Techniques d'enquête, décembre 1994
Vol. 20, n° 2, pp. 101–110
Statistique Canada



# Estimation par régression généralisée pour un échantillon à deux phases de dossiers fiscaux

JOHN ARMSTRONG et HÉLÈNE ST-JEAN[1]

RÉSUMÉ

Dans cet article, nous déterminons un estimateur par régression généralisé pour domaines ainsi qu'un estimateur approximatif de la variance correspondante suivant un plan d'échantillonnage à deux phases pour stratification avec échantillonnage de Poisson à chaque phase. Ces estimateurs sont une application du modèle général d'estimation par régression pour l'échantillonnage à deux phases élaboré dans Särndal et Swensson (1987) et par Särndal, Swensson et Wretman (1992). Nous étudions l'efficacité empirique de l'estimateur par régression généralisé à l'aide de données de l'échantillon à deux phases de dossiers fiscaux de Statistique Canada, formé annuellement. Enfin, nous comparons à l'estimateur de Horvitz-Thompson trois cas particuliers de l'estimateur par régression généralisé, soit deux estimateurs par régression et un estimateur de stratification a posteriori.

MOTS CLÉS: Estimation fondée sur un modèle; estimation pour domaines; échantillonnage de Poisson.

## 1. INTRODUCTION

Nous étudions dans cet article le problème de l'estimation pour domaines suivant un échantillonnage à deux phases pour stratification lorsque l'échantillonnage de Poisson est utilisé dans les deux phases. Considérons une population de $N$ unités et supposons qu'il faut estimer le total d'un caractère étudié, $y$, pour $L$ domaines disjoints. On peut prédire convenablement, mais non avec exactitude, l'effectif des domaines à l'aide d'une variable auxiliaire, $\Theta$, dont les valeurs ne sont pas observées avant l'échantillonnage. Il est moins coûteux d'obtenir de l'information sur $\Theta$ que sur $y$ et moins coûteux aussi que d'obtenir des données exactes sur l'effectif des domaines. Dans la première phase d'échantillonnage, on prélève un échantillon de Poisson dans la population et l'on observe la valeur de $\Theta$ pour chaque unité échantillonnée. Les unités sont ensuite stratifiées au moyen des valeurs de $\Theta$. Cette opération ressemble à une stratification par domaine. Dans la seconde phase d'échantillonnage, on tire un échantillon de Poisson dans chaque strate et l'on observe la valeur de $y$ pour chaque unité de l'échantillon, tout en recueillant des données exactes sur l'effectif des domaines.

L'estimateur de Horvitz-Thompson du total de $y$ pour le domaine $d$ est $\hat{Y}_{H-T}(d) = \sum_{i \in s2} y_i(d)/(p_{1i}p_{2i})$, où $y_i(d)$ prend la valeur de $y_i$ si l'unité $i$ appartient au domaine $d$ et la valeur zéro dans le cas contraire, $s2$ désigne l'échantillon de seconde phase, et $p_{1i}$ et $p_{2i}$ désignent les probabilités de sélection de l'unité $i$ pour les première et seconde phases, respectivement. Comme la taille des échantillons prélevés au moyen d'un échantillonnage de Poisson est une variable aléatoire, l'estimateur ci-dessus peut n'être pas efficace. (Voir Sunter 1986, ou Särndal, Swensson et Wretman 1992, p. 63.) L'estimateur par régression généralisé peut remplacer l'estimateur de Horvitz-Thompson lorsqu'on dispose d'information supplémentaire. Dans cet article, nous déterminons un estimateur par régression généralisé pour l'échantillonnage de Poisson à deux phases ainsi qu'un estimateur approximatif de la variance correspondante.

Dans la section 2, nous établissons l'estimateur par régression généralisé et l'estimateur approximatif de la variance. Ensuite, nous décrivons dans la section 3 l'application qui a fait ressortir le problème de l'estimation, soit l'échantillon à deux phases de dossiers fiscaux de Statistique Canada formé annuellement. Enfin, dans la section 4, nous présentons les résultats d'une étude empirique où l'estimateur de Horvitz-Thompson est comparé à trois cas particuliers de l'estimateur par régression généralisé, soit l'estimateur de stratification a posteriori utilisé actuellement dans le plan de sondage appliqué aux déclarations et deux estimateurs par régression.

