Calage avec total de contrôle estimé

Mike Hidiroglou et Christian O. Nambeu[1]

RÉSUMÉ

Le calage est une approche d'estimation intensément utilisée en pratique à cause de ses nombreux avantages. Très souvent, les totaux de calage sont des quantités connues au niveau de la population. Dans cet article, nous étudions le cas où certains totaux sont substitués par des totaux estimés. L'estimateur correspondant est analysé et comparé à des estimateurs par calage habituels en termes de biais et d'erreur quadratique moyenne théoriquement et via une étude par simulations.

MOTS CLÉS : Calage; Estimateur par la Régression; Estimateur optimal.

ABSTRACT

Calibration is an approach of estimation intensively used in practise because of its numerous advantages. Very of often, calibration constraints are known totals at the population level. In this paper, we studied the situation when some of these totals are unknown and are substituted with estimated totals. The derived estimator is compared to standard calibration estimators in terms of bias and mean square error in theory and also by means of simulations.

KEY WORDS: Calibration; Regression Estimator; Optimal Estimator.

1. INTRODUCTION

L'estimation par la régression est une technique utilisée pour l'estimation des variables d'enquête dans un contexte de population finie lorsque des variables auxiliaires sont disponibles dans la population. De nombreux estimateurs par régression sont également des estimateurs par calage (Deville et Särndal, 1992; Andersson et Thorburn, 2005). Supposons que nous désirons estimer le total d'une variable *y*, $Y = \sum_{i\in U} y_i$ au sein d'une population *U= {1,..., N}* comprenant *N* unités. Pour ce faire, un échantillon *s* de taille *n* est sélectionné selon un plan de sondage quelconque *p(s)*. À l'absence de variables auxiliaires, l'estimateur Horvitz-Thompson est fréquemment utilisé. Il s'exprime par $\hat{Y}_\pi = \sum_{i\in s} d_i y_i$ (Horvitz et Thompson, 1952) où $d_i = 1/\pi_i$ désigne le poids de sondage associé à l'unité *i* et est égal l'inverse de la probabilité d'inclusion du premier ordre. Cet estimateur est sans biais sous la distribution engendrée par le plan de sondage *p(s)*, c'est-à-dire que $E_p(\hat{Y}_\pi) = Y$. Sous l'hypothèse que l'on dispose d'un vecteur de *p* variables auxiliaires désigné par $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ dont les totaux sont connus au niveau de la population, il est possible de construire un estimateur plus performant que l'estimateur Horvitz-Thompson en la qualité de l'estimateur par régression qui se formule comme

$$\hat{Y}_{REG} = \hat{Y}_\pi + \left(\boldsymbol{X} - \hat{\boldsymbol{X}}_\pi\right)^\top \hat{\boldsymbol{B}} \qquad (1)$$

où $\boldsymbol{X} = \sum_{i\in U} \boldsymbol{x_i}$, $\hat{\boldsymbol{X}}_\pi = \sum_{i\in s} d_i \boldsymbol{x_i}$ et $\hat{\boldsymbol{B}}$ est un vecteur de *p* paramètres estimés.
Deux approches sont fréquemment utilisées pour le calcul de l'estimateur $\hat{\boldsymbol{B}}$. Il s'agit de l'approche assistée d'un modèle (Särndal et al. 1992) et l'approche à variance-minimale (Montanari, 1987)**.** Sous l'approche assistée d'un modèle, les propriétés de base (biais et variance sous le plan) sont valides même si le modèle n'est pas correctement spécifié. Toutefois, l'efficacité de l'estimateur par régression est améliorée si le modèle est correctement spécifié. Par contre, pour être capable de calculer l'estimateur par régression (1), il faut connaître les *p* totaux du vecteur de variables $\boldsymbol{x}$.
Dans bien des cas en pratique, la variable d'intérêt peut être corrélée à certaines variables auxiliaires dont le total n'est pas connu. Sous l'approche assistée d'un modèle, les praticiens d'enquête vont se restreindre à un modèle ne comportant que les variables dont les totaux sont connus. La question que nous tentons de répondre est de savoir s'il est possible d'incorporer certaines variables dont le total est inconnu dans le modèle afin d'améliorer l'efficacité de l'estimateur par

[1] Mike Hidiroglou. Division de la Recherche et de l'Innovation en Statistique, Statistique Canada, Ottawa, Canada. Mike.hidiroglou@statcan.gc.ca
Christian Nambeu. Division des Méthodes d'Enquêtes Sociales, Statistique Canada, Ottawa, Canada christianolivier.nambeu@statcan.gc.ca

régression. Autrement dit, est-il possible de produire un estimateur par calage plus efficace dont les totaux de contrôle sont estimés?
Singh, (2004) a proposé l'estimateur par calage dont l'une de variables, la taille de la population est inconnue. Elle est ensuite estimée par un estimateur du type Horvitz-Thompson. En observant de plus près, le modèle de travail présumé consiste à un modèle de régression linéaire avec ordonnée à l'origine. Ce modèle est utilisé pour estimer les paramètres inconnus du modèle. En revanche, l'estimateur par régression qu'ils proposent ne dépend pas de la taille de la population. Dans cet article, nous comparons cet estimateur aux estimateurs par régression habituels en termes de biais relatif et d'erreur quadratique moyenne.

