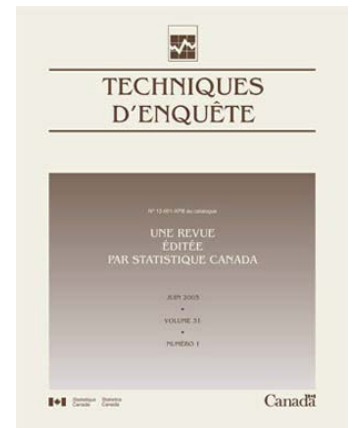


Techniques d'enquête

Poids de calage « optimaux » dans les cas de non-réponse des unités lors de l'échantillonnage d'enquête

par Per Gösta Andersson

Date de diffusion : le 17 décembre 2019



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.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

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

- | | |
|---|----------------|
| • Service de renseignements statistiques | 1-800-263-1136 |
| • Service national d'appareils de télécommunications pour les malentendants | 1-800-363-7629 |
| • Télécopieur | 1-514-283-9350 |

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 sous « Contactez-nous » > « [Normes de service à la clientèle](#) ».

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.

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2019

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.

Poids de calage « optimaux » dans les cas de non-réponse des unités lors de l'échantillonnage d'enquête

Per Gösta Andersson¹

Résumé

Un taux de non-réponse élevé est un problème très courant dans les enquêtes-échantillons de nos jours. Du point de vue statistique, nous entretenons des inquiétudes au sujet du biais et de la variance accrus des estimateurs de chiffres de population comme les totaux ou les moyennes. Diverses méthodes ont été proposées pour compenser ce phénomène. En gros, nous pouvons les diviser en imputation et calage, et c'est sur la dernière méthode que nous nous concentrons ici. La catégorie des estimateurs par calage offre un large éventail de possibilités. Nous examinons le calage linéaire, pour lequel nous suggérons d'utiliser une version de non-réponse de l'estimateur de régression optimal fondé sur le plan. Nous faisons des comparaisons entre cet estimateur et un estimateur de type GREG. Les mesures de la distance jouent un rôle très important dans l'élaboration des estimateurs par calage. Nous démontrons qu'un estimateur de la propension moyenne à répondre (probabilité) peut être inclus dans la mesure de la distance « optimale » dans les cas de non-réponse, ce qui aide à réduire le biais de l'estimateur ainsi obtenu. Une étude en simulation a été réalisée pour illustrer de manière empirique les résultats obtenus de façon théorique pour les estimateurs proposés. La population se nomme KYBOK et se compose de municipalités administratives de la Suède, pour lesquelles les variables comprennent des mesures financières et de la taille. Les résultats sont encourageants pour l'estimateur « optimal » combiné à la propension estimative moyenne à répondre, où le biais a été réduit pour la plupart des cas d'échantillonnage de Poisson faisant partie de l'étude.

Mots-clés : Non-réponse des unités; poids de calage; échantillonnage de Poisson.

1 Introduction

Dans une enquête, le mécanisme de réponse (non-réponse) des unités est en fait inconnu. Pour éviter de définir une mesure de probabilité appropriée qui pourrait n'être ni significative ni réaliste, les cas de non-réponse sont habituellement traités en fonction de la propension d'une unité à participer. Pour tenir compte de l'effet possible de la non-réponse sur les estimateurs, la pratique courante consiste cependant à traiter les propensions comme des probabilités à estimer (par exemple, scores de propension), ce qui peut se faire pour des unités individuelles, pour des groupes d'unités ou en tant que « moyenne » pour l'ensemble des réponses.

Par exemple, dans Haziza et Lesage (2016), deux approches principales sont examinées : la pondération par calage avec et sans la pondération fondée sur les scores de propension susmentionnée, le premier cas comprenant l'estimation basée sur un modèle. Les auteurs font une mise en garde à propos des effets négatifs possibles sur le biais et la variance des estimateurs obtenus quand les propensions ne sont pas prises en compte (Ces deux options de pondération sont appelées par les auteurs procédures en deux étapes et en une étape, à ne pas confondre respectivement avec le calage en deux étapes et en une seule étape défini par Särndal et Lundström (2005).) Toutefois, dans l'étude en simulation de Haziza et Lesage (2016), le plan d'échantillonnage ne joue aucun rôle, puisque $n = N$ et que l'accent est mis uniquement sur le lien entre,

1. Per Gösta Andersson, professeur agrégé, Département de la statistique, Université de Stockholm, Stockholm. Courriel : per.gosta.andersson@stat.su.se.

d'une part, les renseignements auxiliaires et, d'autre part, la variable à l'étude et le mécanisme de non-réponse.