## 2. ESTIMATION PAR RÉGRESSION GÉNÉRALISÉE

L'estimation par régression généralisée n'est pas nouvelle. Deming et Stephan (1940) décrivent un estimateur par régression généralisé pour un plan d'échantillonnage à une phase. Lemaître et Dufour (1987) et Bankier, Rathwell et Majkowski (1992) décrivent des applications récentes de l'estimation par régression généralisée à Statistique Canada. Hidiroglou, Särndal et Binder (1993) traitent en profondeur l'utilisation des estimateurs par régression généralisés dans les enquêtes-entreprises.

On peut aborder la question des estimateurs par régression généralisés du point du vue de l'échantillonnage fondé sur un modèle (Särndal, Swensson et Wretman 1992) ou du point de vue du calage (Deville et Särndal 1992). Soient

[1] John Armstrong, Division des études sociales et économiques, 24 - Immeuble R.H. Coats, et Hélène St-Jean, Division des méthodes d'enquêtes-entreprises, 11 - Immeuble R.H. Coats, Statistique Canada, Parc Tunney, Ottawa (Ontario), K1A 0T6.

$U = \{u\}$ et $V = \{v\}$ les ensembles de "strates a posteriori" respectivement de première et de seconde phase. Dans la pondération de l'échantillon de première phase pour l'estimation par régression généralisée, on corrige les poids initiaux $1/p_{1i}$ de manière à obtenir des poids $w_{1i} = g_{1i}/p_{1i}$ qui satisfont les équations de calage

$$\sum_{i \in s1 \cap v} w_{1i} x_i = X_u,$$

pour chaque "strate a posteriori" de première phase $u$, où $x_i$ est un vecteur $L_1 \times 1$ de variables auxiliaires connu pour toutes les unités de la population et $X_u$ est le vecteur de totaux de variables auxiliaires pour la strate a posteriori $u$. Les poids corrigés ont pour effet de minimiser la mesure de distance $\sum_{i \in s1} (g_{1i} - 1)^2/p_{1i}$. On peut obtenir les mêmes poids avec un modèle en se servant de l'équation

$$E_\xi(y_i) = x_i' \beta_u, \; i \in u$$

$$V_\xi(y_i) = \sigma^2,$$

où $y_i$ est la valeur de la variable étudiée pour l'unité $i$, et $E_\xi(\cdot)$ et $V_\xi(\cdot)$ désignent respectivement l'espérance et la variance de modèle.

En ce qui concerne les estimateurs par régression généralisés qui nous intéressent, la pondération de l'échantillon de seconde phase comprend une opération de calage qui dépend des résultats de la pondération en première phase. On corrige les poids initiaux, $w_{1i}/p_{2i}$, de manière à obtenir les poids finals $w_i = g_{2i} w_{1i}/p_{2i}$, qui satisfont les équations de calage

$$\sum_{i \in s2 \cap v} w_i z_i = \tilde{Z}_v,$$

pour chaque strate a posteriori de seconde phase $v$, où $z_i$ est un vecteur $L_2 \times 1$ de variables auxiliaires connu pour toutes les unités de l'échantillon de première phase et $\tilde{Z}_v = \sum_{i \in s1 \cap v} w_{1i} z_i$ est l'estimation du vecteur de totaux de variables auxiliaires pour la strate a posteriori $v$, calculée à l'aide des poids corrigés de la première phase, $w_{1i}$. Notons que ces équations de calage diffèrent sensiblement des exemples contenus dans Särndal et Swensson (1987, p. 284-288) et Särndal, Swensson et Wretman (1992, p. 359-366) du fait qu'elles renferment des poids de première phase corrigés au lieu de poids initiaux. Les poids finals ont pour effet de minimiser la mesure de distance $\sum_{i \in s2} w_{1i}(g_{2i} - 1)^2/p_{2i}$. On peut obtenir les mêmes poids avec un modèle en se servant de l'équation

$$E_\xi(w_{1i} y_i) = w_{1i} z_i' \beta_v, \; i \in v$$

$$V_\xi(w_{1i} y_i) = w_{1i} \sigma^2.$$

Il y a deux grands avantages à utiliser des poids de première phase corrigés au lieu de poids initiaux dans les équations de calage de seconde phase. Premièrement, on peut définir l'estimateur par régression généralisé pour le domaine $d$ par l'expression

$$\hat{Y}_{\mathrm{GREG}}(d) = \sum_{i \in s2} y_i(d) g_{1i} g_{2i}/p_{1i} p_{2i},$$

qui utilise des "poids $g$" de première et de seconde phase. Deuxièmement, supposons que des variables auxiliaires servent au calage dans les deux phases de pondération. On peut alors utiliser les poids finals pour obtenir, pour ces variables, des estimations de totaux de population qui soient égales au total réel.

