

## Techniques d'enquête 41-1

# Une ou deux étapes ? Pondération par calage à partir d'une base liste complète en présence de non-réponse

par Phillip S. Kott et Dan Liao

Date de diffusion : le 29 juin 2015



Statistique  
Canada

Statistics  
Canada

Canada

---

## Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à [www.statcan.gc.ca](http://www.statcan.gc.ca).

Vous pouvez également communiquer avec nous par :

**Courriel** à [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)

**Téléphone** entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

**Programme des services de dépôt**

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site [www.statcan.gc.ca](http://www.statcan.gc.ca) sous « À propos de nous » > « Notre organisme » > « [Offrir des services aux Canadiens](#) »

## Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

## Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

- . indisponible pour toute période de référence
- .. indisponible pour une période de référence précise
- ... n'ayant pas lieu de figurer
- 0 zéro absolu ou valeur arrondie à zéro
- 0<sup>s</sup> valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
- <sup>p</sup> provisoire
- <sup>r</sup> révisé
- x confidentiel en vertu des dispositions de la *Loi sur la statistique*
- <sup>E</sup> à utiliser avec prudence
- F trop peu fiable pour être publié
- \* valeur significativement différente de l'estimation pour la catégorie de référence ( $p < 0,05$ )

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2015

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

Une [version HTML](#) est aussi disponible.

*This publication is also available in English.*

---

# Une ou deux étapes ? Pondération par calage à partir d'une base liste complète en présence de non-réponse

Phillip S. Kott et Dan Liao<sup>1</sup>

## Résumé

Quand un échantillon aléatoire tiré d'une base liste complète souffre de non-réponse totale, on peut faire appel à la pondération par calage sur des totaux de population pour éliminer le biais de non-réponse sous un modèle hypothétique de réponse (sélection) ou de prédiction (résultat). De cette façon, la pondération par calage peut non seulement procurer une double protection contre le biais de non-réponse, mais aussi réduire la variance. En employant une astuce simple, on peut estimer simultanément la variance sous le modèle hypothétique de prédiction et l'erreur quadratique moyenne sous la combinaison du modèle hypothétique de réponse et du mécanisme d'échantillonnage probabiliste. Malheureusement, il existe une limite pratique aux types de modèle de réponse que l'on peut supposer lorsque les poids de sondage sont calés sur les totaux de population en une seule étape. En particulier, la fonction de réponse choisie ne peut pas toujours être logistique. Cette limite ne gêne pas la pondération par calage lorsqu'elle est effectuée en deux étapes : de l'échantillon de répondants à l'échantillon complet pour éliminer le biais de réponse, et puis de l'échantillon complet à la population pour réduire la variance. Des gains d'efficacité pourraient découler de l'utilisation de l'approche en deux étapes, même si les variables de calage employées à chaque étape représentent un sous-ensemble des variables de calage de l'approche en une seule étape. L'estimation simultanée de l'erreur quadratique moyenne par linéarisation est possible, mais plus compliquée que lorsque le calage est effectué en une seule étape.

**Mots-clés :** Échantillonnage probabiliste; modèle de réponse; modèle de prédiction; double protection; estimation simultanée des variances.

## 1 Introduction

Le sondage est un outil utilisé surtout pour estimer les paramètres d'une population finie en se basant sur un échantillon de ses membres tiré aléatoirement. Les échantillons probabilistes sont assortis de poids de sondage (d'échantillonnage) qui sont souvent les inverses des probabilités de sélection des membres individuels. À condition que chaque élément de la population possède une probabilité de sélection positive, il est simple de produire un estimateur du total de population de la variable étudiée qui est sans biais par rapport au mécanisme d'échantillonnage probabiliste. Le ratio de deux estimateurs sans biais des totaux, ou toute autre fonction lisse des totaux estimés, n'est pas forcément sans biais, mais est asymptotiquement sans biais et souvent convergent puisque sa variance relative, comme son biais relatif, tend vers zéro quand la taille de l'échantillon devient arbitrairement grande.

Deville et Särndal (1992) ont introduit la pondération par calage comme outil d'ajustement des poids de sondage de façon que les sommes pondérées de certaines variables de « calage » soient égales à leurs totaux de population connus (ou mieux estimés). Si ces *équations de calage* sont vérifiées, l'erreur-type d'un total estimé pour une variable dont le total de population est inconnu est souvent réduite, tandis que l'estimation demeure quasi (c'est-à-dire asymptotiquement) sans biais sous le mécanisme d'échantillonnage probabiliste.

Bien qu'elle ait été élaborée au départ pour réduire les erreurs-types, la pondération par calage a souvent été utilisée pour éliminer le biais de sélection résultant de la non-réponse totale sous certaines

1. Phillip S. Kott, statisticien-chercheur principal, RTI International, Rockville, Maryland 20852, États-Unis. Courriel : pkott@rti.org; Dan Liao, statisticien-chercheur, RTI International, Rockville, Maryland 20852, États-Unis.

hypothèses (par exemple, Folsom 1991; Fuller, Loughin et Baker 1994; Lundström et Särndal 1999; Folsom et Singh 2000). À cette fin, on traite le fait qu'un élément sélectionné dans l'échantillon répond (ou non) à une enquête comme une phase additionnelle de l'échantillonnage aléatoire de Poisson avec probabilités de sélection inconnues, mais positives. La pondération par calage estime ces probabilités de sélection de Poisson implicitement et produit des totaux estimés qui sont presque sans biais sous le mécanisme combiné de sélection de l'échantillon et des répondants, qui est souvent appelé le « quasi-plan d'échantillonnage ». Voir Oh et Scheuren (1983).

Une *mise en garde* importante est que, si le mécanisme de sélection de l'échantillon est entièrement sous le contrôle du statisticien, le mécanisme de sélection des réponses est inconnu. Une hypothèse est émise quant à la forme particulière du mécanisme de réponse, et si cette hypothèse n'est pas vérifiée, les estimateurs peuvent être biaisés.

Une autre justification de la pondération par calage s'appuie sur un type de modélisation différent. Il est facile de montrer que la pondération par calage produit un estimateur qui est sans biais sous un modèle de prédiction (résultat) linéaire si la valeur prévue de la variable étudiée sous le modèle de prédiction est une fonction linéaire des variables de calage pourvu que les mécanismes d'échantillonnage et de réponse soient ignorables, c'est-à-dire que l'on puisse appliquer le même modèle de prédiction que l'élément de la population soit ou non échantillonné ou qu'il réponde ou non lorsqu'il est échantillonné.

Contrairement au modèle de sélection qui régit le mécanisme de réponse, il est possible que le modèle de prédiction linéaire soit vérifié pour une variable étudiée et non pour une autre. C'est la raison pour laquelle la plupart des échantillonneurs préfèrent émettre l'hypothèse d'un *modèle de sélection* lorsqu'ils corrigent la non-réponse totale. Néanmoins, il est rassurant de savoir que si *l'un ou l'autre* modèle est correct, le total estimé est quasi sans biais (c'est-à-dire qu'il possède un biais relatif qui s'évanouit asymptotiquement), une propriété que Kim et Park (2006) ont appelée « double protection » contre le biais de non-réponse.

Il est possible de simultanément éliminer le biais de sélection et réduire l'erreur-type sous le mécanisme d'échantillonnage probabiliste en une seule étape en ajustant les poids de sondage des unités répondantes afin que les totaux estimés pour un ensemble de variables de calage soient égaux aux totaux de population connus de ces unités. Néanmoins, il existe des raisons de préférer l'approche de pondération par calage en deux étapes, même quand les ensembles de variables de calage utilisés aux deux étapes sont les mêmes ou sont un sous-ensemble des variables de calage de l'approche en une étape : la première étape, de l'échantillon de répondants à l'échantillon original, élimine le biais de sélection et la deuxième étape, de l'échantillon original à la population, réduit la variance des estimateurs résultants.

Bien que Folsom et Singh (2000) et d'autres aient souligné que la pondération par calage peut aussi être utilisée pour éliminer le biais de sélection dû à une sous-couverture ou une surcouverture de la base de sondage, nous nous concentrons ici sur un échantillon à un degré tiré d'une base liste complète sans enregistrements en double. Autrement dit, nous supposons que la base de sondage est identique à la population cible (c'est-à-dire que chaque unité de la population est énumérée sur la liste de la base de sondage).

La présentation de l'article est la suivante. À la section 2, nous passons en revue certains éléments de théorie sur la pondération par calage. À la section 3, nous présentons un estimateur de variance légèrement nouveau qui, comme l'estimateur de variance décrit dans Kott (2006), peut être utilisé pour mesurer à la fois l'erreur quadratique moyenne d'un estimateur pondéré par calage sous le quasi-plan

d'échantillonnage et la variance sous le modèle de prédiction ou la combinaison du modèle de prédiction et du mécanisme d'échantillonnage original, ce qui rend sans doute la double protection contre le biais de non-réponse plus utile pour l'inférence. L'estimateur de variance donné dans Kott s'applique seulement lorsque le calage se fait sur les valeurs de population. Ici, à l'instar de Folsom et Singh (2000), nous donnons la possibilité d'effectuer le calage sur l'échantillon original.

