Composante du produit n° 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimateurs par la régression généralisée d'un total de population finie utilisant la méthode de Box-Cox

par Yan Li

Juin 2008

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 87-98
Statistique Canada, N° 12-001-X au catalogue

Estimateurs par la régression généralisée d'un total de population finie utilisant la méthode de Box-Cox

Yan Li [1]

Résumé

Un nouvel estimateur par la régression généralisée d'un total de population finie basé sur la méthode de transformation de Box-Cox et son estimateur de la variance sont proposés sous un plan général de sondage à probabilités inégales. En étant convergent par rapport au plan de sondage, l'estimateur proposé retient la propriété de robustesse de l'estimateur GREG, même si le modèle sous-jacent est défaillant. En outre, la méthode de Box-Cox permet de trouver automatiquement une transformation raisonnable de la variable dépendante en se servant des données. La robustesse et l'efficacité du nouvel estimateur sont évaluées analytiquement et par des études en simulation de Monte Carlo.

Mots clés : Estimateur par la régression généralisée (GREG); méthode de Box-Cox; pseudo maximum de vraisemblance (PMV).

1. Introduction

Les estimateurs par la régression généralisée (GREG) de totaux et de moyennes de population finie sont établis au moyen de modèles de régression appropriés. Bien que l'on se serve de modèles pour construire ces estimateurs, la sélection des échantillons et l'évaluation des propriétés statistiques des stratégies d'estimation résultantes doivent se faire sous randomisation. Des exemples peuvent être consultés dans Särndal, Swensson et Wretman (1992), Estevao, Hidiroglou et Särndal (1995), Fuller, Loughin et Baker (1994), ainsi que Jayasuriya et Valliant (1996). Le choix d'un bon modèle est essentiel si l'on veut limiter la variabilité d'un estimateur *assisté par modèle* tel que l'estimateur GREG. Si le modèle hypothétique décrit bien la population finie, l'utilisation de l'estimateur GREG au lieu de l'estimateur d'Horvitz-Thompson (Horvitz et Thompson 1952) peut réduire considérablement la variance. Une discussion générale de l'estimation par la régression est présentée dans Fuller (2002). Särndal et coll. (1992) offrent une description détaillée du cadre de travail assisté par modèle pour la construction d'estimateurs de sondage.

Les études portant sur l'estimateur GREG ont été réalisées pour la plupart dans le contexte de la modélisation par la régression linéaire. Essentiellement, l'estimateur GREG intègre les variables auxiliaires pertinentes par la voie de leurs totaux de population de contrôle connus, même si les variables auxiliaires sont connues pour chaque unité de la population (Cassel, Särndal et Wretman 1976; Särndal 1980; Deville et Särndal 1992; Särndal et coll. 1992; Jiang et Lahiri 2006). De nos jours, il est assez fréquent que l'on dispose d'information auxiliaire complète : les données de recensement, les registres administratifs, les données de télédétection et les enquêtes antérieures constituent une mine de renseignements précieux qui peuvent être utilisés pour accroître la précision de la procédure d'estimation (Montanari et Ranalli 2003). Par conséquent, des modèles complexes et des méthodes souples s'appuyant sur l'information auxiliaire complète ont été introduits dans les sondages ces dernières années. Les techniques des splines pénalisées ont été adaptées pour construire des estimateurs fondés sur un modèle (Zheng et Little 2004) et assistés par un modèle (Breidt, Claeskens et Opsomer 2005) pour un total de population finie en se basant sur des données d'enquête complexe. Breidt et Opsomer (2000) ont considéré un estimateur par la régression non paramétrique, assisté par modèle, recourant à la régression par polynômes locaux et ont montré que la régression non paramétrique peut accroître significativement l'efficacité des estimateurs quand les modèles paramétriques sont spécifiés incorrectement. Leurs travaux ont été étendus du modèle à une seule covariable au cas du modèle additif semi-paramétrique. Wu et Sitter (2001) ajustent un modèle de travail général, qui pourrait comprendre à la fois des composantes linéaires et non linéaires, puis le calent sur les valeurs prédites résultantes par régression linéaire simple. Montanari et Ranalli (2005) combinent l'estimation par calage fondé sur un modèle et des méthodes non paramétriques, et proposent des estimateurs non paramétriques assistés par modèle pour une moyenne de population finie.

Alors qu'en statistique classique, on procède souvent à une transformation appropriée de la variable dépendante dans le modèle hypothétique pour obtenir la normalité, la linéarité et l'homoscédasticité (Carroll et Ruppert 1988), la littérature sur les transformations dans le contexte de l'inférence en population finie n'est pas très riche. L'intérêt croît cependant pour l'élaboration de méthodes utilisant une transformation appropriée et des données de sondage. Dans

1. Yan Li, University of Maryland at College Park. Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS; Executive Plaza South; Room 8014, 6120 Executive Boulevard, MSC7244, Bethesda, MD 20892, États-Unis. Courriel : lisherry@mail.nih.gov.

certaines applications des sondages, particulièrement les enquêtes auprès des entreprises et des établissements, les variables étudiées sont souvent continues, positives et fortement asymétriques (par exemple, le revenu). Pour estimer le total de population finie, l'application d'un modèle linéaire à la variable étudiée pourrait ne pas convenir, mais être raisonnable pour une transformation strictement monotone de cette variable. Chen et Chen (1996) ont envisagé l'utilisation de données d'enquête transformées pour améliorer la précision de l'approximation normale. Korn et Graubard (1998) ont comparé divers intervalles de confiance, y compris des intervalles fondés sur une transformation logit, pour l'estimation de proportions dans le cas d'un petit nombre prévu de dénombrements positifs. Karlberg (2000) a proposé un estimateur basé sur un modèle de superpopulation lognormal-logistique pour prédire le total de population finie d'une variable étudiée à distribution fortement asymétrique. Les résultats de simulation ont indiqué que l'estimateur fondé sur le modèle lognormal-logistique offre une alternative sensée à d'autres estimateurs, particulièrement quand la taille de l'échantillon est faible. Chambers et Dorfman (2003) ont discuté de l'estimation d'une moyenne de population finie sous certaines transformations générales, mais connues, des données continues.

Les chercheurs jugent la technique de transformation fructueuse pour l'analyse des données d'enquête. L'étape clé consiste à trouver une transformation appropriée, bien adaptée à ces données. Dans de nombreuses applications, la forme de la transformation est déterminée subjectivement. Malheureusement, les renseignements *a priori* ou la théorie ne suggèrent pas nécessairement celle qu'il convient d'utiliser. Le cas échéant, il serait commode d'identifier la transformation de manière adaptative, en utilisant les données d'enquête.

Les travaux de Box et Cox (1964) ont mené à l'élaboration de méthodes du type « les données décident de la transformation » pour la construction de modèles avec erreurs indépendantes et identiquement distribuées. Leur article et d'autres traitant du sujet, dont Tukey (1957), John et Draper (1980), ainsi que Bickel et Doksum (1981), ont inspiré une foule d'études appliquées. Spitzer (1976) a examiné la relation entre la demande d'argent et la trappe à liquidité à l'aide d'un modèle de Box-Cox généralisé. Dans le contexte de la recherche sur la malaria, Newman (1977) a conclu que la spécification fonctionnelle de Box-Cox était supérieure aux spécifications antérieures. Miner (1982), ainsi que Davison, Arnade et Hallahan (1989) ont étudié la modélisation des fonctions du rendement du soja et de l'exportation de soja par les États-Unis, respectivement. Ils ont conclu que la transformation de Box-Cox produit des termes d'erreur qui suivent approximativement une loi normale. Une bibliographie sur la transformation de Box-Cox peut être consultée dans un article de synthèse publié par Sakia (1992). Pour une application de la méthodologie de Box-Cox à un modèle linéaire mixte, voir Gurka (2004, 2006).