Désignons par $\check{X}_u = \sum_{i \in s1 \cap u} x_i/p_{1i}$ le vecteur $L_1 \times 1$ des estimations Horvitz-Thompson de totaux de variables auxiliaires pour la strate a posteriori de première phase $u$. Le poids $g$ de première phase est défini par l'équation

$$g_{1i} = 1 + \lambda_u' x_i,$$

où $\lambda_u' = (X_u - \check{X}_u)' M_u^{-1}$ et $M_u^{-1} = (\sum_{i \in s1 \cap u} x_i x_i'/p_{1i})^{-1}$. Pour la strate a posteriori de seconde phase $v$, désignons l'estimation de $\tilde{Z}_v$ fondée sur des poids initiaux de seconde phase par $\dot{Z}_v = \sum_{i \in s2 \cap v} w_{1i} z_i/p_{2i}$. Le poids $g$ de seconde phase est défini par l'équation

$$g_{2i} = 1 + \lambda_v' z_i,$$

où $\lambda_v' = (\tilde{Z}_v - \dot{Z}_v)' M_v^{-1}$ et $M_v^{-1} = (\sum_{i \in s2 \cap v} w_{1i} z_i z_i'/p_{2i})^{-1}$.

La variance approximative de $\hat{Y}_{\mathrm{GREG}}(d)$ est définie par l'équation

$$V(\hat{Y}_{\mathrm{GREG}}(d)) \approx \sum_i \frac{1 - p_{1i}}{p_{1i}} Q_{1i}^2 + E_1\left[\sum_{i \in s2} \frac{1 - p_{2i}}{p_{2i}} (w_{1i} Q_{2i})^2\right],$$

où $E_1(\cdot)$ désigne l'espérance par rapport à la première phase d'échantillonnage, $Q_{1i} = y_i(d) - x_i' B_u$ pour chaque unité contenue dans la strate a posteriori de première phase $u$, et $B_u$, le vecteur des coefficients estimés de la régression de $y(d)$ par rapport à $x$ que l'on obtiendrait si $y(d)$ était connue pour toutes les unités de la strate a posteriori de première phase $u$, est défini par l'équation

$$B_u = \left(\sum_{i \in u} x_i x_i'\right)^{-1} \left(\sum_{i \in u} x_i y_i(d)\right).$$

De même, $Q_{2i} = y_i(d) - z_i' B_v$ pour chaque unité contenue dans la strate a posteriori de seconde phase $v$ et $B_v$, le vecteur des coefficients estimés de la régression de $y(d)$ par rapport à $z$ que l'on obtiendrait – moyennant le calage de la première phase – si $y(d)$ était connue pour toutes les unités de l'échantillon de première phase contenues dans la strate a posteriori de seconde phase $v$, est défini par l'équation

$$B_v = \left(\sum_{i \in s1 \cap v} w_{1i} z_i z_i'\right)^{-1} \left(\sum_{i \in s1 \cap v} w_{1i} z_i y_i(d)\right).$$

On peut définir l'estimateur de la variance approximative de $\hat{Y}_{\mathrm{GREG}}(d)$ par l'équation

$$\hat{V}(\hat{Y}_{\mathrm{GREG}}(d)) = \sum_i \frac{1 - p_{1i}}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2 + \sum_i \frac{1 - p_{2i}}{(p_{1i} p_{2i})^2} (g_{1i} g_{2i} q_{2i})^2.$$

Comme $y(d)$ n'est connue que pour les unités de $s2$, les estimations de $B_u$ et $B_v$ sont calculées au moyen des équations

$$\hat{B}_u = \left( \sum_{i \in s2 \cap u} w_i x_i x_i' \right)^{-1} \left( \sum_{i \in s2 \cap u} w_i x_i y_i(d) \right),$$

$$\hat{B}_v = \left( \sum_{i \in s2 \cap v} w_i z_i z_i' \right)^{-1} \left( \sum_{i \in s2 \cap v} w_i z_i y_i(d) \right).$$