Dans le présent document, nous proposons d'utiliser une version de non-réponse pour ce qui, dans les cas de réponse complète, porte le nom d'estimateur de régression optimal (fondé sur le plan). La mesure sous-jacente de la distance est une forme quadratique dont la structure est plus complexe (voir Andersson et Thorburn (2015)) que celle qui donne l'estimateur GREG (voir Deville et Särndal (1992)). Il s'avère qu'il y a également place à l'amélioration en ce qui concerne la propension moyenne à répondre (probabilité) lors de l'élaboration de la mesure de la distance dans les cas de non-réponse, ce qui donne un estimateur « optimal » modifié.

1.1 Aperçu du document

La section 2 commence par une introduction à l'idée du calage dans les cas de réponse complète avant de s'attacher aux cas de non-réponse. Nous nous intéressons surtout à trois estimateurs du total de la population : l'estimateur GREG et deux versions de l'estimateur « optimal ». Certains résultats théoriques du biais qui en découle sont ensuite présentés. La section 3 porte sur une étude en simulation dans laquelle l'échantillonnage aléatoire simple et l'échantillonnage de Poisson sont utilisés à des fins d'illustration. Le plan d'échantillonnage de Poisson nous permet d'élaborer et d'étudier une situation où les renseignements auxiliaires font partie du plan et du mécanisme de non-réponse. Nous illustrons également les risques liés à l'utilisation du mauvais modèle pour estimer des propensions individuelles. Nous terminons en faisant des observations finales à la section 4.

1.2 Notation et configuration

Nous commençons par une population U de taille N dans laquelle nous prélevons un échantillon probabiliste s de taille n_s avec des probabilités d'inclusion $\pi_1, \dots, \pi_N$. La non-réponse signifie que nous observons uniquement l'ensemble des réponses r de taille n_r . Notre but est estimer le total de la variable à l'étude $t_y = \sum_U y_k$. Nous supposons l'accès à un vecteur de variable auxiliaire $\mathbf{x}$ de dimension J , où soit $\mathbf{x} = \mathbf{x}^*$ et $(\mathbf{x}_k^*)_{k \in U}$ sont connus (le niveau de la population), soit $\mathbf{x} = \mathbf{x}^o$ et $(\mathbf{x}_k^o)_{k \in s}$ sont connus (le niveau de l'échantillon) ou possiblement une combinaison de ces cas : $\mathbf{x} = (\mathbf{x}^{*'}, \mathbf{x}^{o'})'$.

2 Estimation par calage

2.1 Estimateurs par calage dans les cas de réponse complète

En commençant par les cas de réponse complète ($r = s$) et en suivant la procédure établie par Deville et Särndal (1992), l'estimateur par calage est défini comme

$$\hat{t}_{y\text{cal}} = \sum_s w_{ks} y_k,$$

où les poids dépendants de l'échantillon w_{ks} sont choisis pour que

$$\sum_s w_{ks} \mathbf{x}_k = \mathbf{t}_x, \quad (\text{l'équation du calage}) \quad (2.1)$$

tout en réduisant la mesure de la distance quadratique

$$(\mathbf{w}_s - \mathbf{w}_{0s})' \mathbf{R} (\mathbf{w}_s - \mathbf{w}_{0s}),$$

où $\mathbf{w}_s = (w_{ks})_{k \in s}$, $\mathbf{w}_{0s} = (1/\pi_k)_{k \in s} = (d_k)_{k \in s}$ et $\mathbf{R}$ sont des diagonales. (D'autres mesures de la distance sont envisagées dans Deville et Särndal (1992) et Haziza et Lesage (2016).)

En d'autres termes, étant donné la contrainte (2.1), w_{ks} devrait être « le plus près possible » des poids déterminés par le plan d'échantillonnage d_k , ce qui est souhaitable puisque $\sum_s d_k y_k$ est un estimateur sans biais de t_y .

Les poids qui en découlent sont

$$\mathbf{w}_s = \mathbf{w}_{0s} + \mathbf{R}^{-1} \mathbf{x}' (\mathbf{X} \mathbf{R}^{-1} \mathbf{X}')^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_x).$$

Il s'ensuit que l'estimateur GREG homoscédastique assisté par modèle $\hat{t}_{yr}$ (Särndal, Swensson et Wretman (1992)) est un estimateur par calage pour lequel

$$\mathbf{R} = (\mathbf{w}_{0s} \mathbf{I}_{n_s})^{-1},$$

où $\mathbf{I}_{n_s}$ est la matrice diagonale unitaire de taille n_s .