À la section suivante, une définition des termes utilisés dans ce document est donnée. À la section 3, des comparaisons algébriques entre différents estimateurs par régression sont présentées. À la section 4, une étude par simulations est présentée pour illustrer la performance des estimateurs étudiés en termes de biais relatif et d'erreur quadratique moyenne Monte Carlo et la section 5, les résultats de cette étude par simulations sont discutés. Finalement, à la section 6 les conclusions sont présentées.

2. TERMINOLOGIE ET DÉFINITIONS

Sous des conditions générales de régularité, (Isaki et Fuller, 1982; Montanari, 1987) ont montré que l'estimateur par la régression (1) peut être approximé par

$$\tilde{Y}_{REG} = \hat{Y}_{\pi} + \left(\boldsymbol{X} - \widehat{\boldsymbol{X}}_{\pi}\right)^{\top} \boldsymbol{B} \tag{2}$$

où $\boldsymbol{B}$ est la limite en probabilité de l'estimateur $\widehat{\boldsymbol{B}}$ lorsque la taille de l'échantillon et celle de la population tendent simultanément vers l'infini. La performance de l'estimateur par la régression (1) pour des grands échantillons peut donc être étudiée par le biais de l'estimateur (2). L'estimateur $\tilde{Y}_{REG}$ est sans biais sous le plan et peut être formulé par :

$$\tilde{Y}_{REG} = \boldsymbol{X}^{\top}\boldsymbol{B} + \sum_{i \in s} \frac{E_i}{\pi_i} \tag{3}$$

où $E_i = y_i - \boldsymbol{x_i^{\top}}\boldsymbol{B}$**.** En utilisant l'expression (3), la variance de l'estimateur (2) peut être calculée par

$$V_p(\tilde{Y}_{REG}) = \sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{E_i}{\pi_i}\frac{E_j}{\pi_j} \tag{4}$$

où $\Delta_{ij} = \pi_{ij} - \pi_i\,\pi_j$ et π_{ij} représente la probabilité de second ordre associée aux unités i et j. L'estimateur (2) est une approximation de l'estimateur (1) ; Pour des tailles d'échantillon suffisamment grandes, l'estimateur (1) est approximativement sans biais et une approximation de sa variance est exprimée par l'équation (4).
L'approche décrite par Särndal et al. (1992) suppose un modèle de travail décrivant la variable d'intérêt et les variables auxiliaires dont les totaux dont connus au niveau de la population. Ce modèle sert à calculer l'estimateur $\widehat{\boldsymbol{B}}$. Plus formellement, il s'exprime comme suit:

$$m\colon y_i = \boldsymbol{x_i^{\top}}\boldsymbol{\beta} + \epsilon_i \tag{5}$$

où $\boldsymbol{\beta}$ est un vecteur de p paramètres inconnus et $E_m(\epsilon_i\,|x_i) = 0$, $V_m(\epsilon_i\,|x_i) = \sigma_i^2\ et\ Cov_m(\epsilon_i, \epsilon_j\,|x_i, x_j) = 0, i \neq j$. Sous cette approche, la quantité $\boldsymbol{B}$ est l'estimateur des moindre carrées ordinaires de $\boldsymbol{\beta}$ au niveau de la population et est donnée par :

$$\boldsymbol{B} = \left(\sum_{i \in U} \frac{\boldsymbol{x_i}\boldsymbol{x_i^{\top}}}{\sigma_i^2}\right)^{-1}\left(\sum_{i \in U} \frac{\boldsymbol{x}_i y_i}{\sigma_i^2}\right) \tag{6}$$

Puisque la variable y n'est pas observée pour toutes les unités de la population, $\boldsymbol{B}$ n'est pas calculable. En utilisant les estimateurs Horvitz-Thompson de chacune de ses composantes, on obtient l'estimateur suivant :

$$\widehat{\boldsymbol{B}}_{\boldsymbol{GREG}} = \left(\sum_{i\in S}\frac{\boldsymbol{x_i}\boldsymbol{x_i}^{\top}}{\pi_i\sigma_i^2}\right)^{-\mathbf{1}}\left(\sum_{i\in S}\frac{\boldsymbol{x_i}y_i}{\pi_i\sigma_i^2}\right) \tag{7}$$