Les résidus d'échantillon requis pour calculer l'estimateur de la variance sont $q_{1i} = y_i(d) - x_i' \hat{B}_u$ et $q_{2i} = y_i(d) - z_i' \hat{B}_v$. L'annexe A donne plus de détails sur le calcul de la variance approximative de $\hat{Y}_{\mathrm{GREG}}(d)$ et de l'estimateur de cette variance.

Si $y$ est fortement corrélée avec $x$ et $z$, la variance de l'estimateur par régression généralisé du total de population de $y$ sera plutôt faible. Cependant, il convient de souligner que l'existence d'une forte corrélation entre $y$, d'une part, et $x$ et $z$, d'autre part, ne signifie pas nécessairement que la variance de l'estimateur du total de $y$ pour un domaine en particulier sera assez peu élevée, puisque $y(d)$ peut être faiblement corrélée avec $x$ et $z$ dans les strates a posteriori qui contiennent au moins une unité de l'échantillon appartenant au domaine $d$.

La corrélation de $y(d)$ avec $x$ et $z$ dans une strate à posteriori contenant au moins une unité de l'échantillon appartenant au domaine $d$ pourra être faible si la strate a posteriori compte des unités échantillonnées qui n'appartiennent pas au domaine $d$. Ce cas peut se produire souvent si l'on ne connaît pas les totaux de domaines pour les variables auxiliaires ou si l'on ne peut obtenir de données exactes sur l'effectif des domaines en ce qui concerne les unités de l'échantillon de première phase. Dans le cas de l'échantillonnage à deux phases pour stratification, on ne connaît rien sur l'effectif des domaines avant le tirage de l'échantillon de première phase. Si chaque strate a posteriori de première phase est formée par le regroupement de plusieurs strates d'échantillonnage de première phase, par exemple, la plupart des strates a posteriori contiendront plus d'un domaine. La variable $\Theta$ utilisée pour prédire l'effectif des domaines durant la stratification de l'échantillon de première phase n'est pas un prédicteur exact. Si les strates a posteriori de seconde phase sont formées par le regroupement de strates d'échantillonnage de la seconde phase, chaque domaine peut être réparti entre un certain nombre de strates a posteriori de la seconde phase.

Suivant le type d'information supplémentaire qui sera utilisée, les poids $g$ rattachés à l'estimateur par régression généralisé et, par conséquent, aux estimations correspondantes pourraient être négatifs.

## 3. APPLICATION: ÉCHANTILLONNAGE À DEUX PHASES DE DOSSIERS FISCAUX

L'échantillonnage à deux phases de dossiers fiscaux s'inscrit dans une stratégie globale de Statistique Canada concernant la production d'estimations annuelles sur l'activité économique au Canada. Pour les grandes entreprises, les données sont recueillies au moyen d'enquêtes postales; pour les petites, elles proviennent d'un échantillon de dossiers fiscaux. La combinaison des estimations d'enquête et des estimations fondées sur l'échantillon de dossiers fiscaux permet d'estimer la valeur de variables financières pour l'univers des entreprises. En ce qui regarde le calcul d'estimations pour les petites entreprises seulement, on préfère utiliser les données fiscales plutôt que les données d'enquête afin de réduire les coûts et le fardeau de réponse.

L'échantillonnage à deux phases a été rendu nécessaire parce qu'il fallait faire des estimations pour des domaines définis selon le code à quatre chiffres de la Classification type des industries (CTI) (Statistique Canada 1980). Les deux premiers chiffres du code (CTI2) permettent de diviser l'activité économique en 76 groupes dans un premier temps. Ensuite, les deux derniers chiffres permettent d'obtenir une classification plus détaillée à l'intérieur de chaque groupe (CTI4). Par exemple, une entreprise pourrait être classée dans l'industrie du transport d'après son code CTI2 et dans l'industrie du camionnage de vrac liquide d'après son code CTI4.

