Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

# Remarque concernant l'estimation par régression lorsque la taille de la population est inconnue


par Michael A. Hidiroglou, Jae Kwang Kim et
Christian Olivier Nambeu

Date de diffusion : le 22 juin 2016

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 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 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.

## 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 |
| $0^s$ | valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie |
| $^p$ | provisoire |
| $^r$ | révisé |
| x | confidentiel en vertu des dispositions de la *Loi sur la statistique* |
| $^E$ | 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) |



**Une version HTML est aussi disponible.**

*This publication is also available in English*.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 131-145
Statistique Canada, nº 12-001-X au catalogue


# Remarque concernant l'estimation par régression lorsque la taille de la population est inconnue

**Michael A. Hidiroglou, Jae Kwang Kim et Christian Olivier Nambeu[1]**

Résumé

L'estimateur par régression est utilisé de façon intensive en pratique, car il peut améliorer la fiabilité de l'estimation des paramètres d'intérêt tels que les moyennes ou les totaux. Il utilise les totaux de contrôle des variables connues au niveau de la population qui sont incluses dans le modèle de régression. Dans cet article, nous examinons les propriétés de l'estimateur par régression qui utilise les totaux de contrôle estimés à partir de l'échantillon, ainsi que ceux connus au niveau de la population. Cet estimateur est comparé aux estimateurs par régression qui utilisent uniquement les totaux connus du point de vue théorique et par simulation.

**Mots-clés :** Estimateur optimal; échantillonnage; pondération.

# 1 Introduction

Les grands organismes statistiques utilisent de plus en plus l'estimation par régression afin d'améliorer la fiabilité des estimateurs des paramètres d'intérêt (comme les totaux et les moyennes) lorsque des variables auxiliaires sont disponibles à l'échelle de la population. Cassel, Särndal et Wretman (1976) ainsi que Fuller (2009), entre autres, donnent un aperçu détaillé de l'estimateur par régression dans le contexte de l'échantillonnage. Nous montrons comment utiliser l'estimateur par régression pour estimer le total, $Y = \sum_{i \in U} y_i$ où $U = \{1, \ldots, N\}$ désigne la population cible. Un échantillon $s$ de taille attendue $n$ est sélectionné selon un plan de sondage $p(s)$ de $U$, où $\pi_i$ est la probabilité d'inclusion du premier ordre. En l'absence de variables auxiliaires, nous utilisons l'estimateur d'Horvitz-Thompson donné par $\hat{Y}_\pi = \sum_{i \in s} d_i y_i$ (Horvitz et Thompson 1952), où $d_i = 1/\pi_i$ est le poids de sondage associé à l'unité $i$. L'estimateur par régression est donné par

$$\hat{Y}_{\text{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\text{T}} \hat{\mathbf{B}}, \tag{1.1}$$

où $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\text{T}}$, et $\hat{\mathbf{B}}$ est un vecteur de coefficients de régression estimés de dimension $p$, s'exprimant comme une fonction des variables observées $\left(y_i, \mathbf{x}_i^{\text{T}}\right)^{\text{T}}$ dans l'échantillon $s$.

Il est à noter que les composantes du vecteur du total de population $\mathbf{X}$ sont connues pour chacune des variables correspondantes du vecteur $\mathbf{x}_i = \left(1, x_{2i}, \ldots, x_{pi}\right)^{\text{T}}$ utilisé pour calculer $\hat{\mathbf{B}}$. Cependant, il arrive parfois qu'il y ait plus de variables auxiliaires observées dans l'échantillon que dans la population. Supposons que l'échantillon comprend $q$ variables observées $(q > p)$, et que les $p$ variables de la

1. Michael A. Hidiroglou, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, (Ontario), Canada K1A 0T6. Courriel : hidirog@yahoo.ca; Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. Courriel : jkim@iastate.edu; Christian Olivier Nambeu, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, (Ontario), Canada K1A 0T6. Courriel : christianolivier.nambeu@canada.ca.

population sont un sous-ensemble des $q$ variables observées dans l'échantillon. Supposons par ailleurs que certaines des $q - p$ variables supplémentaires de l'échantillon sont bien corrélées avec la variable d'intérêt $y$. Ces variables supplémentaires peuvent-elles être incorporées dans l'estimateur par régression afin d'en accroître l'efficacité ? Singh et Raghunath (2011) ont essayé de répondre à cette question lorsque $q = p + 1$. La variable supplémentaire de l'échantillon était l'ordonnée à l'origine, qu'ils ont utilisée pour estimer la taille de population inconnue $N$ au moyen de l'équation $\hat{N} = \sum_{i \in s} d_i$.

Dans cet article, nous comparons l'estimateur proposé par Singh et Raghunath (2011) à d'autres estimateurs par régression lorsque $N$ est connu et lorsqu'il ne l'est pas. Dans la section 2, nous décrivons les estimateurs par régression standard pour l'estimation des totaux lorsque $N$ est connu, ainsi que par la régression proposée par Singh et Raghunath (2011) lorsque $N$ est inconnu. Dans la section 3, nous proposons un autre estimateur lorsque $N$ est inconnu. Dans la section 4, nous procédons à une étude par simulations afin d'illustrer la performance des différents estimateurs examinés en termes de biais et d'erreur quadratique moyenne. Enfin, dans la section 5, nous présentons nos conclusions et recommandations générales.

# 2 Estimateurs par régression

Sous des conditions générales de régularité (Isaki et Fuller 1982; Montanari 1987), une approximation de l'estimateur par régression (1.1) est

$$\tilde{Y}_{\text{REG}} = \hat{Y}_{\pi} + \left(\mathbf{X} - \hat{\mathbf{X}}_{\pi}\right)^{\text{T}} \mathbf{B}, \tag{2.1}$$

où $\mathbf{B}$ est la limite en probabilité de $\hat{\mathbf{B}}$ lorsque la taille de l'échantillon et celle de la population tendent vers l'infini. Pour de grands échantillons, la variance de l'estimateur par régression (1.1) peut être étudiée avec (2.1). Notons que $\tilde{Y}_{\text{REG}}$ est sans biais sous le plan de sondage $p(s)$ et peut être réexprimé sous la forme :

$$\tilde{Y}_{\text{REG}} = \mathbf{X}^{\text{T}}\mathbf{B} + \sum_{i \in s} d_i E_i, \tag{2.2}$$

où $E_i = y_i - \mathbf{x}_i^{\text{T}}\mathbf{B}$.

Une approximation de la variance par rapport au plan de $\hat{Y}_{\text{REG}}$ peut être donnée par

$$\text{AV}_p\left(\hat{Y}_{\text{REG}}\right) = \sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}, \tag{2.3}$$

où $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$ et $\pi_{ij}$ est la probabilité d'inclusion du second ordre pour les unités $i$ et $j$. Notons que $\mathbf{B}$ peut être estimée selon l'approche assistée par modèle (Särndal, Swensson et Wretman 1992) et l'approche de la variance optimale (Montanari 1987). Les deux méthodes permettent d'obtenir des estimateurs approximativement sans biais. Dans le cas de l'approche assistée par modèle, les propriétés de base (biais et variance) sont valides même lorsque le modèle n'est pas spécifié correctement. Sous l'approche de la variance optimale, aucune hypothèse n'est formulée au sujet de la variable d'intérêt.