À la section 4, nous discutons des limites de la pondération par calage en une seule étape et élaborons une théorie pour l'approche en deux étapes. Bien que notre principal objectif ici soit de faire valoir les avantages de l'utilisation de deux étapes, même lorsque des ensembles similaires de variables de calage sont employés aux deux étapes, l'estimateur par calage que nous traitons dans cette section est plus général. À la section 5, nous décrivons les résultats de certaines expériences par simulation, tandis qu'à la section 6, nous tirons quelques conclusions.

## 2 Pondération par calage en une étape

### 2.1 Pondération par calage et non-réponse totale

En l'absence de non-réponse (ou d'erreurs de base de sondage), la pondération par calage est une méthode d'ajustement des poids d'échantillonnage en vue de créer un ensemble de poids  $\{w_k; k \in S\}$ , asymptotiquement proche des poids de sondage originaux,  $d_k = 1/\pi_k$ , qui satisfont à un ensemble d'équations de calage (une pour chaque composante de  $\mathbf{z}_k$ ) :

$$\sum_S w_k \mathbf{z}_k = \sum_U \mathbf{z}_k,$$

où  $S$  désigne l'échantillon,  $\pi_k$  désigne la probabilité de sélection dans l'échantillon de l'unité  $k$ ,  $U$  désigne la population de taille  $N$ ,  $\mathbf{z}_k$  est un vecteur comprenant  $P$  composantes ayant chacune un total de population connu, et  $\sum_A$  signifie  $\sum_{k \in A}$ .

Kott (2009) décrit un ensemble prudent de conditions faibles sous lesquelles  $t_y = \sum_S w_k y_k$  est un estimateur quasi sans biais du total de population  $T_y = \sum_U y_k$  (c'est-à-dire que le biais relatif de  $t_y$  est asymptotiquement nul). Fait plus important, on suppose que chaque probabilité  $\pi_k N/n$  possède une borne inférieure positive égale à  $N$  et que la taille d'échantillon (prévue),  $n$ , devient arbitrairement grande (nous ajoutons entre parenthèses le terme « prévue » au cas où la taille d'échantillon est aléatoire).

En outre, on suppose que les quatre premiers moments de population centrés de chaque composante de  $\mathbf{z}_k$  possèdent une borne supérieure, tandis que  $N^{-1} \sum_U \mathbf{z}_k \mathbf{z}_k^T$  converge vers une matrice définie positive.

L'utilisation de la pondération par calage aura tendance à réduire l'erreur quadratique moyenne par rapport à l'estimateur à facteur d'extension (*expansion estimator*),  $t_y^E = \sum_S d_k y_k$ , quand  $y_k$  est corrélée à certaines composantes de  $\mathbf{z}_k$ . Cependant, il ne faut pas perdre de vue que, dans la plupart des enquêtes, les variables étudiées  $y_k$  sont nombreuses.

Un moyen simple de calculer les poids de calage consiste à le faire linéairement en utilisant la formule suivante :

$$\begin{aligned} w_k &= d_k \left[ 1 + \left( \sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j \right)^T \left( \sum_S d_j \mathbf{z}_j \mathbf{z}_j^T \right)^{-1} \mathbf{z}_k \right] \\ &= d_k \left[ 1 + \mathbf{g}^T \mathbf{z}_k \right]. \end{aligned}$$

Fuller et coll. (1994) et plus tard Lundström et Särndal (1999) ont soutenu que ce calage linéaire peut aussi être utilisé pour traiter la non-réponse totale. L'échantillon  $S$  est remplacé par l'échantillon de répondants  $R$ , tandis que

$$\mathbf{g} = \left[ (1 - \theta) \left( \sum_U \mathbf{z}_j - \sum_R d_j \mathbf{z}_j \right)^T + \theta \left( \sum_S d_j \mathbf{z}_j - \sum_R d_j \mathbf{z}_j \right)^T \right] \left( \sum_R d_j \mathbf{z}_j \mathbf{z}_j^T \right)^{-1},$$

selon que l'échantillon de répondants est *calé sur la population* ( $\theta = 0$ ) ou *calé sur l'échantillon original* ( $\theta = 1$ ). Dans l'un et l'autre cas, l'estimation est quasi sans biais sous le quasi-plan d'échantillonnage qui traite la réponse comme une deuxième phase d'échantillonnage aléatoire à condition que la probabilité de réponse de chaque unité soit de la forme :

$$p_k = 1 / (1 + \boldsymbol{\gamma}^T \mathbf{z}_k), \quad (2.1)$$

et  $\mathbf{g}$  est un estimateur convergent du vecteur de paramètres inconnus  $\boldsymbol{\gamma}$  dans l'équation (2.1).

Le problème en ce qui concerne la fonction de réponse donnée par l'équation (2.1) est que l'estimateur implicite de  $p_k$ ,  $\hat{p}_k = 1 / (1 + \mathbf{g}^T \mathbf{z}_k)$  peut être négatif. Une forme non linéaire de la pondération par calage permettant d'éviter cette possibilité a été proposée par Kott et Liao (2012) qui se sont fondés sur la forme exponentielle généralisée de Folsom et Singh (2000). Cette forme de calage fait appel à la méthode de Newton (approximations itératives du développement en série de Taylor) pour trouver un  $\mathbf{g}$  tel que l'équation de calage (à partir d'ici, nous utilisons le terme équation de calage pour faire référence au vecteur des équations de calage des composantes) :

$$\sum_R w_k \mathbf{z}_k = \sum_R d_k \alpha(\mathbf{g}^T \mathbf{z}_k) \mathbf{z}_k = (1 - \theta) \sum_U \mathbf{z}_k + \theta \sum_S d_k \mathbf{z}_k \quad (2.2)$$

est vérifiée, où  $\theta = 0$  ou  $1$ ,

$$\alpha(\mathbf{g}^T \mathbf{z}_k) = \frac{\ell + \exp(\mathbf{g}^T \mathbf{z}_k)}{1 + \exp(\mathbf{g}^T \mathbf{z}_k)/u}, \quad (2.3)$$

$\ell$ , la borne inférieure de  $\alpha(\cdot)$ , est non négative (de sorte que les poids de calage sont également non négatifs), et la borne supérieure de  $\alpha(\cdot)$ ,  $u > \ell$ , peut être finie ou infinie.

Bien que la *fonction d'ajustement des poids*  $\alpha(\mathbf{g}^T \mathbf{z}_k)$  puisse prendre d'autres formes raisonnables, nous nous limiterons aux fonctions de la forme de l'équation (2.3). Il s'agit d'une généralisation du ratissage (*raking*) où  $\ell = 0, u = \infty$ , ainsi que de l'estimation implicite d'un modèle de réponse logistique, où  $\ell = 1, u = \infty$ . Dans l'algorithme d'ajustement proportionnel itératif original de Deming et Stephan (1940) pour le ratissage, les composantes de  $\mathbf{z}_k$  ont été restreintes à des fonctions indicatrices. Nous utilisons ici le terme « ratissage » de manière plus générale pour désigner une pondération par calage avec une fonction d'ajustement des poids de la forme  $\alpha(\mathbf{g}^T \mathbf{z}_k) = \exp(\mathbf{g}^T \mathbf{z}_k)$ .

Quand  $\ell < 1$ , l'équation (2.3) devient l'ajustement par calage généralisé introduit dans Deville et Särndal (1992) et discuté plus en détail dans Deville, Särndal et Sautory (1993). Le calage généralisé permet non seulement que les composantes de  $\mathbf{z}_k$  soient continues, mais aussi que l'étendue des  $\alpha(\mathbf{g}^T \mathbf{z}_k)$  soit contrainte entre une valeur positive  $\ell$  et une valeur (possiblement) finie  $u$ .

Deville et Särndal (1992) posaient comme condition que  $\alpha(0) = \alpha'(0) = 1$ . Puisqu'ils ne s'intéressaient pas à des échantillons avec non-réponse (ou à des bases de sondage incorrectes),  $\mathbf{g}^T \mathbf{z}_k$  devait converger vers 0 et  $\alpha(\mathbf{g}^T \mathbf{z}_k)$  vers 1 quand la taille d'échantillon (prévue) devenait arbitrairement grande. Cependant, lorsqu'on ajuste les poids de sondage pour corriger la non-réponse, poser que  $\ell \geq 1$  est une stratégie plus raisonnable afin que la probabilité de réponse estimée implicite ne soit pas supérieure à 1.

Tandis que la définition originale de la pondération par calage donnée dans Deville et Särndal (1992) comprenait la minimisation des écarts dans  $R$  entre les  $w_k$  et  $d_k$ , mesurés par une certaine fonction de perte, des formulations ultérieures (par exemple, Estevao et Särndal 2000) ont éliminé la fonction de perte de la définition. Forcer  $w_k$  et  $d_k$  à être proches a peu de sens quand la pondération par calage est utilisée pour corriger la non-réponse totale, puisque si une unité  $k$  échantillonnée a une probabilité relativement faible de réponse, l'écart entre  $w_k$  et  $d_k$  doit être relativement grand.

