 et J_2 sont les nombres respectifs de classes, et où le « moins un » a pour but d'éviter une singularité dans le calcul des poids calés sur les deux matrices des dénombrements marginaux ; 4) l'extension de 3) à plus de deux variables catégoriques. Les vecteurs de type 3) et 4) sont particulièrement importants pour la production de statistiques par les organismes statistiques (le choix $\mathbf{x}_k = x_k$, non couvert par (2.3), mène à l'estimateur de la non réponse par le ratio qui a la réputation d'être habituellement un mauvais choix pour contrôler le biais de non réponse comparativement à $\mathbf{x}_k = (1, x_k)'$, si bien qu'exclure l'estimateur par le ratio ne constitue pas une grande perte).

L'estimateur par calage de $Y = \sum_U y_k$, calculé sur les données y_k pour $k \in r$, est

$$\hat{Y}_{\text{CAL}} = \sum_r w_k y_k \tag{2.4}$$

avec $w_k = d_k\{1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k\}$. Les poids w_k sont calés sur deux types d'information : $\sum_r w_k \mathbf{x}_k = \mathbf{X}$, qui implique que $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$ et $\sum_r w_k \mathbf{x}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$. Nous supposons tout au long de l'exposé que la matrice symétrique $\sum_r d_k \mathbf{x}_k \mathbf{x}_k'$ est non singulière (pour des raisons de calcul, il est prudent d'imposer une contrainte plus forte : la matrice ne doit pas être mal conditionnée, ou presque singulière). Étant donné (2.3), nous avons $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$ avec les poids $w_k = d_k v_k$, où $v_k = \mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$. Les poids satisfont $\sum_r d_k v_k \mathbf{x}_k = \mathbf{X}$, où $\mathbf{X}$ contient l'une des composantes de (2.2), ou les deux.

Un estimateur par calage étroitement apparenté est celui basé sur le même vecteur à deux composantes $\mathbf{x}_k$, mais avec le calage uniquement au niveau de l'échantillonnage :

$$\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k \tag{2.5}$$

où

$$m_k = \left(\sum_s d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k. \tag{2.6}$$

L'équation de calage se lit alors $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$, où $\mathbf{x}_k$ comprend les deux composantes données en (2.2). Le vecteur auxiliaire $\mathbf{x}_k$ sert à réaliser deux objectifs : obtenir une faible variance et un faible biais de non réponse. Du point de vue de la variance uniquement, $\hat{Y}_{\text{CAL}}$ est habituellement préféré à $\tilde{Y}_{\text{CAL}}$, parce que le premier bénéficie de l'apport d'un total de population connu $\sum_U \mathbf{x}_k^*$. Toutefois, comme la présente étude porte sur le biais, cela revient pour ainsi dire au même d'utiliser $\hat{Y}_{\text{CAL}}$ ou $\tilde{Y}_{\text{CAL}}$, et nous nous concentrons sur le second. Sous des conditions libérales, la différence entre le biais de $N^{-1}\hat{Y}_{\text{CAL}}$ et celui de $N^{-1}\tilde{Y}_{\text{CAL}}$ est d'ordre n^{-1}, et a donc peu de conséquences pratiques, même pour des tailles d'échantillon n modestes, comme il est discuté, par exemple, dans Särndal et Lundström (2005).

L'estimateur (2.5) peut aussi s'exprimer sous la forme

$$\tilde{Y}_{\text{CAL}} = \left(\sum_s d_k \mathbf{x}_k\right)' \mathbf{B}_{\mathbf{x}} \tag{2.7}$$

où

$$\mathbf{B}_{\mathbf{x}} = \mathbf{B}_{\mathbf{x}|r;d} = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k \mathbf{x}_k y_k \tag{2.8}$$

est le vecteur des coefficients de régression résultant de l'ajustement par les moindres carrés (pondérés par d_k) sur les données $(y_k, \mathbf{x}_k)$ pour $k \in r$.

Remarque concernant la notation : Lorsque cela est nécessaire pour insister sur un fait, un symbole possède deux indices séparés par un point virgule. Le premier indique l'ensemble d'unités sur lequel la quantité est calculée et le second, la pondération, comme dans $\mathbf{B}_{\mathbf{x}|r;d}$ donné par (2.8), et dans les moyennes pondérées telles que $\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k$. Si la pondération est uniforme, le second indice est abandonné, comme dans $\bar{y}_U = (1/N)\sum_U y_k$.

3. Points de référence

Le choix de vecteur le plus élémentaire est la valeur unitaire constante, $\mathbf{x}_k = 1$ pour tout k. Bien que ce vecteur soit inefficace en ce qui concerne la réduction du biais de non réponse, il sert de valeur repère. Alors, $m_k = 1/P$ pour tout k, où P est le taux de réponse à l'enquête (2.1) et $\tilde{Y}_{\text{CAL}}$ est l'estimateur à facteur d'extension (*expansion estimator*) :

$$\tilde{Y}_{\text{EXP}} = (1/P)\sum_r d_k y_k = \hat{N}\,\bar{y}_{r;d} \tag{3.1}$$

où $\hat{N} = \sum_s d_k$ est sans biais sous le plan pour la taille de population N. Le biais de $\tilde{Y}_{\text{EXP}}$ peut être important.

À l'autre extrémité du spectre de biais se trouvent les estimateurs sans biais, ou presque sans biais, qui peuvent être obtenus sous réponse complète, quand $r = s$. Il s'agit d'estimateurs hypothétiques, non calculables en présence de non réponse. Parmi eux figure l'estimateur GREG avec les poids calés sur le total de population connu $\sum_U \mathbf{x}_k^*$,

$$\hat{Y}_{\text{FUL}} = \sum_s d_k g_k y_k$$

où $g_k = 1 + (\sum_U \mathbf{x}_k^* - \sum_s d_k \mathbf{x}_k^*)'(\sum_s d_k \mathbf{x}_k^* \mathbf{x}_k^{*'})^{-1}\mathbf{x}_k^*$, et FUL désigne la réponse complète (*full response*). L'estimateur HT sans biais (obtenu quand $g_k = 1$ pour tout k) s'écrit

$$\tilde{Y}_{\text{FUL}} = \sum_s d_k y_k = \hat{N}\,\bar{y}_{s;d}. \tag{3.2}$$

Il ne tient pas compte de l'information $\sum_U \mathbf{x}_k^*$, qui peut être importante pour la réduction de la variance. Cependant, pour l'étude du biais exposée ici, nous pouvons utiliser indifféremment $\hat{Y}_{\text{FUL}}$ ou $\tilde{Y}_{\text{FUL}}$. La différence de biais entre les deux a peu d'importance, même pour des tailles d'échantillon modestes. Nous pouvons donc nous concentrer sur $\tilde{Y}_{\text{FUL}}$.

4. Le ratio des biais

Pour un résultat donné (s, r), considérons les estimations $\tilde{Y}_{\text{CAL}}$, $\tilde{Y}_{\text{EXP}}$ et $\tilde{Y}_{\text{FUL}}$, données par (2.5), (3.1) et (3.2), comme trois points sur un axe horizontal. Les estimations $\tilde{Y}_{\text{EXP}}$ (produite par le vecteur élémentaire $\mathbf{x}_k = 1$) et $\tilde{Y}_{\text{CAL}}$ (produite par un meilleur vecteur $\mathbf{x}$) sont calculables, mais biaisées. À mesure que s'améliore le vecteur $\mathbf{x}$, $\tilde{Y}_{\text{CAL}}$ s'écarte de $\tilde{Y}_{\text{EXP}}$ et peut se rapprocher de l'estimation $\tilde{Y}_{\text{FUL}}$ sans biais idéale, mais non réalisée. Nous considérons par conséquent trois écarts : $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$, $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}$ et

$\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$, parmi lesquels seul celui du milieu est calculable. L'« écart total » inconnu, $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$, peut être décomposé en un « écart expliqué » (par le vecteur **x** choisi) plus un « écart restant » :

$$\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) + (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}). \qquad (4.1)$$

S'ils étaient calculables, l'écart $\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$ serait particulièrement intéressant en tant qu'estimation du biais persistant dans $\tilde{Y}_{\text{CAL}}$ (et dans $\hat{Y}_{\text{CAL}}$), tandis que l'écart $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$ estimerait le biais habituellement beaucoup plus grand de l'estimation repère, $\tilde{Y}_{\text{EXP}}$. Le ratio des biais pour un résultat donné (*s*, *r*) détermine le biais estimé de $\tilde{Y}_{\text{CAL}}$ par rapport à celui de $\hat{Y}_{\text{EXP}}$:

$$\text{ratio des biais} = \frac{\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}}{\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}}. \qquad (4.2)$$

Nous normalisons les trois écarts en utilisant comme facteur la taille estimée de la population $\hat{N} = \sum_s d_k$ et utilisons la notation $\Delta_T = \Delta_A + \Delta_R$, où *T* signifie « total », *A* signifie « expliqué (*accounted for*) » et *R* signifie « restant ». En notant que $\sum_r d_k (y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}}) = 0$, nous avons

$$\Delta_T = \hat{N}^{-1}(\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}) = \bar{y}_{r;d} - \bar{y}_{s;d} \ ;$$

$$\Delta_R = \hat{N}^{-1}(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) = \bar{\mathbf{x}}_{s;d}' \mathbf{B}_{\mathbf{x}} - \bar{y}_{s;d}$$

$$\Delta_A = \hat{N}^{-1}(\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}}$$

où $\bar{\mathbf{x}}_{s;d} = \sum_s d_k \mathbf{x}_k / \sum_s d_k$, $\bar{\mathbf{x}}_{r;d} = \sum_r d_k \mathbf{x}_k / \sum_r d_k$, et $\bar{y}_{r;d}$ et $\bar{y}_{r;d}$ sont les moyennes définies de manière analogue pour la variable *y*. Alors, (4.2) prend la forme :

$$\text{ratio des biais} = \frac{\Delta_R}{\Delta_T} = 1 - \frac{\Delta_A}{\Delta_T} = 1 - \frac{(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}}}{\bar{y}_{r;d} - \bar{y}_{s;d}}. \qquad (4.3)$$

Pour le vecteur élémentaire $\mathbf{x}_k = 1$, le ratio des biais =1. Idéalement, nous voulons que le vecteur auxiliaire $\mathbf{x}_k$ pour $\tilde{Y}_{\text{CAL}}$ donne un ratio des biais ≈ 0. Pour un résultat donné (*s*, *r*) et une variable *y* donnée, nous progressons dans cette direction en trouvant un vecteur **x** qui rend le numérateur calculable $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}}$ grand (en valeur absolue), ce qui est réalisable. Cependant, quel que soit le vecteur **x** que nous choisissons finalement, le biais restant dans $\tilde{Y}_{\text{CAL}}$ est inconnu. Même en utilisant le vecteur **x** le meilleur possible, un biais important peut persister. Nous avons donc tenté de trouver la meilleure solution possible, dans des conditions peut-être défavorables.

En résumé, pour un résultat donné (*s*, *r*) et une variable *y* donnée, les trois écarts possèdent les caractéristiques suivantes : i) $\Delta_T = \bar{y}_{r;d} - \bar{y}_{s;d}$ est une valeur constante inconnue, qui dépend des valeurs de *y* non observées ainsi qu'observées, ii) Δ_A est calculable ; sa valeur dépend de y_k pour $k \in r$ et des valeurs de $\mathbf{x}_k$ pour $k \in s$ pour le vecteur **x** choisi, iii) Δ_R ne peut pas être calculé ; sa valeur dépend des valeurs non observées y_k, et de $\mathbf{x}_k$ pour $k \in s$.