L'estimateur assisté par modèle de Särndal et coll. (1992) suppose un modèle de travail entre la variable d'intérêt $(y)$ et les variables auxiliaires $(\mathbf{x})$. Le modèle de travail est désigné par $m$ : $y_i = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta} + \varepsilon_i$ où

$\boldsymbol{\beta}$ est un vecteur de $p$ paramètres inconnus, $E_m(\varepsilon_i \mid \mathbf{x}_i) = 0$, $V_m(\varepsilon_i \mid \mathbf{x}_i) = \sigma_i^2$, et $\mathrm{Cov}_m(\varepsilon_i, \varepsilon_j \mid \mathbf{x}_i, \mathbf{x}_j) = 0, i \neq j$. Sous cette approche, $\mathbf{B}$ dans l'équation (2.1) est l'estimateur des moindres carrés ordinaires de $\boldsymbol{\beta}$ dans la population et est donné par

$$\mathbf{B}_{\mathrm{GREG}} = \left(\sum_{i \in U} c_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i \in U} c_i \mathbf{x}_i y_i\right), \tag{2.4}$$

où $c_i = \sigma_i^{-2}$. Cela donne l'estimateur suivant pour le total $Y$

$$\hat{Y}_{\mathrm{GREG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{GREG}}, \tag{2.5}$$

où

$$\hat{\mathbf{B}}_{\mathrm{GREG}} = \left(\sum_{i \in s} c_i d_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i \in s} c_i d_i \mathbf{x}_i y_i\right). \tag{2.6}$$

L'estimateur optimal de Montanari (1987), obtenu en minimisant la variance par rapport au plan de

$$\tilde{Y}_{\mathrm{REG}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \mathbf{B},$$

est

$$\tilde{Y}_{\mathrm{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \mathbf{B}_{\mathrm{OPT}}, \tag{2.7}$$

où

$$\begin{aligned} \mathbf{B}_{\mathrm{OPT}} &= \left\{V\left(\hat{\mathbf{X}}_\pi\right)\right\}^{-1} \mathrm{Cov}\left(\hat{\mathbf{X}}_\pi, \hat{Y}_\pi\right) \\ &= \left(\sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\mathrm{T}}}{\pi_j}\right)^{-1} \left(\sum_{i \in U}\sum_{j \in U} \Delta_{ij} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \end{aligned} \tag{2.8}$$

L'estimateur optimal pour le total $Y$ est estimé par

$$\hat{Y}_{\mathrm{OPT}} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{OPT}}, \tag{2.9}$$

où

$$\hat{\mathbf{B}}_{\mathrm{OPT}} = \left(\sum_{i \in s}\sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{\mathbf{x}_j^{\mathrm{T}}}{\pi_j}\right)^{-1} \left(\sum_{i \in s}\sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathbf{x}_i}{\pi_i} \frac{y_j}{\pi_j}\right). \tag{2.10}$$

Il est à noter que, pour que nous puissions calculer les vecteurs de régression, la première composante qui les définit doit être inversible. Nous pouvons nous assurer qu'elle l'est en réduisant le nombre de variables auxiliaires qui entrent dans la régression si l'efficience de l'estimateur par régression qui en découle n'en souffre pas trop. Par contre, si la perte d'efficience est importante, nous pouvons inverser ces matrices singulières en utilisant des inverses généralisés.

Comme il est mentionné dans l'introduction, les totaux de population ne sont pas nécessairement connus pour toutes les composantes du vecteur auxiliaire $\mathbf{x}$. La régression utilise normalement les variables auxiliaires pour lesquelles un total de population correspondant est connu. En décomposant $\mathbf{x}_i$ en $\left(1, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ où $\mathbf{x}_i^* = \left(x_{2i}, \ldots, x_{pi}\right)^{\mathrm{T}}$, Singh et Raghunath (2011) ont proposé un estimateur semblable au GREG qui suppose une régression fondée sur une ordonnée à l'origine et la variable $\mathbf{x}^*$, même si seul le total de population de $\mathbf{x}^*$ est connu.

Si $N$ est inconnu et que le total de population de $\mathbf{x}^*$ est connu, leur estimateur est

$$\hat{Y}_{\mathrm{SREG}} = \hat{Y}_\pi + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}}_{2,\mathrm{GREG}}, \tag{2.11}$$

où $\mathbf{X}^* = \sum_{i \in U} \mathbf{x}_i^*$ et $\hat{\mathbf{X}}_\pi^* = \sum_{i \in s} d_i \mathbf{x}_i^*$. Le vecteur de régression des coefficients estimés $\hat{\mathbf{B}}_{2,\mathrm{GREG}}$ est obtenu à partir de $\hat{\mathbf{B}}_{\mathrm{GREG}} = \left(\hat{B}_{1,\mathrm{GREG}}, \hat{\mathbf{B}}_{2,\mathrm{GREG}}^{\mathrm{T}}\right)^{\mathrm{T}}$ donné par (2.6). La variance approximative par rapport au plan de $\hat{Y}_{\mathrm{SREG}}$ prend la même forme que l'équation (2.3), où $E_i = y_i - \mathbf{x}_i^{*\mathrm{T}} \mathbf{B}_{2,\mathrm{GREG}}$, et

$$\mathbf{B}_{2,\mathrm{GREG}} = \left\{\sum_{i \in U} c_i \left(\mathbf{x}_i^* - \bar{\mathbf{X}}_N^*\right)\left(\mathbf{x}_i^* - \bar{\mathbf{X}}_N^*\right)^{\mathrm{T}}\right\}^{-1} \sum_{i \in U} c_i \left(\mathbf{x}_i^* - \bar{\mathbf{X}}_N^*\right) y_i$$

et $\bar{\mathbf{X}}_N^* = \sum_{i \in U} \mathbf{x}_i^* \big/ N$.

Nous pouvons obtenir les propriétés de (2.11) en notant que

$$\begin{aligned} \hat{Y}_{\mathrm{SREG}} - Y &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}}_{2,\mathrm{GREG}} \\ &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \mathbf{B}_{2,\mathrm{GREG}} + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \left(\hat{\mathbf{B}}_{2,\mathrm{GREG}} - \mathbf{B}_{2,\mathrm{GREG}}\right). \end{aligned}$$

Étant donné que $\hat{\mathbf{B}}_{2,\mathrm{GREG}} - \mathbf{B}_{2,\mathrm{GREG}} = O_p\left(n^{-1/2}\right)$ sous certaines conditions de régularité examinées dans Fuller (2009, chapitre 2), le dernier terme est d'ordre plus faible. Ainsi, en ignorant les termes d'ordre plus faible, nous obtenons l'approximation

$$\hat{Y}_{\mathrm{SREG}} - Y \cong \sum_{i \in s} d_i E_i - \sum_{i \in U} E_i, \tag{2.12}$$

où $E_i = y_i - \mathbf{x}_i^{*\mathrm{T}} \mathbf{B}_{2,\mathrm{GREG}}$. Par conséquent, $\hat{Y}_{\mathrm{SREG}}$ est approximativement sans biais sous le plan. Nous pouvons calculer la variance asymptotique en utilisant

$$V\left\{\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right\} = E\left\{\left(\sum_{i \in s} d_i E_i - \sum_{i \in U} E_i\right)^2\right\}.$$

Comme nous pouvons le voir, la variance asymptotique peut être assez importante à moins que $\sum_{i \in U} E_i = 0$.