Au lieu de supposer un modèle de réponse ayant une forme fonctionnelle particulière, une autre justification de l'utilisation de la pondération par calage comme moyen d'éliminer le biais de non-réponse totale consiste à émettre l'hypothèse d'un modèle de prédiction dans lequel la variable étudiée  $y_k$  est elle-même une variable aléatoire telle que  $E(y_k | \mathbf{z}_k) = \mathbf{z}_k^T \boldsymbol{\beta}$  pour un  $\boldsymbol{\beta}$  inconnu, que l'unité  $k$  soit échantillonnée ou non ou qu'elle réponde ou non quand elle est échantillonnée. Kott (2006) et d'autres ont observé que l'estimateur pondéré par calage de  $T_y = \sum_U y_k$  sera quasi sans biais sous le modèle de prédiction quand le calage est effectué sur la population (quand  $\theta = 0$  dans l'équation (2.2)), et sous la combinaison du modèle de prédiction et du mécanisme de sélection de l'échantillon original quand le calage est effectué sur l'échantillon original (quand  $\theta = 1$ ).

La propriété faisant qu'un estimateur pondéré par calage est dans un certain sens quasi sans biais quand un modèle hypothétique de réponse ou un modèle hypothétique de prédiction est vérifié a été appelée « double protection contre le biais de non-réponse » par Kim et Park (2006). Elle est appelée « double robustesse » dans la littérature biostatistique (Bang et Robins 2005) et attribuée à Robins, Rotnitzky et Zhao (1994), qui ont traité la non-réponse partielle plutôt que totale.

On suppose souvent que la distribution de  $y_k | \mathbf{z}_k$  sous le modèle de prédiction est la même pour les membres de la population échantillonnés et non échantillonnés. Autrement dit, le mécanisme d'échantillonnage est considéré comme étant *ignorable*. En outre, on suppose souvent que la distribution de  $y_k | \mathbf{z}_k$  est la même qu'un membre de la population réponde ou non quand il est échantillonné, c'est-à-dire que le mécanisme de réponse est également considéré comme étant ignorable (Little et Rubin 2002). Ici, nous faisons des hypothèses analogues plus faibles sous le modèle de prédiction, nommément que  $E(y_k | \mathbf{z}_k)$  ne dépend pas du fait que l'unité  $k$  est échantillonnée ou non ou qu'elle répond ou non quand elle est échantillonnée. Disons que les mécanismes d'échantillonnage et de réponse sont considérés comme étant « ignorable au premier moment ».

## 2.2 Variables instrumentales

Deville (2000) a observé que l'on peut utiliser le calage avec des variables instrumentales pour corriger le biais de non-réponse possible en émettant l'hypothèse d'un modèle de réponse qui dépend de  $\mathbf{x}_k$ ,

$$p_k = [\alpha(\boldsymbol{\gamma}^T \mathbf{x}_k)]^{-1} = \frac{1 + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)/u}{\ell + \exp(\boldsymbol{\gamma}^T \mathbf{x}_k)}, \quad (2.4)$$

mais en ajustant les équations de calage avec  $\mathbf{z}_k$  :

$$\sum_R w_k \mathbf{z}_k = \sum_R d_k \alpha(\mathbf{g}^T \mathbf{x}_k) \mathbf{z}_k = (1 - \theta) \sum_U \mathbf{z}_k + \theta \sum_S d_k \mathbf{z}_k, \quad (2.5)$$

où le  $\mathbf{g}$  satisfaisant l'équation (2.5) avec  $\theta = 0$  ou 1 est un estimateur convergent du vecteur de paramètres inconnus  $\boldsymbol{\gamma}$  dans l'équation (2.4). Certaines conditions faibles sont nécessaires ici. Les conditions qui suivent sont suffisantes :  $N^{-1} \sum_R d_k \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) \mathbf{z}_k$  est un estimateur convergent et borné pour  $N^{-1} [(1 - \theta) \sum_U \mathbf{z}_k + \theta \sum_S d_k \mathbf{z}_k]$ ,  $\alpha(\phi)$  est partout deux fois dérivable, et  $N^{-1} \sum_R d_k \alpha'(\phi) \mathbf{z}_k \mathbf{x}_k^T$  est toujours inversible et borné quand l'échantillon devient arbitrairement grand.

Soit  $R_k = 1$  quand  $k \in R, 0$  autrement. Il n'est pas difficile de montrer que

$$\begin{aligned} \mathbf{g} - \boldsymbol{\gamma} &= -\left(\sum_S d_k R_k \alpha'(c_k) \mathbf{z}_k \mathbf{x}_k^T\right)^{-1} \left\{ \sum_S d_k R_k \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) \mathbf{z}_k - [(1 - \theta) \sum_U \mathbf{z}_k + \theta \sum_S d_k \mathbf{z}_k] \right\} \\ &\quad - \left(N^{-1} \sum_S d_k R_k \alpha'(c_k) \mathbf{z}_k \mathbf{x}_k^T\right)^{-1} \left\{ N^{-1} \sum_S d_k R_k \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) \mathbf{z}_k - N^{-1} [(1 - \theta) \sum_U \mathbf{z}_k + \theta \sum_S d_k \mathbf{z}_k] \right\} \end{aligned}$$

pour un certain  $c_k$  compris entre  $\mathbf{g}^T \mathbf{x}_k$  et  $\boldsymbol{\gamma}^T \mathbf{x}_k$ , comme l'ont démontré Kott et Liao (2012) quand  $\mathbf{x}_k = \mathbf{z}_k$ .

Deville note également que les composantes de  $\mathbf{x}_k$  peuvent être des variables étudiées dont les valeurs ne sont connues que pour les répondants. Chang et Kott (2008) ont étendu la notion de la pondération par calage afin de permettre que la dimension du vecteur  $\mathbf{z}_k$  soit plus grande que celle du vecteur  $\mathbf{x}_k$ . Nous ne traiterons ni l'une ni l'autre possibilité dans les sections qui suivent.

Kim et Shao (2013), en traitant la non-réponse non ignorable, désignent par « variables instrumentales » les composantes de  $\mathbf{z}_k$  qui ne sont pas entièrement des fonctions des composantes de  $\mathbf{x}_k$ . Pour limiter toute confusion future, nous utiliserons donc le terme « variables du modèle » pour désigner les composantes de  $\mathbf{x}_k$ .

### 3 Estimation de la variance de l'estimateur par calage en une étape

À la présente section, nous posons que

$$t_y = \sum_R w_k y_k = \sum_R d_k \alpha(\mathbf{g}^T \mathbf{x}_k) y_k$$

est l'estimateur pondéré par calage de  $T_y$ , où  $w_k = d_k \alpha(\mathbf{g}^T \mathbf{x}_k)$  quand  $k \in R$  est le poids de calage, et  $w_k$  est défini de façon commode comme étant égal à 0 quand  $k \notin R$ . La fonction d'ajustement des poids  $\alpha(\cdot)$  est définie implicitement par l'équation (2.4), et  $\mathbf{g}$  est de nouveau choisi de façon que l'équation de calage (2.5) soit vérifiée pour  $\theta = 0$  ou 1.

Nous proposons l'estimateur suivant de la variance de  $t_y$  :

$$v(t_y) = \sum_{k,j \in S} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right) \left[ d_k (\theta \mathbf{z}_k^T \mathbf{b} + \alpha_k e_k) \right] \left[ d_j (\theta \mathbf{z}_j^T \mathbf{b} + \alpha_j e_j) \right] + \sum_{k \in R} d_k (\alpha_k^2 - \alpha_k) e_k^2, \quad (3.1)$$

où  $\pi_{kj}$  est la probabilité de sélection conjointe de  $k$  et  $j$  sous le plan d'échantillonnage original,  $\pi_{kk} = \pi_k = 1/d_k$ ,  $\pi_k = \alpha(\mathbf{g}^T \mathbf{x}_k)$  quand  $k \in R$  et 0 autrement,

$$\mathbf{b} = \left[ \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{x}_k \mathbf{z}_k^T \right]^{-1} \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{x}_k y_k, \quad (3.2)$$

et  $e_k = y_k - \mathbf{z}_k^T \mathbf{b}$ . Nous montrerons que  $v(t_y)$  dans l'équation (3.1) peut être quasi sans biais dans un certain sens si *soit* un modèle de réponse (section 3.1) *soit* un modèle de prédiction est vérifié (section 3.2).

L'estimateur de variance dans l'équation (5.2) de Kott (2006) est identique à  $v(t_y)$  dans l'équation (3.1) quand  $\theta = 0$ . L'estimateur de variance dans Kim et Haziza (2014) est également similaire. Leur modèle de prédiction est plus général que le modèle de prédiction linéaire considéré ici.