Afin de suivre l'évolution des estimations quand le vecteur **x** s'améliore, considérons un résultat donné (*s*, *r*). L'écart Δ_T peut avoir n'importe quel signe. Supposons que $\Delta_T > 0$, qui indique un biais positif dans $\tilde{Y}_{\text{EXP}}$, comme dans le cas où de grandes unités manifestent une plus grande propension à répondre que les petites. À mesure que le vecteur **x** dans $\tilde{Y}_{\text{CAL}}$ devient plus puissant grâce à l'inclusion d'un nombre croissant de variables *x*, Δ_A a tendance à croître et à s'écarter de zéro et, idéalement, s'approchera de Δ_T, indiquant une proximité souhaitée de $\tilde{Y}_{\text{CAL}}$ et de l'estimation $\tilde{Y}_{\text{FUL}}$ sans biais. Aussi longtemps que le vecteur **x** demeure relativement faible, il est vraisemblable que l'inégalité $\Delta_A < \Delta_T$ soit vérifiée. À mesure que la puissance du vecteur **x** s'accroît, Δ_A se rapproche de l'écart fixe Δ_T, signe que le biais devient presque nul. Il pourrait même « aller au-delà », de sorte qu'un « surajustement », $\Delta_A > \Delta_T$, se produira. Cette situation n'est pas nuisible ; quoique $\Delta_R = \Delta_T - \Delta_A$ est alors négatif, il est ordinairement petit (l'analyste ne peut travailler qu'avec Δ_A ; il sait pas quand Δ_A et Δ_T sont proches, ni si le surajustement $\Delta_A > \Delta_T$ a eu lieu). La simulation décrite à la section 10 illustre ces points. Si $\Delta_T < 0$, ces tendances sont inversées.

La forme de (4.3) peut évoquer un argument susceptible d'être incorrect. Supposons que l'on ait proposé un vecteur $\mathbf{x}_k$, contenant des variables considérées comme efficaces et que l'on ait émis l'hypothèse que $y_k = \boldsymbol{\beta}' \mathbf{x}_k + \varepsilon_k$, où ε_k est un petit résidu. Alors, $\bar{y}_{r;d} - \bar{y}_{s;d} \approx (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}} \approx (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \boldsymbol{\beta}$, et conséquemment, le ratio des biais ≈ 0, ce qui communique le message, souvent faux, que le vecteur postulé $\mathbf{x}_k$ est efficace. L'une des faiblesses de l'argument découle du fait bien connu que la non réponse (à moins qu'elle soit entièrement aléatoire) induira un biais dans $\mathbf{B}_{\mathbf{x}}$ pour un vecteur de régression qui décrit la relation de *y* en fonction de **x** dans la population. D'autres commentaires à ce sujet sont donnés à la section 8.

Enfin, il faut tenir compte du fait qu'en pratique, une enquête porte habituellement sur de nombreuses variables *y*. À chaque variable *y* correspond un estimateur par calage et un ratio des biais donné par (4.3). Le vecteur **x** idéal est celui qui serait capable de contrôler le biais dans tous ces estimateurs, ce qui est habituellement impossible sans compromis, comme nous en discutons plus loin.

5. Expression de l'écart expliqué

L'unité répondante k reçoit le poids $d_k m_k$ dans l'estimateur $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$. Le facteur de correction de la non réponse $m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ accroît le poids de sondage d_k. Nous pouvons considérer m_k comme la valeur d'une variable dérivée, définie pour un résultat (r, s) et un choix de $\mathbf{x}_k$, particuliers, indépendante de toutes les variables y d'intérêt, et calculable pour $k \in s$ (mais utilisée dans $\tilde{Y}_{\text{CAL}}$ uniquement pour $k \in r$). En utilisant (2.3), nous avons

$$\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k \; ; \sum_r d_k m_k = \sum_s d_k \; ; \quad \sum_r d_k m_k^2 = \sum_s d_k m_k. \tag{5.1}$$

Deux moyennes pondérées sont nécessaires :

$$\bar{m}_{r\,;\,d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} = \frac{1}{P} \; ; \bar{m}_{s\,;\,d} = \frac{\sum_s d_k m_k}{\sum_s d_k} \tag{5.2}$$

où P est le taux de réponse (2.1). Donc, le facteur de correction moyen dans $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$ est $1/P$, quel que soit le choix du vecteur $\mathbf{x}$. La capacité que possèdera un vecteur $\mathbf{x}$ choisi de réduire efficacement ou non le biais dépendra des moments d'ordre plus élevé de m_k. La variance pondérée de m_k est donnée par

$$S_m^2 = S_{m|r\,;\,d}^2 = \sum_r d_k (m_k - \bar{m}_{r\,;\,d})^2 \Big/ \sum_r d_k. \tag{5.3}$$

Nous utiliserons la notation simplifiée S_m^2. Un développement de (5.3), et l'utilisation de (5.1) et de (5.2) donnent

$$S_m^2 = \bar{m}_{r\,;\,d}(\bar{m}_{s\,;\,d} - \bar{m}_{r\,;\,d}). \tag{5.4}$$

Le coefficient de variation de m_k est

$$\text{cv}_m = \frac{S_m}{\bar{m}_{r\,;\,d}} = \sqrt{\frac{\bar{m}_{s\,;\,d}}{\bar{m}_{r\,;\,d}} - 1}. \tag{5.5}$$

La variance pondérée de la variable étudiée y est donnée par

$$S_y^2 = S_{y|r\,;\,d}^2 = \sum_r d_k (y_k - \bar{y}_{r\,;\,d})^2 \Big/ \sum_r d_k \tag{5.6}$$

(quand les probabilités de réponse ne sont pas toutes égales, $S_y^2 = S_{y|r\,;\,d}^2$ n'est pas sans biais pour la variance de population $S_{y|U}^2$, mais il ne s'agit pas d'un problème pour les dérivations qui suivent). Nous avons besoin de la covariance

$$\text{Cov}(y,m) = \text{Cov}(y,m)_{r\,;\,d} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r\,;\,d})(y_k - \bar{y}_{r\,;\,d}) \tag{5.7}$$

et du coefficient de corrélation $R_{y,m} = \text{Cov}(y,m)/(S_y S_m)$, satisfaisant $-1 \le R_{y,m} \le 1$.

L'écart $\Delta_A = (\bar{\mathbf{x}}_{r\,;\,d} - \bar{\mathbf{x}}_{s\,;\,d})'\mathbf{B}_{\mathbf{x}}$ est une composante essentielle du ratio des biais (4.3). Nous recherchons un vecteur $\mathbf{x}$ qui rend Δ_A grand. Les facteurs qui déterminent Δ_A figurent dans les expressions (5.8) à (5.10). Les outils de calcul (indicateurs) destinés à faciliter la recherche des variables x efficaces sont donnés en (5.11) et (5.12). Leur dérivation par l'algèbre linéaire est reportée à la section 7, que peuvent omettre les lecteurs qui s'intéressent surtout à l'utilisation pratique de ces outils pour trouver les variables x, comme l'illustre empiriquement les sections 9 et 10. Nous pouvons décomposer Δ_A/S_y en facteurs comme il suit

$$\Delta_A/S_y = -R_{y,m} \times \text{cv}_m. \tag{5.8}$$

Deux facteurs multiplicatifs simples déterminent Δ_A/S_y : le coefficient de variation cv_m, qui est indépendant de y_k et calculé sur le vecteur $\mathbf{x}_k$ connu uniquement, et le coefficient de corrélation (positif ou négatif) $R_{y,m}$. Une autre décomposition en facteurs basée sur des concepts simples est

$$\Delta_A / S_y = F \times R_{y,\mathbf{x}} \times \text{cv}_m \tag{5.9}$$

où $R_{y,\mathbf{x}} = \sqrt{R_{y,\mathbf{x}}^2}$ est le coefficient de corrélation multiple entre y et $\mathbf{x}$, $R_{y,\mathbf{x}}^2$ est la proportion de la variance S_y^2 de y expliquée par le prédicteur $\mathbf{x}$, et $F = -R_{y,m}/R_{y,\mathbf{x}}$ (la formule (7.8) donne l'expression précise pour $R_{y,\mathbf{x}}^2$). Comme nous le montrons également à la section 7, $|R_{y,m}| \le R_{y,\mathbf{x}}$ pour tout vecteur $\mathbf{x}$ et toute variable y ; par conséquent, $-1 \le F \le 1$.

Dans (5.8) et (5.9), cv_m et $R_{y,\mathbf{x}}$ sont des termes non négatifs, tandis que $R_{y,m}$ et F peuvent prendre n'importe quel signe (ou éventuellement être nuls). Donc,

$$|\Delta_A| / S_y = |R_{y,m}| \times \text{cv}_m = |F| \times R_{y,\mathbf{x}} \times \text{cv}_m. \tag{5.10}$$

Les termes $S_y, \text{cv}_m, R_{y,\mathbf{x}}, R_{y,m}$ et F se calculent tous facilement d'après les données d'enquête. Les termes cv_m et $R_{y,\mathbf{x}}$ augmentent tous deux (ou demeurent éventuellement invariables) quand des variables x supplémentaires sont ajoutées au vecteur $\mathbf{x}$; $R_{y,m}$ ne possède pas cette propriété.

Illustrons cela à l'aide de chiffres assez habituels. Si $F = 0,5$; $R_{y,\mathbf{x}} = 0,6$ et $\text{cv}_m = 0,4$, alors $\Delta_A/S_y = 0,12$, ce qui implique que $\tilde{Y}_{\text{CAL}}/N = \tilde{Y}_{\text{EXP}}/N - 0,12 \times S_y$. Autrement dit, la moyenne de y estimée $\tilde{Y}_{\text{CAL}}/\hat{N}$ a été rajustée à la baisse d'un facteur égal à 0,12 écart-type par rapport à l'estimation élémentaire $\tilde{Y}_{\text{EXP}}/\hat{N}$. La correction peut être importante comparativement à l'écart-type de la moyenne estimée de y, surtout quand la taille de l'échantillon est de

l'ordre des milliers. Il reste à savoir si cette correction a éliminé ou non la plupart du biais dû à la non réponse.

Il découle de (5.8) que $0 \le |\Delta_A|/S_y \le \text{cv}_m$, quelle que soit la variable y. L'inégalité $|\Delta_A|/S_y \le R_{y,\mathbf{x}} \times \text{cv}_m$ est meilleure, mais elle dépend de la variable y. En outre, si le ratio de corrélation F demeure approximativement constant quand le vecteur $\mathbf{x}$ change, de sorte que $F \approx F_0$, alors $|\Delta_A|/S_y \approx |F_0| \times R_{y,\mathbf{x}} \times \text{cv}_m$.