Finalement, l'estimateur GREG est obtenu en remplaçant dans l'équation (1), la quantité $\widehat{\boldsymbol{B}}$ par l'expression obtenue en (7). On obtient alors :

$$\hat{Y}_{GREG} = \hat{Y}_{\pi} + \left(\boldsymbol{X} - \widehat{\boldsymbol{X}}_{\pi}\right)^{\top}\widehat{\boldsymbol{B}}_{\boldsymbol{GREG}} \tag{8}$$

L'estimateur $\hat{Y}_{GREG}$ représente une famille d'estimateurs par régression regroupant de nombreux estimateurs utilisés en pratique. Par exemple, l'estimateur par le ratio est un cas particulier de l'estimateur (8) lorsque $\boldsymbol{x_i} = x_i$ et $\sigma_i^2 = \sigma^2 x_i$. Il peut être exprimé comme suit:

$$\hat{Y}_{RAT} = X\frac{\hat{Y}_{\pi}}{\hat{X}_{\pi}} \tag{9}$$

L'estimateur $\hat{Y}_{RAT}$ peut présenter un gain important d'efficacité lorsque la relation liant la variable d'intérêt et la variable auxiliaire *x* est linéaire et passe par l'origine. Toutefois, pour être calculé l'estimateur $\hat{Y}_{RAT}$ exige que le total de la variable *x* soit connu au niveau de la population. Sa variance approximative peut être obtenue en remplaçant dans la formule (4) les résidus E_i par les résidus appropriés. Le résidu associé à l'unité *i* dans ce cas est égal à $E_{i,RAT} = y_i - x_i R$. On obtient alors :

$$V_p\left(\hat{Y}_{RAT}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_{i,RAT}}{\pi_i}\frac{E_{j,RAT}}{\pi_j}$$

Un autre cas particulier de l'estimateur $\hat{Y}_{GREG}$ peut être obtenu en posant $\boldsymbol{x_i} = (1, x_i)^{\top}$ et $\sigma_i^2 = \sigma^2$ et s'exprimer comme suit:

$$\hat{Y}_{GREG2} = \hat{Y}_{\pi} + \left(N - \widehat{N}_{\pi}\right)\hat{B}_{1,GREG2} + \left(X - \hat{X}_{\pi}\right)\hat{B}_{2,GREG2} \tag{10}$$

où $\hat{B}_{1,GREG2} = \hat{\bar{Y}}_{\pi} - \hat{B}_{2,GREG2}\,\hat{\bar{X}}_{\pi}$et $\hat{B}_{2,GREG2} = \frac{\sum_{i\in s}\frac{1}{\pi_i}(x_i - \hat{\bar{X}}_{\pi})(y_i - \hat{\bar{Y}}_{\pi})}{\sum_{i\in s}\frac{1}{\pi_i}(x_i - \hat{\bar{X}}_{\pi})^2}$, $\hat{\bar{X}}_{\pi} = \hat{X}_{\pi} / \widehat{N}_{\pi}$ et $\hat{\bar{Y}}_{\pi} = \hat{Y}_{\pi} / \widehat{N}_{\pi}$. Le modèle de travail sous-jacent à l'estimateur donné en (10) suppose que la variable d'intérêt *y* et la variable auxiliaire *x* est linéaire et que l'ordonnée à l'origine est significative. Par ailleurs, le calcul de l'estimateur (10) nécessite de connaître les totaux *N* et *X*. L'expression de la variance approximative de l'estimateur $\hat{Y}_{GREG2}$ peut être obtenue en remplaçant dans l'équation (4) les résidus E_i par les résidus suivants $E_{i,GREG2} = y_i - B_{1,GREG2} - B_{2,GREG2}x_i$. Cette variance peut être exprimée par la formule suivante :

$$V_p\left(\hat{Y}_{GREG2}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_{i,GREG2}}{\pi_i}\frac{E_{j,GREG2}}{\pi_j}$$

où $B_{1,GREG2} = \bar{Y} - B_{2GREG2}\bar{X}$, $B_{1GREG2} = \frac{\sum_{i\in U}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i\in U}(x_i - \bar{X})^2}$ et tels que $\bar{Y} = \frac{Y}{N}$ et $\bar{X} = \frac{X}{N}$.

Il est fréquent en pratique que certains totaux du vecteur de variables $\boldsymbol{x}$ soient inconnus. Dans ce contexte, le modèle de travail est très souvent utilisé avec les variables dont les totaux sont connus pour estimer les paramètres du modèle. Par exemple l'estimateur présenté en (10) suppose que le modèle de travail le plus adéquat est un modèle linéaire entre la variable *y* et la variable *x* et dont l'ordonnée à l'origine est statistiquement différente de zéro. Si la taille de la population *N* est inconnue il n'y a souvent aucun autre choix que de se contenter d'un estimateur comme l'estimateur par le ratio donné par (9). Ce qui correspond à utiliser un modèle de travail (5) ne comportant qu'une unique variable auxiliaire, soit *x* dans ce cas, dont le total est connu au niveau de la population. Singh (2004) et Singh et Raghunath (2011) ont proposé un estimateur alternatif qui utilise un modèle de travail comportant l'ordonnée à l'origine ainsi que la variable auxiliaire *x*. Dans ce cas simple, leur estimateur prend la forme suivante :

$$\hat{Y}_{SREG1} = \hat{Y}_{\pi} + \left(X - \hat{X}_{\pi}\right)\hat{B}_{2,GREG2} \tag{11}$$

où $\hat{B}_{2,GREG2} = \frac{\sum_{i\in s} d_i(x_i - \hat{\bar{X}}_\pi)(y_i - \hat{\bar{Y}}_\pi)}{\sum_{i\in s} d_i(x_i - \hat{\bar{X}}_\pi)^2}$.