**Remarque 2.1** *Si* $y_i = a + bx_i$, *nous avons* $\hat{Y}_{\mathrm{SREG}} - Y = \left(\hat{N}_\pi - N\right)a$, *ce qui implique que* $V\left(\hat{Y}_{\mathrm{SREG}}\right) = a^2 V\left(\hat{N}_\pi\right)$. *Cela signifie que si* $V\left(\hat{N}_\pi\right) > 0$, *nous pouvons accroître artificiellement* $a^2 V\left(\hat{N}_\pi\right)$, *la variance de* $\hat{Y}_{\mathrm{SREG}}$, *en choisissant des valeurs élevées de* $a$.

Il est à noter que l'estimateur par régression optimal obtenu en utilisant $\mathbf{x}^* = \left(x_2, \ldots, x_p\right)^{\mathrm{T}}$ est lui aussi approximativement sans biais sous le plan, car

$$\begin{aligned}\hat{Y}_{\mathrm{OPT}}^* - Y &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{OPT}}^* \\ &= \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \mathbf{B}_{\mathrm{OPT}}^* + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \left(\hat{\mathbf{B}}_{\mathrm{OPT}}^* - \mathbf{B}_{\mathrm{OPT}}^*\right),\end{aligned}$$

où $\mathbf{B}_{\mathrm{OPT}}^*$ est obtenu en remplaçant $\mathbf{x}_i$ par $\mathbf{x}_i^*$ dans l'équation (2.8). Étant donné que $\hat{\mathbf{B}}_{\mathrm{OPT}}^* - \mathbf{B}_{\mathrm{OPT}}^* = O_p\left(n^{-1/2}\right)$ sous certaines conditions de régularité examinées dans Fuller (2009, chapitre 2), en ignorant les termes d'ordre plus faible, nous obtenons

$$\hat{Y}_{\mathrm{OPT}}^* - Y \cong \hat{Y}_\pi - Y + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \mathbf{B}_{\mathrm{OPT}}^*.$$

La variance asymptotique de $\hat{Y}_{\mathrm{OPT}}^*$ est plus faible que celle de $\hat{Y}_{\mathrm{SREG}}$, car l'estimateur optimal minimise la variance asymptotique dans la classe d'estimateurs de la forme

$$\hat{Y}_B = \hat{Y}_\pi + \left(\mathbf{X}^* - \hat{\mathbf{X}}_\pi^*\right)^{\mathrm{T}} \hat{\mathbf{B}} \tag{2.13}$$

indexée par $\hat{\mathbf{B}}$.

# 3 Estimateur par régression alternatif

Nous examinons maintenant un estimateur alternatif qui n'utilise pas l'information sur la taille de population $(N)$. Il utilise plutôt les probabilités d'inclusion connues $\pi_i$, à condition qu'elles soient connues pour chaque unité de la population. Étant donné que $\sum_{i \in U} \pi_i = n$, nous pouvons utiliser $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ comme données auxiliaires dans le modèle

$$y_i = \mathbf{z}_i^{\mathrm{T}} \beta + e_i,$$

où $e_i \overset{\text{ind}}{\sim} \left(0, \sigma^2 \pi_i\right)$. Cela signifie que l'introduction de la structure de variance $c_i$ de l'erreur dans le vecteur de régression est donnée par $c_i = d_i / \sigma^2$. L'estimateur qui en découle est donné par

$$\hat{Y}_{\mathrm{KREG}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{KREG}}, \tag{3.1}$$

où $\mathbf{Z} = \sum_{i \in U} \mathbf{z}_i$, $\hat{\mathbf{Z}} = \sum_{i \in s} d_i \mathbf{z}_i$ et

$$\hat{\mathbf{B}}_{\mathrm{KREG}} = \left(\sum_{i \in s} c_i d_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right)^{-1} \sum_{i \in s} c_i d_i \mathbf{z}_i y_i. \tag{3.2}$$

Cet estimateur correspond exactement à celui fourni par Isaki et Fuller (1982).

**Remarque 3.1** *Par construction,*

$$\sum_{i \in s} d_i^2 \left(y_i - \mathbf{z}_i^{\mathrm{T}} \hat{\mathbf{B}}_{\mathrm{KREG}}\right) \mathbf{z}_i = \mathbf{0}$$

*et, comme* $\pi_i$ *est une composante de* $\mathbf{z}_i$*, nous avons* $\sum_{i\in s} d_i \left(y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{KREG}}\right) = 0$*, ce qui aboutit à*

$$\hat{Y}_{\mathrm{KREG}} = \mathbf{Z}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{KREG}}.$$

*Ainsi,* $\hat{Y}_{\mathrm{KREG}}$ *est le meilleur prédicteur linéaire sans biais de* $Y = \sum_{i=1}^{N} y_i$ *sous le modèle*

$$y_i = \pi_i\beta_1 + \mathbf{x}_i^{*\mathrm{T}}\boldsymbol{\beta}_2 + e_i,$$

*où* $e_i \sim \left(0, \sigma^2\pi_i\right)$.

Il est à noter que nous pouvons exprimer $\hat{\mathbf{B}}_{\mathrm{KREG}}$ sous la forme $\hat{\mathbf{B}}_{\mathrm{GREG}}$ en posant que $c_i = d_i/\sigma^2$ et $\mathbf{x}_i = \mathbf{z}_i$. L'estimateur par régression proposé peut donc être considéré comme un cas spécial de l'estimateur GREG. En utilisant un argument semblable à (2.12), nous obtenons

$$\hat{Y}_{\mathrm{KREG}} - Y \cong \sum_{i\in s} d_i E_i^* - \sum_{i\in U} E_i^*, \tag{3.3}$$

où $E_i^* = y_i - \mathbf{z}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{KREG}}$ et

$$\mathbf{B}_{\mathrm{KREG}} = \left(\sum_{i\in U} c_i\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right)^{-1}\sum_{i\in U} c_i\mathbf{z}_i y_i.$$

L'estimateur proposé est approximativement sans biais, et sa variance asymptotique

$$V\left\{\sum_{i\in s} d_i\left(y_i - \mathbf{z}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{KREG}}\right)\right\} = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_i^*}{\pi_i}\frac{E_j^*}{\pi_j}$$

est souvent plus faible que celle de l'estimateur de Singh et Raghunath (2011).

La version optimale de $\hat{Y}_{\mathrm{KREG}}$ utilise $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ en tant que données auxiliaires. Elle est donnée par

$$\hat{Y}_{\mathrm{KOPT}} = \hat{Y}_\pi + \left(\mathbf{Z} - \hat{\mathbf{Z}}_\pi\right)^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{KOPT}}, \tag{3.4}$$

où $\hat{\mathbf{B}}_{\mathrm{KOPT}}$ est obtenu en remplaçant $\mathbf{x}_i$ par $\mathbf{z}_i$ dans l'équation (2.10).