Bien qu'il soit calculable pour tout vecteur $\mathbf{x}$ et tout résultat (s, r), Δ_A ne révèle pas la valeur du ratio des biais. Cependant, Δ_A suggère des outils de calcul, appelés indicateurs, pour comparer divers vecteurs $\mathbf{x}$. Partant de (5.8), posons que

$$H_0 = \Delta_A / S_y = -R_{y,m} \times \text{cv}_m. \qquad (5.11)$$

Comme le prouve la théorie exposée à la section 8 et les travaux empiriques présentés à la section 10, sur une longue série de résultats (s, r), la moyenne de H_0 concorde avec l'écart moyen $\tilde{Y}_{\text{CAL}} - Y$ (qui mesure le biais de $\tilde{Y}_{\text{CAL}}$) de manière presque parfaitement linéaire quand le vecteur $\mathbf{x}$ change. Il en est ainsi indépendamment de la distribution des réponses qui génère r à partir de s. Puisque H_0 peut prendre l'un ou l'autre signe, il est pratique de travailler avec sa valeur absolue désignée par H_1 ; en plus, nous considérons deux autres indicateurs, H_2 et H_3, inspirés de (5.9) et (5.10) :

$$H_1 = |\Delta_A|/S_y = |R_{y,m}| \times \text{cv}_m \ ;$$

$$H_2 = R_{y,\mathbf{x}} \times \text{cv}_m \ ; H_3 = \text{cv}_m. \qquad (5.12)$$

Nos principales alternatives sont H_1 et H_3. De celles-ci, H_1 est motivé par son lien direct avec Δ_A, que nous voulons rendre plus grand, pour une variable y donnée. Une bonne raison de considérer H_3 est son indépendance à l'égard de toutes les variables y de l'enquête. L'indicateur H_2 est une alternative *ad hoc* ; bien qu'il contienne un concept connu, le coefficient de corrélation multiple $R_{y,\mathbf{x}}$, il est moins approprié que H_1, parce que le ratio des coefficients de corrélation $F = -R_{y,m}/R_{y,\mathbf{x}}$ peut varier considérablement d'un vecteur $\mathbf{x}$ à l'autre. Aussi bien H_2 que H_3 augmentent quand d'autres variables x sont ajoutées au vecteur $\mathbf{x}$, ce qui n'est généralement pas vérifié pour H_1. L'utilisation de ces indicateurs est illustrée aux sections 9 et 10 décrivant les travaux empiriques.

6. Classement préférentiel des vecteurs auxiliaires

Les méthodes décrites dans le présent article sont destinées à être utilisées principalement avec les grands échantillons caractéristiques des enquêtes gouvernementales. Habituellement, la taille de l'échantillon est beaucoup plus grande que la dimension du vecteur $\mathbf{x}$. La variance des estimations est généralement faible comparativement au carré du biais. Néanmoins, pour les variables auxiliaires catégoriques, aucune taille de groupe « trop petite » ne devrait être permise. Il est recommandé que toutes les tailles de groupe valent au moins 30, voire au moins 50, afin d'éviter l'instabilité. Le croisement des variables catégoriques (pour permettre les interactions) comporte un certain risque qu'il existe des petits groupes. Il est préférable de caler sur les dénombrements marginaux, plutôt que sur les fréquences pour les cellules croisées de faible fréquence.

Dans un certain nombre de pays, les nombreux registres administratifs existants constituent une riche source d'information auxiliaire, particulièrement pour les enquêtes auprès des particuliers et des ménages. Ces registres contiennent de nombreuses variables x possibles parmi lesquelles choisir. Un grand nombre de vecteurs $\mathbf{x}$ différents peuvent donc être composés. Les indicateurs donnés par (5.12) offrent des outils de calcul pour obtenir un classement préférentiel des vecteurs $\mathbf{x}$ possibles, l'objectif étant de réduire autant que possible le biais qui persiste dans l'estimateur par calage.

Scénario 1 : Concentrons nous sur une variable y particulière. Le biais persistant dans l'estimateur par calage dépend de la variable y ; certaines variables sont plus sujettes au biais que d'autres. Nous identifions une variable y particulière considérée comme étant très importante dans l'enquête, et nous cherchons à déterminer un vecteur $\mathbf{x}$ qui réduit autant que possible le biais pour cette variable (s'il faut prendre en considération plus d'une variable y, un compromis doit être trouvé, ce que suggère le scénario 2 qui suit). Dans le présent exemple, nous utilisons l'indicateur $H_1 = |\Delta_A|/S_y = |R_{y,m}| \times \text{cv}_m$ qui dépend de la variable y et choisissons le vecteur $\mathbf{x}$ de façon à rendre H_1 grand. Une alternative *ad hoc* consiste à utiliser l'indicateur $H_2 = R_{y,\mathbf{x}} \times \text{cv}_m$ et à s'efforcer de le rendre aussi grand que possible.

Scénario 2 : L'objectif est de déterminer un vecteur $\mathbf{x}$ d'usage général, efficace pour toutes les variables y de l'enquête, ou pour la plupart d'entre elles. Cela suggère d'opter pour $H_3 = \text{cv}_m$ comme indicateur de compromis et de choisir le vecteur $\mathbf{x}$ qui maximise H_3. Dans ce même sens, Särndal et Lundström (2005, 2008) ont utilisé l'indicateur $S_m^2 = H_3^2 / P^2$. Ils ont montré que la variable dérivée m_k dans (2.6) peut être considérée comme un prédicteur de l'inverse de la probabilité de réponse inconnue et que choisir le vecteur $\mathbf{x}$ de manière à rendre S_m^2 grand signale une réduction du biais dans l'estimateur par calage, indépendamment de la variable y.

Pour chaque scénario, nous distinguons deux procédures :

Procédure englobant tous les vecteurs : Une liste de vecteurs **x** possibles est dressée en se basant sur un jugement approprié. Nous calculons l'indicateur choisi pour chaque vecteur **x** possible et choisissons celui qui donne la valeur la plus élevée de l'indicateur. Le vecteur **x** résultant ne sera pas nécessairement le même pour H_1 (qui a pour cible une variable *y* particulière) que pour H_3 (qui recherche un compromis pour l'ensemble des variables *y* de l'enquête).

Procédure pas à pas (*ou Stepwise*) : Il existe un réservoir de variables *x* disponibles. Nous construisons le vecteur **x** par sélection pas à pas ascendante (ou sélection pas à pas descendante) parmi les variables *x* disponibles, une variable à la fois, en nous basant sur les variations successives (si elles sont considérées suffisamment importantes) de la valeur de l'indicateur choisi pour déterminer l'inclusion (ou l'exclusion) d'une variable *x* donnée à une étape particulière. En général, les indicateurs H_1, H_2 et H_3 ne produisent pas la même sélection de variables. Considérons deux vecteurs **x**, $\mathbf{x}_{1k}$ et $\mathbf{x}_{2k}$, tels que $\mathbf{x}_{2k}$ est composé de $\mathbf{x}_{1k}$ et d'un vecteur supplémentaire $\mathbf{x}_{+k}$: $\mathbf{x}_{2k} = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$. Le passage de $\mathbf{x}_{1k}$ à $\mathbf{x}_{2k}$ accroîtra la valeur de H_2 et de H_3. À chaque étape d'une méthode de sélection ascendante, nous sélectionnons la variable qui produit l'accroissement le plus important de H_2 ou de H_3. Toutefois, la transition ne garantit pas un accroissement de la valeur de l'indicateur le plus approprié, H_1. Néanmoins, H_1 peut être utilisé dans la sélection pas à pas de la manière décrite à la section 9.

7. Démonstrations

Pour une variable *y* et un résultat (*s*, *r*) donnés, nous recherchons un vecteur **x** qui rend grand, en valeur absolue, le numérateur calculable $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_\mathbf{x}$ du ratio des biais (4.3). À la présente section, nous prouvons les décompositions en facteurs $\Delta_A / S_y = -R_{y,m} \times \mathrm{cv}_m = F \times R_{y,\mathbf{x}} \times \mathrm{cv}_m$ données par (5.8) et (5.9). Nous commençons par noter que cv_m^2 est une forme quadratique dans le vecteur qui contraste la moyenne de **x** dans l'ensemble de réponses *r* avec la moyenne de **x** dans l'échantillon *s*. Soit

$$\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d} \; ; \; \mathbf{\Sigma} = \sum_r d_k \mathbf{x}_k \mathbf{x}'_k \Big/ \sum_r d_k . \qquad (7.1)$$

Alors, avec *P* donné par (2.1),

$$\mathrm{cv}_m^2 = P^2 \times S_m^2 = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}. \qquad (7.2)$$

Cette expression découle de (5.3) et d'une conséquence de (2.3), à savoir

$$\bar{\mathbf{x}}'_{r;d}\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}'_{r;d}\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{s;d} = 1. \qquad (7.3)$$

Le vecteur des covariances avec la variable étudiée *y* est donné par

$$\mathbf{C} = \Big(\sum_r d_k (\mathbf{x}_k - \bar{\mathbf{x}}_{r;d})(y_k - \bar{y}_{r;d})\Big) \Big/ \Big(\sum_r d_k\Big). \qquad (7.4)$$

Nous pouvons alors écrire Δ_A sous une forme bilinéaire :

$$\Delta_A = \mathbf{D}'\mathbf{B}_\mathbf{x} = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C} \qquad (7.5)$$

en utilisant le fait que $\mathbf{D}'\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'$ $\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = 0$ en vertu de (7.3).

Une perspective utile de Δ_A nous est fournie par l'interprétation géométrique de **C** et **D** figurant dans (7.5) en tant que vecteurs dans l'espace dont la dimension est celle de $\mathbf{x}_k$. Nous avons

$$\Delta_A = \Lambda\, (\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2} (\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2} \qquad (7.6)$$

où

$$\Lambda = \frac{\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C}}{(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2} (\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2}}. \qquad (7.7)$$

Pour une variable *y* particulière et un vecteur **x** particulier, les quantités scalaires $(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2}$ et $(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2}$ représentent les longueurs vectorielles respectives de **D** et **C** (à la suite d'une transformation octogonale basée sur les vecteurs propres et les valeurs propres de $\mathbf{\Sigma}^{-1}$). La quantité scalaire Λ représente le cosinus de l'angle formé par **D** (qui est indépendant de *y*) et **C** (qui dépend de *y*) ; d'où $-1 \le \Lambda \le 1$.

Quand le vecteur auxiliaire $\mathbf{x}_k$ peut s'agrandir par ajout de variables *x* supplémentaires, les deux longueurs vectorielles $(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2}$ et $(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2}$ augmentent. La variation de l'angle Λ peut se faire dans l'une ou l'autre direction ; Si $|\Lambda|$ demeure approximativement constant, (7.6) montre que $|\Delta_A|$ augmentera.

Une deuxième perspective utile concernant Δ_A découle de la décomposition de la variabilité totale de la variable étudiée *y*, $\sum_r d_k (y_k - \bar{y}_{r;d})^2 = (\sum_r d_k)\, S_y^2$. Nous devons examiner l'ajustement de deux régression, celle de *y* sur les vecteurs auxiliaires **x**, et celle de *y* sur la variable dérivée *m* définie par (2.6). À chaque ajustement correspond une décomposition de S_y^2 en une variation expliquée de *y* et une variation résiduelle de *y*. Les deux parties expliquées possèdent des liens importants avec le ratio des biais (4.3). Le résultat 7.1 résume les deux décompositions.

Résultat 7.1. Pour un résultat d'enquête donné (*s*, *r*), soit **D**, $\mathbf{\Sigma}$ et **C** donnés par (7.1) et (7.4). Alors, la proportion de la variance S_y^2 de *y* expliquée par la régression de *y* sur **x** est

$$R^2_{y,\mathbf{x}} = (\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/S^2_y. \qquad (7.8)$$

Le coefficient de corrélation entre y et le prédicteur univarié m est

$$R_{y,m} = -(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/[(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2} \times S_y]. \qquad (7.9)$$

Par conséquent, la proportion de S^2_y expliquée par m est

$$R^2_{y,m} = (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2/[(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}) \times S^2_y]. \qquad (7.10)$$

Les proportions $R^2_{y,\mathbf{x}}$ et $R^2_{y,m}$ satisfont $R^2_{y,m} \le R^2_{y,\mathbf{x}} \le 1$.