Une variance approximative de l'estimateur $\hat{Y}_{SREG1}$ peut être obtenue de manière analogue à celle de l'estimateur $\hat{Y}_{GREG2}$. En utilisant la technique des résidus, l'expression de la variance suivante est obtenue :

$$V_p(\hat{Y}_{SREG1}) = \sum_{i\in U}\sum_{i\in U} \Delta_{ij} \frac{E_{i,SREG1}}{\pi_i}\frac{E_{i,SREG1}}{\pi_j}$$

où $E_{i,SREG1} = y_i - x_i B_{2,GREG2}$.

Plus généralement, considérons le vecteur de p variables auxiliaires $\boldsymbol{x_i} = (1, x_{2i} \dots, x_{pi})^\top$ associée à l'unité i. L'estimateur (11) devient alors:

$$\hat{Y}_{SREG} = \hat{Y}_\pi + (\boldsymbol{X}^* - \hat{\boldsymbol{X}}_{\boldsymbol{\pi}}^*)^\top \hat{\boldsymbol{B}}_{\boldsymbol{2,GREG}} \tag{12}$$

où $\boldsymbol{X}^* = \sum_{i\in U} \boldsymbol{x_i}^*$, $\hat{\boldsymbol{X}}_{\boldsymbol{\pi}}^* = \sum_{i\in s} d_i \boldsymbol{x_i^*}$ et $\hat{\boldsymbol{B}}_{\boldsymbol{2,GREG}}$ est tels que $\hat{\boldsymbol{B}}_{\boldsymbol{GREG}} = (\hat{B}_{1,GREG}, \hat{\boldsymbol{B}}_{\boldsymbol{2,GREG}}^{\boldsymbol{T}})^\top$ et $\boldsymbol{x}_i^* = (x_{2i} \dots, x_{pi})^T$. La variance de l'estimateur $\hat{Y}_{SREG}$ peut être obtenue en remplaçant dans l'équation (4) les résidus E_i par les résidus appropriés. On obtient alors la formulation de la variance suivante :

$$V_p(\hat{Y}_{SREG}) = \sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{E_{i,SREG}}{\pi_i}\frac{E_{j,SREG}}{\pi_j}$$

où $E_{i,SREG} = y_i - \boldsymbol{x_i^{*\top}} \boldsymbol{B_{2,GREG}}$.

Sous l'approche à variance minimale proposée par Montanari (1987), la quantité $\boldsymbol{B}$ dans l'expression (2) est calculée de sorte que la variance (4) soit minimale. Pour le vecteur $\boldsymbol{x} = (1, x_2, \dots, x_p)^T$, la valeur optimale de $\boldsymbol{B}$ est égale à

$$\boldsymbol{B_{OPT}} = \left(\sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{\boldsymbol{x}_i}{\pi_i}\frac{\boldsymbol{x}_j^T}{\pi_j}\right)^{-1} \left(\sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{\boldsymbol{x}_i}{\pi_i}\frac{y_j}{\pi_j}\right) \tag{13}$$

La quantité (12) est inconnue car certaines valeurs de la variable y ne sont pas observées. L'estimateur optimal est obtenu en remplaçant dans l'équation (1) l'estimateur $\hat{\boldsymbol{B}}$ par la quantité

$$\hat{\boldsymbol{B}}_{\boldsymbol{OPT}} = \left(\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\boldsymbol{x}_i}{\pi_i}\frac{\boldsymbol{x}_j^T}{\pi_j}\right)^{-1} \left(\sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\boldsymbol{x}_i}{\pi_i}\frac{y_j}{\pi_j}\right) \tag{14}$$

et on obtient alors

$$\hat{Y}_{OPT} = \hat{Y}_\pi + (\boldsymbol{X} - \hat{\boldsymbol{X}}_\pi)^\top \hat{\boldsymbol{B}}_{\boldsymbol{OPT}} \tag{15}$$

Sous des conditions générales de régularité (Isaki et Fuller, 1982; Montanari, 1987), l'estimateur (15) peut être approximé par la quantité suivante :

$$\tilde{Y}_{OPT} = \hat{Y}_\pi + (\boldsymbol{X} - \hat{\boldsymbol{X}}_\pi)^\top \boldsymbol{B}_{\boldsymbol{OPT}} \tag{16}$$