**Remarque 3.2** *Pour les plans de sondage de taille fixe, nous avons* $V_p\left(\sum_{i\in s} d_i\pi_i\right) = 0$*. Dans ce cas, le vecteur du coefficient de régression optimal* $\mathbf{B}_{\mathrm{KOPT}} = V_p\left(\hat{\mathbf{Z}}_\pi\right)^{-1}\mathrm{Cov}_p\left(\hat{\mathbf{Z}}_\pi, \hat{Y}_\pi\right)$ *ne peut pas être calculé, car la matrice de variances-covariances* $V_p\left(\hat{\mathbf{Z}}_\pi\right)$ *n'est pas inversible. En conséquence, l'estimateur optimal où* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ *se réduit à l'estimateur optimal (2.9) seulement si nous utilisons* $\mathbf{x}_i^*$.

**Remarque 3.3** *Pour les plans de sondage de taille aléatoire,* $V_p\left(\sum_{i\in s} d_i\pi_i\right) \geq 0$*. Dans ce cas, toutes les composantes de* $\mathbf{z}_i = \left(\pi_i, \mathbf{x}_i^{*\mathrm{T}}\right)^{\mathrm{T}}$ *peuvent être utilisées dans l'estimateur par régression optimal sous le plan (2.9).*

Une difficulté liée à l'utilisation de l'estimateur optimal $\hat{Y}_{\mathrm{KOPT}}$ est qu'il faut calculer les probabilités d'inclusion conjointe $\pi_{ij}$, ce qui peut s'avérer difficile sous certains plans de sondage. Nous pouvons obtenir

un estimateur qui ne nous oblige pas à calculer les probabilités d'inclusion conjointe en supposant que $\pi_{ij} = \pi_i \pi_j$. Nous donnons à cet estimateur le nom d'estimateur pseudo-optimal $\hat{Y}_{\text{POPT}}$. Il est donné par

$$\hat{Y}_{\text{POPT}} = \hat{Y}_{\pi} + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{\pi}\right)^{\text{T}} \hat{\mathbf{B}}_{\text{POPT}}, \tag{3.5}$$

où

$$\hat{\mathbf{B}}_{\text{POPT}} = \left(\sum_{i \in s} c_i d_i \mathbf{z}_i \mathbf{z}_i^{\text{T}}\right)^{-1} \sum_{i \in s} c_i d_i \mathbf{z}_i y_i$$

et

$$c_i = d_i - 1.$$

En général, l'estimateur pseudo-optimal $\hat{Y}_{\text{POPT}}$ devrait produire des estimations proches de celles produites par $\hat{Y}_{\text{KREG}}$ lorsque la fraction de sondage est faible. Il est à noter que $\hat{Y}_{\text{POPT}}$ est exactement égal à l'estimateur optimal $\hat{Y}_{\text{KOPT}}$ dans le cas d'un plan de sondage de Poisson. Sous ce plan, les probabilités d'inclusion des unités de l'échantillon sont indépendantes. La variance approximative par rapport au plan pour $\hat{Y}_{\text{KREG}}$, $\hat{Y}_{\text{KOPT}}$ et $\hat{Y}_{\text{POPT}}$ a la même forme que celle donnée par l'équation (2.3) où les $E_i$ sont donnés respectivement par $y_i - \mathbf{z}_i^{\text{T}} \hat{\mathbf{B}}_{\text{KREG}}$, $y_i - \mathbf{z}_i^{\text{T}} \hat{\mathbf{B}}_{\text{KOPT}}$ et $y_i - \mathbf{z}_i^{\text{T}} \hat{\mathbf{B}}_{\text{POPT}}$.

# 4 Simulations

Nous avons réalisé deux études par simulations. La première utilisait un ensemble de données fourni dans l'ouvrage de Rosner (2006), tandis que la deuxième se fondait sur une population artificielle créée selon un modèle de régression linéaire simple. La première simulation évaluait la performance de tous les estimateurs sous les différents plans de sondage, alors que la deuxième mettait l'accent sur l'impact de la modification de la valeur de l'ordonnée à l'origine dans le modèle.

Le paramètre d'intérêt pour ces deux simulations est le total de la variable d'intérêt $y$: $Y = \sum_{i \in U} y_i$. Tous les estimateurs $\left(\hat{Y}_{\text{GREG}}, \hat{Y}_{\text{OPT}}, \hat{Y}_{\text{POPT}}, \hat{Y}_{\text{SREG}}, \hat{Y}_{\text{KREG}} \text{ et } \hat{Y}_{\text{KOPT}}\right)$ ont été utilisés avec les données auxiliaires disponibles. Le tableau 4.1 résume les données auxiliaires et la structure de variance des erreurs (s'il y a lieu) qui sont associées aux estimateurs utilisés dans les deux études.

**Tableau 4.1**
**Estimateurs utilisés dans l'étude de simulation**

| ***N* connu** | ***N* inconnu** |
|---|---|
| $\hat{Y}_{\text{GREG2}}$ défini par (2.5) où $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ et $c_i = c$ | $\hat{Y}_{\text{SREG1}}$ défini comme étant un cas spécial de (2.11) où $\mathbf{x}_i^* = (x_{2i})$ |
| $\hat{Y}_{\text{OPT2}}$ défini par (2.9) où $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ | $\hat{Y}_{\text{OPT1}}$ défini par (2.9) où $\mathbf{x}_i = (x_{2i})$ |
| $\hat{Y}_{\text{OPT3}}$ défini par (2.9) où $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$ | $\hat{Y}_{\text{KREG2}}$ défini par (3.1) où $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ et $c_i = d_i / \sigma^2$ |
| $\hat{Y}_{\text{POPT3}}$ défini par (3.5) où $\mathbf{z}_i = (1, \pi_i, x_{2i})^{\text{T}}$ et $c_i = d_i - 1$ | $\hat{Y}_{\text{KOPT2}}$ défini par (3.4) où $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ |
| | $\hat{Y}_{\text{POPT2}}$ défini par (3.5) où $\mathbf{z}_i = (\pi_i, x_{2i})^{\text{T}}$ et $c_i = d_i - 1$ |

La performance de tous les estimateurs a été évaluée en fonction du biais relatif, de l'efficacité relative de Monte Carlo et de l'efficacité relative approximative. Des expressions de ces quantités sont présentées ci-dessous.

1. *Biais relatif :*

$$\mathrm{RB}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{100}{R}\sum_{i=1}^{R}\frac{\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)}{Y}, \tag{4.1}$$

où $\hat{Y}_{\mathrm{EST}(r)}$ représente un des estimateurs présentés au tableau 4.1 tel que calculé dans le $r^{\mathrm{e}}$ échantillon de Monte Carlo.

2. *Efficacité relative Monte Carlo*

$$\mathrm{RE}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.2}$$

où

$$\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{\mathrm{EST}(r)} - Y\right)^2.$$

RE mesure l'efficacité relative de l'estimateur $\hat{Y}_{\mathrm{EST}}$ en ce qui concerne $\hat{Y}_{\mathrm{GREG2}}$.