Il y a deux types de déclarants: ceux qui produisent une formule T1 (déclarants T1) et ceux qui produisent une T2 (déclarants T2). Les premiers sont des particuliers, qui peuvent posséder entièrement ou partiellement une ou plusieurs entreprises non constituées en société, tandis que les seconds sont des entreprises constituées en société. C'est Revenu Canada, le ministère du gouvernement canadien chargé de la perception de l'impôt, qui fournit à Statistique Canada les fichiers administratifs contenant de l'information restreinte sur tous les déclarants qui ont des intérêts dans des entreprises. Ces fichiers servent à établir une base de sondage. Cette base ne contient cependant aucune information sur le nombre d'entreprises que possèdent les déclarants T1 ni sur le degré de participation de ces personnes. Mais elle contient des données géographiques et des données sur le revenu brut et le bénéfice net d'entreprise pour les deux types de déclarants. Elle contient aussi des données sur quelques autres variables financières importantes, dont la rémunération et les stocks, pour les déclarants T2. On cherche à établir des estimations sur quelque 35 variables financières qui figurent dans les déclarations de revenus et les états financiers qui y sont annexés, mais non dans les fichiers administratifs fournis par Revenu Canada.

Les déclarants qui ont des intérêts dans des entreprises sont classés par Revenu Canada selon la CTI. En règle générale, l'activité économique mentionnée sur la déclaration de revenus permet de déterminer assez précisément le code à deux chiffres du déclarant. Revenu Canada élargit ce code à trois ou à quatre chiffres pour la plupart des déclarants. Cependant, les déclarants n'ont pas tous un code à quatre chiffres, et les deux derniers chiffres du code CTI4 attribué par Revenu Canada sont assez imprécis. L'échantillonnage à deux phases de dossiers fiscaux a pour but de permettre une estimation plus précise de la production économique au niveau d'agrégation à quatre chiffres.

La section 3.1 contient une brève description du plan d'échantillonnage à deux phases. Armstrong, Block et Srinath (1993) traitent plus en détail ce type d'échantillonnage. Les sections 3.2 et 3.3 portent sur l'estimation dans un contexte d'échantillonnage à deux phases: la section 3.2 traite de l'estimateur de Horvitz-Thompson et la section 3.3 examine un estimateur de stratification a posteriori.

### 3.1 Plan d'échantillonnage

Les données administratives qui servent à constituer la base de sondage pour une année d'imposition particulière sont tirées des déclarations de revenus que traite Revenu Canada sur deux années civiles. L'échantillonnage de Poisson offre donc de nets avantages opérationnels parce qu'on peut commencer l'échantillonnage avant même de disposer d'une base de sondage complète.

La population visée par l'échantillonnage de dossiers fiscaux est l'univers des entreprises dont le revenu brut dépasse 25,000$, à l'exception des grandes entreprises visées par les enquêtes postales. L'échantillon de première phase est un échantillon longitudinal de déclarants. Les strates sont définies selon le niveau à 2 chiffres de la Classification type des industries (CTI2), la province et la taille de l'entreprise (revenu brut). Tous les déclarants qui font partie de l'échantillon de première phase pour l'année d'imposition $T$ et dont le dossier peut être échantillonné à nouveau pour l'année d'imposition $T + 1$ demeurent dans l'échantillon pour l'année $T + 1$. Chaque année, on peut ajouter des déclarants dans l'échantillon de première phase afin d'accroître la précision de certaines estimations et de remplacer les déclarants échantillonnés les années précédentes mais qui ne font plus partie du champ de l'enquête.

Pour effectuer l'échantillonnage de Poisson en première phase, on attribue à chaque déclarant un nombre pseudo-aléatoire (nombre factice) qui se situe dans l'intervalle (0,1). Ce nombre est généré par une fonction de calcul qui utilise le numéro d'identification du déclarant comme paramètre, puis il est comparé à l'intervalle d'échantillonnage pour la strate correspondante. Si un nombre factice particulier se situe dans l'intervalle de sondage en question et que le déclarant correspondant n'est pas déjà inclus dans l'échantillon de première phase, alors ce déclarant est ajouté à l'échantillon. Comme les identificateurs de déclarants ne changent pas, l'échantillonnage de Poisson facilite la formation d'un échantillon longitudinal de première phase.