Il est aisé de voir que l'estimateur (16) est un estimateur sans biais sous le plan de sondage. De ce fait on déduit que l'estimateur $\hat{Y}_{OPT}$ est approximativement sans biais. La variance de $\hat{Y}_{OPT}$ peut être approximée par celle de $\tilde{Y}_{OPT}$ dont la formule est donnée par :

$$V_p(\tilde{Y}_{OPT}) = \sum_{i\in U}\sum_{j\in U} \Delta_{ij} \frac{E_{i,OPT}}{\pi_i}\frac{E_{j,OPT}}{\pi_j} \tag{17}$$

où $E_{i,OPT} = y_i - \boldsymbol{x_i^\top B_{OPT}}$.

Selon la disponibilité de l'information auxiliaire, de nombreux estimateurs peuvent découler de (15). Par exemple, dans le cas où $\boldsymbol{x_i} = x_i$, l'estimateur $\hat{Y}_{OPT}$ se réduit à l'estimateur suivant

$$\hat{Y}_{OPT1} = \hat{Y}_{\pi} + \left(X - \hat{X}_{\pi}\right)\hat{B}_{OPT} \tag{18}$$

où $X = \sum_{i\in U} x_i$, $\hat{X}_{\pi} = \sum_{i\in s} d_i x_i$, et

$$\begin{aligned}\hat{B}_{OPT} &= \left(\sum_{i\in s}\sum_{i\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{x_i}{\pi_i}\frac{x_j}{\pi_j}\right)^{-1}\left(\sum_{i\in s}\sum_{i\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{\boldsymbol{x_i}}{\pi_i}\frac{y_j}{\pi_j}\right)\\ &= v_p\left(\hat{X}_{\pi}\right)^{-1} cov_p(\hat{X}_{\pi}, \hat{Y}_{\pi})\end{aligned} \tag{19}$$

tels que $v_p(\hat{X}_{\pi}) = \sum_{i\in s}\sum_{i\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{x_i}{\pi_i}\frac{x_j}{\pi_j}$ et $cov_p\left(\hat{X}_{\pi}, \hat{Y}_{\pi}\right) = \sum_{i\in s}\sum_{i\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{\boldsymbol{x_i}}{\pi_i}\frac{y_j}{\pi_j}$.

D'autre part si la variable $\boldsymbol{x_i} = (1, x_i)^{\top}$, l'estimateur (15) peut être réduit à un estimateur ayant la même forme que l'estimateur (10) et s'écrire comme suit :

$$\hat{Y}_{OPT2} = \hat{Y}_{\pi} + \left(N - \hat{N}_{\pi}\right)\hat{B}_{1,OPT}\ \left(X - \hat{X}_{\pi}\right)\hat{B}_{2,OPT} \tag{20}$$

où

$$\hat{B}_{1,OPT} = \left[v_p\left(\hat{X}_{\pi}\right) cov_p\left(\hat{Y}_{\pi}, \hat{N}_{\pi}\right) - \ cov_p\left(\hat{X}_{\pi}, \hat{N}_{\pi}\right) cov_p\left(\hat{X}_{\pi}, \hat{Y}_{\pi}\right)\right]/\Delta \tag{21}$$

$$\hat{B}_{2,OPT} = \left[v_p\left(\hat{N}_{\pi}\right) cov_p\left(\hat{X}_{\pi}, \hat{Y}_{\pi}\right) - \ cov_p\left(\hat{X}_{\pi}, \hat{N}_{\pi}\right) cov_p\left(\hat{Y}_{\pi}, \hat{N}_{\pi}\right)\right]/\Delta \tag{22}$$

et tel que $\Delta = v_p\left(\hat{N}_{\pi}\right) v_p\left(\hat{X}_{\pi}\right) - cov_p\left(\hat{X}_{\pi}, \hat{N}_{\pi}\right)^2$.

L'estimateur (19) exige que les totaux *N* et *X* soient connus au niveau de la population. Lorsque le total de la population est inconnu, la pratique consiste habituellement à se restreindre à l'estimateur optimal (18).

3. COMPARAISON DES ESTIMATEURS

Dans la présente section quelques résultats sont présentés comparant les estimateurs par régression présentés dans les sections précédentes. Sans perte de généralité, on se restreint au cas deux variables auxiliaires, c'est-à-dire $\boldsymbol{x_i} = (1, x_i)^{\top}$. Notons que ces résultats peuvent être généralisés sans difficulté au cas de *p* variables auxiliaires le cas échéant.