Un autre estimateur par calage est l'estimateur de régression optimal $\hat{t}_{y\text{opt}}$ (voir, par exemple, Rao (1994) et Montanari (1998)), selon lequel

$$\mathbf{R} = \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} \right)_{k, l \in s}^{-1},$$

comme le montrent Andersson et Thorburn (2005).

D'une façon asymptotique, cet estimateur a une variance minimale (dans le sens qu'il est fondé sur le plan) parmi les estimateurs de régression linéaire.

2.2 Estimateurs par calage dans les cas de non-réponse

Dans les cas de non-réponse, un estimateur par calage possible est

$$\sum_r w_{kr} y_k,$$

où on devrait supposer que

$$\sum_r w_{kr} \mathbf{x}_k = \mathbf{X}, \quad (2.2)$$

où $\mathbf{X} = \sum_U \mathbf{x}_k^*$, si les renseignements auxiliaires sont connus jusqu'au niveau de la population. Autrement, $\mathbf{X} = \sum_s d_k \mathbf{x}_k^o$, l'estimateur sans biais de $\mathbf{t}_x$. (Nous pouvons également combiner les deux types de renseignements dans la contrainte $\mathbf{X}$.)

Dans de nombreux cas, des poids qui remplissent l'exigence (2.2) sont présentés, par exemple, par Särndal et Lundström (2005). À l'aide de l'approche directe, où tous les renseignements sont utilisés dans un seul calage, nous obtenons

$$w_{kr} = d_k \left(1 + \mathbf{x}_k' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right) \right). \quad (2.3)$$

L'estimateur ainsi obtenu sera dorénavant désigné comme $\hat{t}_{y\text{cal}}$. (D'autres approches, notamment les procédures en deux étapes, sont présentées et examinées, par exemple, par Andersson et Särndal (2016).)

Il faut se poser une question évidente : quelle mesure sous-jacente de la distance génère ces poids ? Särndal et Lundström (2005) ne font pas de commentaires sur cette question précise mais, selon Lundström et Särndal (1999), nous devrions choisir « w_k "le plus près possible" de d_k », ce qui ne semble pas être tout à fait approprié dans les cas de non-réponse. Si nous retournons à Lundström (1997), nous pouvons constater que la mesure correspondante de la distance *est* effectivement

$$(\mathbf{w}_r - \mathbf{w}_{0r})' (\mathbf{w}_{0r} \mathbf{I}_{n_r})^{-1} (\mathbf{w}_r - \mathbf{w}_{0r}),$$

où $\mathbf{w}_r = (w_{kr})_{k \in r}$ et $\mathbf{w}_{0r} = (d_k)_{k \in r}$.

Si un mécanisme aléatoire génère l'ensemble des réponses r de l'échantillon s avec des probabilités θ_k d'inclusion, nous pouvons envisager les cas de non-réponse comme un plan de sondage à deux phases et c'est exactement l'hypothèse que nous ferons ci-après. Ensuite, nous devrions réduire la distance entre w_{kr} et $d_k \cdot (1/\theta_k)$. Grâce à la modélisation, θ_k peut être estimé à l'aide de $\hat{\theta}_k$, afin d'être utilisé dans la minimisation de la distance. Toutefois, dans le présent document, nous n'irons pas dans la direction de l'inférence basée sur un modèle. Pour réduire l'effet du biais dans les cas de non-réponse, dans la mesure de la distance, nous pourrions plutôt penser à comparer w_{kr} non pas avec d_k , mais bien avec $d_{k,\text{alt}} = d_k \cdot c$, où c est une constante supérieure à 1, afin de compenser l'effet « moyen » de la non-réponse.