Li et Lahiri (2007) appliquent la transformation de Box-Cox à la variable étudiée pour obtenir des prédicteurs fondés sur un modèle d'un total de population finie qui sont robustes. Ils mentionnent aussi les estimateurs assistés par modèle dans une sous-section (section 2.6), mais n'étudient pas leurs propriétés. Le présent article fournit cette analyse. Nous employons la technique de Box-Cox pour ajuster une régression de la variable étudiée sur un ensemble de variables auxiliaires. Puis, nous utilisons la régression ajustée pour prédire les valeurs de la variable étudiée pour les unités non observées de la population finie qui, à leur tour, fournissent un estimateur de type régression adaptative dans le cadre de travail assisté par un modèle.

La présentation de l'article est la suivante. À la section 2, nous proposons un nouvel estimateur et examinons les propriétés analytiques de ce dernier en fonction du plan de sondage. À la section 3, pour mieux évaluer la robustesse et l'efficacité de l'estimateur proposé, nous le comparons aux estimateurs GREG fondés sur les modèles de travail linéaire et loglinéaire sous-jacents par simulations de Monte Carlo. À la section 4, nous discutons des résultats. Enfin, à la section 5, nous tirons certaines conclusions au sujet des domaines dans lesquels les travaux de recherche pourraient se poursuivre.

2. Un nouvel estimateur du total de population finie

Supposons que la quantité d'intérêt est le total de population finie

$$T = \sum_{i \in U} y_i,$$

où $U = \{1, ..., N\}$ dénote une population finie de taille connue N, et $y_i > 0$ est la valeur de la variable étudiée associée à l'unité i. Écrivons $\mathbf{Y} = (y_1, ..., y_N)'$. Pour estimer T, nous tirons un échantillon s de taille n de la population finie selon un plan d'échantillonnage probabiliste. Soit w_i le poids d'échantillonnage de l'unité i. Ce poids est simplement l'inverse de la probabilité d'inclusion de l'unité i, dénotée par $\pi_i = P(i \in s)$ $(i = 1, ..., N)$. Nous supposons que nous avons de l'information sur $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)'$, où $\mathbf{x}_i = (1, x_{i1}, ..., x_{ik})'$ est un vecteur colonne de k variables auxiliaires pour l'unité i. Pour tout échantillon s de taille n, nous redéfinissons $\mathbf{Y}$ et $\mathbf{X}$ de façon que les n premières lignes de $\mathbf{Y}$ et $\mathbf{X}$ correspondent à celles figurant dans l'échantillon. Écrivons

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix},$$

où

- $\mathbf{y}_s$ est un vecteur colonne de dimensions $n \times 1$ de valeurs observées de la variable étudiée;
- $\mathbf{y}_r$ est un vecteur colonne de dimensions $(N - n) \times 1$ de valeurs non observées de la variable étudiée;
- $\mathbf{X}_s$ est une matrice de dimensions $n \times (k + 1)$ de variables auxiliaires connues dans l'échantillon;
- $\mathbf{X}_r$ est une matrice de dimensions $(N - n) \times (k + 1)$ de variables auxiliaires connues hors de l'échantillon.

Tout au long de l'article, nous utilisons E_d et V_d pour dénoter l'espérance, ou valeur prédite, et la variance sous le plan d'échantillonnage.

2.1 Estimateurs GREG de totaux de population finie

L'estimateur GREG est défini ici comme étant

$$\hat{T}_G = \sum_{i \in U} \hat{y}_{i,w} + \sum_{i \in s} (y_i - \hat{y}_{i,w}) / \pi_i,$$

où $\hat{y}_{i,w}$ est le prédicteur de y_i fondé sur un modèle. Quel que soit le degré avec lequel le modèle sous-jacent décrit bien la population, les estimateurs GREG du total de population finie sont convergents par rapport au plan sous des contraintes faibles (Särndal et coll. 1992). Le modèle utilisé le plus fréquemment est le modèle de régression linéaire standard, donné par

$$\mathbf{M}_1 : \ \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$, une loi de probabilité à N variables de vecteur moyen $\mathbf{0}$ et de matrice de variance-covariance $\sigma^2\mathbf{I}$, et $\mathbf{I}$ est la matrice d'identité de dimensions $N \times N$ (rien ne serait perdu dans ce contexte en remplaçant $\sigma^2\mathbf{I}$ par une matrice définie positive plus générale). Dans cette équation, $\boldsymbol{\beta}$ est un vecteur colonne de dimensions $(k + 1) \times 1$ de coefficients de régression. Les paramètres σ^2 et $\boldsymbol{\beta}$ sont l'un et l'autre des paramètres de superpopulation inconnus. Un prédicteur sans biais pour la i^{e} unité est donné par

$$\hat{y}_{i,w} = \mathbf{x}_\mathbf{i}' \hat{\boldsymbol{\beta}}_w, \tag{1}$$

où $\hat{\boldsymbol{\beta}}_w$ est l'estimateur des moindres carrés pondérés (MCP) de $\boldsymbol{\beta}$ sous $\mathbf{M_1}$ et

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{i \in s} w_i \, \mathbf{x}_\mathbf{i} \, \mathbf{x}_\mathbf{i}' \right)^{-1} \left(\sum_{i \in s} w_i \, \mathbf{x}_i \, y_i \right).$$

Dans certaines applications, surtout dans les enquêtes auprès des entreprises et les enquêtes agricoles, un modèle linéaire n'est pas nécessairement approprié pour y, mais peut être raisonnable pour une transformation strictement monotone de y. Pour l'ensemble de données utilisé dans Royall et Cumberland (1981), Chen et Chen (1996) ont observé que la distribution de la population finie était fortement asymétrique et que la transformation logarithmique aidait à la rendre symétrique. Le besoin d'effectuer la transformation logarithmique et les avantages de celle-ci étaient évidents. Par conséquent, nous considérons le modèle de régression loglinéaire, où la transformation logarithmique est appliquée à la variable dépendante

$$\mathbf{M}_2 : \ \log\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. Un prédicteur évident pour la i^{e} unité est donné par

$$\hat{y}_{i,w} = e^{\mathbf{x}_\mathbf{i}' \hat{\boldsymbol{\beta}}_{lw}}, \tag{2}$$

où

$$\hat{\boldsymbol{\beta}}_{lw} = \left(\sum_{i \in s} w_i \, \mathbf{x}_\mathbf{i} \, \mathbf{x}_\mathbf{i}' \right)^{-1} \left(\sum_{i \in s} w_i \, \mathbf{x}_i \log y_i \right).$$

Le modèle $\mathbf{M_2}$ requiert une spécification subjective de la transformation appliquée à la variable étudiée, ce qui peut être raisonnable dans des situations où nous savons quelle est la transformation appropriée grâce à des données empiriques *a priori* ou à la théorie. En l'absence de toute connaissance *a priori* au sujet de la transformation, il est prudent de choisir cette dernière parmi une famille flexible de transformations en utilisant les données.

Tukey (1957) a examiné la famille de transformations puissance :

$$y^{(\lambda)} = \begin{cases} y^\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

où $y > 0$. Afin d'éliminer la discontinuité à $\lambda = 0$, Box et Cox (1964) ont proposé la famille de transformations suivante :

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

où $y > 0$. Le paramètre λ détermine la nature de la transformation. Par exemple, les valeurs $\lambda = 1, 0, 0{,}5, -1$ correspondent à l'absence de transformation, une transformation logarithmique, une transformation racine carrée et une transformation inverse, respectivement. Le paramètre de transformation λ est estimé d'après les données. L'analyse de Box-Cox peut mener à une transformation logarithmique, mais tout aussi bien aboutir à une autre transformation de la famille susmentionnée, selon les données réelles observées.

Nous considérons le modèle de superpopulation suivant pour la variable étudiée transformée :

$$\mathbf{M}_3 : \mathbf{Y}^{(\lambda)} = \mathbf{X}\beta + \boldsymbol{\varepsilon},$$

où les résidus $\boldsymbol{\varepsilon}$ suivent une loi approximativement normale de moyenne $\mathbf{0}$ et de matrice de variance $\sigma^2\mathbf{I}$.