Preuve. La preuve de (7.8) s'appuie sur la régression par les moindres carrés pondérés de y sur $\mathbf{x}$ ajustée sur r. Les résidus sont $y_k - \hat{y}(\mathbf{x})_k$, où $\hat{y}(\mathbf{x})_k = \mathbf{x}'_k\mathbf{B}_\mathbf{x}$ avec $\mathbf{B}_\mathbf{x}$ donné par (2.8). La décomposition est

$$\sum_r d_k (y_k - \bar{y}_{r;d})^2 = \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 + \sum_r d_k (y_k - \hat{y}(\mathbf{x})_k)^2.$$

Le terme mixte est nul. Un développement du terme représentant la « variation expliquée » donne $\sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 = (\sum_r d_k)\, \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$. Par conséquent, la proportion expliquée de la variance est $R^2_{y,\mathbf{x}} = \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2/[(\sum_r d_k)\, S^2_y] = \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}/S^2_y$, comme l'affirme (7.8). Pour démontrer (7.9), nous notons que la covariance (5.7) peut s'écrire, avec l'aide de (7.5), sous la forme

$$\mathrm{Cov}(y, m) = -\Delta_A/P = -\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}/P.$$

Il découle alors de (7.2) que $R_{y,m} = \mathrm{Cov}(y, m)/(S_y S_m)$ a pour expression (7.9). Les résidus de la régression (avec constante) de y sur la variable explicative univariée m sont $\hat{y}(m)_k = \bar{y}_{r;d} + B_m(m_k - \bar{m}_{r;d})$ avec $B_m = \mathrm{Cov}(y, m)/S^2_m = -P\,(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})$. La proportion expliquée de la variance est $\sum_r d_k (\hat{y}(m)_k - \bar{y}_{r;d})^2/[(\sum_r d_k)\, S^2_y]$, qui par développement donne l'expression de $R^2_{y,m}$ correspondant à (7.10). Enfin, $R^2_{y,m} \le R^2_{y,\mathbf{x}}$ découle de l'inégalité de Cauchy Schwarz pour une forme bilinéaire : $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2 \le (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})$.

L'inégalité $R^2_{y,m} \le R^2_{y,\mathbf{x}} \le 1$ peut également être déduite du fait que, parmi toutes les prédictions $\hat{y}_k = \mathbf{x}'_k\boldsymbol{\beta}$ linéaires dans le vecteur $\mathbf{x}$, celles qui maximisent la variance expliquée sont $\hat{y}(\mathbf{x})_k = \mathbf{x}'_k\mathbf{B}_\mathbf{x}$, si bien que les prédictions $\hat{y}(m)_k$, qui sont linéaires dans $\mathbf{x}_k$ par la voie de m_k, ne peuvent pas produire une plus grande variance expliquée que ce maximum.

Or, de (7.9), (7.2) et (7.5), il découle que $-R_{y,m}\mathrm{cv}_m = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}/S_y = \Delta_A/S_y$, comme l'affirme la formule (5.8). De surcroît, (7.7), (7.8) et (7.9) impliquent que $-R_{y,m}/R_{y,\mathbf{x}} = \Lambda$, de sorte que le ratio des coefficients de corrélation F dans (5.9) est égal à l'angle Λ défini par (7.7).

8. Commentaires : qualité de l'ajustement, propriétés du biais et procédures de sélection connexes

Trois problèmes sont examinées à la présente section : i) la relation entre le biais et la qualité de l'ajustement des régressions, ii) la relation linéaire entre la valeur prévue de $\Delta_A = \hat{N}^{-1}(\tilde{Y}_{\mathrm{EXP}} - \tilde{Y}_{\mathrm{CAL}})$ et le biais de $\tilde{Y}_{\mathrm{CAL}}$ ou $\hat{Y}_{\mathrm{CAL}}$, et iii) la méthode alternative de sélection des variables auxiliaires proposée par Schouten (2007).

En ce qui concerne le problème (i), rappelons que l'écart total donné à la section 4 est $\Delta_T = \Delta_A + \Delta_R$, où Δ_A est calculable, mais Δ_T et Δ_R ne le sont pas. S'il était calculable, $\hat{N}\Delta_R = \tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}}$ serait une estimation du biais de $\tilde{Y}_{\mathrm{CAL}}$ (et de celui de $\hat{Y}_{\mathrm{CAL}}$). Une faible valeur de Δ_R est souhaitable. La question qui se pose est celle de savoir si cela est réalisé quand le modèle $y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$ (pour un vecteur $\mathbf{x}_k$ donné) est bien ajusté aux données. Nous devons distinguer deux aspects : a) l'ajustement calculable aux données $(y_k, \mathbf{x}_k)$ observé pour $k \in r$, et b) l'ajustement hypothétique aux données $(y_k, \mathbf{x}_k)$ pour $k \in s$, certaines étant observées et d'autres pas.

Un bon ajustement pour les répondants, $k \in r$, ne garantit pas un petit Δ_R : l'ajustement par les moindres carrés pondérés en utilisant les données observées $(y_k, \mathbf{x}_k)$ pour $k \in r$ donne les résidus $e_{k|r;d} = y_k - \mathbf{x}'_k\mathbf{B}_{\mathbf{x}|r;d}$, calculables pour $k \in r$, avec la propriété $\sum_r d_k e_{k|r;d} = 0$ (ici, la notation détaillée $\mathbf{B}_{\mathbf{x}|r;d}$ spécifiée dans (2.8) est préférable à la notation simplifiée $\mathbf{B}_\mathbf{x}$). Pour $k \in s - r$, $e_{k|r;d}$ n'est pas calculable ; il possède une moyenne non nulle inconnue $\bar{e}_{s-r;d} = \sum_{s-r} d_k e_{k|r;d}/\sum_{s-r} d_k$. Nous avons

$$\Delta_R = (\tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}})/\hat{N} = -(1-P)\,\bar{e}_{s-r;d} \neq 0. \qquad (8.1)$$

Que l'ajustement soit bon (petits résidus $e_{k|r;d}$; $R^2_{y,\mathbf{x}}$ proche de un) ou mauvais (grands résidus $e_{k|r;d}$; $R^2_{y,\mathbf{x}}$ proche de zéro), il se peut que l'écart Δ_R donné par (8.1) soit grand et que $\tilde{Y}_{\mathrm{CAL}}$ soit loin d'être sans biais. Même si l'ajustement est parfait pour les répondants ($e_{k|r,d} = 0$ pour tout $k \in r$, et $R^2_{y,\mathbf{x}} = 1$), rien ne garantit que le biais sera faible.

Une inadéquation comparable affecte l'imputation basée sur les données fournies par les répondants. Si des imputations par la régression $\hat{y}_k = \mathbf{x}'_k\mathbf{B}_{\mathbf{x}|r;d}$ sont utilisées pour remplacer les valeurs y_k manquantes pour $k \in s - r$, l'estimateur imputé est donné par

$$\hat{Y}_{\mathrm{imp}} = \sum_r d_k y_k + \sum_{s-r} d_k \hat{y}_k.$$

Alors, $\hat{Y}_{\mathrm{imp}} = \tilde{Y}_{\mathrm{CAL}}$, de sorte que l'exposition au biais est la même pour $\hat{Y}_{\mathrm{imp}}$ que pour $\tilde{Y}_{\mathrm{CAL}}$, ce qui est facile à comprendre : quand la non réponse cause une sélection biaisée des valeurs de y, les valeurs imputées calculées en se basant sur cette sélection représenteront incorrectement

les valeurs inconnues de y qui caractérisent l'échantillon s ou la population U.

Considérons maintenant l'aspect (b) de l'ajustement, c'est à dire l'ajustement hypothétique de la régression par les moindres carrés pondérés aux données $(y_k, \mathbf{x}_k)$ pour $k \in s$. Le vecteur des coefficients de régression sera $\mathbf{B}_{\mathbf{x}|s;d} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_s d_k \mathbf{x}_k y_k$, et les résidus $e_{k|s;d} = y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}|s;d}$ pour $k \in s$ satisferont $\sum_s d_k e_{k|s;d} = 0$. Étant donné que $\sum_r d_k m_k \mathbf{x}_k / \hat{N} = \overline{\mathbf{x}}_{s;d}$ et $\sum_r d_k m_k y_k / \hat{N} = \overline{\mathbf{x}}_{s;d}' \mathbf{B}_{\mathbf{x}|r;d}$, nous avons

$$\Delta_R = \hat{N}^{-1}(\tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}}) = (1/\hat{N}) \sum_r d_k m_k e_{k|s;d}. \quad (8.2)$$

Supposons que le modèle est « vrai pour l'échantillon s », avec un ajustement parfait, de sorte que $e_{k|s;d} = 0$ pour tout $k \in s$. Alors, en vertu de (8.2), nous avons $\Delta_R = 0$, de sorte que l'estimateur corrigé de la non-réponse $\tilde{Y}_{\mathrm{CAL}}$ concorde avec l'estimateur sans biais $\tilde{Y}_{\mathrm{FUL}}$. L'opinion que le biais est faible repose sur une hypothèse non vérifiable.

Penchons nous maintenant sur la question (ii) et expliquons la relation essentiellement linéaire entre le biais de $\tilde{Y}_{\mathrm{CAL}}$ et la valeur prévue de l'indicateur $H_0 = \Delta_A / S_y = (\tilde{Y}_{\mathrm{EXP}} - \tilde{Y}_{\mathrm{CAL}}) / \hat{N} S_y$. Pour un résultat (s, r) donné, une variable y fixe et un vecteur $\mathbf{x}$ fixe, nous avons

$$(\tilde{Y}_{\mathrm{CAL}} - Y) / \hat{N} S_y = (\tilde{Y}_{\mathrm{EXP}} - Y) / \hat{N} S_y - H_0.$$

Soit E_{pq} l'opérateur d'espérance par rapport à tous les résultats (s, r), c'est à dire $E_{pq}(\cdot) = E_p(E_q(\cdot|s))$, où $p(s)$ et $q(r|s)$ sont, respectivement, le plan d'échantillonnage connu et la distribution des réponses inconnue. Nous écrivons $\mathrm{biais}(\tilde{Y}_{\mathrm{CAL}}) = E_{pq}(\tilde{Y}_{\mathrm{CAL}}) - Y$, $\mathrm{biais}(\tilde{Y}_{\mathrm{EXP}}) = E_{pq}(\tilde{Y}_{\mathrm{EXP}}) - Y$ et $C = E_{pq}(\hat{N} S_y)$. En recourant au remplacement habituel en grand échantillon de la valeur prévue d'un ratio par le ratio des valeurs prévues, nous obtenons $E_{pq}[(\tilde{Y}_{\mathrm{CAL}} - Y) / \hat{N} S_y] \approx [E_{pq}(\tilde{Y}_{\mathrm{CAL}}) - Y] / E_{pq}(\hat{N} S_y)$ et procédons de manière analogue pour $\tilde{Y}_{\mathrm{EXP}}$, de sorte que

$$\mathrm{biais}(\tilde{Y}_{\mathrm{CAL}}) \approx \mathrm{biais}(\tilde{Y}_{\mathrm{EXP}}) - C \times E(H_0). \quad (8.3)$$

Ici $\mathrm{biais}(\tilde{Y}_{\mathrm{EXP}})$ et C ne dépendent pas du choix du vecteur $\mathbf{x}$, tandis que $\mathrm{biais}(\tilde{Y}_{\mathrm{CAL}})$ et $E(H_0)$ en dépendent. Donc, à mesure que le vecteur $\mathbf{x}$ change, $\mathrm{biais}(\tilde{Y}_{\mathrm{CAL}})$ et $E(H_0)$ sont reliés de manière essentiellement linéaire. Aucune forme particulière de $p(s)$ et de $q(r|s)$ ne doit être spécifiée pour que l'expression (8.3) soit vérifiée. Par conséquent, quand deux vecteurs auxiliaires, $\mathbf{x}_{1k}$ et $\mathbf{x}_{2k}$, sont comparés, la différence de biais est, de façon étroitement approximative, proportionnelle à la variation de la valeur prévue de H_0 :

$$\mathrm{biais}(\tilde{Y}_{\mathrm{CAL}}(\mathbf{x}_{1k})) - \mathrm{biais}(\tilde{Y}_{\mathrm{CAL}}(\mathbf{x}_{2k})) \approx -C(E_1 - E_2) \quad (8.4)$$

où $E_i = E_{pq}(H_0(\mathbf{x}_{ik}))$ pour $i = 1, 2$. Les propriétés (8.3) et (8.4) sont validées par l'étude de Monte Carlo à la section 10.