Or, Lundström (1997) montre que, dans bien des cas importants, notamment quand on peut trouver un vecteur $\boldsymbol{\mu}$ pour lequel $\boldsymbol{\mu}' \mathbf{x}_k = 1$, pour tous les k , l'augmentation multiplicative de $d_{k,\text{alt}}$ implique les mêmes poids de calage w_{kr} ainsi obtenus. C'est ce qu'on obtient si $\boldsymbol{\mu}' \mathbf{x}_k = 1$, pour tous les $k \in U$, nous pouvons simplifier l'expression (2.3) de w_{kr} comme ceci

$$w_{kr} = d_k \mathbf{x}_k' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{X}.$$

Par conséquent, nous avons une propriété d'invariance pour les poids. Ce résultat demeure valide quand la population est fractionnée en groupes et les poids initiaux sont gonflés à l'aide d'une constante dans chaque groupe. Il faut souligner que, si nous incluons une constante, par exemple, « 1 », comme première composante du vecteur auxiliaire $\mathbf{x}_k$, nous pouvons simplement supposer que $\boldsymbol{\mu}' = (1, 0, \dots, 0)$ pour obtenir $\boldsymbol{\mu}' \mathbf{x}_k = 1$.

Dans ce contexte, nous proposons d'utiliser d'autres poids « optimaux » découlant de la mesure de la distance

$$(\mathbf{w}_r - \mathbf{w}_{0r})' \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} \right)_{k,l \in r}^{-1} (\mathbf{w}_r - \mathbf{w}_{0r}),$$

ce qui donne $\hat{t}_{y\text{opt}}$. (π_{kl} désigne la probabilité d'inclusion de la paire (k, l) .)

Il faut observer que, comme dans les cas de réponse complète, il arrive parfois que les poids « optimaux » sont identiques à (2.3), par exemple, sous échantillonnage aléatoire simple.

L'emploi de guillemets autour du mot *optimal* est délibéré mais, dans le contexte de la réponse complète, le mot *optimal* a un sens très clair. Tel que susmentionné, l'estimateur de régression optimal a une variance minimale asymptotique parmi les estimateurs de régression linéaire. En raison de l'ajout de la non-réponse là où le mécanisme de non-réponse est au moins partiellement inconnu, il est plus difficile de bien définir les critères d'optimalité.

Pour cette mesure « optimale », il pourrait être utile de remplacer d_k par $d_{k,\text{alt}}$, où nous incluons dans $d_{k,\text{alt}}$ la réciproque d'une estimation de la probabilité de réponse moyenne $\bar{\theta}_U = \sum_U \theta_k / N$. Voici un candidat simple :

$$\hat{\bar{\theta}}_U = n_r / n_s,$$

ce qui donne $d_{k,\text{alt}} = d_k \cdot (n_s / n_r)$. Voici un autre choix naturel :

$$\hat{\bar{\theta}}_U = \sum_r d_k / \sum_s d_k, \quad (2.4)$$

puisque $E(\sum_s d_k) = N$ et $E(\sum_r d_k) = \sum_U \theta_k = N\bar{\theta}$, ce qui donne $E(\sum_r d_k / \sum_s d_k) \approx \bar{\theta}_U$. L'estimateur modifié ainsi obtenu est désigné par $\hat{t}_{y\text{optm}}$. (il faut aussi observer que $E(n_r / n_s) \approx \sum_U (\theta_k / d_k) / \sum_U (1 / d_k)$.)

Dans l'étude en simulation suivante, nous nous concentrons sur un plan d'échantillonnage où, en général, $\hat{t}_{y\text{cal}} \neq \hat{t}_{y\text{opt}}$, notamment l'échantillonnage de Poisson. L'indépendance des dessins simplifie la mesure de la distance « optimale » :

$$\sum_r \frac{\pi_k^2}{1 - \pi_k} (w_{kr} - d_k)^2 = \sum_r \frac{(w_{kr} - d_k)^2}{d_k (d_k - 1)}$$

et la minimisation donne

$$w_{kr} = d_k \left(1 + (d_k - 1) \mathbf{x}'_k \left(\sum_r d_k (1 - d_k) \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right) \right).$$

Pour l'estimateur « optimal » modifié, d_k est remplacé par $d_{\text{kalt}} = d_k \cdot (1/\hat{\theta}_U)$, avec $\hat{\theta}_U$, comme dans (2.4).