Schlesselman (1971) a montré que les estimateurs du maximum de vraisemblance des paramètres du modèle de Box-Cox sont indépendants de l'échelle, si bien que le rééchelonnement de la variable y originale aboutit à la même fonction de log-vraisemblance à condition que le modèle de régression contienne un terme constant. Inspirés par cette étude et par la plupart des articles sur les modèles de Box-Cox, nous incluons un terme constant dans le modèle.

Sous $\mathbf{M_3}$, le prédicteur pour la i^{e} unité non observée s'obtient par une simple rétrotransformation à partir de la transformée de Box-Cox :

$$\hat{y}_{i,w} = g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) = (\hat{\lambda}_w \mathbf{x}_i' \hat{\boldsymbol{\beta}}_w + 1)^{1/\hat{\lambda}_w}, \qquad (3)$$

où $\hat{\boldsymbol{\beta}}_w$ et $\hat{\lambda}_w$ sont les estimateurs des paramètres du modèle. La méthode d'estimation est expliquée à la section suivante. Les équations (2) et (3) ne fournissent pas de prédicteur sans biais de y_i pour les modèles sous-jacents respectifs. Li et Lahiri (2007) ont montré que, si la variance des résidus est faible, c'est-à-dire si le modèle est très bien ajusté aux données, les deuxièmes membres des équations (2) et (3) sont de bonnes solutions de rechange pour les prédicteurs sans biais. Pour simplifier et pour réduire le fardeau des calculs, nous traiterons les deuxièmes membres des équations (2) et (3) comme des alternatives appropriées pour les prédicteurs sans biais. Rappelons que notre but est de décrire des estimateurs assistés par modèle pour des totaux de population finie. Le modèle sous-jacent est utilisé uniquement pour suggérer un estimateur, qui sera évalué dans le cadre de randomisation. Même si les prédicteurs des y_i individuels sont biaisés, nous pouvons encore construire des estimateurs convergents par rapport au plan pour des totaux de population finie, contrairement aux estimateurs strictement fondés sur un modèle proposé par Li et Lahiri (2007).

Nous dénotons les estimateurs GREG sous les trois modèles $\mathbf{M_1}$ **à** $\mathbf{M_3}$ par $\hat{T}_{G_L}$, $\hat{T}_{G_\text{LOG}L}$, et $\hat{T}_{G_BC}$, respectivement. L'estimateur $\hat{T}_{G_BC}$ diffère des estimateurs $\hat{T}_{G_L}$ et $\hat{T}_{G_\text{LOG}L}$, parce que les données dictent la transformation qu'il faut utiliser.

Il est possible d'intégrer des transformations de Box-Cox sur la variable y ainsi que les variables x. Par le passé, diverses formes fonctionnelles du modèle de Box-Cox ont été étudiées. Khan et Ross (1977), Spitzer (1976), Zarembka (1968), Boylan, Cuddy et O'Muircheartaigh (1980), et d'autres ont considéré un cas particulier du modèle général de Box-Cox quand on émet l'hypothèse d'un paramètre de transformation commun pour la variable y et les variables x. Gemmill (1980), Boylan, Cuddy et O'Muircheartaigh (1982), et d'autres ont appliqué le modèle général de Box-Cox avec divers paramètres de transformation à la variable y et aux variables x. Li et Lahiri (2007) ont également utilisé le modèle général de Box-Cox pour prédire le total de population finie sous un cadre fondé sur un modèle. Dans l'avenir, cette méthode pourrait être étendue à diverses formes fonctionnelles de la transformation de Box-Cox.

Ici, nous discutons uniquement d'une transformation Box-Cox de la variable y, ce qui permet une comparaison plus équitable des trois estimateurs GREG.

2.2 Estimation des paramètres du modèle et de la transformation $\boldsymbol{\varphi} = (\boldsymbol{\beta}, \lambda, \sigma^2)'$ par la méthode du pseudo maximum de vraisemblance (PMV)

Afin de faciliter l'estimation de λ en utilisant les procédures de calcul existantes, il faut remplacer $\mathbf{Y}^{(\lambda)}$ dans le modèle $\mathbf{M_3}$ par une transformation à l'échelle $\mathbf{Y}^{*(\lambda)}$. Pour la i^{e} unité,

$$y_i^{*(\lambda)} = \begin{cases} (y_i^{\lambda} - 1)/\lambda\tilde{y}^{\lambda-1} & \lambda \neq 0, \\ \tilde{y}\log(y_i) & \lambda = 0, \end{cases}$$

où $\tilde{y}$ est la moyenne géométrique des y. Les calculs qui suivent seront fondés sur le nouveau modèle à l'échelle :

$$\mathbf{M}_4 : \mathbf{Y}^{*(\lambda)} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,$$

où les $\boldsymbol{\varepsilon}^*$ suivent une loi approximativement normale de moyenne $\mathbf{0}$ et de matrice de variance $\sigma_e^{*2}\mathbf{I}$. Soit $\boldsymbol{\varphi}^* = (\boldsymbol{\beta}^*, \lambda, \sigma_e^{*2})'$.

L'estimateur du maximum de vraisemblance (EMV) de $\boldsymbol{\varphi}^*$ maximise la log-vraisemblance

$$l(\boldsymbol{\varphi}^*) = \sum_i \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}),$$

où

$$f(y_i; \boldsymbol{\varphi}^*, \tilde{y}) = (2\pi\sigma_e^{*2})^{-1/2} \exp\left\{-(2\sigma_e^{*2})^{-1}(y_i^{*(\lambda)} - \mathbf{x}_i'\boldsymbol{\beta}^*)^2\right\} \cdot (y_i/\tilde{y})^{\lambda-1}.$$

Skinner, Holt et Smith (1989) redéfinissent $\boldsymbol{\varphi}^*$ comme la valeur de $\tilde{\boldsymbol{\varphi}}^*$ qui maximise

$$l(\tilde{\boldsymbol{\varphi}}^*) = \sum_{i \in U} \log f(y_i, \tilde{\boldsymbol{\varphi}}^*),$$

la somme étant calculée sur l'ensemble des unités de la population finie. Donc, parmi tous les modèles possibles $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, on choisit celui qui « est le mieux ajusté » à la population finie. Si nous choisissons mal la famille $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, cet ajustement le meilleur restera médiocre, mais dans notre inférence, nous le traitons comme la cible que nous essayons d'atteindre avec nos données d'échantillon. Par conséquent, il est important de choisir les options appropriées pour $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$.

Pour la population finie, $\boldsymbol{\varphi}^*$ satisfait

$$\dot{l}_U(\boldsymbol{\varphi}^*) = \sum_{i\in U}\left[\partial \log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y})/\partial\boldsymbol{\varphi}^*\right] = 0,$$

où

$$\tilde{y} = \prod_{i=1}^{N} y_i^{1/N}.$$

Pour $\boldsymbol{\varphi}^*$ donné, posons que $\dot{l}_U(\boldsymbol{\varphi}^*)$, sommation de la dérivée première de la log-vraisemblance par rapport à $\boldsymbol{\varphi}^*$, est un paramètre de population finie. Nous tirons un échantillon et, en approximant $\log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y})$ pour chaque unité i de l'échantillon par $\log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y}_w)$, nous estimons le total de population, $\dot{l}_U(\boldsymbol{\varphi}^*)$, par $\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}})$:

$$\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}) = \sum_{i\in s} w_i\left[\partial \log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y}_w)/\partial\boldsymbol{\varphi}^*\right]_{\boldsymbol{\varphi}^*=\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}},$$