3. *Efficacité relative approximative*

$$\mathrm{AR}\left(\hat{Y}_{\mathrm{EST}}\right) = \frac{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right)}{\mathrm{AV}_p\left(\hat{Y}_{\mathrm{GREG2}}\right)}, \tag{4.3}$$

où

$$\mathrm{AV}_p\left(\hat{Y}_{\mathrm{EST}}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_i}{\pi_i}\frac{E_j}{\pi_i},$$

est la variance approximative de $\hat{Y}_{\mathrm{EST}}$ où $E_i = y_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{\mathrm{EST}}$. L'efficacité relative approximative $(\mathrm{AR})$ mesure le gain d'efficacité relatif de $\hat{Y}_{\mathrm{EST}}$ par rapport à $\hat{Y}_{\mathrm{GREG2}}$ en utilisant le résidu de population obtenu au moyen de la technique de linéarisation de Taylor. On s'attend à ce que RE et AR donnent des résultats comparables. Cependant, comme nous allons le voir, ce n'est pas nécessairement le cas.

## 4.1 Simulation 1

La population était l'ensemble de données (FEV.DAT) disponible sur le CD qui accompagne l'ouvrage de Rosner (2006). Le fichier de données contient 654 enregistrements tirés d'une étude réalisée à Boston sur les maladies respiratoires des enfants. Les variables du fichier étaient l'âge, la taille, le sexe (masculin ou féminin), le tabagisme (c'est-à-dire si la personne fume ou non) et le volume expiratoire maximal (VEM). Singh et Raghunath (2011) ont utilisé le même ensemble de données. Le paramètre d'intérêt est la taille totale $(y)$ de la population. La variable âge $(x_1)$ a été utilisée comme variable auxiliaire dans la régression. La variable VEM $(x_2)$ a été choisie comme variable de taille pour calculer les probabilités de sélection sous les plans de sondage examinés dans cette simulation. Les variables sexe et tabagisme ont été écartées. Le tableau 4.2 résume les mesures de la tendance centrale des trois variables dans la population. La moyenne et la médiane étaient similaires pour chaque variable, ce qui indique une répartition symétrique des trois variables.

**Tableau 4.2**
**Statistiques descriptives de $y, x_1$ et $x_2$**

| | Minimum | Q1 | Médiane | Moyenne | Q3 | Maximum |
|---|---|---|---|---|---|---|
| $y$ | 46 | 57 | 61,5 | 61,14 | 65,5 | 74 |
| $x_1$ | 3 | 8 | 10 | 9,931 | 12 | 19 |
| $x_2$ | 0,79 | 1,98 | 2,55 | 2,64 | 3,12 | 5,79 |

La figure 4.1 illustre la relation entre la variable d'intérêt $y$ et la variable auxiliaire $x_1$. La relation entre la taille $(y)$ et l'âge $(x_1)$ semble linéaire, mais ne passe pas par l'origine. Le coefficient de corrélation de Pearson entre $y$ et $x_1$ était de 0,79.

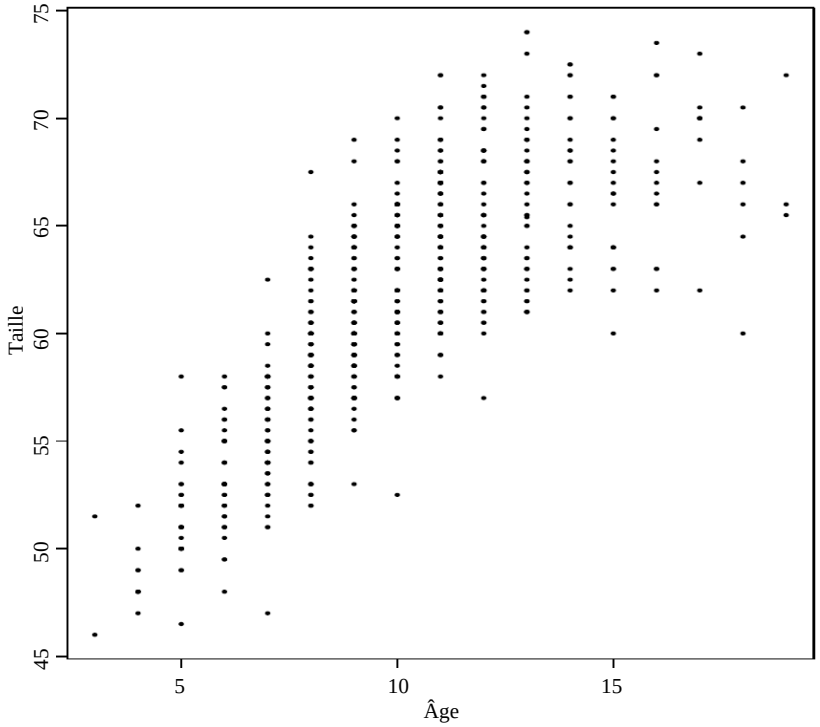


**Figure 4.1 Relation entre la variable d'intérêt *Taille* et la variable auxiliaire *Âge*.**

L'objectif de cette étude par simulations était d'évaluer la performance des estimateurs présentés au tableau 4.1 en utilisant différents plans de sondage. Nous avons examiné les plans de sondage de Midzuno, de Sampford et de Poisson. La variable $x_2$ a été utilisée comme mesure de taille sous les trois plans de sondage pour calculer les probabilités d'inclusion. Ces plans de sondage se présentent comme suit :

1. *Plan de sondage de Midzuno* (voir Midzuno 1952) : La première unité est échantillonnée avec la probabilité $p_i$ et les $n-1$ unités restantes sont sélectionnées par échantillonnage aléatoire simple sans remise parmi les $N-1$ unités restantes de la population. Les probabilités de

sélection $p_i$ pour l'unité $i$ sont données par $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. La probabilité d'inclusion de premier ordre pour l'unité $i$ est donnée par $\pi_i = (N-1)^{-1}\left[(N-n)\, p_i + (n-1)\right]$.

2. *Plan de sondage de Sampford* (voir Sampford 1967) : Dans l'algorithme de sélection de l'échantillon, la première unité est sélectionnée avec la probabilité $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$ tandis que les $n-1$ unités restantes sont sélectionnées avec remise et avec la probabilité $\lambda_i = (1 - np_i)^{-1}\, p_i$. S'il y a des unités qui ont été sélectionnées plus d'une fois, la procédure est répétée jusqu'à ce que tous les éléments de l'échantillon soient différents. La probabilité d'inclusion de premier ordre est donnée par $\pi_i = np_i$.
3. *Plan de sondage de Poisson* : Chaque unité est sélectionnée indépendamment, ce qui donne une taille d'échantillon aléatoire. La probabilité de sélection de l'unité $i$ est $p_i = x_{2i} \Big/ \sum_{i \in U} x_{2i}$. La probabilité d'inclusion associée à l'unité $i$ est $\pi_i = np_i$. Une bonne description de cette procédure figure dans l'ouvrage de Särndal et coll. (1992).

Le paramètre d'intérêt était le total de $Y = \sum_{i \in U} y_i$. En nous basant sur chacun de ces plans de sondage, nous avons sélectionné $R = 2\,000$ échantillons Monte Carlo de taille $n = 50$. Nous avons ensuite calculé les estimateurs du tableau 4.1 pour chaque échantillon, puis avons évalué leur performance en utilisant le biais relatif, l'efficacité relative Monte Carlo et l'efficacité relative approximative tels que décrits dans les équations (4.1), (4.2) et (4.3) respectivement.