2.2.1 Biais des estimateurs par calage dans les cas de non-réponse

Nous pouvons représenter $\hat{t}_{\text{ycal}}$ comme suit

$$\hat{t}_{\text{ycal}} = \sum_r d_k y_k + \hat{\mathbf{B}}_{U;\theta} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right), \quad (2.5)$$

où $\hat{\mathbf{B}}_{U;\theta} = \left(\sum_r d_k \mathbf{x}'_k y_k \right) \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1}$. Afin d'obtenir une équation d'approximation du biais de $\hat{t}_{\text{ycal}}$ et par la suite de $\hat{t}_{\text{yopt}}$ et $\hat{t}_{\text{yoptm}}$, nous suivons le calcul présenté dans Särndal et Lundström (2005) et nous constatons d'abord que $\hat{t}_{\text{ycal}}$ peut être réécrit comme suit :

$$\hat{t}_{\text{ycal}} = \sum_r d_k y_k + \mathbf{B}_{U;\theta} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right) + (\hat{\mathbf{B}}_{U;\theta} - \mathbf{B}_{U;\theta}) \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right),$$

où $\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}'_k y_k \right) \left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1}$.

Si nous supposons que $\hat{t}_{\text{ycal}} - t_y = A_1 + A_2$, où $A_1 = \sum_r d_k y_k - t_y + \mathbf{B}_{U;\theta} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right)$ et $A_2 = (\hat{\mathbf{B}}_{U;\theta} - \mathbf{B}_{U;\theta}) \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right)$, cela peut par ailleurs démontrer que

$$A_1 = \sum_r d_k e_{\theta k} - \sum_U e_{\theta k} + \mathbf{B}_{U;\theta}^o \left(\sum_s d_k \mathbf{x}_k^o - \sum_U \mathbf{x}_k^o \right),$$

où $e_{\theta k} = y_k - \mathbf{B}_{U;\theta} \mathbf{x}_k$ et $\mathbf{B}_{U;\theta}^o = \left(\sum_U \theta_k \mathbf{x}_k^o \mathbf{x}_k^{o'} \right)^{-1} \sum_U \theta_k \mathbf{x}_k^o y_k$.

Alors,

$$E(\hat{t}_{\text{ycal}}) - t_y \approx E(A_1) = \sum_U \theta_k e_{\theta k} - \sum_U e_{\theta k} = -\sum_U (1 - \theta_k) e_{\theta k},$$

puisque nous pouvons dire que $\hat{\mathbf{B}}_{U;\theta}$ est un estimateur convergent de $\mathbf{B}_{U;\theta}$ et donc $E(A_2) \approx 0$.

L'approximation du biais de $\hat{t}_{\text{ycal}}$ porte le nom de quasi-biais :

$$\text{quasi-biais}(\hat{t}_{\text{ycal}}) = -\sum_U (1 - \theta_k) e_{\theta k}.$$

Le quasi-biais de $\hat{t}_{\text{ycal}}$ est zéro si $\theta_k = 1$, pour tous les $k \in U$ et/ou $y_k = \mathbf{B}_{U;\theta} \mathbf{x}_k$, pour tous les $k \in U$.

Alors, si nous considérons $\hat{t}_{\text{yopt}}$, nous supposons que

$$\hat{t}_{\text{yopt}} = \sum_r d_k y_k + \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right) \hat{\mathbf{C}}_{U;\theta}, \quad (2.6)$$

où

$$\hat{\mathbf{C}}_{U;\theta} = \left(\sum_{k \in r} \sum_{l \in r} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{\mathbf{x}'_k}{\pi_k} \frac{y_l}{\pi_l} \right) \left(\sum_{k \in r} \sum_{l \in r} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{\mathbf{x}_k}{\pi_k} \frac{\mathbf{x}'_l}{\pi_l} \right)^{-1}.$$

Puisque $\hat{t}_{y_{\text{opt}}}$ peut être représenté comme (2.6), qui prend la même forme que $\hat{t}_{y_{\text{cal}}}$ sous (2.5), nous obtenons encore une fois l'expression du quasi-biais

$$\text{quasi-biais}(\hat{t}_{y_{\text{opt}}}) = - \sum_U (1 - \theta_k) e_{\theta k}, \quad (2.7)$$

où $e_{\theta k} = y_k - \mathbf{C}_{U;\theta} \mathbf{x}_k$ et θ_{kl} désignent la probabilité de réponse de la paire : (k, l)

$$\mathbf{C}_{U;\theta} = \left(\sum_{k \in U} \sum_{l \in U} \theta_{kl} (\pi_{kl} - \pi_k \pi_l) \frac{\mathbf{x}'_k y_l}{\pi_k \pi_l} \right) \left(\sum_{k \in U} \sum_{l \in U} \theta_{kl} (\pi_{kl} - \pi_k \pi_l) \frac{\mathbf{x}_k \mathbf{x}'_l}{\pi_k \pi_l} \right)^{-1}.$$