où

$$\tilde{y}_w = \prod_{i\in s} y_i^{w_i/\sum_{i\in s} w_i},$$

la moyenne géométrique pondérée des y dans l'échantillon et $\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}$ est l'estimateur du pseudo maximum de vraisemblance de $\boldsymbol{\varphi}^*$, qui satisfait $\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}) = 0$ (Wu et Sitter 2001). L'estimateur PMV, $\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}} = (\hat{\boldsymbol{\beta}}^*_w,\ \hat{\lambda}_w,\ \hat{\sigma}^{*2}_{e,w})'$, peut être obtenu par une méthode de balayage de l'espace des solutions. Autrement dit, calculer et tracer les valeurs de log-vraisemblance pondérées,

$$\log L(\boldsymbol{\varphi}^*) = \sum_{i\in s} w_i \log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y}_w) \qquad (4)$$

en fonction de l'ensemble de valeurs de λ nous permettra de localiser l'estimation PMV, $\hat{\lambda}_w$, du paramètre de transformation. Lorsque nous évaluons la fonction de log-vraisemblance à chaque valeur fixe de λ dans le contexte de l'échantillonnage, $\boldsymbol{\beta}^*$ et σ_e^{2*} sont estimés en intégrant les poids d'échantillonnage comme il suit :

$$\hat{\boldsymbol{\beta}}_w^{\ *} = \left(\sum_{i\in s} w_i\,\mathbf{x_i}\,\mathbf{x_i'}\right)^{-1}\left(\sum_{i\in s} w_i\,\mathbf{x}_i\,y_i^{*(\lambda)}\right),$$

$$\hat{\sigma}^{*2}_{e,w} = \sum_{i\in s} w_i\,(y_i^{*(\lambda)} - \mathbf{x_i'}\,\hat{\boldsymbol{\beta}}^*_w)^2 \Big/ \sum_{i\in s} w_i\,.$$

Puisque c'est le modèle $\mathbf{M_3}$ qui nous intéresse, la reconversion de $\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}$ qui maximise (4) à $\hat{\boldsymbol{\varphi}}_{\mathrm{PMV}}$ dans le modèle $\mathbf{M_3}$ est nécessaire et $\hat{\boldsymbol{\beta}}_w = \tilde{y}_w^{\hat{\lambda}_w-1}\hat{\boldsymbol{\beta}}^*_w$, $\hat{\sigma}^2_{e,w} = \tilde{y}_w^{2(\hat{\lambda}_w-1)}\,\hat{\sigma}^{*2}_{e,w}$.

Puisque nous prévoyons utiliser un estimateur convergent par rapport au plan pour le total de population, il est raisonnable de nous demander pourquoi nous avons besoin des poids d'échantillonnage pour estimer les paramètres du modèle. Sverchkov et Pfeffermann (2004) soutiennent que l'intégration des poids peut produire de meilleures estimations des paramètres du modèle quand la spécification de ce dernier est correcte dans la population, mais incorrecte dans l'échantillon. Cette situation peut se produire quand les probabilités ou les sélections sont corrélées à l'intérieur des ε_i.

Soit B^* et Λ les deux premières composantes de la solution en échantillon complet de la maximisation de $\sum_U \log f(y_i,\ \tilde{\boldsymbol{\varphi}}^*)$, et $\mathbf{B} = \tilde{y}^{\Lambda-1}\mathbf{B}^*$. Cela nous permet de définir $\boldsymbol{\theta}_N = (\mathbf{B},\ \Lambda)$ comme étant les paramètres de population finie, quelle que soit la validité du modèle.

Nous pouvons montrer que $\hat{\boldsymbol{\theta}}_w = (\hat{\beta}_w,\ \hat{\lambda}_w)$ est un estimateur convergent par rapport au plan de $\boldsymbol{\theta}_N$, c'est-à-dire que

$$\hat{\boldsymbol{\beta}}_{\mathbf{w}} \to \mathbf{B} \text{ en probabilité et } \hat{\lambda}_w \to \Lambda \text{ en probabilité}$$

sous certaines conditions de régularité en utilisant des arguments semblables à ceux de Binder (1983), Wu (1999), ainsi que Wu et Sitter (2001). Ici, la convergence probabiliste est celle par rapport au plan d'échantillonnage.

2.3 Propriété de convergence de l'estimateur $\hat{T}_{G_BC}$

Il est bien connu que l'estimateur $\hat{T}_{G_L}$ a la propriété désirable de convergence par rapport au plan sous des contraintes faibles (Särndal et coll. 1992). Autrement dit, l'écart relatif entre l'estimateur et la grandeur qu'il estime converge en probabilité vers 0 à mesure que l'échantillon devient arbitrairement grand, que tienne ou non le modèle de travail sur lequel il est fondé. Cette propriété peut être conservée pour l'estimateur $\hat{T}_{G_BC}$.

Définissons

$$\hat{T}_{D_G_BC} = \sum_{i\in U} g_i(\boldsymbol{\theta}_N) + \sum_{i\in s}\left(y_i - g_i(\boldsymbol{\theta}_N)\right)/\pi_i,$$

et

$$\hat{T}_{G_BC} = \sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) + \sum_{i\in s}(y_i - g_i(\hat{\boldsymbol{\theta}}_w))/\pi_i.$$

Théorème : Sous les hypothèses suivantes, l'estimateur GREG fondé sur la transformation de Box-Cox $\hat{T}_{G_BC}$ est convergent par rapport au plan pour T, en ce sens que $N^{-1}(\hat{T}_{G_BC} - T) = O_p(1/\sqrt{n})$. De surcroît, la variance asymptotique de $\hat{T}_{G_BC}$ est donnée par

$$AV_d(\hat{T}_{G_BC}) =$$
$$\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\left(y_i - g_i(\boldsymbol{\theta}_N)\right)\left(y_j - g_j(\boldsymbol{\theta}_N)\right)/(\pi_i\,\pi_j),$$

qui peut être estimée par

$$\hat{V}_d(\hat{T}_{G_BC}) =$$
$$\sum_{i\in s}\sum_{j\in s}((\pi_{ij} - \pi_i\pi_j)/\pi_{ij}\quad)(y_i - g_i(\hat{\boldsymbol{\theta}}_w))(y_j - g_j(\hat{\boldsymbol{\theta}}_w))\big/(\pi_i\pi_j). \quad (5)$$

Hypothèse 1 : $\hat{\boldsymbol{\theta}}_w = \boldsymbol{\theta}_N + O_p(1/\sqrt{n})$.

Hypothèse 2 : Pour chaque $\mathbf{x}_i$, $\partial g_i(\mathbf{t})/\partial \mathbf{t}$, est continu en $\mathbf{t}$ et $|\partial g_i(\mathbf{t})/\partial \mathbf{t}| \le h(\boldsymbol{\theta})$ pour $\mathbf{t}$ dans le voisinage de $\boldsymbol{\theta}$.

Hypothèse 3 : Les estimateurs d'Horvitz-Thompson calculés avec les poids de sondage de base pour certains totaux de population suivent une loi asymptotiquement normale.

Hypothèse 4 : Pour chaque $\mathbf{x}_i$, la dérivée seconde de $g_i(\mathbf{t})$ par rapport à $\mathbf{t}$ est continue et bornée dans le voisinage de $\boldsymbol{\theta}$.

Preuve : (voir l'annexe).

L'estimateur proposé de la variance représenté par l'équation (5) est basé sur des approximations en grand échantillon. Pour un niveau nominal donné $1 - \alpha$, l'intervalle de confiance habituel fondé sur l'approximation normale pour l'estimateur de la variance donne approximativement un taux de couverture de $100(1 - \alpha)$ % dans les grands échantillons répétés. Malheureusement, dans certains cas, on a constaté que les propriétés de couverture de ce type d'estimateur de la variance peuvent être médiocres pour certains choix du modèle auxiliaire (*assisting model*) pour les estimateurs $\hat{T}_{G_L}$ (Särndal 1982; Särndal, Swensson et Wretman 1989; *etc.*). Les études théoriques et empiriques de la propriété de couverture de l'estimateur proposé de la variance devront être approfondies.