Cet estimateur de variance  $v(t_y)$  présuppose que le plan d'échantillonnage original est tel que chaque élément ne peut être tiré qu'une seule fois. À la section 3.1, nous voyons que, quand les probabilités de réponse sont indépendantes (Poisson), alors sous des hypothèses faibles,  $v(t_y)$  est un estimateur quasi sans biais de l'erreur quadratique moyenne de  $t_y$  sous le quasi-plan d'échantillonnage, que le modèle de prédiction,  $E(y_k | \mathbf{x}_k, \mathbf{z}_k) = \mathbf{z}_k^T \boldsymbol{\beta}$ , soit vérifié ou non.

À la section 3.2, nous montrons que  $v(t_y)$  est un estimateur quasi sans biais pour le modèle de prédiction combiné à la variance sous le plan d'échantillonnage original de  $t_y$  en tant qu'estimateur de  $T_y$ , que le modèle de réponse donné par l'équation (2.4) soit vérifié ou non. Donc,  $v(t_y)$  peut être appelé un « estimateur simultané des variances ».

### 3.1 Estimation de la variance sous le modèle de réponse

Pour simplifier l'exposé, nous supposons que le modèle de réponse donné par l'équation (2.4) avec une borne supérieure  $u$  finie est vérifié. Les conditions suffisantes pour que  $v(t_y)$  soit un estimateur quasi sans biais de l'erreur quadratique moyenne de  $t_y$  (en vertu desquelles le biais converge vers 0 quand la taille de l'échantillon devient arbitrairement grande) sont

$$\pi_{kj} \geq B_0 > 0 \quad (3.3)$$

$$\sum_{j=1}^N \left| \frac{\pi_{kj}}{\pi_k \pi_j} - 1 \right| \leq B_1 < \infty \text{ pour chaque } k, \quad (3.4)$$

$$\frac{\sum_{j=1}^N \psi_j^r}{N} \leq B_2 < \infty \text{ où } \psi_j \text{ est } y_j \text{ ou toute composante de } \mathbf{x}_j \text{ ou } \mathbf{z}_j, \text{ tandis que } r = 1 \text{ ou } 2, \quad (3.5)$$

et  $N^{-1} \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{z}_k \mathbf{x}_k^T$  est de plein rang et est bornée en probabilité quand la taille de l'échantillon devient arbitrairement grande.

En vertu de cela, de  $\alpha'(\phi) = (1 - \alpha(\phi)/u) \exp(\phi)/[(1 + \exp(\phi)/u)]$  étant bornée quand  $u$  est finie, et de l'inégalité de Cauchy-Schwarz  $(\sum a_k b_k)^2 \leq \sum a_k^2 \sum b_k^2$ , il n'est pas difficile de voir non seulement que  $\mathbf{g}$  est un estimateur convergent de  $\boldsymbol{\gamma}$ , mais aussi que  $\mathbf{b}$  dans l'équation (3.2) (qui peut être rendue sous la forme  $\mathbf{b} = [N^{-1} \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{x}_k \mathbf{z}_k^T]^{-1} N^{-1} \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{x}_k y_k$ ) possède une limite en probabilité, que nous appellerons  $\mathbf{b}^*$ , que le modèle de prédiction soit vérifié ou non. En outre,  $\mathbf{b} - \mathbf{b}^*$  ainsi que  $\mathbf{g} - \boldsymbol{\gamma}$  sont  $O_p(1/\sqrt{n})$ .

Observons que

$$\begin{aligned} (t_y - T_y)/N &= \theta(\sum_S d_k \mathbf{z}_k^T \mathbf{b}^* - \sum_U \mathbf{z}_k^T \mathbf{b}^*)/N \\ &+ [\sum_R d_k \alpha(\mathbf{g}^T \mathbf{x}_k) e_k^* - \sum_R d_k \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) e_k^*]/N \\ &+ [\sum_R d_k \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) e_k^* - \sum_U e_k^*]/N, \end{aligned}$$

où  $e_k^* = y_k - \mathbf{z}_k^T \mathbf{b}^*$ . L'insertion de  $\alpha'(\cdot)$  dans le « coefficient de régression »  $\mathbf{b}$  nous permet d'ignorer la contribution du deuxième terme de cette somme,  $Q = \sum_R d_k [\alpha(\mathbf{g}^T \mathbf{x}_k) - \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k)] e_k^*/N$ , à l'erreur quadratique moyenne sous le quasi-plan d'échantillonnage. Il en est ainsi parce que  $\sum_R d_k \alpha'(\boldsymbol{\gamma}^T \mathbf{x}_k) \mathbf{x}_k e_k^* = 0$  est vraie par définition, ce qui implique que  $\sum_R d_k \alpha'(\boldsymbol{\gamma}^T \mathbf{x}_k) \mathbf{x}_k e_k^*$  est  $O_p(1/\sqrt{n})$  sous nos hypothèses. En outre, puisque  $\alpha(\mathbf{g}^T \mathbf{x}_k) - \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) = \alpha'(c_k)(\mathbf{g} - \boldsymbol{\gamma})^T \mathbf{x}_k$  est aussi  $O_p(1/\sqrt{n})$ ,  $Q = (\mathbf{g} - \boldsymbol{\gamma})^T \sum_R d_k \alpha'(c_k) \mathbf{x}_k e_k^*$  est  $O_p(1/n)$ , qui est asymptotiquement ignorable par rapport aux deux composantes  $O_p(1/\sqrt{n})$  de  $(t_y - T_y)/N$ .

La contribution de  $Q$  étant éliminée, un estimateur sans biais idéalisé, mais incalculable, de l'erreur quadratique moyenne sous le quasi-plan d'échantillonnage de  $t_y$  est donné par

$$v_{I1}(t_y) = \sum_{k,j \in S} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right) [d_k (\theta \mathbf{z}_k^T \mathbf{b}^* + e_k^*)] [d_j (\theta \mathbf{z}_j^T \mathbf{b}^* + e_j^*)] + \sum_{k \in R} \left(\frac{d_k e_k^*}{p_k}\right)^2 (1 - p_k), \quad (3.6)$$

où le premier terme du deuxième membre estime l'erreur quadratique moyenne avant la non-réponse (s'il y en a une) et le deuxième terme estime la variance ajoutée par la non-réponse.

Un estimateur quasi sans biais idéalisé de l'erreur quadratique moyenne de rechange, plus près d'être calculable, est donné par

$$v_{I2}(t_y) = \sum_{k,j \in S} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right) \left[ d_k \left( \theta \mathbf{z}_k^T \mathbf{b}^* + \frac{R_k}{p_k} e_k^* \right) \right] \left[ d_j \left( \theta \mathbf{z}_j^T \mathbf{b}^* + \frac{R_j}{p_j} e_j^* \right) \right] + \sum_{k \in R} d_k \left( \frac{e_k^*}{p_k} \right)^2 (1 - p_k), \quad (3.7)$$

où de nouveau  $R_k = 1$  quand  $k \in R, 0$  autrement. Puisque les  $(R_k/p_k) e_k^*$  sont indépendants sous le modèle de réponse et sont de moyenne  $e_k^*$  et de variance  $(e_k^*/p_k)^2 p_k(1 - p_k)$ ,  $E[(R_k/p_k) e_k^* (R_j/p_j) e_j^*] = e_k^* e_j^*$  quand  $k \neq j$ . Par contre, l'expression qui suit est vérifiée quand  $k = j$  :

$$\begin{aligned}
(1 - \pi_k) E \left[ \left( d_k \frac{R_k}{p_k} e_k^* \right)^2 \right] &= (1 - \pi_k) \left[ (d_k e_k^*)^2 + \left( \frac{d_k e_k^*}{p_k} \right)^2 p_k (1 - p_k) \right] \\
&= (1 - \pi_k) (d_k e_k^*)^2 + \left( \frac{d_k e_k^*}{p_k} \right)^2 p_k (1 - p_k) - d_k \left( \frac{e_k^*}{p_k} \right)^2 p_k (1 - p_k).
\end{aligned}$$

La première sommation dans le deuxième membre de l'équation (3.7) contient des termes où  $k \neq j$  et des termes où  $k = j$ , les derniers faisant que la deuxième sommation dans (3.7) diffère de la deuxième sommation dans le deuxième membre de l'équation (3.6). Notons que l'espérance sous le modèle de réponse de  $\sum_R d_k (e_k^*/p_k)^2 (1 - p_k)$  dans la deuxième sommation dans le deuxième membre de (3.7) est  $\sum_S d_k (e_k^*/p_k)^2 p_k (1 - p_k)$ .

Enfin,  $v_{I2}(t_y)$  peut être remplacé par l'estimateur  $v(t_y)$  asymptotiquement identique, mais calculable, dans l'équation (3.1) puisque  $\sum_{j \in S} (1 - \pi_k \pi_j / \pi_{kj})$  est borné pour tout  $k$  sous les hypothèses (3.3) et (3.4), ce qui permet de substituer  $e_k$  et  $\alpha_k$  à  $e_k^*$  et  $1/p_k$  inconnus, respectivement (parce que  $e_k^* - e_k$  et  $\alpha_k - 1/p_k$  sont  $O_p(1/\sqrt{n})$  pour tout  $k$ ).