Résultat 1 :

L'estimateur $\hat{Y}_{SREG1}$ est asymptotiquement plus efficace (plus petite variance approximative) que l'estimateur l'estimateur $\hat{Y}_{GREG2}$ si et seulement si l'inégalité suivante est satisfaite:

$$I_1 = B_{1,GREG2}^2 V_p\left(\hat{N}_{\pi}\right) - 2B_{1,GREG2} Cov_p(\hat{E}_{SREG1}, \hat{N}_{\pi}) > 0$$

où $\hat{E}_{SREG1} = \sum_{i\in s} d_i E_{i,SREG1}$, $B_{1,GREG2} = \bar{Y} - \bar{X} B_{2,GREG2}$.

Résultat 2 :

L'estimateur $\hat{Y}_{SREG1}$ est asymptotiquement plus efficace que l'estimateur par $\hat{Y}_{RAT}$ si et seulement si l'inégalité qui suit est vérifié:

$$I_2 = \left(B_{2,GREG2} - R\right)^2 V_p\left(\hat{X}_{\pi}\right) - 2\left(B_{2,GREG2} - R\right) Cov_p(\hat{E}_{RAT}, \hat{X}_{\pi}) < 0$$

où $\hat{E}_{RAT} = \sum_{i\in s} d_i E_{i,RAT}$, $R = \bar{Y}/\bar{X}$ et $\hat{X}_{\pi} = \sum_{i\in s} d_i x_i$.

Résultat 3

Pour des échantillons de taille suffisante, l'estimateur optimal $\hat{Y}_{OPT1}$ est toujours plus efficace que l'estimateur $\hat{Y}_{SREG1}$.(voir, Montanari ,1998)

Résultat 4

L'estimateur $\hat{Y}_{RAT}$ est asymptotiquement plus efficace que l'estimateur $\hat{Y}_{GREG2}$ si et seulement si l'inégalité suivante est satisfaite

$$I_3 = B_{1,GREG2}^2 V_p(\hat{N}_\pi) + \left(B_{2,GREG2} - R\right)^2 V_p(\hat{X}_\pi) - 2B_{1,GREG2}\, Cov_p(\hat{E}_{RAT}, \hat{N}_\pi) \\ -2\left(B_{2,GREG2} - R\right) Cov_p(\hat{E}_{RAT}, \hat{X}_\pi) + 2B_{1,GREG2}\left(B_{2,GREG2} - R\right) Cov_p(\hat{N}_\pi, \hat{X}_\pi) > 0$$

Résultat 5

Pour des échantillons de taille suffisante, l'estimateur optimal $\hat{Y}_{OPT2}$ est toujours plus efficace que l'estimateur $\hat{Y}_{GREG2}$. Une preuve de ce résultat peut être retrouvée dans Montanari (1987, 1998); Rao (1994).

Résultat 6

L'estimateur optimal $\hat{Y}_{OPT2}$ est asymptotiquement plus efficace que l'estimateur par le ratio $\hat{Y}_{RAT}$ si et seulement si l'inégalité suivante est satisfaite,

$$I_4 = \left(B_{2,OPT} - R\right)^2 V_p(\hat{X}_\pi) + B_{1,OPT}^2 V_p(\hat{N}_\pi) - 2\left(B_{2,OPT} - R\right) Cov_p(\hat{E}_{RAT}, \hat{X}_\pi) \\ +2B_{1,OPT}\left(B_{2,OPT} - R\right) Cov_p(\hat{X}_\pi, \hat{N}_\pi) - 2B_{1,OPT2} Cov_p(\hat{E}_{RAT}, \hat{N}_\pi) < 0$$

Le résultat 6 montre que si on a plusieurs totaux connus dans le vecteur de variables auxiliaires, l'estimateur optimal (Montanari, 1987) n'est pas nécessairement plus efficace qu'un estimateur par régression utilisant un ensemble réduit de variables auxiliaires.

Résultat 7

L'estimateur optimal $\hat{Y}_{OPT1}$ est toujours plus efficace que l'estimateur ratio $\hat{Y}_{RAT}$. La preuve générale de ce résultat peut être retrouvée dans Montanari (1987, 1998); Rao (1994).

4. SIMULATIONS

Dans la présente section, la performance des estimateurs est étudiée en termes de biais relatif et d'efficacité relative. Les données utilisées proviennent du manuel de Rosner (2007). Le fichier de données est intitulé FEV.data et comprend 6 variables parmi lesquelles nous avons sélectionné comme variable d'intérêt la variable $y = Height$, comme variables auxiliaires les variables : $x = Age$ et $t = FEV$.