Notons que la formule (8.3) ne garantit pas que le biais de $\tilde{Y}_{\mathrm{CAL}}$ basé sur un certain vecteur $\mathbf{x}_k$ sera nul ou presque nul. Elle ne précise pas qu'une valeur comparativement grande de $|\Delta_A|$ garantit un petit biais dans $\tilde{Y}_{\mathrm{CAL}}$. Ce que dit (8.3) est que $\mathrm{biais}(\tilde{Y}_{\mathrm{CAL}})$ est relié linéairement à l'espérance de l'indicateur $H_0 = \Delta_A / S_y$. Par conséquent, l'évaluation des vecteurs $\mathbf{x}$ disponibles en fonction de l'indicateur H_0 (ou de l'indicateur $H_1 = |\Delta_A| / S_y$) est conforme à l'objectif de réduction du biais.

Passons maintenant au problème (iii) et commentons la méthode alternative de sélection des variables auxiliaires proposée par Schouten (2007). L'indicateur de ce dernier pour la sélection pas à pas des variables diffère des nôtres et ne produit habituellement pas exactement le même ensemble de variables. Dans une liste de, disons, 30 variables x catégoriques disponibles, les dix premières sélectionnées ne seront pas les mêmes que les dix retenues par nos indicateurs H_0 à H_3. L'ordre dans lequel les variables sont sélectionnées ne sera pas nécessairement le même non plus. Nous avons comparé, dans certains de nos travaux empiriques, nos résultats au choix de variables réalisé selon la méthode de Schouten. Dans certains cas, nous avons constaté une congruence importante entre les deux séries de « dix premières variables » choisies selon les deux procédures.

Le meilleur moyen d'apprécier la différence entre les deux approches consiste à comparer leur contexte et leur démonstration mathématique. Nos indicateurs H_0 et H_1 reposent sur la notion de séparation (ou de distance), pour un résultat donné (s, r), entre l'estimateur corrigé $\tilde{Y}_{\mathrm{CAL}}$ et l'estimateur élémentaire, $\tilde{Y}_{\mathrm{EXP}}$, et sur l'idée que cette séparation augmentera habituellement quand le vecteur $\mathbf{x}$ devient plus puissant. Le plan d'échantillonnage probabiliste est pris en considération et aucune hypothèse n'est formulée au sujet de la distribution des réponses.

Schouten s'appuie sur un argument en superpopulation et ne semble pas tenir compte des poids d'échantillonnage. Il trouve qu'une expression du biais prévu sous le modèle d'un estimateur de la moyenne de population est proportionnelle à la corrélation (au niveau de la population) entre la variable y et l'indicateur $0-1$ de réponse. Il montre que cette corrélation (et, conséquemment, le biais) peut être bornée à l'intérieur d'un intervalle. En particulier, il considère l'estimateur par la régression généralisée et montre que son biais absolu maximum est égal à la largeur de l'intervalle de biais. Cette largeur dépend du vecteur réel inconnu de régression $\boldsymbol{\beta}$ pour la régression (au niveau de la population) de y en fonction de $\mathbf{x}$. Ce vecteur inconnu $\boldsymbol{\beta}$ est remplacé par une estimation basée sur les répondants, donc sujette à un certain biais, à cause de la non-réponse. Schouten met l'accent sur le fait qu'il n'est pas nécessaire d'émettre l'hypothèse que les données manquent au hasard

pour cette méthode, qui est, à cet égard, semblable à la nôtre.

9. Choix des variables auxiliaires pour l'enquête pilote suédoise sur les jeux de hasard et le jeu compulsif

Nous avons utilisé un ensemble de données d'enquête réelles pour illustrer l'utilisation des indicateurs H_1, H_2 et H_3 en vue de construire le vecteur **x**. En 2008, l'Institut national de santé publique de la Suède (*Svenska Folkhälsoinstitutet*) a réalisé une enquête pilote pour étudier la portée de la participation à des jeux de hasard et les caractéristiques des joueurs compulsifs. L'échantillonnage et le calage des poids ont été effectués par Statistics Sweden. Nous illustrons l'utilisation des indicateurs dans cette enquête, pour laquelle un échantillon aléatoire simple stratifié s de $n = 2\,000$ personnes a été tiré du registre de la population totale (RPT) de la Suède. Les strates ont été définies par recoupement de la région de résidence et du groupe d'âge. Chacune des six régions a été définie comme une grappe de zones postales considérées comme étant semblables en ce qui a trait à des variables telles que le niveau de scolarité, le pouvoir d'achat, le type de logement, et l'origine étrangère. Les quatre groupes d'âge utilisés étaient 16 à 24 ans, 25 à 34 ans, 35 à 64 ans et 65 à 84 ans.

Le taux de réponse global pondéré était de 50,8 %. La non réponse, plus ou moins prononcée selon le domaine d'intérêt, interfère avec l'objectif de précision établi. Un important réservoir de variables auxiliaires possibles était disponible pour l'enquête, y compris des variables du RPT, des variables du registre de l'éducation et un sous ensemble de celles comprises dans une autre grande base de données de Statistics Sweden, appelée LISA. Pour le présent exemple, nous avons créé un fichier de données dans lequel nous avons sélectionné 13 variables catégoriques. Nous avons utilisé 12 d'entre elles comme des variables x et la treizième, la variable dichotomique *Occupé*(e), comme variable étudiée. Les valeurs de toutes les variables sont disponibles pour toutes les unités $k \in s$. La réponse $(k \in r)$ ou la non réponse $(k \in s - r)$ à l'enquête est également indiquée dans le fichier de données.

Les variables de nature continue sont utilisées comme des variables groupées, de sorte que les 12 variables x sont catégoriques et de type $\mathbf{x}_k^\circ$, comme il est défini à la section 2 (comme la plupart des variables sont disponibles pour la totalité de la population, elles pourraient être de type $\mathbf{x}_k^*$, mais puisque l'effet du biais a peu d'importance, nous les avons utilisées comme des variables $\mathbf{x}_k^\circ$). La valeur de la variable étudiée, $y_k = 1$ si l'unité k est *occupée* et $y_k = 0$ autrement, est connue pour $k \in s$, de sorte que l'estimation sans biais $\tilde{Y}_{\text{FUL}}$ définie par (3.2) peut être calculée et utilisée comme référence. Nous avons également calculé $\tilde{Y}_{\text{EXP}}$ définie par (3.1), ainsi que $\tilde{Y}_{\text{CAL}}$, définie par (2.5), pour divers vecteurs **x** construits par sélection pas à pas à partir du réservoir de 12 variables x à l'aide des indicateurs H_1, H_2 et H_3 définis par (5.12).

Nous avons effectué la sélection de la façon suivante : à l'étape 0, le vecteur auxiliaire est le vecteur élémentaire $\mathbf{x}_k = 1$, et l'estimateur est $\tilde{Y}_{\text{EXP}}$. À l'étape 1, la valeur de l'indicateur est calculée pour chacune des 12 variables auxiliaires possibles ; la variable produisant la valeur la plus grande de l'indicateur est sélectionnée. À l'étape 2, la valeur de l'indicateur est calculée pour chacun des 11 vecteurs de dimension deux contenant la variable sélectionnée à l'étape 1 et l'une des variables restantes. La variable qui donne la plus grande valeur de l'indicateur est sélectionnée à l'étape 2, et ainsi de suite, aux étapes suivantes. Chaque nouvelle variable est jointe à celles déjà incluses dans le vecteur selon un schéma « côte à côte » (ou « + »). Par conséquent, il est fait abstraction des interactions. L'ordre de sélection est différent pour chaque indicateur.

Les valeurs de H_2 et H_3 qui déterminent la prochaine variable qui sera incluse dans le vecteur sont, par nécessité mathématique, plus grandes à chaque étape. Il n'en est toutefois pas ainsi pour H_1. À une certaine étape j, nous avons utilisé la règle consistant à inclure la variable x possédant la plus grande des valeurs calculées de H_1. Cette valeur peut être plus faible que la valeur de H_1 qui a déterminé la variable sélectionnée à l'étape précédente, $j-1$. La série de valeurs de H_1 pour l'inclusion dans le vecteur augmente jusqu'à une certaine étape, puis commence à diminuer, comme l'illustre le tableau 9.1.

L'estimation sans biais est $\tilde{Y}_{\text{FUL}} = 4\,265$; l'estimation élémentaire est $\tilde{Y}_{\text{EXP}} = 4\,719$ (toutes les deux en milliers). Cela suggère un grand biais positif dans $\tilde{Y}_{\text{EXP}}$, dont l'écart relatif (en %) par rapport à l'estimation sous réponse complète $\tilde{Y}_{\text{FUL}}$ est $\text{ERC} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}) / \tilde{Y}_{\text{FUL}} \times 10^2 = 10,7$. L'ajout de variables x catégoriques, une à une, dans le vecteur **x** modifiera successivement cet écart, quoique après avoir introduit quelques variables, la variation ne se produit pas toujours dans le sens d'une valeur plus faible. À chaque étape, nous avons calculé l'indicateur, $\tilde{Y}_{\text{CAL}}$ et $\text{ERC} = (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) / \tilde{Y}_{\text{FUL}} \times 10^2$.

Le tableau 9.1 donne la sélection pas à pas avec l'indicateur H_1 (le nombre de catégories est indiqué entre parenthèses pour chaque variable sélectionnée). La première variable sélectionnée est celle de la *catégorie de revenu*, qui produit une réduction importante de ERC, qui passe de 10,7 à 4,5. Les cinq sélections suivantes ont lieu pour des valeurs croissantes de H_1, et la valeur de ERC est réduite, mais d'une quantité successivement plus petite. L'étape 6, où la variable d'*état matrimonial* est sélectionnée, correspond à un point de virage, indiqué par la double ligne dans le tableau 9.1 : la valeur de H_1 commence alors à diminuer, et

$\tilde{Y}_{\mathrm{CAL}}$ et ERC commencent à augmenter. À l'étape 6, ERC atteint sa valeur la plus faible, soit 0,5, puis commence à augmenter, illustrant le fait que l'ajout de toutes les variables x disponibles pourrait ne pas représenter la meilleure approche. Le point de virage de H_1 et le point auquel ERC est le plus proche de zéro coïncident dans le présent exemple, mais il n'en est généralement pas ainsi. En outre, dans des conditions réelles d'enquête, ERC est inconnu, de même que l'étape à laquelle il est le plus proche de zéro.