### 3.2 Estimation de la variance sous le modèle de prédiction

Les choses sont un peu plus simples quand nous supposons qu'un modèle de prédiction est vérifié mais que le modèle de réponse de l'équation (2.4) ne l'est pas nécessairement. Supposons que  $E(y_k | \mathbf{x}_k, \mathbf{z}_k) = \mathbf{z}_k^T \boldsymbol{\beta}$ , peu importe que l'unité  $k$  soit échantillonnée ou non ou qu'elle réponde ou non quand elle est échantillonnée, et que les  $\varepsilon_k = y_k - \mathbf{z}_k^T \boldsymbol{\beta}$  sont des variables aléatoires non corrélées de variance égale à  $\sigma_k^2 = \mathbf{z}_k^T \boldsymbol{\eta}$ , où  $\boldsymbol{\eta}$  ne nécessite pas d'autres spécifications que le fait d'avoir des composantes finies.

L'erreur quadratique moyenne de  $t_y$  en tant qu'estimateur de  $T_y$  sous le modèle de prédiction est égale à la somme de la variance de prédiction de  $t_y$  en tant qu'estimateur de  $T_y$ ,  $\sum_R (w_k^2 - w_k) \sigma_k^2$  (voir, par exemple, Kott 2009, page 69), et du carré du biais,  $(\sum_S \mathbf{x}_k^T \boldsymbol{\beta} - \sum_U \mathbf{x}_k^T \boldsymbol{\beta})^2$ , ce dernier étant égal à zéro quand  $\theta = 0$ . La variance combinée de  $t_y$  en tant qu'estimateur de  $T_y$  sous le modèle de prédiction et le plan d'échantillonnage original est donnée par

$$V_C = \theta \text{Var}_D \left( \sum_S \mathbf{x}_k^T \boldsymbol{\beta} \right) + E_D \left[ \sum_S (w_k^2 - w_k) \sigma_k^2 \right],$$

où l'indice inférieur  $D$  indique que l'opération (variance ou espérance) est effectuée par rapport au plan d'échantillonnage original. Rappelons que  $w_k = 0$  pour  $k \neq R$ .

Pour voir que  $v(t_y)$  dans l'équation (3.1) donne un estimateur quasi sans biais de  $V_C$ , observons d'abord que

$$e_k = y_k - \mathbf{z}_k^T \mathbf{b} = \varepsilon_k - \mathbf{z}_k^T \left[ N^{-1} \sum_R d_j \alpha'(\mathbf{g}^T \mathbf{x}_j) \mathbf{x}_j \mathbf{z}_j^T \right]^{-1} N^{-1} \sum_R d_j \alpha'(\mathbf{g}^T \mathbf{x}_j) \mathbf{x}_j \varepsilon_j.$$

Soit  $\delta_{kj} = 1$  quand  $k = j$  et 0 autrement. Parce que les  $\varepsilon_k$  ne sont pas corrélés, et que  $E(\varepsilon_k^2) = \sigma_k = \mathbf{z}_k^T \boldsymbol{\eta}$ , il est maintenant facile de montrer que  $E(e_k e_j) = \delta_{kj} \sigma_k^2 + O(1/n)$  pour presque chaque paire  $k, j$  sous le modèle de prédiction quand  $N^{-1} \sum_R d_k \boldsymbol{\alpha}'(\mathbf{g}^T \mathbf{x}_k) \mathbf{z}_k \mathbf{x}_k^T$  converge vers une matrice inversible, et que les hypothèses (3.3), (3.4), et

$$\frac{\sum_{j=1}^N \psi_j^r}{N} \leq B_2 < \infty \text{ où } \psi_j \text{ désigne toute composante de } \mathbf{x}_j \text{ ou } \mathbf{z}_j, \text{ et } r = 1, 2, 3 \text{ ou } 4, \quad (3.8)$$

sont vérifiées. Observons que le changement provenant des hypothèses dans (3.5) à (3.8) fait que le biais relatif de  $v(t_y)$  est un estimateur de  $V_C$  (ou  $\sum_R (w_k^2 - w_k) \sigma_k^2$  quand  $\theta = 0$ )  $O(1/n)$  plutôt que  $O(1/\sqrt{n})$ .

## 4 Pondération par calage en deux étapes

### 4.1 Pondération par calage en deux étapes

En pratique, les composantes de  $\mathbf{x}_k$  sont souvent des identificateurs d'appartenance à un groupe de type 0/1, et les groupes sont mutuellement exclusifs et exhaustifs. Dans cette situation,  $\mathbf{g}^T \mathbf{x}_k$  ne peut prendre que  $P$  valeurs. *Presque* toute fonction d'ajustement des poids,  $\alpha(\mathbf{g}^T \mathbf{x}_k)$ , donnera des résultats équivalents. La fonction linéaire,  $\alpha(\mathbf{g}^T \mathbf{x}_k) = 1 + \mathbf{g}^T \mathbf{x}_k$ , de Lundström et Särndal (1999) en est un exemple.

Une fonction d'ajustement des poids d'usage répandu qui, parfois, *ne peut pas* être utilisée (noter le mot « presque » en italiques dans le paragraphe précédent) est  $\alpha(\mathbf{g}^T \mathbf{x}_k) = 1 + \exp(\mathbf{g}^T \mathbf{x}_k)$ , qui suppose que la réponse est une fonction logistique de  $\mathbf{x}_k$ . Le problème est que cette fonction d'ajustement des poids ne peut pas retourner des valeurs plus petites que l'unité. Nous avons mentionné à la section précédente que, parfois, on peut avoir besoin que  $\alpha_k$  soit plus petit que 1. Une routine qui essaie d'utiliser  $\alpha(\mathbf{g}^T \mathbf{x}_k) = 1 + \exp(\mathbf{g}^T \mathbf{x}_k)$  et d'ajuster les équations de calage échouera.

Cela peut poser problème en particulier quand on émet l'hypothèse d'un modèle de réponse logistique et que l'on essaie de le caler sur la population en une seule étape. Il pourrait exister une composante de  $\mathbf{z}_k$ , disons  $z_{ka}$ , qui est toujours non négative, mais l'échantillon original et l'ensemble de réponses sont tels que  $\sum_R d_k z_{ka} > \sum_U z_{ka}$  même si  $\sum_R d_k z_{ka}$  ne peut pas excéder  $\sum_S d_k z_{ka}$ . Donc, le calage sur la population échouera toujours, parce qu'aucun  $\alpha_k$  ne peut être plus petit que 1.

Le calage sur l'échantillon original, par contre, ne doit pas échouer, puisque  $\sum_R d_k z_{ka} \leq \sum_S d_k z_{ka}$ . Cela suggère que l'on effectue d'abord le calage sur l'échantillon original, ce qui élimine le biais de réponse si le modèle hypothétique de réponse est vérifié, puis sur la population, ce qui élimine le biais de réponse si le modèle de prédiction est vérifié. Estevao et Särndal (2002) discutent de divers moyens de procéder au calage par étapes, mais nous nous concentrons sur une seule méthode ici.

Un deuxième avantage de la pondération par calage en deux étapes tient au fait qu'elle peut être réalisée même si les variables de calage utilisées aux deux étapes sont les mêmes ou sont un

sous-ensemble de celles utilisées dans la méthode en une seule étape. Cela se produit quand le modèle de réponse est vérifié et que le modèle de prédiction linéaire n'est qu'approximativement vrai. Une certaine version ou estimation « optimale » peut alors être utilisée à la deuxième étape de pondération par calage pour accroître l'efficacité. Rao (1994) a introduit la notion d'estimateur par la régression optimal. Il a été mis sous forme de pondération par calage et discuté plus en détail dans Bankier (2002) et dans Kott (2009, section 4.2). Des renseignements détaillés sur la façon dont cela peut être fait sont fournis aux sections 4.2 et 5.

## 4.2 Estimation et estimation de la variance sous calage en deux étapes

À la présente sous-section, nous commençons par décrire un estimateur par calage en deux étapes assez général d'un total, puis nous abordons l'estimation de sa variance. La première étape de pondération par calage, qui est effectuée sur l'échantillon original, emploie  $\mathbf{x}_{1k}$  comme vecteur des variables du modèle de réponse et  $\mathbf{z}_{1k}$  comme vecteur de calage. Chacun possède  $P_1$  composantes. La fonction d'ajustement des poids est de la forme décrite à l'équation (2.4) où  $\mathbf{g}_1$  remplace maintenant  $\mathbf{g}$ . L'équation de calage est  $\sum_R d_k \alpha (\mathbf{g}_1^T \mathbf{x}_{1k}) \mathbf{z}_{1k} = \sum_S d_k \mathbf{z}_{1k}$ .