La quantité à estimer est le total de la population $Y = \sum_{i \in U} y_i$. On procède à la sélection de R = 2,000 échantillons Monte Carlo de taille n = 50 selon les plans de sondage suivant :

i. Le plan de sondage de Midzuno : consiste à sélectionner la première unité de l'échantillon avec probabilité de sélection p_i et les *n-1* unités restantes selon un plan aléatoire simple sans remise parmi les *N-1* unités restantes de la population. Sous ce plan de sondage, la variable *t* est utilisée comme variable de taille. La probabilité de sélection de l'unité *i* est donnée par : $p_i = t_i / \sum_{i \in U} t_i$. La probabilité d'inclusion de premier ordre de l'unité *i* est ensuite calculée à l'aide de la formule $\pi_i = [(N-n)p_i + (n-1)]/(N-1)$.

ii. Le plan de Sampford : Comme dans le cas du plan de Midzuno, la variable *t* est utilisée comme variable de taille lors du processus de sélection de l'échantillon. La probabilité de sélection est donnée par : $p_i = t_i / \sum_{i \in U} t_i$. L'algorithme de sélection de l'échantillon peut être décrit comme suit : À la première étape, sélectionner la première unité de l'échantillon avec probabilité p_i et les n-1 unités subséquentes avec $\lambda_i = \frac{p_i}{1-np_i}$. À la deuxième étape, répéter l'étape 1 jusqu'à ce que tous les éléments de l'échantillon soient distincts. La probabilité d'inclusion de premier ordre est donnée par : $\pi_i = np_i$

iii. Le plan de Poisson : est un plan de sondage à taille aléatoire où la variable de taille choisie est la variable *t*. La probabilité de sélection de l'unité *i* est ensuite calculée selon la formule suivante $p_i = t_i / \sum_{i \in U} t_i$. Finalement, la probabilité d'inclusion associée à l'unité i est ensuite calculée par $\pi_i = np_i$. La probabilité de sélection d'un échantillon *s* donnée est $p(s) = \prod_{i \in s} \pi_i \prod_{i \in U \setminus s} (1 - \pi_i)$.

Dans chacun des échantillons sélectionnés, nous avons calculés les estimateurs $\hat{Y}_{RAT}$, $\hat{Y}_{SREG1}$, $\hat{Y}_{OPT1}$, $\hat{Y}_{GREG2}$ et $\hat{Y}_{OPT2}$. Le vecteur de variables $\boldsymbol{x} = (1, x)^\top$ est utilisé lors du calcul des trois premiers estimateurs tandis que la variable *x* est utilisé pour le calcul des deux derniers. Pour mesurer la performance de ces estimateurs, nous avons calculé le biais relatif de tous les estimateurs comme suit

$$RB(\hat{Y}) = \frac{100}{R} \sum_{r=1}^{R} \frac{(\hat{Y}_{(r)} - Y)}{Y} \tag{23}$$

où $\hat{Y}$ est un terme générique qui désigne l'un des cinq estimateurs mentionnés précédemment et $\hat{Y}_{(r)}$ désigne l'estimateur $\hat{Y}$ obtenu dans le r ième échantillon Monte Carlo. La deuxième statistique utilisée pour mesurer l'efficacité des estimateurs est l'efficacité relative qui est le ratio de l'erreur quadratique moyenne Monte Carlo de chacun des cinq estimateurs sur l'erreur quadratique moyenne Monte Carlo de l'estimateur de référence que nous avons comme l'estimateur $\hat{Y}_{GREG2}$. Il s'ensuit que :

$$RE1(\hat{Y}) = \frac{MSE_{MC}(\hat{Y})}{MSE_{MC}(\hat{Y}_{GREG2})} \tag{24}$$

où $MSE_{MC}(\hat{Y}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{Y}_{(r)} - Y)^2$.

La dernière statistique utilisée est également une mesure d'efficacité relative. Cette dernière est calculée comme le ratio de la variance approximative de chacun des cinq estimateurs sur la variance approximative de l'estimateur de référence ($\hat{Y}_{GREG2}$). Soit,

$$RE2(\hat{Y}) = \frac{AV_p(\hat{Y})}{AV_p(\hat{Y}_{GREG2})} \tag{25}$$

où $AV_p(\hat{Y}) = V_p(\tilde{Y})$ est une approximation de la variance de l'estimateur $\hat{Y}$et $\tilde{Y}$ est une approximation par linéarisation de Taylor de $\hat{Y}$.

5. RÉSULTATS

Les résultats de l'étude par simulations sont présentés dans le tableau 1. Le biais relatif et les efficacités relatives sont utilisés pour évaluer la performance des estimateurs étudiés pour différents plans de sondage.

Tableau 1 : Comparaison des estimateurs en termes de biais relatifs et d'efficacité relative

		RAT	OPT1	SREG1	GREG2	OPT2
Midzuno	RB	0,16	0,06	0,06	0,06	0,06
	RE1	15,41	0,94	0,93	1,00	8,95
	RE2	16,40	0,94	0,93	1,00	0,55
Sampford	RB	-0,02	-0,04	0,03	0,13	0,12
	RE1	15,81	14,80	13,87	1,00	0,59
	RE2	16,80	15,77	14,39	1,00	0,55
Poisson	RB	0.16	0.07	0.11	0.12	0.07
	RE1	15.34	15.34	176.50	1,00	0.97
	RE2	16.73	16.73	180.36	1.00	0.96

Les résultats sont examinés selon deux scénarios. Dans le premier la taille de la population est connue et dans le second elle ne l'est pas.