## 4.2 Résultats de la simulation 1

Les résultats de la simulation sont présentés au tableau 4.3. Tous les estimateurs étudiés sont approximativement sans biais, et leur biais relatif est inférieur à 1 %. Nous aborderons séparément l'efficacité relative approximative (AR) et l'efficacité relative (RE) des estimateurs lorsque la taille de la population $N$ est connue et lorsqu'elle est inconnue.

**Cas 1 :** *La taille de la population $N$ est connue*

Nous comparons les efficacités AR et RE des estimateurs $\hat{Y}_{\text{GREG2}}$, $\hat{Y}_{\text{OPT2}}$, $\hat{Y}_{\text{OPT3}}$ et $\hat{Y}_{\text{POPT3}}$ pour chacun des trois plans de sondage. Nous pouvons le faire pour presque tous ces estimateurs sauf $\hat{Y}_{\text{OPT3}}$ sous les plans de sondage de Midzuno et de Sampford. En l'occurrence, nous ne pouvons pas calculer $\mathbf{B}_{\text{OPT3}}$ pour une raison semblable à celle décrite dans la remarque 3.2.

Selon les efficacités AR et RE, l'estimateur pseudo-optimal $\hat{Y}_{\text{OPT3}}$ est l'estimateur le plus fiable, quel que soit le plan de sondage. Il est proche de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ seulement pour AR. Les efficacités RE et AR de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ n'étaient pas aussi proches que prévu dans le plan de sondage de Midzuno. Montanari (1998) a lui aussi observé la faible efficacité relative de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$. La figure 4.2 montre ce qui se passe. Nous pouvons observer que la plupart des estimations obtenues au moyen de l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ pour les 2 000 échantillons Monte Carlo sont proches de la moyenne. Cependant, dans certains échantillons, les estimations sont très éloignées de la moyenne. Cela contraste avec

$\hat{Y}_{\text{POPT3}}$, où les valeurs sont concentrées autour de la moyenne. Il est à noter que les efficacités RE et AR associées sont très proches l'une de l'autre.

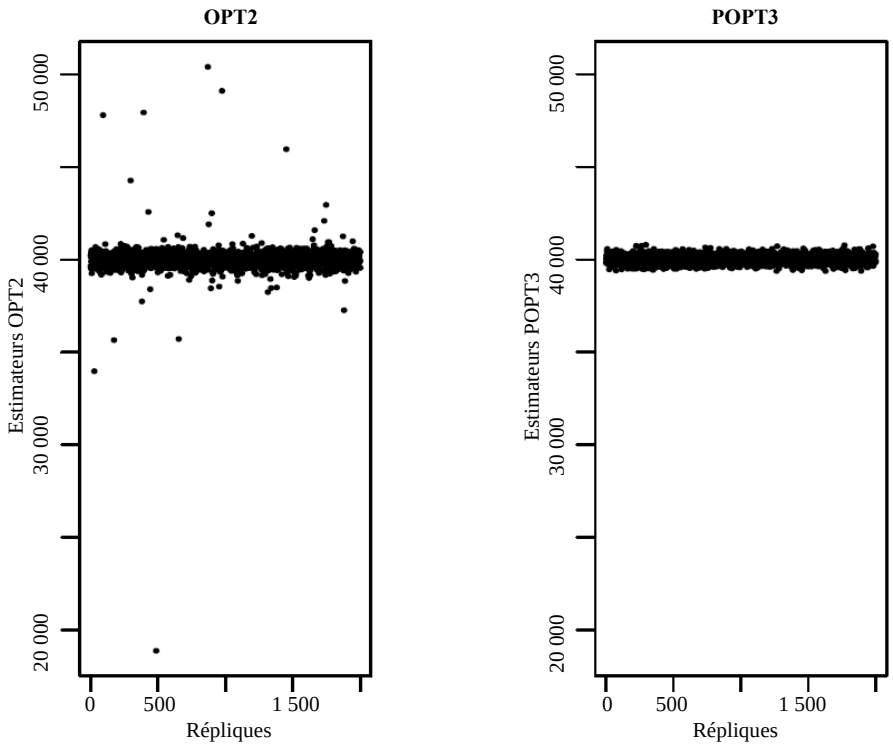


**Figure 4.2 Nuages de points des estimateurs Monte Carlo sous le plan de sondage de Midzuno.**

L'estimateur optimal $\hat{Y}_{\text{OPT3}}$ est équivalent à l'estimateur pseudo-optimal $\hat{Y}_{\text{POPT3}}$ sous le plan de sondage de Poisson. Il faut se rappeler que l'estimateur optimal $\hat{Y}_{\text{OPT2}}$ utilisait $\mathbf{x}_i = (1, x_{2i})^{\text{T}}$ comme données auxiliaires, tandis que l'estimateur optimal $\hat{Y}_{\text{OPT3}}$ utilisait $\mathbf{x}_i = (1, \pi_i, x_{2i})^{\text{T}}$. L'ajout de $\pi_i$ a beaucoup amélioré l'efficacité de l'estimateur optimal sous le plan de sondage de Poisson.

Singh et Raghunath (2011) utilisaient $\hat{Y}_{\text{SREG1}}$ lorsque $N$ était connu, mais ne l'incluaient pas comme total de contrôle. Ils ont néanmoins observé que $\hat{Y}_{\text{SREG1}}$ était assez comparable à $\hat{Y}_{\text{GREG2}}$ en ce qui a trait aux efficacités AR et RB sous le plan de sondage de Midzuno. Pourquoi ? Parce que ce plan de sondage ressemble beaucoup à l'échantillonnage aléatoire simple sans remise. Cependant, si nous utilisons ces deux mesures, $\hat{Y}_{\text{SREG1}}$ est de loin le pire estimateur sous les deux autres plans de sondage.

**Cas 2 :** *La taille de la population $N$ est inconnue*

Cinq estimateurs sont présentés au tableau 4.3 pour ce cas. Toutefois, comme $\hat{Y}_{\text{KREG2}}$ est très proche de $\hat{Y}_{\text{KOPT2}}$ et $\hat{Y}_{\text{POPT2}}$, nous commentons les résultats obtenus pour $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ et $\hat{Y}_{\text{KREG2}}$. Les estimateurs $\hat{Y}_{\text{SREG1}}$, $\hat{Y}_{\text{OPT1}}$ et $\hat{Y}_{\text{KREG2}}$ sont très semblables pour ce qui est de l'efficacité relative et de l'efficacité relative approximative sous le plan de sondage de Midzuno. Sous le plan de sondage de Sampford, $\hat{Y}_{\text{OPT1}}$, $\hat{Y}_{\text{KREG2}}$ et $\hat{Y}_{\text{POPT2}}$ étaient comparables et donnaient des résultats légèrement meilleurs que ceux de l'estimateur $\hat{Y}_{\text{SREG1}}$. Sous le plan de sondage de Poisson, $\hat{Y}_{\text{OPT1}}$ et $\hat{Y}_{\text{KREG2}}$ étaient plus efficaces que $\hat{Y}_{\text{SREG1}}$. Nous constatons également que $\hat{Y}_{\text{SREG1}}$ était très inefficace, son efficacité relative étant au moins 10 fois plus

élevée que celles associées à $\hat{Y}_{\text{KREG2}}$ et $\hat{Y}_{\text{POPT2}}$. Notons que $\hat{Y}_{\text{KREG2}}$ donnait de meilleurs résultats que $\hat{Y}_{\text{OPT1}}$, ce qui est raisonnable puisque $\hat{Y}_{\text{KREG2}}$ utilise deux variables auxiliaires, tandis que $\hat{Y}_{\text{OPT1}}$ utilise seulement la variable auxiliaire $x_{2i}$.