La deuxième étape de la pondération par calage, qui est effectuée sur la population, emploie  $\mathbf{x}_{2k}$  et  $\mathbf{z}_{2k}$ , chacun ayant  $P_2$  composantes. Le biais de non-réponse sous le modèle de réponse est éliminé à la première étape. Comme fonction d'ajustement des poids pour la deuxième étape, nous proposons d'utiliser

$$h_k(\mathbf{g}_2^T \mathbf{x}_{2k}) = \frac{\ell_k + \exp(\mathbf{g}_2^T \mathbf{x}_{2k})}{1 + \exp(\mathbf{g}_2^T \mathbf{x}_{2k})/u_k}, \quad (4.1)$$

où l'on peut fixer  $u_k > \ell_k > 0$  presque à sa guise (mais voir plus bas). Le deuxième membre de l'équation (4.1) peut varier sur les unités  $k$  (et peut donc dépendre de  $d_k$  et  $\alpha_k$ ), pourtant  $h_k(0) = h'_k(0) = 1$ , ce qui la rend asymptotiquement indistinguable de la fonction linéaire :  $1 + \mathbf{g}_2^T \mathbf{x}_{2k}$ . Pour simplifier, nous désignerons  $h_k(\mathbf{g}_2^T \mathbf{x}_{2k})$  et  $h'_k(\mathbf{g}_2^T \mathbf{x}_{2k})$ , par  $h_k$  et  $h'_k$ , respectivement. Du point de vue d'un quasi-plan d'échantillonnage, les deux fonctions sont asymptotiquement identiques à l'unité. La deuxième équation de calage est  $\sum_S d_k h_k(\mathbf{g}_2^T \mathbf{x}_{2k}) \mathbf{z}_{2k} = \sum_U \mathbf{z}_{2k}$ . Comme cette équation doit être vérifiée, il existe des limites aux choix disponibles pour  $u_k$  et  $\ell_k$  dans l'équation (4.1).

Un bon estimateur simultané des variances pour  $t_y = \sum_R w_k y_k = \sum_R d_k \alpha (\mathbf{g}_1^T \mathbf{x}_{1k}) h_k(\mathbf{g}_2^T \mathbf{x}_{2k}) y_k$  est (comme nous le verrons)

$$v(t_y) = \sum_{k,j \in S} \left( 1 - \frac{\pi_k \pi_j}{\pi_{kj}} \right) [d_k (\mathbf{z}_{1k}^T \mathbf{b}_1 + \alpha_k h_k e_{1k})] [d_j (\mathbf{z}_{1j}^T \mathbf{b}_1 + \alpha_j h_j e_{1j})] + \sum_{k \in R} d_k (h_k^2 \alpha_k^2 - h_k \alpha_k) e_{1k}^2, \quad (4.2)$$

où

$$e_{2k} = y_k - \mathbf{z}_{2k}^T \left( \sum_S d_j \alpha_j h'_j \mathbf{x}_{2j} \mathbf{z}_{2j}^T \right)^{-1} \sum_S d_j \alpha_j h'_j \mathbf{x}_{2j} y_j, \quad (4.3)$$

$$\mathbf{b}_1 = \left( \sum_S d_f \alpha'_f \mathbf{x}_{1f} \mathbf{z}_{1f}^T \right)^{-1} \sum_S d_f \alpha'_f h_f \mathbf{x}_{1f} e_{2f}, \quad (4.4)$$

et

$$e_{1k} = e_{2k} - \mathbf{x}_{1k}^T \mathbf{b}_1. \quad (4.5)$$

Soit maintenant  $\mathbf{x}_k$  le vecteur composé des composantes non en double de  $\mathbf{x}_{1k}$  et  $\mathbf{x}_{2k}$ , et définissons  $\mathbf{z}_k$  de manière analogue. Les conditions suffisantes pour que (4.2) soit un estimateur simultané des variances comprennent les composantes correspondantes de l'équation (4.1) selon que le modèle de réponse de l'équation (2.4) est vérifié avec  $\mathbf{x}_{1k}$  remplaçant  $\mathbf{x}_k$  ou que le modèle de prédiction est  $E(y_k | \mathbf{x}_k, \mathbf{z}_k) = \mathbf{z}_{2k}^T \boldsymbol{\beta}_2$ , que l'unité  $k$  soit ou non échantillonnée ou réponde ou non si elle est échantillonnée, et les  $\varepsilon_{2k} = y_k - \mathbf{z}_{2k}^T \boldsymbol{\beta}_2$  sont des variables aléatoires non corrélées de variances égales à  $\sigma_{2k}^2 = \mathbf{z}_{2k}^T \boldsymbol{\eta}_2$ , où  $\boldsymbol{\eta}_2$  ne doit pas être spécifié outre le fait que ses composantes doivent être finies. Maintenant,  $N^{-1} \sum_R d_k \alpha' (\mathbf{g}_1^T \mathbf{x}_{1k}) \mathbf{z}_{1k} \mathbf{x}_{1k}^T$  ainsi que  $N^{-1} \sum_R d_k h'_k (\mathbf{g}_2^T \mathbf{x}_{2k}) \mathbf{z}_{2k} \mathbf{x}_{2k}^T$  sont considérées comme étant de plein rang et bornées quand la taille de l'échantillon devient arbitrairement grande.

L'estimateur de variance donné par l'équation (4.2) est presque le même que l'estimateur donné en (3.1) :  $\mathbf{x}_k$  a été remplacé par  $\mathbf{x}_{1k}$  et  $\mathbf{z}_k$ , par  $\mathbf{z}_{1k}$ , tandis que  $h_k e_{2k}$  se substitue à  $y_k$  (nous parlerons sous peu d'une petite différence). Observons que  $e_{2k}$  est effectivement une expression du « résidu » de la deuxième étape de pondération par calage. Ce résidu est multiplié par la fonction d'ajustement des poids  $h_k$ , qui est asymptotiquement égale à l'unité dans la perspective fondée sur le quasi-plan d'échantillonnage et à une constante du point de vue du modèle de prédiction. Le produit est alors utilisé pour créer le « coefficient de régression » de la première étape  $\mathbf{b}_1$  dans l'équation (4.4) et ses « résidus » connexes  $e_{1k}$  dans l'équation (4.5). Nous effectuons la régression de la deuxième étape pour commencer, parce que  $t_y - T_y = \sum_R w_k y_k - \sum_U y_k = \sum_R w_k e_{2k} - \sum_U e_{2k}$ .

C'est pour estimer le modèle de prédiction de  $t_y$  en tant qu'estimateur de  $T_y, \sum_S (w_k^2 - w_k) \sigma_{2k}^2$ , que la dernière apparition de  $h_k$  dans le deuxième membre de l'équation (4.2) n'est pas élevée au carré, comme elle le serait si  $h_k e_{2k}$  se substituait à  $y_k$  partout. Du point de vue d'un quasi-plan,  $h_k$  est asymptotiquement identique à l'unité, de sorte que, qu'elle soit élevée au carré ou non ne fait asymptotiquement aucune différence.

Notons que les  $h'_j$  ont été insérées dans l'équation (4.3) pour la même raison que  $\alpha'$  a été inséré dans  $\mathbf{b}$  dans l'équation (3.1). Cependant, comme les  $h'_j$  sont asymptotiquement égales à l'unité, elles ne sont pas vraiment nécessaires (et ne remplissent aucune fonction du point de vue d'un modèle de prédiction). Un argument similaire s'applique aux  $h_f$  dans l'équation (4.4) : elles sont asymptotiquement égales à l'unité du point de vue du quasi-plan d'échantillonnage (et font partie d'une estimation de 0 du point de vue du modèle de prédiction).

## 5 Quelques simulations

Comme dans Kott et Liao (2012), nous avons créé une population synthétique,  $U$ , d'hôpitaux à partir du fichier de données à grande diffusion DAWN de 2008. Après avoir créé  $U$ , nous avons tiré indépendamment 3 600 échantillons aléatoires simples stratifiés de taille 400 de  $U$  en utilisant les

définitions des strates du fichier de données à grande diffusion. Ces définitions incorporent l'information sur l'emplacement et la propriété de l'hôpital (publique ou privée) qui n'est pas fournie directement dans le fichier.

Nous avons fixé les tailles des échantillons de strate de façon qu'elles soient approximativement proportionnelles à une mesure de taille  $q_k$ , mais jamais inférieures à quatre. Pour  $q_k$ , nous avons utilisé le nombre annuel de visites au service d'urgence associées à la consommation de drogues, qui était toujours positif. Dans le fichier DAWN, une variable de taille est en fait associée à chaque hôpital figurant dans la base de sondage, à savoir le nombre de visites au service d'urgence durant une année antérieure selon l'*American Hospital Association*. Malheureusement, cette variable n'était pas incluse dans le fichier de données à grande diffusion. Dans nos simulations, les poids de sondage variaient entre 4,375 et 48, ce qui nous a permis de traiter les facteurs de correction pour population finie comme étant ignorables dans l'estimation de la variance.