On peut remarquer que tous les estimateurs sont approximativement sans biais comme on pouvait s'y attendre. Le biais relatif est toujours inférieur à 1% pour tous les estimateurs présentés dans le tableau 1. Dans la suite de cette discussion, nous examinerons la performance des estimateurs sous l'angle de l'efficacité relative.

Cas1 : Taille de la population connue.

Lorsque la taille de la population *N* est connue, tous les cinq estimateurs peuvent être calculés. Du point de vue de l'efficacité relative Monte Carlo, les résultats obtenus sous le plan de Midzuno montrent que l'estimateur $\hat{Y}_{SREG1}$ est plus efficace que les estimateurs $\hat{Y}_{RAT}$ et $\hat{Y}_{GREG2}$. On constate que dans ce cas, l'efficacité relative de $\hat{Y}_{SREG1}$ est quasiment similaire à celle de l'estimateur $\hat{Y}_{OPT2}$.
Les résultats obtenus sous les deux autres plans de sondage indiquent que l'estimateur $\hat{Y}_{GREG2}$ est toujours plus efficace que les estimateurs $\hat{Y}_{SREG1}$, $\hat{Y}_{RAT}$ et $\hat{Y}_{OPT2}$ en se basant sur l'efficacité relative Monte Carlo et sur l'efficacité relative basée sur les variances approximatives.

Cas 2 : Taille de population inconnue.

Si la taille de la population *N* est inconnue, il n'est pas possible de calculer les estimateurs $\hat{Y}_{GREG2}$ et $\widehat{Y}_{OPT2}$. Dans cette situation, il est coutume en pratique d'utiliser l'estimateur par le ratio $\hat{Y}_{RAT}$ ou d'utiliser l'estimateur à variance-minimale $\hat{Y}_{OPT1}$. Les résultats obtenus sous le plans de Sampford indiquent que l'estimateur $\hat{Y}_{OPT1}$ est plus efficace que les estimateurs $\hat{Y}_{SREG1}$ et $\hat{Y}_{RAT}$ comme on pouvait s'y attendre. Par ailleurs on peut également remarquer que l'estimateur $\hat{Y}_{SREG1}$ est plus efficace que l'estimateur $\hat{Y}_{RAT}$.

Par contre, sous le plan de poisson l'estimateur $\hat{Y}_{SREG1}$ est nettement moins efficace que tous les autres estimateurs en particulier $\hat{Y}_{RAT}$ et $\hat{Y}_{OPT1}$ avec une efficacité relative d'au moins 10 fois inférieure à celle de ces concurrents.

6. CONCLUSIONS

Dans cet article, nous avons comparés plusieurs estimateurs par calage en fonction des deux approches d'estimation (l'approche assistée d'un modèle et l'approche à variance minimale) les plus utilisées en pratique. Le but était de comparé l'estimateur par calage initialement proposé par Singh(2004) aux estimateurs par calage habituellement utilisés en pratique. Dans l'étude par simulation, nous avons évalué l'impact du plan de sondage sur la performance de tous les estimateurs. Il en ressort que l'estimateur $\hat{Y}_{SREG1}$ peut être plus efficace que les estimateurs standards notamment sous le plan de Midzuno mais beaucoup moins efficace sous un plan similaire au plan de Poisson. En général, l'estimateur optimal restait l'estimateur le plus efficace mais semblait souvent moins instable que les estimateurs de type régression.

RÉFÉRENCES

Andersson, P.G. and Thorburn, D. (2005). « An Optimal Calibration Distance Leading to the Optimal Regression Estimator ». *Survey Methodology*, **31,** 95-99.

Deville, J-C. and Tille, Y. (1998). « Unequal probability sampling without replacement through a splitting method». *Biometrika*, **85**, 89-101

Isaki, C.T. and Fuller, W.A. (1982). « Survey design under the regression superpopulation model ». *Journal of the American Statistical Association*, **77**, 89-96.

Montanari, G.E. (1987). « Post-sampling efficient QR-prediction in large-scale surveys». *International Statistical Review*, 55, 191-202

Montanari, G. E. (1998). « On Regression Estimation of Finite Population Means ». *Survey Methodology*, **24,** 69-77.

Midzuno, H. (1952). « On the sampling system with probability proportional to sum of size». *Annals of the institute of statistical Mathematics*, **3**, 99-107

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., Swensson, B. and Wretman, J. (1992). *Model assisted Survey Sampling*. New York: Springer Verlag.

Singh, S. (2004). Golden and Silver Jubilee Year-2003 of the linear regression estimators, presented at the joint Statistical Meeting, Toronto, 4382-4389.

Singh, S. and Raghunath, A. (2011). « On Calibration of design weights ». *METRON International Journal of Statistics*, vol.LXIX, **2**, 185-205.