L'estimateur $\hat{T}_{G_BC}$ est convergent par rapport au plan pour le total de population finie T sous l'approche de randomisation, et la technique de Box-Cox permet qu'une transformation raisonnable de la variable dépendante soit déterminée automatiquement par les données à partir d'une grande famille de fonctions, donc permet d'accroître l'efficacité.

3. Une étude par simulation

L'objectif de cette étude par simulation est d'évaluer les propriétés de divers estimateurs GREG pour un total de population finie. Dans cet exercice de simulation, nous produisons une population finie à partir de l'Australian Agricultural and Grazing Industries Survey (AAGIS). Ces données d'enquête contiennent des renseignements sur le nombre de têtes de bétail (y) et la superficie agricole (x) pour chacune des 431 fermes.

Nous considérons une population finie de taille $N =$ 4 000, générée à partir du modèle suivant :

$$\mathbf{M_5}: y_i^{(\lambda)} = (y_i^{\lambda} - 1)/\lambda = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où les ε_i sont indépendants et de loi approximativement normale $N(0, \sigma^2)$, et x_i est le logarithme d'une valeur produite à partir d'une distribution exponentielle de moyenne μ_x et d'erreur-type σ_x. Afin d'imiter une situation réelle, nous choisissons $\lambda = 0{,}1$, $\beta_0 = 4{,}20$ et $\beta_1 = 2{,}66$ qui sont les estimations obtenues en ajustant les données réelles d'enquête au modèle $\mathbf{M_3}$. Nous fixons $\mu_x = 1\,040$ et $\sigma_x = 1\,000$ pour être certain que $y_i > 0$ pour presque toutes les unités i. Strictement parlant, nous avons une distribution normale tronquée de y, puisque nous écartons toutes les valeurs négatives. L'effet en est négligeable, car moins de 0,1 % des valeurs produites de y doivent être supprimées. Le même phénomène a été observé par Taylor (1986).

La simulation est fondée sur un échantillonnage répété à partir de la population finie générale. Nous étudions deux plans d'échantillonnage, à savoir l'échantillonnage aléatoire simple (EAS) et l'échantillonnage aléatoire simple stratifié (EASS). Lorsqu'un échantillon est sélectionné par EASS, nous appliquons des probabilités de sélection inégales entre les diverses strates. Nous définissons deux strates en utilisant comme valeur limite la médiane des valeurs de y dans la population finie. Pour la strate h de taille N_h, nous sélectionnons un échantillon aléatoire simple de taille n_h. Définissons les probabilités de sélection p_1 et p_2 pour la strate 1 et la strate 2, respectivement. Nous spécifions $p_1 = 2 \times p_2$. Pour une taille d'échantillon fixe n, $n_1 = N_1 \times p_1$, et $n_2 = N_2 \times p_2$.

Nous souhaitons estimer le total de population finie

$$T = \sum_{i \in U} y_i.$$

Dans cette étude par simulation, nous étudions les propriétés des estimateurs $\hat{T}_{G_BC}$, $\hat{T}_{G_L}$, et $\hat{T}_{G_\mathrm{LOG}L}$, ainsi que l'estimateur d'Horvitz-Thompson fondé sur le plan $(\hat{T}_D)$, où les indices inférieurs « -L », « -LOGL » et « -BC » dénotent les modèles linéaire, loglinéaire et Box-Cox sous-jacents, respectivement.

Un millier d'échantillons sont tirés de la population finie simulée pour chaque taille d'échantillon $n \in (30, 80, 150)$. Quatre estimateurs sont produits pour chaque échantillon sélectionné. L'estimateur du paramètre de transformation en population finie Λ est également produit pour chaque échantillon. Pour les besoins de la comparaison, nous utilisons deux méthodes pour estimer Λ. Soit $\hat{\lambda}(\hat{\lambda}_w)$ les estimateurs MCO/MV (PMV) de Λ. Sur l'ensemble des 1 000 échantillons, nous calculons les biais relatifs empiriques en pourcentage (BiaisRel) et les racines carrées des erreurs quadratiques moyennes (REQM) pour évaluer ces estimateurs en utilisant les formules suivantes :

$$\mathrm{BiaisRel} = B^{-1} \sum_{b=1}^{B} (\hat{\varpi}_b - \varpi)/\varpi,$$

et

$$\text{REQM} = \sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\varpi}_b - \varpi)^2},$$

où B est le nombre de répétitions dans la simulation de Monte Carlo et $\hat{\varpi}$ représente une estimation arbitraire du paramètre de population finie ϖ.

4. Résultats

Au tableau 1, nous présentons le biais relatif et la racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon quand $\sigma = 0{,}5$. Les quatre estimateurs donnent tous un biais relatif proche de 0 [le maximum des valeurs absolues du biais relatif (|BiaisRel|) est inférieur à 0,01 dans le tableau 1]. Parmi ces estimateurs, $\hat{T}_{G_BC}$ est celui dont le biais relatif absolu et la REQM sont les plus faibles sur diverses tailles d'échantillon et plans d'échantillonnage. Par conséquent, la technique de Box-Cox protège $\hat{T}_{G_BC}$, qui voit son efficacité accrue comparativement à d'autres estimateurs GREG.

Tableau 1
Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon (λ = 0,1)

		$\hat{T}_D^1$	$\hat{T}_{G_L}^2$	$\hat{T}_{G_\text{LOG}L}^3$	$\hat{T}_{G_BC}^4$
Échantillonnage aléatoire simple					
BiaisRel($\times 10^{-3}$)	n = 30	4,37	-9,82	5,46	3,62
	n = 80	1,43	-3,67	1,24	0,65
	n = 150	-2,60	-1,24	1,22	0,53
REQM($\times 10^7$)	n = 30	7,17	3,54	2,02	1,94
	n = 80	4,26	2,09	1,18	1,09
	n = 150	3,20	1,58	0,88	0,79
Échantillonnage aléatoire simple stratifié					
BiaisRel($\times 10^{-3}$)	n = 30	6,01	-5,82	3,84	1,92
	n = 80	9,93	-1,01	3,04	0,98
	n = 150	1,75	-1,85	1,06	0,42
REQM($\times 10^7$)	n = 30	5,63	3,67	2,29	2,11
	n = 80	3,51	2,31	1,43	1,28
	n = 150	2,49	1,59	1,01	0,90

[1] $\hat{T}_D$: estimateur d'Horvitz-Thompson fondé sur le plan;
[2] $\hat{T}_{G_L}$: estimateur GREG avec modèle linéaire sous-jacent;
[3] $\hat{T}_{G_\text{LOG}L}$: estimateur GREG avec modèle loglinéaire sous-jacent;
[4] $\hat{T}_{G_BC}$: estimateur GREG avec modèle de Box-Cox sous-jacent.

Nous étudions aussi le biais relatif et la REQM sous les mêmes conditions quand $\sigma = 1$ et $\sigma = 2$. La figure 1 donne la REQM pour les trois estimateurs GREG ($\hat{T}_{G_L}$, $\hat{T}_{G_\text{LOG}L}$ et $\hat{T}_{G_BC}$) lorsqu'on utilise différents plans d'échantillonnage. Nous notons que $\hat{T}_{G_BC}$ possède systématiquement la plus petite REQM quand $\sigma = 0{,}5$ et 1. Donc, un modèle robuste choisi par la méthode de Box-Cox réduit la REQM, surtout quand le modèle est approprié pour un petit σ.

Nous pourrions aussi nous intéresser aux propriétés de $\hat{T}_{G_L}$ comparativement à $\hat{T}_{G_BC}$ quand $\lambda = 1$, c'est-à-dire la situation qui favorise $\hat{T}_{G_L}$. Si nous examinons le tableau 2, nous voyons que $\hat{T}_{G_BC}$ est relativement comparable à $\hat{T}_{G_L}$, surtout quand la taille d'échantillon est grande. Ce résultat implique que le fait d'utiliser $\hat{T}_{G_BC}$ alors que nous devrions utiliser $\hat{T}_{G_L}$ n'entraîne pas une grande perte en ce qui concerne le biais relatif et la REQM.