Le tableau 9.2 donne la sélection pas à pas avec l'indicateur H_3. La valeur de ce dernier augmente à chaque étape, mais à une vitesse qui finit par plafonner et les variations successives de $\tilde{Y}_{\mathrm{CAL}}$ deviennent négligeables. Ces résultats donnent à penser qu'il faut s'arrêter après six étapes, au moment où ERC = 2,8. Dans aucune des 12 étapes ERC ne s'approche autant de zéro que la valeur de 0,5 obtenue avec H_1 après six étapes. À cet égard, H_1 est meilleur que H_3, dans le présent exemple. Quand les 12 variables x sont toutes sélectionnées, ERC atteint la valeur finale de 2,6 dans les deux tableaux.

Tableau 9.1
Sélection pas à pas ascendante, indicateur H_1, variable étudiée dichotomique *Occupé(e)*. Valeurs successives de $H_1 \times 10^3$, de $\tilde{Y}_{\mathrm{CAL}}$ en milliers et de ERC = $(\tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}}) / \tilde{Y}_{\mathrm{FUL}} \times 10^2$. Pour comparaison, $\tilde{Y}_{\mathrm{EXP}} \times 10^{-3} = 4\,719$; $\tilde{Y}_{\mathrm{FUL}} \times 10^{-3} = 4\,265$

Variable auxiliaire ajoutée	$H_1 \times 10^3$	$\tilde{Y}_{\mathrm{CAL}} \times 10^{-3}$	ERC
Catégorie de revenu (3)	76	4 458	4,5
Niveau de scolarité (3)	107	4 350	2,0
Présence d'enfants (2)	114	4 326	1,4
Logement urbain (2)	118	4 310	1,1
Sexe (2)	123	4 296	0,7
État matrimonial (2)	125	4 286	0,5
Jours de chômage (3)	121	4 301	0,9
Mois de prestations de maladie (3)	120	4 305	1,0
Niveau d'endettement (3)	115	4 322	1,3
Grappe de codes postaux (6)	109	4 343	1,8
Pays de naissance (2)	103	4 363	2,3
Groupe d'âge (4)	99	4 377	2,6

Tableau 9.2
Sélection pas à pas ascendante, indicateur H_3, variable dichotomique *Occupé(e)*. Valeurs successives de $H_3 \times 10^3$, de $\tilde{Y}_{\mathrm{CAL}}$ en milliers, de ERC = $(\tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}}) / \tilde{Y}_{\mathrm{FUL}} \times 10^2$. Pour comparaison, $\tilde{Y}_{\mathrm{EXP}} \times 10^{-3} = 4\,719$; $\tilde{Y}_{\mathrm{FUL}} \times 10^{-3} = 4\,265$

Variable auxiliaire ajoutée	$H_3 \times 10^3$	$\tilde{Y}_{\mathrm{CAL}} \times 10^3$	ERC
Niveau de scolarité (3)	186	4 520	6,0
Grappe de codes postaux (6)	250	4 505	5,6
Pays de naissance (2)	281	4 498	5,5
Catégorie de revenu (3)	298	4 369	2,4
Groupe d'âge (4)	354	4 399	3,1
Sexe (2)	364	4 384	2,8
Logement urbain (2)	374	4 378	2,6
Niveau d'endettement (3)	381	4 364	2,3
Mois de prestations de maladie (3)	384	4 380	2,7
Présence d'enfants (2)	387	4 379	2,7
État matrimonial (2)	388	4 379	2,7
Jours de chômage (3)	388	4 377	2,6

L'ensemble des six premières variables sélectionnées avec H_3 contient trois des mêmes variables que l'ensemble correspondant de six variables sélectionnées avec H_1. Les deux profils de sélection assez différents ne sont pas en contradiction, parce que H_1 est axé spécifiquement sur la variable *y Occupé*(*e*), tandis que H_3 est un indicateur de compromis, indépendant de toute variable *y*. Faute d'espace, nous ne présentons pas les résultats de la sélection pas à pas pour l'indicateur H_2. Le profil de sélection de cet indicateur ressemble davantage à celui de H_3 qu'à celui de H_1. Parmi les six premières variables sélectionnées avec H_2, quatre sont parmi les six premières sélectionnées avec H_3. À titre de commentaire général, nous pensons que dans de nombreuses situations pratiques, l'utilisation de plus de six variables est inutile et que la sélection des quelques premières devient très importante.

10. Validation empirique par simulation pour une population synthétique

La théorie présentée aux sections précédentes ne comporte aucune hypothèse quant à la distribution des réponses, qui est inconnue. Le plan d'échantillonnage est arbitraire et les probabilités d'inclusion connue sont prises en compte. Pour l'expérience décrite à la présente section, nous spécifions plusieurs distributions des réponses pour lesquelles nous précisons une valeur positive de la probabilité de réponse θ_k pour chaque $k \in U$. Autrement dit, pour la probabilité spécifiée θ_k, la valeur y_k est enregistrée dans l'expérience et pour la probabilité $1 - \theta_k$, cette valeur est manquante. Nous constatons que les indicateurs H_0 (ou $H_1 = |H_0|$) définis en (5.11) classent les divers vecteurs $\mathbf{x}$ dans l'ordre correct de préférence pour toutes les distributions des réponses prises en considération, conformément aux résultats théoriques (8.3) et (8.4). Nous confirmons que, sur une longue série de résultats (s, r), la moyenne de $H_0 = \Delta_A / S_y = -R_{y,m} \times \mathrm{cv}_m$ suit le biais des estimateurs par calage, mesuré par la moyenne de $\tilde{Y}_{\mathrm{CAL}} - Y$, de manière essentiellement parfaitement linéaire, quand le vecteur $\mathbf{x}$ correspond successivement aux 16 formules différentes. Nous examinons aussi les indicateurs H_2 et H_3 définis en (5.12) et constatons dans cette expérience qu'ils sont fortement reliés au biais de $\tilde{Y}_{\mathrm{CAL}}$.

Nous avons expérimenté plusieurs populations synthétiques et obtenu des conclusions similaires. Nous présentons ici les résultats pour une population synthétique de taille $N = 6\,000$, avec les valeurs créées $(y_k, \mathbf{x}_k, \theta_k)$ pour $k = 1, 2, \ldots, N = 6\,000$, pour 16 formules catégoriques différentes de $\mathbf{x}_k$, et quatre façons distinctes d'attribuer la probabilité θ_k.

Les 16 vecteurs $\mathbf{x}$ de variables catégoriques ont été obtenus en regroupant les valeurs générées x_{1k} et x_{2k} de

deux variables auxiliaires continues, x_1 et x_2. Les valeurs (y_k, x_{1k}, x_{2k}) pour $k = 1, 2, ..., 6\,000$ ont été créées en trois étapes, comme il suit. Étape 1 (variable x_1) : les 6 000 valeurs x_{1k} ont été obtenues comme des résultants indépendants de la variable aléatoire de loi gamma $\Gamma(a, b)$ en donnant aux paramètres les valeurs $a = 2$, $b = 5$. La moyenne et la variance des 6 000 valeurs réalisées x_{1k} étaient de 10,0 et 49,9, respectivement. Étape 2 (variable x_2) : pour l'unité k, avec la valeur x_{1k} fixée à l'étape 1, une valeur x_{2k} est réalisée en tant que résultat de la variable aléatoire gamma dont les paramètres sont tels que l'espérance et la variance conditionnelles de x_{2k} sont $\alpha + \beta x_{1k} + K\,h(x_{1k})$ et $\sigma^2 x_{1k}$, respectivement, où $h(x_{1k}) = x_{1k}(x_{1k} - \mu_{x_1})(x_{1k} - 3\mu_{x_1})$ avec $\mu_{x_1} = 10$. Nous avons utilisé les valeurs $\alpha = 1$, $\beta = 1$, $k = 0,001$ et $\sigma^2 = 25$. Le terme polynomial $K\,h(x_{1k})$ donne une forme légèrement non linéaire au tracé de (x_{2k}, x_{1k}), pour éviter une relation exactement linéaire. La moyenne et la variance des 6 000 valeurs réalisées x_{2k} étaient égales à 11,0 et 210,0, respectivement. Le coefficient de corrélation entre x_1 et x_2, calculé sur les 6 000 couples (x_{1k}, x_{2k}), était de 0,48. Étape 3 (variable étudiée y) : pour l'unité k, avec les valeurs x_{1k} et x_{2k} fixées aux étapes 1 et 2, une valeur y_k est réalisée en tant que résultat de la variable aléatoire gamma dont les paramètres sont tels que l'espérance et la variance conditionnelles de y_k sont $c_0 + c_1 x_{1k} + c_2 x_{2k}$ et $\sigma_0^2 (c_1 x_{1k} + c_2 x_{2k})$, respectivement. Nous avons utilisé $c_0 = 1$, $c_1 = 0,7$, $c_2 = 0,3$ et $\sigma_0^2 = 2$. La moyenne et la variance des 6 000 valeurs réalisées y_k étaient égales à 11,4 et 86,5, respectivement. Le coefficient de corrélation entre y et x_1, calculé sur les 6 000 couples (y_k, x_{1k}), était de 0,76 ; celui entre y et x_2, calculé sur les 6 000 couples (y_k, x_{2k}), était de 0,73.

Chacune des deux variables x a été transformée ensuite en quatre variantes de mode de groupement, désignées 8G, 4G, 2G et 1G, produisant $4 \times 4 = 16$ vecteurs auxiliaires $\mathbf{x}_k$. différents. Nous avons classé les 6 000 valeurs x_{1k} de la variable x_1 par ordre de grandeur et avons formé huit groupes de taille égale. Le premier groupe comprenait les 750 unités ayant les valeurs les plus grandes de x_{1k}, le deuxième, les 750 unités suivantes du classement par ordre de grandeur, et ainsi de suite, jusqu'au huitième groupe. Dans ce mode de groupement 8G de x_1, à l'unité k est affectée la valeur vectorielle $\boldsymbol{\gamma}_{(x_1;\,8)k}$, de dimension huit avec sept entrées « 0 » et une seule entrée « 1 » pour coder l'appartenance de k au groupe. Ensuite, des fusions successives de groupes sont effectuées, de sorte que deux groupes adjacents définissent toujours un nouveau groupe, chaque fois en doublant la taille du groupe. Donc, pour le mode 4G, la fusion des groupes 1 et 2 place les unités ayant les 1 500 plus grandes valeurs x_{1k} dans un premier nouveau groupe ; la fusion des groupes 3 et 4 produit le deuxième nouveau groupe de 1 500, et ainsi de suite ; la valeur vectorielle associée à l'unité k est $\boldsymbol{\gamma}_{(x_1;4)k}$. Dans le mode 2G, l'unité k possède la valeur vectorielle $\boldsymbol{\gamma}_{(x_1;2)k} = (1,0)'$ pour les 3 000 unités possédant les plus grandes valeurs x_1 et $\boldsymbol{\gamma}_{(x_1;2)k} = (0,1)'$ pour les autres. Dans le dernier mode 1G, les 6 000 unités sont toutes regroupées, toute l'information contenue dans x_1 est abandonnée, et $\boldsymbol{\gamma}_{(x_1;\,1)k} = 1$ pour tout k. Les 6 000 valeurs x_{2k} ont été regroupées selon la même procédure en les quatre modes 8G, 4G, 2G et 1G. L'appartenance correspondante de l'unité k à ces groupes est codée par les vecteurs $\boldsymbol{\gamma}_{(x_2;8)k}$, $\boldsymbol{\gamma}_{(x_2;4)k}$, $\boldsymbol{\gamma}_{(x_2;2)k}$ et $\boldsymbol{\gamma}_{(x_2;\,1)k} = 1$. Les $4 \times 4 = 16$ vecteurs auxiliaires $\mathbf{x}_k$ différents tiennent compte des deux sortes d'information de groupe ; les deux vecteurs $\boldsymbol{\gamma}$ sont placés côte à côte (par opposition à croisés), le résultat étant un calage sur deux marges, comme indiqué par le signe « + ». Donc, pour le cas désigné 8G + 8G, l'unité k possède la valeur vectorielle auxiliaire $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;\,8)k}, \boldsymbol{\gamma}'_{(x_2;\,8)k})'_{(-1)}$, où (-1) indique qu'une catégorie est exclue dans $\boldsymbol{\gamma}_{(x_1;8)k}$ ou dans $\boldsymbol{\gamma}_{(x_2;8)k}$ pour éviter une matrice singulière dans les calculs, ce qui donne à $\mathbf{x}_k$ la dimension $8 + 8 - 1 = 15$. Le cas 8G + 8G est celui dont le contenu informationnel est le plus important. À l'autre extrême, le cas 1G + 1G exclut toute l'information contenue dans x et $\mathbf{x}_k = 1$ pour tout k. Il existe 14 cas intermédiaires de contenu informationnel. Par exemple, 4G + 2G possède le vecteur $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;4)k}, \boldsymbol{\gamma}'_{(x_2;2)k})'_{(-1)}$ de dimension $4 + 2 - 1 = 5$; 4G + 1G possède $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;\,4)k}, 1)'_{(-1)} = \boldsymbol{\gamma}_{(x_1;\,4)k}$ de dimension 4 (il existe une interaction non négligeable entre x_1 et x_2 dans cette expérience, mais nous limitons cette dernière aux vecteurs $\mathbf{x}$ sans interaction, ce qui évite le risque d'obtenir des groupes dans lesquels la fréquence est faible).