Si nous utilisons la pondération de rechange $d_{k,\text{alt}} = d_k \cdot (1/\hat{\theta}) = d_k \cdot (\sum_s d_k / \sum_r d_k)$, nous obtenons ceci

$$\text{quasi-biais}(\hat{t}_{y_{\text{optm}}}) = E \left(\sum_r d_{k,\text{alt}} e_{\theta k} - \sum_U e_{\theta k} \right) \approx \sum_U \frac{\theta_k}{\bar{\theta}_U} e_{\theta k} - \sum_U e_{\theta k} = - \sum_U \left(1 - \frac{\theta_k}{\bar{\theta}_U} \right) e_{\theta k},$$

où $\sum_U (1 - (\theta_k / \bar{\theta}_U)) = 0$, à comparer avec (2.7), où $\sum_U (1 - \theta_k) = N (1 - \bar{\theta}_U)$.

À moins que $\mathbf{\mu}' \mathbf{x}_k = 1$, pour tous les $k \in U$, nous pouvons obtenir une expression équivalente pour $\hat{t}_{y_{\text{cal}}}$. Par contre, si la restriction $\mathbf{\mu}' \mathbf{x}_k = 1$, pour tous les $k \in U$, *demeure valide*, nous pouvons démontrer (Särndal et Lundström (2005)) que

$$\text{quasi-biais}(\hat{t}_{y_{\text{cal}}}) = - \sum_U e_{\theta k},$$

ce qui demeure valide, indépendamment du plan d'échantillonnage et qui est un résultat complètement conforme à la propriété d'invariance susmentionnée des poids de calage.

3 Une étude en simulation

Les propriétés des estimateurs ont été étudiées au moyen d'une simulation Monte Carlo. Nous nous sommes servis d'une population authentique, appelée KYBOK, composée de $N = 832$ municipalités administratives en Suède en 1992. (Cette population a également été utilisée à des fins de simulation dans Särndal et Lundström (2005) et Andersson et Särndal (2016).)

La variable à l'étude y_k est celle des « dépenses d'administration et d'entretien » ($t_y = 1\,023\,983$). La population se divise en quatre groupes pour ce qui est de la taille, du plus petit au plus grand. La taille des groupes est $N_1 = 218$, $N_2 = 272$, $N_3 = 290$ et $N_4 = 52$. Le vecteur marqué d'un cercle est $\mathbf{x}_k^o = (x_{1k}^o, \dots, x_{4k}^o)'$, où $x_{ik}^o = 1$ si l'unité k fait partie du groupe de population i et, autrement, 0, $i = 1, \dots, 4$. La variable quantitative marquée d'un astérisque x_k^* est la racine carrée des « recettes facturées par anticipation », dont la corrélation avec y_k est fortement positive.

La taille de l'échantillon/taille de l'échantillon prévue était de 300 et nous avons utilisé la probabilité de réponse exponentielle

$$\theta_k = 1 - \exp(-c \cdot x_k^*), \quad k \in U, \quad (3.1)$$

où c est retenu en fonction de la probabilité de réponse moyenne souhaitée; dans cette étude, elle variait entre 0,60 et 0,86 (la dernière valeur étant la probabilité de réponse retenue, par exemple, dans Särndal et Lundström (2005)). Deux plans d'échantillonnage ont été envisagés séparément : l'échantillonnage aléatoire simple et l'échantillonnage de Poisson. Dans le dernier cas, $\pi_k \propto x_k^*$. Pour chaque combinaison du plan, taille de l'échantillon/taille de l'échantillon prévue et probabilité de réponse moyenne, 10 000 échantillons ont été produits. Pour cet échantillon s , un ensemble de réponses r a été créé à l'aide de schémas indépendants de Bernoulli, avec une probabilité θ_k de succès, $k \in s$.

Les estimateurs qui nous intéressent surtout sont $\hat{t}_{y\text{cal}}$, $\hat{t}_{y\text{opt}}$ et $\hat{t}_{y\text{optm}}$, mais dans cette étude en simulation, nous inclurons aussi un exemple de modélisation de la propension paramétrique reposant sur $\hat{t}_{y\text{cal}}$. Un choix simple est le modèle logistique (logit)

$$\theta_k = \frac{\exp(\mathbf{x}'\mathbf{B})}{1 + \exp(\mathbf{x}'\mathbf{B})}, \quad (3.2)$$

où nous supposons que $\mathbf{x}' = (1 \ x^*)$. Pour chaque échantillon accompagné de la non-réponse observée, une estimation du maximum de vraisemblance a permis d'obtenir $\hat{\mathbf{B}}$, ce qui donne des estimations $(\hat{\theta})_{k \in r}$. Pour obtenir $\hat{t}_{y\text{cal logit}}$, les poids déterminés par le plan d'échantillonnage d_k sont alors remplacés par $d_k (1/\hat{\theta}_k)$ avant le calage. Il y a une erreur de spécification dans le modèle logit (3.2) puisque la vraie probabilité de réponse est déterminée par (3.1).