Afin de mieux évaluer la robustesse et l'accroissement de l'efficacité de $\hat{T}_{G_BC}$ comparativement à $\hat{T}_{G_L}$ et à $\hat{T}_{G_\text{LOG}L}$, nous produisons une population finie de taille $N = 4\,000$, non plus conformément au modèle de Box-Cox $\mathbf{M_5}$, mais à partir de

$$\mathbf{M_6}: y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + z_i + \varepsilon_i,$$

où $z_i = x_i^2$. Nous utilisons les mêmes x_i et les mêmes valeurs de paramètres que celles spécifiées à la section 3 pour générer les y_i. Nous étudions les quatre mêmes estimateurs basés sur la nouvelle population finie $\{(x_i, y_i)$ $i = 1, ..., 4{,}000\}$. Il s'agit de la situation non idéale pour n'importe lequel des estimateurs GREG. Les résultats sont présentés au tableau 3. Les avantages de l'utilisation de $\hat{T}_{G_BC}$ sont manifestes en ce qui concerne le biais relatif et la REQM.

Le tableau 4 donne le biais relatif et la REQM pour $\hat{\lambda}$ et $\hat{\lambda}_w$ sous échantillonnage EASS avec diverses tailles d'échantillon et les σ fondés sur les valeurs de population produites à partir de $\mathbf{M_5}$. Puisque nous avons stratifié sur la variable y, l'utilisation des pondérations devrait avoir un impact, du moins sur le biais de l'estimation des paramètres. Quand σ est petit, c'est-à-dire quand les données simulées sont bien ajustées au modèle hypothétique, $\hat{\lambda}_w$ donne un biais relatif plus proche de 0, mais $\hat{\lambda}_w$ et $\hat{\lambda}$ donnent l'un et l'autre d'aussi bons résultats en ce qui concerne la REQM. Par contre, quand σ est grand, $\hat{\lambda}_w$ donne systématiquement des valeurs absolues plus petites du biais relatif et de la REQM que $\hat{\lambda}$, quoique les REQM restent proches. En effet, en ce qui concerne l'estimation de T, ni l'une ni l'autre approche ne l'emporte pour ce qui est du biais empirique ou de la racine carrée de l'erreur quadratique moyenne, quelle que soit la taille de l'échantillon ou la valeur choisie de σ (données non présentées). Cela pourrait tenir au fait que l'estimateur de T est biaisé sous le modèle quelle que soit la qualité de l'estimation de λ.

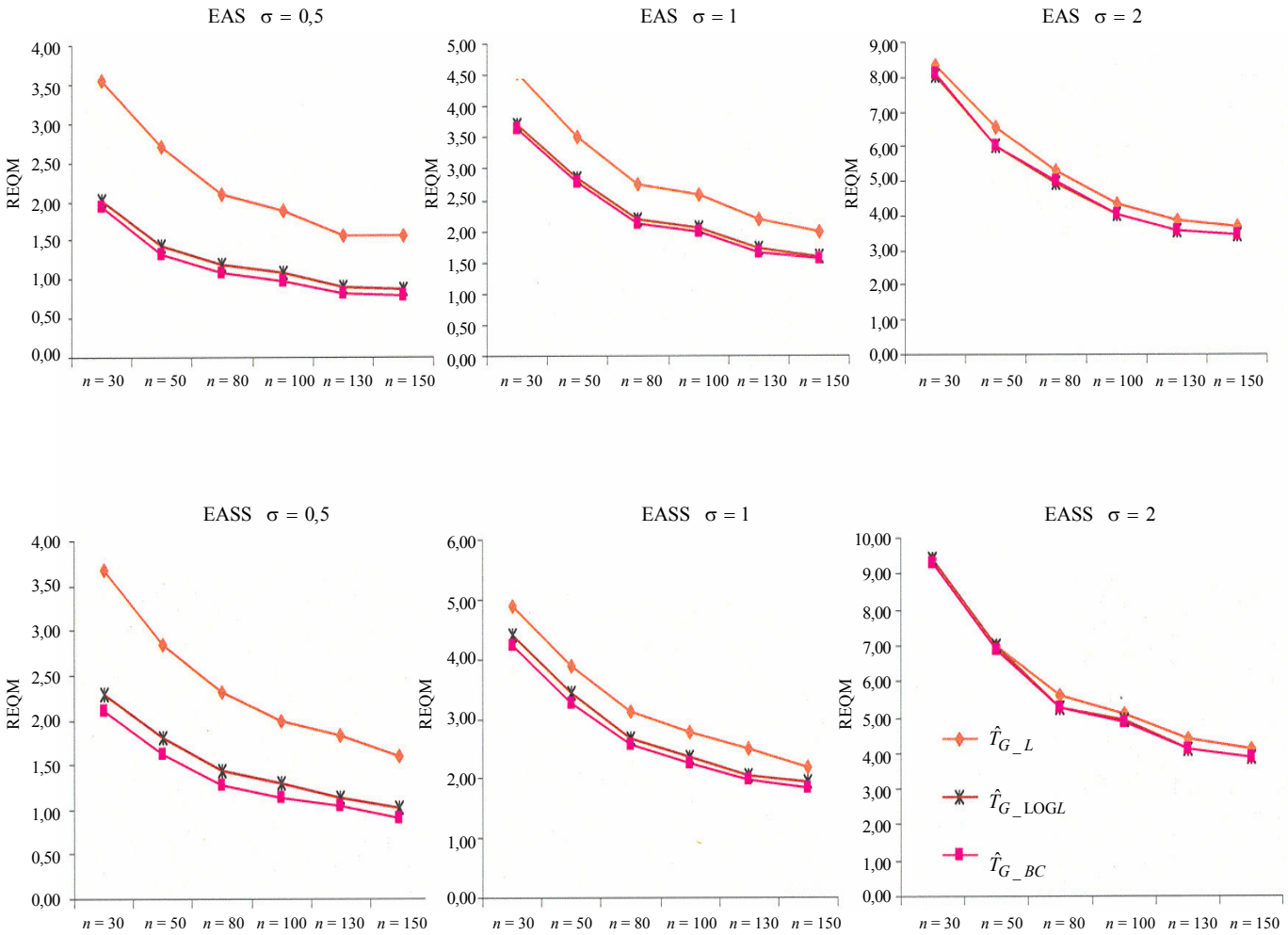

Figure 1 Comparaison de la racine carrée de l'erreur quadratique moyenne de $\hat{T}_{G_L}$, $\hat{T}_{G_LOGL}$ et $\hat{T}_{G_BC}$ pour divers plans d'échantillonnage, tailles d'échantillon et écarts-types

Tableau 2
Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon ($\lambda = 1$)

		$\hat{T}_D^1$	$\hat{T}_{G_L}^2$	$\hat{T}_{G_LOGL}^3$	$\hat{T}_{G_BC}^4$
Échantillonnage aléatoire simple					
BiaisRel($\times 10^{-3}$)	$n = 30$	0,31	0,16	0,63	0,24
	$n = 80$	-0,37	-0,09	0,03	-0,09
	$n = 150$	-0,25	0,03	0,10	0,02
REQM($\times 10^{7}$)	$n = 30$	20,91	3,51	3,98	3,63
	$n = 80$	12,22	2,12	2,36	2,13
	$n = 150$	8,98	1,57	1,75	1,57
Échantillonnage aléatoire simple stratifié					
BiaisRel($\times 10^{-3}$)	$n = 30$	0,51	0,27	0,56	0,22
	$n = 80$	3,97	-0,11	-0,11	-0,13
	$n = 150$	-0,23	0,04	0,06	0,04
REQM($\times 10^{7}$)	$n = 30$	12,54	3,79	4,20	3,91
	$n = 80$	8,39	2,27	2,61	2,29
	$n = 150$	5,48	1,67	1,90	1,67

[1] $\hat{T}_D$: estimateur d'Horvitz-Thompson fondé sur le plan;
[2] $\hat{T}_{G_L}$: estimateur GREG avec modèle linéaire sous-jacent;
[3] $\hat{T}_{G_LOGL}$: estimateur GREG avec modèle loglinéaire sous-jacent;
[4] $\hat{T}_{G_BC}$: estimateur GREG avec modèle de Box-Cox sous-jacent.