**Tableau 4.3**
**Comparaison des estimateurs en ce qui concerne le biais relatif et les efficacités relatives**

| | | Taille de population connue | | | | Taille de population inconnue | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{Y}_{\text{GREG2}}$ | $\hat{Y}_{\text{OPT2}}$ | $\hat{Y}_{\text{OPT3}}$ | $\hat{Y}_{\text{POPT3}}$ | $\hat{Y}_{\text{SREG1}}$ | $\hat{Y}_{\text{OPT1}}$ | $\hat{Y}_{\text{KREG2}}$ | $\hat{Y}_{\text{KOPT2}}$ | $\hat{Y}_{\text{POPT2}}$ |
| **Midzuno** | RB (en %) | 0,08 | 0,04 | | 0,07 | 0,07 | 0,07 | 0,07 | | 0,07 |
| | RE | 1,00 | 5,84 | | 0,54 | 0,94 | 0,93 | 0,93 | | 0,93 |
| | AR | 1,00 | 0,55 | | 0,55 | 0,94 | 0,93 | 0,93 | | 0,93 |
| **Sampford** | RB (en %) | 0,11 | 0,11 | | 0,07 | -0,01 | 0,07 | 0,02 | | 0,02 |
| | RE | 1,00 | 0,59 | | 0,58 | 14,72 | 13,69 | 13,55 | | 13,56 |
| | AR | 1,00 | 0,55 | | 0,56 | 15,77 | 14,39 | 14,39 | | 14,40 |
| **Poisson** | RB (en %) | 0,11 | 0,11 | 0,08 | 0,08 | 0,09 | 0,14 | 0,16 | 0,16 | 0,16 |
| | RE | 1,00 | 0,96 | 0,57 | 0,57 | 160,47 | 15,49 | 13,85 | 13,85 | 13,85 |
| | AR | 1,00 | 0,96 | 0,55 | 0,56 | 180,36 | 16,73 | 14,40 | 14,39 | 15,73 |

Note : Nous n'avons pas produits des résultats pour les estimateurs $\hat{Y}_{\text{OPT3}}$ et $\hat{Y}_{\text{KOPT2}}$ sous les plans de Midzuno et de Sampford car la matrice de variance-covariance n'était pas inversible.

## 4.3 Simulation 2

La performance des estimateurs a été évaluée pour différentes valeurs de l'ordonnée à l'origine dans le modèle. Nous nous sommes limités au plan de sondage de Poisson afin d'illustrer la remarque 2.1 de la section 2, à savoir que l'efficacité de $\hat{Y}_{\text{SREG}}$ se détériore au fur et à mesure que l'ordonnée à l'origine augmente. La population a été générée selon le modèle suivant :

$$y_i = a + x_i + e_i. \tag{4.4}$$

Les valeurs $e_i$ ont été générées à partir de la loi normale de moyenne 0 et de variance $\sigma_i^2 = 1$. Les valeurs $x$ ont été générées suivant une loi du chi-carré à un degré de liberté. Trois populations de taille $N = 5\,000$ ont été générées à l'aide de l'équation (4.4) avec différentes valeurs de l'ordonnée à l'origine $a$. Il est à noter que les valeurs $x$ ont été générées à nouveau pour chaque population. Les trois populations étaient désignées A, B et C selon l'ordonnée à l'origine utilisée. Les valeurs de l'ordonnée à l'origine ont été fixées à 3, 5 et 10 respectivement pour les populations A, B et C. Dans chacune de ces populations, nous avons prélevé $R = 2\,000$ échantillons Monte Carlo d'une taille prévue $n = 50$ en utilisant le plan de sondage de Poisson. La première probabilité d'inclusion était égale à $\pi_i = nz_i \Big/ \sum_{i \in U} z_i$ pour chaque unité $i$. Les valeurs $z$ ont été générées suivant le modèle

$$z_i = 0{,}5y_i + u_i,$$

où $u_i$ est une erreur aléatoire générée selon la loi exponentielle de moyenne $k$ égale à 0,5 ou 1.

## 4.4 Résultats de la simulation 2

Les résultats numériques sont présentés au tableau 4.4 pour $k = 1$ et au tableau 4.5 pour $k = 0{,}5$. Tous les estimateurs sont approximativement sans biais, les biais relatifs étant inférieurs à 1 %.

**Cas 1 :** *La taille de la population $N$ est connue*

Comme prévu, les estimateurs optimaux $\hat{Y}_{\text{OPT2}}$ et $\hat{Y}_{\text{OPT3}}$ sont plus efficaces que $\hat{Y}_{\text{GREG2}}$. L'estimateur optimal $\hat{Y}_{\text{OPT2}}$ fondé sur $(1, x_{2i})^{\text{T}}$ donne des résultats légèrement meilleurs que ceux de $\hat{Y}_{\text{GREG2}}$. L'inclusion de la variable supplémentaire $\pi_i$ engendrant $\hat{Y}_{\text{OPT3}}$ permet d'améliorer considérablement les efficacités RE et AR : ces gains diminuent au fur et à mesure que l'ordonnée à l'origine augmente. Là encore, $\hat{Y}_{\text{SREG1}}$ est très inefficace et, comme nous le soulignons dans la remarque 2.1, cette inefficacité augmente avec l'ordonnée à l'origine. Les observations qui précèdent restent valables, quelle que soit $k$. L'efficacité des estimateurs optimaux $\hat{Y}_{\text{OPT2}}$ et $\hat{Y}_{\text{OPT3}}$, quant à elle, diminue avec $k$.

**Cas 2 :** *La taille de la population $N$ est inconnue*

L'estimateur le plus efficace est $\hat{Y}_{\text{KREG2}}$. Il surpasse $\hat{Y}_{\text{OPT1}}$, car il utilise plus de variables auxiliaires. L'estimateur $\hat{Y}_{\text{SREG1}}$ est de loin le plus inefficace. Lorsque l'ordonnée à l'origine dans le modèle de population augmente, les efficacités relatives RE et AR restent assez stables pour $\hat{Y}_{\text{KREG2}}$. Par contre, les efficacités relatives associées à $\hat{Y}_{\text{SREG1}}$ et $\hat{Y}_{\text{OPT1}}$ se détériorent rapidement à mesure que l'ordonnée à l'origine dans le modèle de population augmente. L'effet de $k$ sur les efficacités des estimateurs est tel que décrit lorsque la taille de la population est connue.