Les probabilités d'échantillonnage en première phase des déclarants qui font déjà partie de l'échantillon de première phase sont mises à jour chaque année. Cette opération est nécessaire pour deux raisons: i) un déclarant peut changer de strate d'échantillonnage de première phase d'une année d'imposition à l'autre; ii) les fractions de sondage de première phase pour une strate donnée peuvent varier d'une année à l'autre.

Revenu Canada expédie à Statistique Canada une copie des déclarations de revenus et des états financiers des déclarants qui font partie de l'échantillon de première phase. Pour former l'échantillon de seconde phase, on crée des entités statistiques à partir de données sur les entreprises correspondant aux déclarants de l'échantillon de première phase. Posons $J = \{j\}$ comme l'univers des entreprises visées par l'échantillonnage de dossiers fiscaux. Une entité statistique, désignée par $(i,j)$, est créée pour chaque combinaison déclarant-entreprise dans l'échantillon de première phase. Pour chaque déclarant T1, les données sur les entreprises que possède entièrement ou partiellement ce déclarant (y compris son degré de participation) – et qui sont nécessaires à la création des entités statistiques – se trouvent dans les déclarations de revenus et dans les états financiers qui y sont annexés. Comme il y a une correspondance biunivoque entre les entreprises et les déclarants T2, une seule entité statistique est créée pour chaque déclarant T2 de l'échantillon de première phase.

Pour chaque année d'imposition, Statistique Canada attribue un code CTI4 aux entités statistiques qui n'ont jamais fait partie de l'échantillon. Ce code est déterminé au moyen d'informations qui ne se limitent pas à l'activité économique mentionnée sur la déclaration de revenus, et il est plus précis, pour les troisième et quatrième chiffres, que le code attribué par Revenu Canada. Pour les entités statistiques qui ont déjà fait partie de l'échantillon, elles conservent le code CTI4 qui leur avait été attribué.

D'un point de vue conceptuel, l'échantillon de seconde phase est un échantillon d'entreprises. D'un point de vue technique, c'est un échantillon de déclarants prélevé au moyen d'entités statistiques. Celles-ci sont stratifiées selon le code CTI4 attribué par Statistique Canada et selon la province et la taille. Le revenu total de l'entreprise $j$ sert de variable de taille pour l'entité statistique $(i,j)$. Si une entité statistique rattachée à un déclarant T1 en particulier est échantillonnée en seconde phase, alors toutes les entités statistiques qui se rapportent à ce déclarant seront échantillonnées. Par conséquent, la probabilité d'échantillonnage de seconde phase pour l'entité statistique $(i,j)$ dépend uniquement de $i$.

L'échantillon de seconde phase est prélevé selon la méthode de Poisson, où l'on utilise des nombres factices générés à partir du numéro d'identification des déclarants. La fonction de calcul utilisée pour l'échantillonnage de seconde phase n'a aucun rapport avec celle qui est employée pour l'échantillonnage de première phase.

Les données relatives à quelque 35 variables financières sont extraites des déclarations de revenus et des états financiers des déclarants échantillonnés dans la seconde phase. Si nécessaire, on met à jour les codes CTI4 attribués par Statistique Canada pour faire en sorte que tous les codes à quatre chiffres utilisés dans le calcul des estimations correspondent à l'année d'imposition pertinente.

### 3.2 Estimateur de Horvitz-Thompson

L'échantillon de seconde phase est un échantillon d'entreprises prélevé à l'aide d'entités statistiques. Comme certaines entreprises sont des sociétés en nom collectif, il se peut que plusieurs entités statistiques se rapportent à la même entreprise. Si l'on veut établir des estimations pour l'univers des entreprises, il faut prévoir une correction qui tienne compte de l'effet de l'existence de sociétés en nom collectif. Si l'entreprise $j$ est une société en nom collectif, elle sera incluse dans l'échantillon de seconde phase si au moins un des déclarants qui ont un intérêt dans cette entreprise est échantillonné. Pour éviter que l'estimateur de Horvitz-Thompson habituel ne produise une surestimation, il faut introduire un facteur de correction qui tienne compte de l'existence de sociétés en nom collectif. Désignons par $\delta_{ij}$ la participation du déclarant $i$ dans l'entreprise $j$ et supposons que l'entité statistique $(i,j)$ est échantillonnée dans la seconde phase. On corrige les chiffres relatifs à l'entreprise $j$ au moyen du facteur $\delta_{ij}$ de manière que seule la part des revenus et des dépenses rattachée au déclarant $i$ entre en ligne de compte dans les estimations. Rao (1968a) décrit une méthode de correction semblable dans un contexte un peu différent.