Tableau 3
Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon (valeurs de y produites à partir d'un modèle M_6)

		$\hat{T}_D^1$	$\hat{T}_{G_L}^2$	$\hat{T}_{G_LOGL}^3$	$\hat{T}_{G_BC}^4$
Échantillonnage aléatoire simple					
BiaisRel($\times 10^{-3}$)	$n = 30$	-16,89	-54,56	28,81	-2,17
	$n = 80$	-5,65	-23,15	11,76	-1,43
	$n = 150$	-13,78	-13,78	10,73	-0,76
REQM ($\times 10^{11}$)	$n = 30$	30,08	24,68	12,11	2,87
	$n = 80$	17,95	13,85	7,25	1,78
	$n = 150$	13,60	10,29	5,50	1,33
Échantillonnage aléatoire simple stratifié					
BiaisRel($\times 10^{-3}$)	$n = 30$	1,11	-36,13	31,95	-7,87
	$n = 80$	5,59	-18,13	13,11	-3,26
	$n = 150$	-2,79	-7,30	7,33	-1,43
REQM($\times 10^{11}$)	$n = 30$	34,10	27,15	14,19	4,37
	$n = 80$	19,61	15,93	8,49	2,61
	$n = 150$	14,60	12,19	6,62	2,02

[1] $\hat{T}_D$: estimateur d'Horvitz-Thompson fondé sur le plan;
[2] $\hat{T}_{G_L}$: estimateur GREG avec modèle linéaire sous-jacent;
[3] $\hat{T}_{G_LOGL}$: estimateur GREG avec modèle loglinéaire sous-jacent;
[4] $\hat{T}_{G_BC}$: estimateur GREG avec modèle de Box-Cox sous-jacent.

Tableau 4
Biais relatif et racine carrée de l'erreur quadratique moyenne de $\hat{\lambda}$[1] et $\hat{\lambda}_w$[2] sous échantillonnage EASS pour diverses tailles d'échantillon et divers écarts-types

	$\sigma = 2$		$\sigma = 1$		$\sigma = 0{,}5$	
	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$
Biais relatif						
$n = 30$	-0,58	0,13	-0,28	0,10	-0,16	-0,01
$n = 80$	-0,42	0,14	-0,19	0,11	-0,13	-0,02
$n = 150$	-0,39	0,10	-0,16	0,09	-0,10	-0,01
Racine de l'erreur quadratique moyenne						
$n = 30$	0,14	0,12	0,11	0,11	0,07	0,07
$n = 80$	0,08	0,07	0,06	0,06	0,04	0,04
$n = 150$	0,06	0,05	0,05	0,04	0,03	0,03

[1] L'estimateur $\hat{\lambda}$ est obtenu par la méthode des moindres carrés ordinaires/méthode du maximum de vraisemblance.
[2] L'estimateur $\hat{\lambda}_w$ est obtenu par la méthode du pseudo maximum de vraisemblance.

5. Conclusion

Nous avons proposé un estimateur par la régression généralisée d'un total de population finie fondé sur une technique de transformation de Box-Cox sous un plan d'échantillonnage à probabilités inégales. Étant convergent par rapport au plan, l'estimateur proposé retient la propriété de robustesse de l'estimateur GREG, même si le modèle sous-jacent est défaillant. Dans de nombreuses situations, une version du modèle $\mathbf{M_3}$ fournira au moins une approximation utile du comportement de la variable dépendante. La technique de Box-Cox permet qu'une transformation raisonnablement ajustée de la variable dépendante soit automatiquement déterminée par les données. Nous avons évalué analytiquement et par simulation de Monte Carlo la robustesse et l'efficacité de l'estimateur proposé.

Lorsqu'on compare un estimateur GREG basé sur un modèle linéaire sous-jacent ($\hat{T}_{G_L}$) à un autre fondé sur un modèle de Box-Cox ($\hat{T}_{G_BC}$), il ne faut pas perdre de vue que $\hat{T}_{G_L}$ ne requiert pas d'information auxiliaire complète. En outre, il peut produire un jeu unique de poids applicable à toutes les variables d'intérêt, contrairement à l'estimateur $\hat{T}_{G_BC}$. Toutefois, pour accroître l'efficacité de l'un et l'autre estimateur, il faut habituellement utiliser des poids distincts pour des variables d'intérêt différentes, parce les meilleures prédictions sont obtenues en ajustant un modèle de travail particulier pour chaque variable étudiée. L'estimateur $\hat{T}_{G_BC}$ peut même être plus efficace que l'estimateur $\hat{T}_{G_L}$, mais au prix de nécessiter des données complètes sur les variables x. Bien que ce genre d'information soit rarement disponible dans les sondages nord-américains auprès des ménages, elle l'est souvent dans les sondages auprès des entreprises.

Les sondages sont peu souvent réalisés pour mesurer une seule variable d'intérêt. La question est de savoir comment estimer les totaux finis pour des sous-population mutuellement exclusives et exhaustives de telle façon que la somme de ces estimations concorde avec l'estimation du total fini pour l'ensemble de la population. Nous devons accorder une attention particulière à ce problème, car les estimateurs de Box-Cox ne sont pas linéaires en y. Ces estimations peuvent être obtenues en utilisant un outil d'étalonnage standard décrit dans Li et Lahiri (2007). Une autre approche consiste à utiliser le calage fondé sur un

Remerciements

Je remercie le rédacteur en chef et trois examinateurs dont plusieurs suggestions constructives m'ont permis d'améliorer considérablement le manuscrit original. Je remercie M. Phil Kott de sa révision copieuse du texte et M. Alan Dorfman d'avoir fourni les données d'enquête utilisées dans l'étude. Je tiens aussi à exprimer ma reconnaissance au professeur Partha Lahiri, mon directeur de thèse de doctorat, et aux professeurs Katherine Abraham, Wolfgang Jank, Fritz Scheuren et Paul Smith, et à M. Paul D. Williams, membres du comité de supervision, pour leur encadrement généreux tout au long de mes travaux de recherche de thèse.

Annexe

Soit

$$\hat{T}_{G_BC} = \sum_{i\in s} y_i/\pi_i + \left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s} g_i(\hat{\boldsymbol{\theta}}_w)/\pi_i\right).$$

Par développement en série de Taylor, sous l'hypothèse (2), nous avons

$$g_i(\hat{\boldsymbol{\theta}}_w) = g_i(\boldsymbol{\theta}_N) + \left(\partial g_i(\mathbf{t})/\partial \mathbf{t}\right)|_{t=\boldsymbol{\theta}^*}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N),$$

où $\boldsymbol{\theta}^* \in (\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}}_w)$ ou $(\hat{\boldsymbol{\theta}}_w, \boldsymbol{\theta}_N)$ et $\left(\partial g_i(\mathbf{t})/\partial \mathbf{t}\right)$ est un vecteur ligne.

En vertu des hypothèses (1) et (2),

$$N^{-1}\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n})$$

$$N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n}).$$

Notons aussi qu'en vertu de l'hypothèse (3),

$$N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N) = N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n}).$$

Donc,

$$N^{-1}\left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w)\right) = O_p(1/\sqrt{n}).$$

En outre, en vertu de l'hypothèse (3),

$$N^{-1}\left(\sum_{i\in U} y_i - \sum_{i\in s} \pi_i^{-1} y_i\right) = O_p(1/\sqrt{n}).$$

Par conséquent, $N^{-1}(\hat{T}_{G_BC} - T) = O_p(1/\sqrt{n})$, *i.e.*, $\hat{T}_{G_BC}$ converge en probabilité vers T à l'ordre $O_p(1/\sqrt{n})$.