Comme dans notre article original, nous avons généré un échantillon de répondants  $R$  pour chaque échantillon simulé selon un tirage de Bernoulli à partir de la fonction logistique :

$$p_k = (1 + \exp(3,735 - 0,4 \log(q_k)))^{-1}, \quad (5.1)$$

Nous avons également créé des échantillons de répondants de rechange en utilisant

$$p_k = (1 + \exp(0,597 - 0,005q_k^{1/2}))^{-1}. \quad (5.2)$$

Les modèles de réponse ont tous deux produit des taux de réponse globaux non pondérés d'environ 54 %, ce qui est similaire à la situation réelle du fichier DAWN, où la réponse prend aussi la forme d'une fonction légèrement croissante de la variable de taille. Notons que  $\alpha_k = 1/p_k$  est borné même si ni l'une ni l'autre probabilité ne peut être exprimée par l'équation (2.4) avec une borne supérieure  $u$  finie.

Comme dans l'étude précédente, nous nous sommes concentrés sur l'estimation des totaux de population pour trois variables étudiées. Les nombres annuels de visites au service d'urgence liées à la consommation de drogues avec réaction pharmaceutique indésirable et de celles résultant en un décès ont été extraits du fichier de données à grande diffusion. Puisque ces variables étaient approximativement linéaires en notre mesure de taille, la troisième variable « étudiée » a été construite artificiellement. Il s'agissait de la mesure de taille (nombre de visites annuelles au service d'urgence liées à la consommation de drogues) élevée à la puissance 1,3.

Nous avons étudié huit estimateurs et estimations de leur variance. Les résultats sont résumés au tableau 5.1. Les deux premiers comportaient le calage sur l'échantillon original seulement (équation (2.5) avec  $\theta = 1$ ), en supposant que la réponse était de forme logistique en le logarithme de la mesure de taille. Nous avons employé l'équation (2.3) avec  $\mathbf{x}_k = (1 \log(q_k))^T$ . Le premier estimateur utilisait  $\mathbf{z}_k = (1 \log(q_k))^T$  comme vecteur de calage, tandis que le deuxième utilisait  $\mathbf{z}_k = (1 q_k)^T$ , qui était davantage en harmonie avec un modèle de prédiction raisonnable, du moins pour les réactions indésirables et les décès.

Nos troisième et quatrième estimateurs comportaient le calage sur l'échantillon et sur la population en une seule étape (équation (2.5) avec  $\theta = 1$ , puis  $\theta = 0$ ) en utilisant  $\mathbf{x}_k = \mathbf{z}_k = (1 \log(q_k) q_k)^T$ . Ils

étaient conçus pour être quasiment sans biais si le modèle de réponse logistique en  $(1 \log(q_k))^T$  ou le modèle de prédiction linéaire en  $(1 q_k)^T$  étaient vérifiés.

**Tableau 5.1**  
**Sommaire de l'exercice de simulation (tous les résultats sont exprimés en pourcentage %)**

Estimateur	$t_{y1}$	$t_{y2}$	$t_{y3}$	$t_{y4}$	$t_{y5}$	$t_{y6}$	$t_{y7}$	$t_{y8}$
<i>Calage sur l'échantillon</i>								
Variables du modèle de réponse : $x_{1k}$	$(1 \log(q_k))^T$	$(1 \log(q_k))^T$	$(1 \log(q_k)q_k)^T$	-	$(1 \log(q_k))^T$	$(1 \log(q_k))^T$	$(1 \log(q_k))^T$	$(1 \log(q_k))^T$
Variables de calage : $z_{1k}$	$(1 \log(q_k))^T$	$(1 q_k)^T$	$(1 \log(q_k)q_k)^T$	-	$(1 \log(q_k))$	$(1 q_k)^T$	$(1 \log(q_k))^T$	$(1 q_k)^T$
<i>Calage sur la population</i>								
Variables du modèle de réponse : $x_{2k}$	-	-	-	$(1 \log(q_k)q_k)^T$	$(1 \log(q_k)q_k)^T$	$(1 \log(q_k)q_k)^T$	$f_k(1 \log(q_k)q_k)^T$	$f_k(1 \log(q_k)q_k)^T$
Variables de calage : $z_{2k}$	-	-	-	$(1 \log(q_k)q_k)^T$	$(1 \log(q_k)q_k)^T$	$(1 \log(q_k)q_k)^T$	$(1 \log(q_k)q_k)^T$	$(1 \log(q_k)q_k)^T$
<i>Réponse vraie : <math>p_k = 1/\{1 + \exp[3,735 + 0,4 \log(q_k)]\}</math></i>								
<i>Réactions indésirables</i>								
Biais relatif de $t_y$	-0,07	0,06	-0,11	-0,13	-0,02	-0,07	0,10	0,09
REQM relative de $t_y$	4,97	3,98	4,01	2,45	2,51	2,57	2,40	2,39
Biais relatif de $v(t_y)$	8,60	12,59	12,52	6,24	6,76	6,16	6,76	6,48
<i>Décès</i>								
Biais relatif de $t_y$	-0,17	0,06	-0,20	-0,26	-0,20	-0,30	0,04	-0,07
REQM relative de $t_y$	11,75	11,39	11,56	11,07	11,28	11,36	10,91	10,91
Biais relatif de $v(t_y)$	-1,34	-0,48	-0,90	-0,76	-1,00	-0,60	-0,12	-0,28
<i>(Taille)<sup>1,3</sup></i>								
Biais relatif de $t_y$	-0,16	-0,05	0,08	0,09	0,04	0,06	-0,02	0,01
REQM relative de $t_y$	6,92	5,07	5,06	0,95	1,05	1,12	0,89	0,89
Biais relatif de $v(t_y)$	10,01	18,49	17,47	-2,26	-3,41	-3,32	0,51	-2,12
<i>Réponse vraie : <math>p_k = 1/\{1 + \exp[0,597 + 0,005 q_k^{1/2}]\}</math></i>								
<i>Réactions indésirables</i>								
Biais relatif de $t_y$	2,87	-0,26	0,08	0,04	0,48	0,53	0,15	0,07
REQM relative de $t_y$	5,90	3,97	4,00	2,35	2,43	2,45	2,33	2,35
Biais relatif de $v(t_y)$	-18,22	11,63	11,95	9,90	8,82	7,35	7,19	6,67
<i>Décès</i>								
Biais relatif de $t_y$	1,24	-1,88	0,47	0,36	1,03	1,20	-0,58	-0,67
REQM relative de $t_y$	11,42	11,01	11,41	10,95	11,18	11,26	10,69	10,72
Biais relatif de $v(t_y)$	5,30	3,00	6,27	6,24	5,65	5,06	6,21	5,90
<i>(Taille)<sup>1,3</sup></i>								
Biais relatif de $t_y$	5,17	1,05	-0,07	-0,05	-0,31	-0,36	0,01	0,08
REQM relative de $t_y$	9,11	5,31	5,05	0,85	0,97	1,01	0,80	0,82
Biais relatif de $v(t_y)$	-26,83	11,70	17,09	8,23	0,29	-3,98	5,17	2,90

$$f_k = d_k \alpha_k - 1 = (d_k / \hat{p}_k) - 1$$

Il n'est pas surprenant de constater que l'erreur quadratique moyenne relative (empirique) du quatrième estimateur est toujours plus faible que celle du troisième. La raison en est assez évidente si l'on examine l'équation (3.1) et que l'on considère la conséquence du fait que  $\theta$  est égal à 0 (calage sur la population) plutôt qu'à 1 (calage sur l'échantillon).

Les cinquième à huitième estimateurs ont été calés en deux étapes. Pour les cinquième et septième estimateurs, on a employé la pondération par calage utilisée pour le premier estimateur à la première étape, tandis que pour les sixième et huitième, on a employé la pondération par calage du deuxième estimateur. Pour les cinquième et sixième estimateurs, on a utilisé  $\mathbf{z}_{2k} = \mathbf{x}_{2k} = (1 \log(q_k) q_k)^T$  à la deuxième étape, tandis que les septième et huitième étaient quasi pseudo-optimaux (Kott 2011) en utilisant  $\mathbf{z}_{2k} = (1 \log(q_k) q_k)^T$  et  $\mathbf{x}_{2k} = (d_k \alpha_k - 1) \mathbf{z}_{2k}$  à la deuxième étape. Pour les quatre estimateurs, on a employé les fonctions d'ajustement des poids individuels suivantes :

$$h_k(\mathbf{g}_2^T \mathbf{x}_{2k}) = \frac{1}{d_k \alpha_k} + \left(1 - \frac{1}{d_k \alpha_k}\right) \exp\left[\frac{\mathbf{g}_2^T \mathbf{x}_{2k}}{1 - \frac{1}{d_k \alpha_k}}\right].$$

Comme l'a montré Kott (2011), ces  $h_k(\mathbf{g}_2^T \mathbf{x}_{2k})$  sont asymptotiquement identiques à la fonction d'ajustement des poids,  $1 + \mathbf{g}_2^T \mathbf{x}_{2k}$ , quand  $\mathbf{g}_2^T \mathbf{x}_{2k} = O_p(1/\sqrt{n})$ , mais empêchent tout poids  $w_k$  de devenir inférieur à l'unité. Chacune est une version de l'équation (4.1) avec  $\ell_k = 1/(d_k \alpha_k)$ ,  $c = 1$ , et  $u = \infty$ .