Nous discutons ici des résultats pour quatre distributions des réponses. Les probabilités de réponse θ_k, $k = 1, 2, ..., N = 6\,000$ de ces distributions ont été spécifiées comme il suit :

IncExp($10 + x_1 + x_2$),	avec $\theta_k = 1 - e^{-c(10 + x_{1k} + x_{2k})}$ où $c = 0,04599$
IncExp($10 + y$),	avec $\theta_k = 1 - e^{-c(10 + y_k)}$ où $c = 0,06217$
DecExp($x_1 + x_2$),	avec $\theta_k = e^{-c(x_{1k} + x_{2k})}$ où $c = 0,01937$
DecExp(y),	avec $\theta_k = e^{-cy_k}$ où $c = 0,03534$.

La constante c était ajustée dans les quatre cas de manière à obtenir une probabilité de réponse moyenne de $\bar{\theta}_U = \sum_U \theta_k / N = 0,70$. Dans les deux premiers, la valeur 10 (plutôt que 0) a été utilisée pour éviter une fréquence élevée de faibles probabilités de réponse θ_k. Ces quatre options représentent des caractéristiques contrastantes pour

les probabilités de réponse : croissantes (IncExp) par opposition à décroissantes (DecExp), dépendantes des valeurs de *x* uniquement par opposition à dépendantes des valeurs de *y* uniquement. Dans la deuxième et la quatrième options, la réponse dépend directement de la variable *y* et pourrait donc être appelée « purement non ignorable ».

Nous avons généré $J = 5\,000$ résultats (s, r), où *s* de taille $n = 1\,000$ est tiré de $N = 6\,000$ par échantillonnage aléatoire simple et, pour chaque *s* donné, l'ensemble de réponses *r* est réalisé par chacune des quatre distributions des réponses. Autrement dit, pour $k \in s$, nous avons effectué un essai de Bernoulli avec la probabilité spécifiée θ_k d'inclusion dans l'ensemble des réponses *r*. Les essais de Bernoulli sont indépendants.

Pour chaque distribution des réponses, pour chacun des 16 vecteurs $\mathbf{x}$, et pour chaque résultat (s, r), nous avons calculé l'écart relatif $\mathrm{ER} = (\hat{Y}_{\mathrm{CAL}} - Y)/Y$, où $\hat{Y}_{\mathrm{CAL}}$ est donné par (2.4) et $Y = \sum_U y_k$ est le total de *y* ciblé, connu dans les conditions de cette expérience (alternativement, nous avons utilisé $\tilde{Y}_{\mathrm{CAL}}$ donné par (2.5), mais, comme prévu, la différence de biais comparativement à $\hat{Y}_{\mathrm{CAL}}$ était négligeable). Nous avons également calculé les indicateurs $H_i, i = 0, 1, 2, 3,$ donnés par (5.11) et (5.12). Nous avons calculé les mesures sommaires suivantes :

$$\text{biais rel} = \mathrm{Moy}(\mathrm{ER}) = \frac{1}{J}\sum_{j=1}^{J} \mathrm{ER}_j;$$

$$\mathrm{Moy}(H_i) = \frac{1}{J}\sum_{j=1}^{J} H_{ij} \quad \text{pour} \quad i = 0,\ 1,\ 2, 3$$

où *j* indique la valeur calculée pour le j^{e} résultat, $j = 1, 2, \ldots, 5\,000 = J$. Pour chaque distribution des réponses, nous avons donc obtenu la valeur *biais rel* (qui est la mesure Monte Carlo du biais relatif $(E_{pq}(\hat{Y}_{\mathrm{CAL}}) - Y)/Y)$ et 16 valeurs de $\mathrm{Moy}(H_i)$ (qui est la mesure Monte Carlo de $E_{pq}(H_i)$), $i = 0, 1,\ 2, 3,$ où *p* désigne l'échantillonnage aléatoire simple et *q* représente l'une des quatre distributions des réponses.

Le tableau 10.1 montre, pour IncExp$(10 + x_1 + x_2)$, *biais rel* en % et $\mathrm{Moy}(H_1) \times 10^3$ pour les 16 vecteurs $\mathbf{x}$. Pour la cellule 1G + 1G, avec le vecteur $\mathbf{x}_k = 1,$ les quatre quantités moyennes (Moy) sont nulles et *biais rel* atteint son niveau le plus élevé, soit 13,2 %. À l'autre extrême, la cellule 8G + 8G représente le niveau le plus élevé d'information ; elle produit la valeur la plus élevée pour $\mathrm{Moy}(H_1),$ et *biais rel* atteint sa valeur la plus faible, soit 0,2 % ; presque tout le biais est éliminé (à part une différence de signe éventuelle, $\mathrm{Moy}(H_0)$ et $\mathrm{Moy}(H_1)$ étaient égaux pour toutes les cellules).

Le résultat (8.4), qui est vérifié pour toute distribution des réponses et tout plan d'échantillonnage, indique que l'indicateur H_0 classera les $4 \times 4 = 16$ vecteurs auxiliaires correctement pour toute distribution des réponses (les probabilités de réponse n'étant pas toutes constantes, comme il est mentionné plus bas). Le tableau 10.1 illustre (8.4), en fonction de $H_1 = |H_0|$: la variation, de n'importe quelle cellule à n'importe quelle autre, de la valeur de $\mathrm{Moy}(H_1)$ (l'estimation Monte Carlo) de la valeur prévue de (H_1) est accompagnée d'une variation proportionnelle de la valeur de *biais rel*. La même proportionnalité a été observée pour les trois autres distributions des réponses. Nous aurions pu choisir d'autres distributions des réponses pour illustrer la même propriété.

Tableau 10.1
***Biais rel* en % et, entre parentheses, la valeur de $\mathrm{Moy}(H_1) \times 10^3$ pour 16 vecteurs auxiliaires $\mathbf{x}_k$. Distribution des réponses IncExp$(10 + x_1 + x_2)$**

Groupes basés sur x_{1k}	Groupes basés sur x_{2k}							
	8G		4G		2G		1G	
8G	0,2	(101)	0,5	(99)	1,3	(93)	3,4	(76)
4G	0,5	(98)	0,9	(96)	1,8	(89)	4,1	(70)
2G	1,5	(91)	1,9	(88)	3,2	(78)	6,5	(52)
1G	4,1	(70)	5,0	(64)	7,3	(46)	13,2	(0)

La distribution des réponses avec une probabilité de réponse constante θ_k pour tout *k* est un cas particulier. L'estimateur par calage $\tilde{Y}_{\mathrm{CAL}}$ basé sur tout vecteur $\mathbf{x}_k$ présente alors un biais nul (quasiment) et cela inclut l'estimateur élémentaire $\tilde{Y}_{\mathrm{EXP}}$ avec $\mathbf{x}_k = 1$. Le résultat 8.3 continue d'être valide, indiquant dans ce cas que $E_{pq}(H_0) \approx \mathrm{biais}(\tilde{Y}_{\mathrm{CAL}}) \approx \mathrm{biais}(\tilde{Y}_{\mathrm{EXP}}) \approx 0$. Dans le contexte de la simulation de la présente section, si $\theta_k = 0,70$ pour tout *k* est considérée comme une distribution supplémentaire des réponses, le tableau 10.1 contenant chacune des 16 cellules montre des valeurs presque nulles de *biais rel en %* et de $\mathrm{Moy}(H_1) \times 10^3,$ de la cellule la plus faible (1G + 1G) jusqu'à la cellule du vecteur $\mathbf{x}$ le plus puissant (8G + 8G). Il n'existe aucun biais devant être éliminé par une amélioration du vecteur $\mathbf{x}$. Si, en pratique, l'indicateur (H_1) ne réagit pas à un agrandissement du vecteur $\mathbf{x}$, il n'y a aucune raison de pousser la recherche au delà de la formule vectorielle la plus simple. Cette situation peut s'interpréter de trois façons : la variable *y* en question ne comporte aucun biais de non réponse, ou la probabilité de réponse est presque constante, ou aucun des vecteurs $\mathbf{x}$ disponibles n'est capable de réduire un biais existant.

Par souci de concision, nous ne montrons pas les tableaux correspondants pour $\mathrm{Moy}(H_2)$ et $\mathrm{Moy}(H_3)$. Par nécessité mathématique, les deux quantités augmentent dans les transitions emboîtées. Nous ne présentons pas non plus les analogues du tableau 10.1 pour les trois autres distributions des réponses, car les profils sont comparables.

Le tableau 10.2 pour IncExp($10 + x_1 + x_2$) et le tableau 10.3 pour IncExp($10 + y$) montrent comment Moy(H_1), Moy(H_2) et Moy(H_3) classent les 16 vecteurs **x**, représentés par leur valeur de *biais rel*. Pour mesurer le succès du classement, nous avons calculé le coefficient de corrélation de rangs de Spearman, désigné par *corrang*, entre *biais rel* et la valeur de l'indicateur, basé sur les 16 valeurs de chacun. Pour Moy(H_1), la ligne inférieure des deux tableaux donne $|corrang| = 1$, pour le classement parfait. Pour les données utilisées, |*corrang*| est aussi presque égal à un pour Moy(H_2) et Moy(H_3) (plus généralement, le classement obtenu avec H_2 et H_3 peut être bon, mais dépend des données).