Un estimateur arbitraire $\hat{t}_y$ est déterminé au moyen du biais empirique (estimé à l'aide d'une simulation) $(\hat{B})$, de la variance $(\hat{V})$ et de l'erreur quadratique moyenne $(\widehat{\text{EQM}})$

$$\begin{aligned} \hat{B} &= \hat{E}(\hat{t}_y) - t_y = \frac{1}{K} \sum_{j=1}^K \hat{t}_{yj} - t_y \\ \hat{V} &= \frac{1}{K} \sum_{j=1}^K (\hat{t}_{yj} - \hat{E}(\hat{t}_y))^2 \\ \widehat{\text{EQM}} &= \hat{B}^2 + \hat{V}, \end{aligned}$$

où $K = 10\ 000$.

À noter que des expressions telles que « le biais a augmenté » devraient être interprétées dans ce qui suit comme une augmentation du biais en valeur absolue.

3.1 Résultats

Puisqu'il s'agit d'un point de repère pour une étude dans laquelle les renseignements auxiliaires ne sont pas utilisés à l'étape du plan, examinons d'abord les résultats de l'échantillonnage aléatoire simple au tableau 3.1. Dans ce cas, $\hat{t}_{y\text{cal}} = \hat{t}_{y\text{opt}}$. (En réalité, pour obtenir l'égalité, les renseignements « marqués d'un astérisque » sont $\mathbf{x}_k^* = (1, x_k^*)'$ pour $\hat{t}_{y\text{cal}}$.) Comme il est établi dans l'ensemble de cette étude, le biais de $\hat{t}_{y\text{cal logit}}$ est bien plus important que le biais de $\hat{t}_{y\text{cal}}$, qui est un effet naturel de l'élaboration de $\hat{t}_{y\text{cal logit}}$ reposant sur un modèle de non-réponse non spécifié. De plus, parmi ces deux estimateurs, $\hat{t}_{y\text{cal logit}}$ comporte toujours la plus grande variance.

Si nous examinons plutôt les résultats de l'échantillonnage de Poisson au tableau 3.2, nous pouvons d'abord observer que, pour $\hat{t}_{y\text{cal}}$, le biais et la variance sont plus importants que lors de l'échantillonnage aléatoire simple. Par contre, $\hat{t}_{y\text{opt}}$ a un biais très réduit lors de l'échantillonnage de Poisson comparativement à l'échantillonnage aléatoire simple, alors qu'il y a une légère augmentation de la variance. Ensuite, si nous nous tournons vers l'estimateur modifié proposé $\hat{t}_{y\text{optm}}$, nous constatons une autre réduction du biais, sauf pour $\bar{\theta}_U = 0,86$. En fait, le biais a un comportement monotone et le signe passe de positif à négatif pour $\bar{\theta}_U \approx 0,64$. Toutefois, comparativement à $\hat{t}_{y\text{opt}}$, la variance de $\hat{t}_{y\text{optm}}$ augmente en raison de l'inclusion de $\hat{\theta}_U$ dans (2.4), ce qui donne un compromis entre le biais et la variance. En outre, nous constatons que, parmi ces deux estimateurs, $\hat{t}_{y\text{optm}}$ présente les valeurs de l'EQM les plus grandes, puisque la majeure partie de l'EQM est la variance pour ces bas niveaux de biais.