Comme le taux de non-réponse n'était pas élevé, nous n'avons pas eu de problème à calculer les troisième et quatrième estimateurs quel qu'était l'échantillon de répondants simulés utilisé. L'erreur quadratique moyenne relative du quatrième estimateur était systématiquement légèrement plus grande que celle des septième et huitième estimateurs, dans lesquels était incorporé un calage quasi pseudo-optimal à la deuxième étape. Curieusement, cela n'était pas le cas pour la comparaison du quatrième estimateur aux cinquième et sixième estimateurs qui, bien que comprenant les deux étapes, n'intégraient pas le calage quasi pseudo-optimal.

Il convient de souligner que, même si le deuxième estimateur possédait systématiquement une plus petite erreur quadratique moyenne relative que le premier, du fait qu'il était davantage en harmonie avec un modèle de prédiction raisonnable (même pour  $q_k^{1,3}$ , la variable étudiée paraissait plus près d'être linéaire en  $q_k$  qu'en  $\log(q_k)$ ), les autres paires analogues (cinquième c. sixième et septième c. huitième) ne présentaient aucun schéma évident de supériorité. Cela tient au fait que ce sont les résidus de la deuxième étape qui sont effectivement modélisés dans l'équation (4.4) et non les valeurs de  $y$ .

La production de la non-réponse au moyen de l'équation (5.2) plutôt que (5.1) ne semble pas avoir beaucoup d'effet sur les résultats, sauf en ce qui concerne les biais relatifs du premier estimateur. Tant pour les réactions indésirables que pour la (taille)<sup>1,3</sup>, le biais relatif de cet estimateur est supérieur à 40 % de l'erreur quadratique moyenne relative. Il en est vraisemblablement ainsi parce que les deux modèles qui pouvaient être utilisés pour justifier cet estimateur (la réponse est logistique en le logarithme de la mesure de taille et la variable étudiée est linéaire en le logarithme de la mesure de taille) n'ont pas tenu. Il n'est donc pas étonnant, puisque le biais relatif représente une telle part de l'erreur quadratique moyenne relative dans ces deux situations, que  $v(t_k)$  sous-estime fortement l'erreur quadratique moyenne. Nulle part ailleurs le biais relatif de  $v(t_k)$  n'est supérieur à 15 %.

Il semble que même notre variable artificielle, (taille)<sup>1,3</sup>, s'approchait suffisamment de la linéarité en la mesure de taille pour que le biais ne soit jamais un problème pour tout autre estimateur que le premier.

Le premier estimateur lui-même avait un biais relatif négligeable quand la réponse était un modèle logistique du logarithme de la mesure de taille, comme on le suppose.

## 6 Conclusion

À la section 4, nous avons mentionné deux raisons de préférer la pondération par calage en deux étapes : rendre l'ajustement implicite d'un modèle de réponse logistique plus facile et intégrer le calage presque quasi-optimal. Un avantage secondaire du calage en deux étapes est une estimation plus efficace du modèle de réponse à la première étape, puisque aucune erreur d'échantillonnage ne fausse l'estimation. Cette propriété est utile si l'on veut analyser les causes de la non-réponse totale en tant que fin en soi.

Nous concédons, cependant, que la réduction de l'erreur quadratique moyenne en utilisant les deux étapes était modeste dans nos expériences par simulation à la section 5. En outre, nous ne pouvons nier l'attrait pratique de la simplicité du calage en une seule étape.

Lorsqu'on utilise la pondération par calage pour corriger la non-réponse quand les réponses ne manquent pas au hasard comme il est décrit dans Chang et Kott (2008) et dans Kott et Chang (2010), des gains d'efficacité vraisemblablement importants découlent d'une deuxième étape où n'interviennent que des variables de calage et des fonctions des variables de calage comme variables du modèle.

Quand les facteurs de correction pour population finie peuvent être ignorés, le rééchantillonnage offre une approche beaucoup plus simple d'estimation de la variance que l'équation (3.7), même si l'on peut laisser tomber la deuxième sommation dans le deuxième membre dans cette situation. Une autre option intéressante est la version « contractée » de l'équation (4.2) qui ignore l'effet de la première étape de calage :

$$\tilde{v}(t_y) = \sum_{k,j \in S} \left( 1 - \frac{\pi_k \pi_j}{\pi_{kj}} \right) [w_k e_{2k}] [w_j e_{2j}] + \sum_{k \in R} d_k (h_k^2 \alpha_k^2 - h_k \alpha_k) e_{2k}^2.$$

Cet estimateur estime manifestement la variance du modèle de prédiction si ce modèle est vérifié. Une version de cet estimateur – avec la deuxième sommation supprimée – a donné de bons résultats dans nos expériences par simulation (résultats non présentés). Une certaine prudence est de rigueur avant de tirer une conclusion trop catégorique de ce résultat, puisque le modèle linéaire n'était jamais très loin d'être vérifié dans nos investigations.

Enfin, un certain nombre d'hypothèses ont été faites pour simplifier l'exposé. Le lecteur que cela intéresse peut étendre les résultats à une  $d_k$  non bornée ou à des fonctions d'ajustement des poids plus générales et qui ne sont pas nécessairement bornées, ou permettre que les erreurs du modèle de prédiction soient corrélées à l'intérieur des unités primaires d'échantillonnage. Quand  $N$  augmente plus rapidement que  $n$ , l'hypothèse selon laquelle  $\sigma_k^2 = \mathbf{z}_k^T \boldsymbol{\eta}$  peut parfois être abandonnée. Voir, par exemple, Kott (2009, page 69).

## Remerciements

Le présent article a été préparé à l'occasion du *Symposium on the Analysis of Survey Data and Small Area Estimation* organisé en l'honneur du 75<sup>e</sup> anniversaire du professeur J.N.K. Rao et parrainé par le

*Fields Institute for Research in Mathematical Sciences*. Les auteurs remercient les organisateurs de la conférence de les avoir invités à présenter cet article et l'Institut de son généreux financement de la conférence sans lequel le présent article n'aurait jamais été rédigé. Ils remercient également plusieurs rédacteurs et examinateurs de leurs commentaires utiles.

## Bibliographie

- Bang, H., et Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.
- Bankier, M. (2002). Regression estimators for the 2001 Canadian Census. Présenté à l'International Conference in Recent Advances in Survey Sampling.
- Chang, T., et Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 557-571.
- Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics*, 11, 427-444.
- Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. Dans *COMPSTAT: Proceedings in Computational Statistics, 14<sup>th</sup> Symposium, Utrecht, The Netherlands*, (Éds., J.G. Bethlehem et P.G.M. Van der Heidjen), Heidelberg : Physica Verlag, 65-76.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 418, 376-382.
- Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.
- Estevao, V.M., et Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.
- Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the American Statistical Association, Social Statistics Section*, 197-202.
- Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the American Statistical Association, Survey Research Methods Section*, disponible en ligne au <http://www.amstat.org/sections/srms/Proceedings/>, 598-603.
- Fuller, W.A., Loughin, 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, 1, 79-89.
- Kim, J.K., et Haziza, D. (2014). Doubly robust inference with missing survey data. *Statistica Sinica*, 24, 375-394.

- Kim, J.K., et Park, H. (2006). Imputation using response probability. *Canadian Journal of Statistics*, 34, 1-12.
- Kim, J.K., et Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*, Londres : Chapman and Hall/CRC.
- 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, 2, 149-160.
- Kott, P.S. (2009). Calibration weighting: Combining probability samples and linear prediction models. Dans *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*, (Éds., D. Pfeffermann et C.R. Rao), New York : Elsevier.
- Kott, P.S. (2011). A nearly pseudo-optimal method for keeping calibration weights from falling below unity in the absence of nonresponse or frame errors. *Pakistan Journal of Statistics*, 27, 391-396.
- Kott, P.S., et Chang, T.C. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265-1275.
- Kott, P.S., et Liao, D. (2012). Comparing weighting methods when adjusting for logistic unit Nonresponse. Présenté au Federal Committee on Survey Methodology Research Conference, disponible en ligne au [http://www.fcs.m.sites.usa.gov/files/2014/05/Kott\\_2012FCSM\\_III-B.pdf](http://www.fcs.m.sites.usa.gov/files/2014/05/Kott_2012FCSM_III-B.pdf).
- Little, R.J., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2<sup>e</sup> Éd.), New York : John Wiley & Sons, Inc.
- Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.
- Oh, H.L., et Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, (Éds., W.G. Madow, I. Olkin et D.B. Rubin), New York : Academic Press, 2.
- Rao, J.N.K. (1994). Estimation of totals and distributing functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- Robins J.M., Rotnitzky A. et Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, p. 846-866.