En plus, sous l'hypothèse (4), une approximation par série de Taylor de deuxième ordre de $g_i(\hat{\boldsymbol{\theta}}_w)$ peut être développée comme il suit :

$$g_i(\hat{\boldsymbol{\theta}}_w) = g_i(\boldsymbol{\theta}_N) + (\partial g_i(\mathbf{t})/\partial \mathbf{t})'\,|_{t=\boldsymbol{\theta}_N}$$
$$(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N)'\left(\partial^2 g_i(\mathbf{t})/\left(\partial \mathbf{t}\partial \mathbf{t}'\right)\right)|_{t=\boldsymbol{\theta}^*}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N),$$

où $\boldsymbol{\theta}^* \in (\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}}_w)$ ou $(\hat{\boldsymbol{\theta}}_w, \boldsymbol{\theta}_N)$. Il découle des hypothèses (1) et (4) que

$$N^{-1}\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N)$$
$$+ N^{-1}\sum_{i\in U}\left(\partial g_i(\mathbf{t})/\partial \mathbf{t}'\right)|_{t=\boldsymbol{\theta}_N}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + O_p(n^{-1}),$$

$$N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N)$$
$$+ N^{-1}\sum_{i\in s} \pi_i^{-1}\left(\partial g_i(\mathbf{t})/\partial \mathbf{t}'\right)|_{t=\boldsymbol{\theta}_N}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + O_p(n^{-1}).$$

En vertu des hypothèses (1) et (3), $\hat{\boldsymbol{\theta}}_w = \boldsymbol{\theta}_N + O_p(1/\sqrt{n})$ et

$$N^{-1}\sum_{i\in s} \pi_i^{-1}\partial g_i(\mathbf{t})/\partial \mathbf{t}\,|_{t=\boldsymbol{\theta}_N} = N^{-1}\sum_{i\in U} \partial g_i(\mathbf{t})/\partial \mathbf{t}\,|_{t=\boldsymbol{\theta}_N}$$
$$+ O_p(1/\sqrt{n}).$$

Donc,

$$N^{-1}\left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w)\right) =$$
$$N^{-1}\left(\sum_{i\in U} g_i(\boldsymbol{\theta}_N) - \sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N)\right) + O_p(n^{-1}).$$

Par conséquent,

$$\hat{T}_{G_BC} = \sum_{i\in s} y_i/\pi_i$$
$$+ \left(\sum_{i\in U} g_i(\boldsymbol{\theta}_N) - \sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N)\right) + O_p(N/n).$$

La variance asymptotique de $\hat{T}_{AG}$ sous le plan est :

$$AV_d(\hat{T}_{G_BC})$$
$$\approx \sum_{i\in U}\sum_{j\in U} (\pi_{ij} - \pi_i\pi_j)(y_i - g_i(\boldsymbol{\theta}_N))(y_j - g_j(\boldsymbol{\theta}_N))/(\pi_i\pi_j),$$

qui peut être estimée par

$$\hat{V}_d(\hat{T}_{G_BC})$$
$$\approx \sum_{i\in s}\sum_{j\in s} (\pi_{ij} - \pi_i\pi_j)/\pi_{ij}\,(y_i - g_i(\hat{\boldsymbol{\theta}}_w))(y_j - g_j(\hat{\boldsymbol{\theta}}_w))/(\pi_i\pi_j).$$

Bibliographie

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *Revue Internationale de Statistique*, 51, 279-92.

Bickel, P.J., et Doksum, K.A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296-311.

Box, G.E., et Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Séries B, 26, 211-252.

Boylan, T.A., Cuddy, M.P. et O'Muircheartaigh, I.G. (1980). The functional form of the aggregate import demand equation: A comparison of three european economies. *Journal of international economics*, 10, 561-566.

Boylan, T.A., Cuddy, M.P. et O'Muircheartaigh, I.G. (1982). Import demand equations: An application of a generalized Box-Cox methodology. *Revue Internationale de Statistique*, 50, 103-112.

Breidt, F.J., Claeskens, G. et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

Carrol, R.J., et Ruppert, D. (1988). *Transformations and weighting in Regression.* Londres : Chapman and Hall.

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

Chambers, R.L., et Dorfman, A.H. (2003). Transformed variables in survey sampling. Rapport technique.

Chen, G., et Chen, J. (1996). Une méthode de transformation applicable à l'échantillonnage de populations finies calée par une méthode de vraisemblance empirique. *Techniques d'enquête*, 22, 139-147.

Davison, C.W., Arnade, C.A. et Hallahan, C.B. (1989). Box-Cox estimation of U.S. soyabean exports. *Journal of Agricultural Economics Research*, 41, 8-15.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Estevao, V., Hidiroglou, M. et Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at statistics Canada. *Journal of Official Statistics*, 11, 181-204.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5–25.

Fuller, W.A., Loughin, M. et Baker, H. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

Gemmil, G. (1980). Using the Box-Cox Form for Estimating Demand: A Comment. *The review of economics and statistics*, 62, 147-148.

Gurka, M. (2004). The Box-Cox transformation in the general linear mixed model for longitudinal data. Ph.D. Dissertation.

Gurka, M. (2006). Expanding the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society*, Séries A, 169, 273-288.

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

Jayasuriya, B., et Valliant, R. (1996). Application de l'estimation par régression restreinte dans une enquête-ménage. *Techniques d'enquête*, 22, 127-138.

Jiang, J., et Lahiri, P. (2006). Estimation of finite population domain means - a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

John, N.R., et Draper, J.A. (1980). An alternative family of transformations. *Applied Statistics*, 29, 190–197.

Karlberg, F. (2000). Survey estimation for highly skewed population in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.

Khan, M.S., et Ross, K.Z. (1977). The functional form of aggregate import demand. *Journal of International Economics*, 7, 149-160.

Korn, E., et Graubard, B. (1998). Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête. *Techniques d'enquête*, 24, 209-218.

Li, Y., et Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102, 664-673.

Miner, A.G. (1982). The contribution of weather and technology to U.S. soybean yield. *Unpublished Dissertation*. University of Minnesota.

Montanari, G.E., et Ranalli, M.G. (2003). Nonparametric model calibration estimation in survey sampling. Manuscript.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429–1442.

Newman, P. (1977). Malaria and Mortality. *Journal of the American Statistical Association*, 72, 257-263.

Royall, R.M., et Cumberland, W.G. (1981). The finite population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Sakia, R.M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, 41, 169-178.

Särndal, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Plan Inference*, 7, 155-170.

Särndal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

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

Schlesselman, J. (1971). Power families: A note on the Box and Cox transformation. *Journal of the Royal Statistical Society*, Séries B, 33, 307-311.

Skinner, C.J., Holt, D. et Smith, T.M.F. (Éds.) (1989). *Analysis of complex surveys.* Chichester : Wiley.

Spitzer, J.J. (1976). The demand for money, the liquidity trap and functional forms. *The International Economics Review*, 17, 220-227.

Sverchkov, M., et Pfeffermann, D. (2004). Prévision des totaux de population finie basée sur la distribution échantillonnale. *Techniques d'enquête*, 30, 87-101.

Taylor, M.J.G. (1986). The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81, 114-118.

Tukey, J.W. (1957). The comparative anatomy of transformation. *Annals of Mathematical Statistics*, 28, 601-632.

Wu, C. (1999). The effective use of complete auxiliary information from survey data. Ph.D. dissertation, Simon Fraser University, Canada.

Wu, C., et Sitter, R.R. (2001). A model-calibration to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, 96, 185-193.

Zarembka, P. (1968). Functional form in the demand for money. *Journal of the American Statistical Association*, 63, 502-511.

Zheng, H., et Little, R.J.A. (2004). Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés. *Techniques d'enquête*, 30, 233-243.