**Tableau 4.4**
**Biais relatif et efficacités relatives des estimateurs pour $k = 1$ sous le plan de sondage de Poisson**

| Ordonnée à l'origine | | Taille de la population connue | | | | Taille de la population inconnue | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{Y}_{\text{GREG2}}$ | $\hat{Y}_{\text{OPT2}}$ | $\hat{Y}_{\text{OPT3}}$ | $\hat{Y}_{\text{POPT3}}$ | $\hat{Y}_{\text{SREG1}}$ | $\hat{Y}_{\text{OPT1}}$ | $\hat{Y}_{\text{KREG2}}$ | $\hat{Y}_{\text{KOPT2}}$ | $\hat{Y}_{\text{POPT2}}$ |
| 3 | RB (en %) | 0,23 | 0,38 | 0,56 | 0,56 | 0,18 | 0,77 | 0,22 | 0,22 | 0,22 |
| | RE | 1,00 | 0,95 | 0,67 | 0,67 | 7,72 | 5,42 | 0,94 | 0,94 | 0,94 |
| | AR | 1,00 | 0,94 | 0,60 | 0,98 | 7,08 | 5,01 | 0,85 | 0,85 | 0,91 |
| 5 | RB (en %) | 0,04 | 0,07 | 0,18 | 0,18 | -0,01 | 0,67 | -0,07 | -0,07 | -0,07 |
| | RE | 1,00 | 0,99 | 0,76 | 0,76 | 23,91 | 16,63 | 1,50 | 1,50 | 1,50 |
| | AR | 1,00 | 0,98 | 0,70 | 0,73 | 23,48 | 16,20 | 1,45 | 1,45 | 1,52 |
| 10 | RB (en %) | -0,01 | -0,02 | 0,06 | 0,06 | -0,57 | 0,79 | -0,02 | -0,02 | -0,02 |
| | RE | 1,00 | 1,00 | 0,80 | 0,80 | 88,30 | 67,47 | 2,20 | 2,20 | 2,20 |
| | AR | 1,00 | 0,99 | 0,73 | 0,74 | 97,92 | 66,13 | 2,15 | 2,15 | 2,20 |

**Tableau 4.5**
**Biais relatif et efficacités relatives des estimateurs pour $k = 0{,}5$ sous le plan de sondage de Poisson**

| Ordonnée à l'origine | | Taille de la population connue | | | | Taille de la population inconnue | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{Y}_{\text{GREG2}}$ | $\hat{Y}_{\text{OPT2}}$ | $\hat{Y}_{\text{OPT3}}$ | $\hat{Y}_{\text{POPT3}}$ | $\hat{Y}_{\text{SREG1}}$ | $\hat{Y}_{\text{OPT1}}$ | $\hat{Y}_{\text{KREG2}}$ | $\hat{Y}_{\text{KOPT2}}$ | $\hat{Y}_{\text{POPT2}}$ |
| 3 | RB (en %) | 0,13 | 0,25 | 0,42 | 0,42 | -0,18 | 0,54 | -0,02 | -0,02 | -0,02 |
| | RE | 1,00 | 0,99 | 0,89 | 0,89 | 8,42 | 5,93 | 1,78 | 1,78 | 1,78 |
| | AR | 1,00 | 0,96 | 0,83 | 0,95 | 8,30 | 5,83 | 1,79 | 1,79 | 2,10 |
| 5 | RB (en %) | 0,03 | 0,09 | 0,22 | 0,22 | 0,72 | 1,49 | 0,18 | 0,18 | 0,18 |
| | RE | 1,00 | 1,00 | 0,91 | 0,91 | 24,35 | 17,39 | 3,26 | 3,26 | 3,26 |
| | AR | 1,00 | 0,98 | 0,88 | 0,94 | 23,83 | 16,41 | 3,15 | 3,15 | 3,54 |
| 10 | RB (en %) | 0,06 | 0,07 | 0,12 | 0,12 | 0,33 | 1,42 | 0,13 | 0,13 | 0,13 |
| | RE | 1,00 | 1,00 | 0,96 | 0,96 | 98,69 | 73,93 | 6,26 | 6,26 | 6,26 |
| | AR | 1,00 | 0,99 | 0,91 | 0,92 | 98,65 | 66,20 | 5,89 | 5,89 | 6,24 |

# 5 Conclusions

L'estimateur par régression peut être très efficace lorsque les données auxiliaires qu'il utilise sont bien corrélées avec la variable d'intérêt. Il faut aussi que les totaux de population correspondant aux variables auxiliaires soient disponibles. Dans cet article, nous avons examiné le comportement de l'estimateur par régression $\left(\hat{Y}_{\mathrm{SREG}}\right)$ proposé par Singh et Raghunath (2011). Cet estimateur utilise le total estimé de la population comme total de contrôle et les totaux de population connus des variables auxiliaires. Nous l'avons comparé à l'estimateur par régression généralisée $\left(\hat{Y}_{\mathrm{GREG}}\right)$, son analogue optimal $\left(\hat{Y}_{\mathrm{OPT}}\right)$, et à un estimateur de rechange $\left(\hat{Y}_{\mathrm{KREG}}\right)$ qui utilise les probabilités d'inclusion de premier ordre et les données auxiliaires pour lesquelles les totaux de population sont connus. Comme l'estimateur par régression optimale nécessite le calcul des probabilités d'inclusion de second ordre, nous avons aussi inclus un estimateur pseudo-optimal $\left(\hat{Y}_{\mathrm{POPT}}\right)$ qui n'utilise pas ces probabilités. Nous avons examiné les propriétés de ces estimateurs en termes de biais et d'efficacité au moyen d'une simulation incluant différents plans de sondage et différentes valeurs de l'ordonnée à l'origine dans le modèle pour une population artificielle générée. Nous avons comparé les résultats obtenus lorsque la taille de la population était connue et inconnue.

Lorsque la taille de population est connue, l'estimateur optimal $\hat{Y}_{\mathrm{OPT}}$ est le plus efficace. Cependant, comme cet estimateur peut être instable, l'estimateur pseudo-optimal $\hat{Y}_{\mathrm{POPT}}$ est un bon substitut. Notre conclusion concorde avec celle de Rao (1994), qui préférait l'estimateur optimal $\hat{Y}_{\mathrm{POPT}}$ à l'estimateur par régression généralisée $\hat{Y}_{\mathrm{GREG}}$. La proposition de Singh et Raghunath (2011), qui recommandaient d'utiliser $\hat{Y}_{\mathrm{SREG}}$, n'est pas viable, car cet estimateur peut être très inefficace. Lorsque la taille de la population est inconnue, l'estimateur de rechange par régression $\hat{Y}_{\mathrm{KREG}}$ donne les meilleurs résultats.

# Remerciements

Les auteurs remercient le rédacteur associé et les arbitres pour leurs suggestions qui ont considérablement améliorées la qualité de cet article.

# Bibliographie

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimators and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Fuller, W.A. (2009). *Sampling Statistics*. New York : John Wiley & Sons, Inc.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

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

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

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

Montanari, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 1, 71-79.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary data information at the estimation stage. *Journal of Official Statistics*, 10(2), 153-165.

Rosner, B. (2006). *Fundamentals of Biostatistics*. Sixième édition, Duxbury Press.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of section. *Biometrika*, 54, 499-513.

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

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