Tableau 3.1

Biais empirique ($\hat{B}$), variance ($\hat{V}$) et erreur quadratique moyenne ($\widehat{\text{EQM}}$) pour $\hat{t}_{y\text{cal}}$ (Cal), $\hat{t}_{y\text{cal logit}}$ (Cal logit) et $\hat{t}_{y\text{opt}}$ (Opt) dans un cas d'échantillonnage aléatoire simple ($n = 300$) dont les probabilités de réponses moyennes sont 0,86; 0,70; 0,65 et 0,60

Échantillonnage aléatoire simple (Cal = Opt)					
$\hat{B} (*10^{-4})$	$\bar{\theta}_U$	0,86	0,70	0,65	0,60
	Cal	-2,44	-4,00	-4,47	-4,89
	Cal logit	4,81	19,4	26,4	35,5
$\hat{V} (*10^{-8})$	$\bar{\theta}_U$	0,86	0,70	0,65	0,60
	Cal	8,40	9,59	10,2	11,3
	Cal logit	10,7	10,9	13,2	16,1
$\widehat{\text{EQM}} (*10^{-9})$	$\bar{\theta}_U$	0,86	0,70	0,65	0,60
	Cal	1,44	2,57	3,01	3,52
	Cal logit	3,38	38,9	71,9	127

Tableau 3.2

Biais empirique ($\hat{B}$), variance ($\hat{V}$) et erreur quadratique moyenne ($\widehat{\text{EQM}}$) pour $\hat{t}_{y\text{cal}}$ (Cal), $\hat{t}_{y\text{cal logit}}$ (Cal logit), $\hat{t}_{y\text{opt}}$ (Opt) et $\hat{t}_{y\text{optm}}$ (Optm) dans un cas d'échantillonnage de Poisson ($E(n) = 300$) dont les probabilités de réponses moyennes sont 0,86; 0,70; 0,65 et 0,60

Échantillonnage de Poisson					
$\hat{B} (*10^{-4})$	$\bar{\theta}_U$	0,86	0,70	0,65	0,60
	Cal	-2,88	-4,71	-5,17	-5,69
	Cal logit	-12,1	-27,5	-32,9	-38,8
	Opt	-0,0732	-0,329	-0,516	-0,810
	Optm	0,690	0,274	0,0536	-0,277
$\hat{V} (*10^{-9})$	$\bar{\theta}_U$	0,86	0,70	0,65	0,60
	Cal	4,46	5,25	5,56	5,81
	Cal logit	5,17	6,60	7,17	7,57
	Opt	1,39	1,63	1,75	1,84
	Optm	2,05	2,89	3,22	3,51
$\widehat{\text{EQM}} (*10^{-9})$	$\bar{\theta}_U$	0,86	0,70	0,65	0,60
	Cal	5,29	7,47	8,23	9,05
	Cal logit	19,8	82,2	115	127
	Opt	1,39	1,64	1,78	1,91
	Optm	2,10	2,90	3,22	3,52

4 Observations finales

La famille des techniques de calage linéaire dans l'échantillonnage d'enquête comporte une multitude de pondérations de rechange dans les cas de réponse complète, notamment des estimateurs GREG et l'estimateur de régression optimal. Comme les cas de non-réponse offrent encore plus d'options et présentent plus de défis, nous avons étudié l'estimateur « optimal » tout en tenant compte des propensions moyennes à répondre (probabilités). L'approche est fondée sur le plan puisque l'estimateur « optimal » modifié peut être motivé par une argumentation asymptotique et, de plus, nous n'avons effectué aucune modélisation des propensions à répondre. Les résultats sont encourageants, surtout en ce qui concerne la réduction du biais de l'estimateur proposé. Les futurs travaux comprendront l'élaboration d'un estimateur de la variance, qui devrait être valide conditionnellement à la taille de l'ensemble des réponses.

Remerciements

Les auteurs sont reconnaissants envers Carl-Erik Särndal pour ses commentaires précieux et ses suggestions au cours de la recherche qui a finalement abouti à cet article. Les auteurs remercient également l'examineur anonyme et le rédacteur adjoint pour leurs commentaires utiles sur l'article.

Bibliographie

- Andersson, P.G., et Särndal, C.-E. (2016). Calibration for nonresponse treatment: In one or two steps? *Statistical Journal of the IAOS*, 32, 375-381.
- Andersson, P.G., et Thorburn, D. (2005). Une distance de calage optimale menant à un estimateur par la régression optimal. *Techniques d'enquête*, 31, 1, 103-107. Article accessible à l'adresse <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005001/article/8092-fra.pdf>.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Haziza, D., et Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1), 129-145.
- Lundström, S. (1997). *Calibration as a Standard Method for Treatment of Nonresponse*, Thèse de doctorat, Department of Statistics, Stockholm University.
- Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 13, 305-327.
- Montanari, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 1, 71-79. Article accessible à l'adresse <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1998001/article/3911-fra.pdf>.
- Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester, Royaume-Uni : Wiley.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.