Tableau 10.2
Valeur, par ordre croissant, de *biais rel* en %, et valeur et rang correspondants de Moy(H_1) × 10^3, Moy(H_2) × 10^3 et Moy(H_3) × 10^3, pour 16 vecteurs auxiliaires. Ligne inférieure : valeur des corrélations de rangs de Spearman, *corrang*. Distribution des réponses IncExp ($10 + x_1 + x_2$)

biais rel	Moy(H_1) × 10^3		Moy(H_2) × 10^3		Moy(H_3) × 10^3	
0,2	101	(1)	127	(1)	232	(1)
0,5	99	(2)	119	(2)	225	(2)
0,5	98	(3)	118	(3)	224	(3)
0,8	96	(4)	109	(4)	217	(4)
1,3	93	(5)	109	(5)	216	(5)
1,5	91	(6)	105	(6)	213	(6)
1,8	89	(7)	98	(7)	207	(7)
1,9	88	(8)	94	(8)	205	(8)
3,2	78	(9)	80	(11)	192	(9)
3,4	76	(10)	90	(9)	188	(11)
4,1	70	(11)	84	(10)	190	(10)
4,1	70	(12)	77	(12)	175	(13)
5,0	64	(13)	70	(13)	179	(12)
6,4	52	(14)	52	(14)	146	(15)
7,3	46	(15)	46	(15)	156	(14)
13,2	0	(16)	0	(16)	0	(16)
Corrang		-1,00		-0,99		-0,99

Il existe un contraste appréciable entre les résultats pour *biais rel* pour les deux distributions des réponses dans les tableaux 10.2 et 10.3. Le meilleur des vecteurs auxiliaires laisse un biais considérablement plus important pour la distribution IncExp($10 + y$) non ignorable que pour la distribution IncExp($10 + x_1 + x_2$). Cela n'est pas étonnant et il est important de noter qu'une réduction importante du biais est obtenue pour le cas non ignorable également.

Dans la simulation, le surajustement mentionné à la section 4, $\Delta_A > \Delta_T > 0$ (quand $(\tilde{Y}_{\text{EXP}})$ présente un biais positif) ou $\Delta_A < \Delta_T < 0$ (quand $\tilde{Y}_{\text{EXP}}$ présente un biais négatif), se produit pour certains résultats (s, r). La fréquence varie selon la force du vecteur auxiliaire et diffère pour les diverses distributions des réponses. La cellule pour laquelle ce surajustement a le plus de chance de se produire est 8G + 8G, qui comprend le plus puissant des 16 vecteurs auxiliaires. Pour IncExp($10 + x_1 + x_2$), le biais est presque entièrement éliminé pour la cellule 8G + 8G ; *biais rel* n'est que de 0,2 %. Donc, $\tilde{Y}_{\text{CAL}}$ est proche de l'estimation sans biais $\tilde{Y}_{\text{FUL}}$, Δ_A est proche de Δ_T, et $\Delta_A > \Delta_T$ se produit dans 45,6 % des résultats (s, r). En revanche, pour le cas non ignorable IncExp($10 + y$), la fréquence de $\Delta_A > \Delta_T$ n'est que de 0,1 % pour la cellule 8G + 8G. Bien que cette cellule donne lieu à une réduction considérable du biais (comparativement au cas élémentaire 1G + 1G), un biais persiste et, par conséquent, la situation $\Delta_A > \Delta_T$ ne se produit presque jamais.

Nous ne montrons pas les tableaux correspondants pour DecExp($x_1 + x_2$) et DecExp(y). La valeur la plus faible de *corrang* était de 0,94, enregistrée pour Moy(H_3) dans le cas de DecExp($x_1 + x_2$).

Une question qui n'est pas traitée dans les tableaux 10.2 et 10.3 est celle de savoir combien de fois, sur une longue série de résultats (s, r), un indicateur donné $H(\mathbf{x}_k)$ réussit à désigner correctement le vecteur **x** préféré. Pour répondre à cette question, comparons les deux vecteurs $\mathbf{x}_{1k}$ et $\mathbf{x}_{2k}$. Si la valeur absolue du biais de $\hat{Y}_{\text{CAL}}(\mathbf{x}_{2k})$ est plus petite que celle du biais de $\hat{Y}_{\text{CAL}}(\mathbf{x}_{1k})$, nous aimerions observer que l'inégalité $H(\mathbf{x}_{2k}) \geq H(\mathbf{x}_{1k})$ est vérifiée pour une grande majorité des résultats (s, r), parce qu'alors, l'indicateur $H(\cdot)$ produira avec une probabilité élevée la décision correcte de préférer $\mathbf{x}_{2k}$. Comme $H(\mathbf{x}_k)$ présente une variabilité d'échantillonnage, son taux de succès (le taux d'indication correcte) dépend de la taille de l'échantillon et nous nous attendons à ce qu'il augmente avec cette taille.

Tableau 10.3
Valeur, par ordre croissant, de *biais rel* en %, et valeur et rang correspondants de Moy(H_1) × 10^3, Moy(H_2) × 10^3 et Moy(H_3) × 10^3, pour 16 vecteurs auxiliaires. Ligne inférieure : valeur des corrélations de rangs de Spearman, *corrang*. Distribution des réponses IncExp ($10 + y$)

biais rel	Moy(H_1) × 10^3		Moy(H_2) × 10^3		Moy(H_3) × 10^3	
3,6	74	(1)	91	(1)	165	(1)
3,9	71	(2)	84	(2)	158	(2)
4,0	71	(3)	83	(3)	156	(3)
4,3	68	(4)	76	(5)	149	(5)
4,4	68	(5)	78	(4)	153	(4)
4,9	64	(6)	68	(8)	142	(8)
4,9	63	(7)	72	(6)	146	(6)
5,3	60	(8)	69	(7)	143	(7)
5,4	60	(9)	64	(9)	137	(9)
6,0	55	(10)	59	(10)	132	(10)
6,2	53	(11)	54	(11)	128	(11)
7,2	46	(12)	54	(12)	122	(12)
7,9	41	(13)	41	(13)	111	(13)
7,9	40	(14)	43	(14)	109	(14)
9,6	27	(15)	27	(15)	90	(15)
13,1	0	(16)	0	(16)	0	(16)
Corrang		-1,00		-0,99		-0,99

Nous apportons certain éclaircissements au sujet de cette question en prolongeant l'expérience de Monte Carlo :

5 000 résultats (s, r) ont été réalisés, d'abord avec un échantillon de taille $n = 1\,000$, puis avec un échantillon de taille $n = 2\,000$ (l'ensemble de réponses r est réalisé conformément à l'une des quatre distributions des réponses, en déclarant l'unité k « répondante » à la suite d'un essai de Bernoulli avec la probabilité spécifiée θ_k). Nous avons calculé le taux de succès comme étant la proportion de l'ensemble des résultats (s, r) dans laquelle l'indication correcte se concrétise dans une confrontation de deux vecteurs **x** différents. Nous avons procédé à plusieurs comparaisons par paire de cette sorte. Des résultats types sont présentés au tableau 10.4 pour IncExp$(10 + x_1 + x_2)$. L'entrée supérieure dans une cellule du tableau montre le taux de succès en % pour $n = 1\,000$, et l'entrée inférieure, le taux pour $n = 2\,000$. La valeur de *biais rel* pour les vecteurs en question est entre parenthèses.

Nous préférons les « tests sévères », c'est à dire les confrontations de vecteurs pour lesquels la différence absolue de *biais rel* est faible, parce que la décision correcte est alors plus difficile à obtenir. Il n'existe a priori aucune raison qu'un des indicateurs donne systématiquement de meilleurs résultats que les autres dans la présente étude. Dans les cinq tests sévères du tableau 10.4, H_1 produit, dans l'ensemble, un meilleur taux de succès que H_2 et H_3. Le taux de succès de H_1 s'améliore si l'on double la taille d'échantillon et a tendance, comme prévu, à être plus élevé quand les valeurs de *biais rel* sont plus écartées. Le cas 4G + 8G *c.* 8G + 8G compare les vecteurs **x** emboîtés, de sorte que l'on sait avant l'expérience que H_2 et H_3 donnent des taux de succès parfaits.

Tableau 10.4
Certaines comparaisons par paire des vecteurs auxiliaires ; pourcentage de résultats avec indication correcte, pour les indicateurs H_1, H_2 et H_3. Entre parenthèses, *biais rel* en %. Entrée supérieure : $n = 1\,000$, entrée inférieure : $n = 2\,000$. Distribution des réponses IncExp $(10 + x_1 + x_2)$

Cellules comparées	Pourtcentage de resultants avec indication correcte		
	H_1	H_2	H_3
4G + 8G(0,5) c.	90,0	100,0	100,0
8G + 8G(0,2)	96,4	100,0	100,0
4G + 2G(1,8) c.	66,8	86,0	70,7
2G + 8G(1,5)	74,2	89,0	67,4
1G + 8G(4,1) c.	74,3	70,3	45,0
8G + 1G(3,4)	82,8	78,0	43,3
4G + 1G(4,1) c.	90,6	61,4	83,9
2G + 2G(3,2)	97,0	68,8	92,3
1G + 2G(7,3) c.	77,4	77,4	34,5
2G + 1G(6,5)	85,9	85,9	28,8

11. Conclusion

Dans le présent article, nous examinons des situations d'enquête dans lesquelles de nombreux vecteurs auxiliaires (vecteurs **x**) possibles peuvent être créés et considérons l'utilisation de l'estimateur par calage $\tilde{Y}_{\mathrm{CAL}}$. Pour tout vecteur **x** donné, un certain biais inconnu persiste dans $\tilde{Y}_{\mathrm{CAL}}$; nous souhaitons, par un choix approprié du vecteur **x**, rendre le biais aussi faible que possible. Donc, nous examinons le ratio des biais défini par (4.2) et (4.3). Nous exprimons, dans (5.8) à (5.10), la composante Δ_A du ratio des biais sous la forme d'un produit de mesures statistiques faciles à interpréter. Cela nous mène à proposer plusieurs variantes d'indicateurs de biais pouvant être utilisées pour évaluer les divers vecteurs **x** en ce qui a trait à leur capacité de réduire efficacement le biais. Nous étudions en particulier l'indicateur H_1 donné par (5.12). Il fonctionne très bien, mais est axé sur une variable étudiée y particulière. Toutefois, une enquête gouvernementale type comprend un grand nombre de variables étudiées et, pour des raisons pratiques, il est souhaitable d'utiliser le même vecteur **x** pour estimer tous les totaux y. Un compromis devient donc nécessaire. Nous soutenons que l'indicateur H_3 donné par (5.12), répond à ce besoin ; il dépend des valeurs $\mathbf{x}_k$, mais ne dépend d'aucune donnée y. L'élaboration d'autres indicateurs (que H_3) pour la « situation comportant de nombreuses variables y » sera le sujet de futurs travaux de recherche. L'examen, pour la sélection pas à pas des variables x avec l'indicateur H_1, d'autres algorithmes que celui utilisé à la section 9 sera aussi le sujet de travaux de recherche à venir.

Remerciements

Les auteurs remercient les examinateurs et le rédacteur associé de leurs commentaires qui ont contribué à l'amélioration du présent article.

Bibliographie

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Eltinge, J., et Yansaneh, I. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenue de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.

Kalton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-98.

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.

Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.

Särndal, C.-E., et Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 4, 251-260.

Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics*, 23, 51-68.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 107-121.

Thomsen, I., Kleven, Ø., Wang, J.H. et Zhang, L.C. (2006). Coping with deceasing response rates in Statistics Norway. Recommended practice for reducing the effect of nonresponse. Rapport 2006/29. Oslo : Statistics Norway.