Soit $y_j$ la valeur de la variable $y$ pour l'entreprise $j$. L'estimateur de Horvitz-Thompson du total de $y$ pour le domaine $d$, y compris la correction pour les sociétés en nom collectif, est défini par l'expression

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} \sum_{j \in J_i} \delta_{ij} y_j(d) / (p_{1i} p_{2i}),$$

où $J_i$ est l'ensemble des indices des entreprises qui appartiennent en tout ou en partie au déclarant $i$. Comme les probabilités d'échantillonnage dépendent uniquement de l'indice $i$, $\hat{Y}_{H-T}(d)$ peut s'écrire

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} y_i(d) / (p_{1i} p_{2i}),$$

où

$$y_i(d) = \sum_{j \in J_i} \delta_{ij} y_j(d).$$

$\hat{Y}_{H-T}(d)$ est un estimateur sans biais du total de $y$ pour les entreprises du domaine $d$. Voir Rao (1968a).

On obtient l'échantillon de seconde phase en soumettant l'échantillon de Poisson prélevé en première phase à un nouvel échantillonnage de Poisson. L'échantillon de seconde phase est donc, lui aussi, un échantillon de Poisson et la variance de $\hat{Y}_{H-T}(d)$ est

$$V(\hat{Y}_{H-T}(d)) = \sum_{i} \left[(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})\right] y_i(d)^2.$$

Un estimateur non biaisé de cette variance est défini par l'expression

$$\hat{V}(\hat{Y}_{H-T}(d)) = \sum_{i \in s2} \left[(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})^2\right] y_i(d)^2.$$

### 3.3 Estimateur de Horvitz-Thompson avec stratification a posteriori

Brewer, Early et Joyce (1972) ont proposé de corriger l'estimateur de Horvitz-Thompson de manière à tenir compte de la différence entre la taille réelle et la taille prévue des échantillons de Poisson. On retrouve ces corrections dans la méthode actuellement utilisée pour calculer des estimations à partir de l'échantillon à deux phases de dossiers fiscaux.

Des coefficients de correction sont appliqués aux estimateurs de strate formée a posteriori durant la pondération des échantillons de première et de seconde phase. Choudhry, Lavallée et Hidiroglou (1989) font une analyse générale de la pondération d'un échantillon de Poisson à deux phases avec des coefficients de correction appliqués dans un contexte de stratification a posteriori. Supposons que la strate a posteriori $u$ de première phase contient $N_u$ déclarants. On peut estimer le nombre de déclarants inclus dans la strate a posteriori $u$ de première phase, d'après l'échantillon de première phase, par l'équation

$$\check{N}_u = \sum_{i \in s1 \cap u} (1/p_{1i}).$$

Le poids de stratification a posteriori du déclarant $i$, $i \in u$, pour la première phase est

$$w_{1i} = (1/p_{1i})(N_u/\check{N}_u).$$

On peut estimer le nombre de déclarants inclus dans la strate a posteriori $v$ de seconde phase, d'après l'échantillon de première phase, par l'équation

$$\tilde{N}_v = \sum_{i \in s1 \cap v} w_{1i}.$$

Un autre estimateur du nombre de déclarants inclus dans la strate a posteriori $v$, basé uniquement sur les unités de l'échantillon de seconde phase, est

$$\dot{N}_v = \sum_{i \in s2 \cap v} w_{1i}/p_{2i}.$$

Le poids de stratification a posteriori de l'entité statistique $(i,j) \in v$, pour la seconde phase est

$$w_{2i} = (1/p_{2i})(\tilde{N}_v/\dot{N}_v)$$

et le poids final est

$$w_i = w_{1i} w_{2i}.$$

L'estimateur de stratification a posteriori du total de $y$ pour le domaine $d$ est défini par l'expression

$$\hat{Y}(d) = \sum_{i \in s2} w_i y_i(d).$$

Choudhry, Lavallée et Hidiroglou (1989) notent que la variance de $\hat{Y}(d)$ est définie approximativement par l'expression

$$V(\hat{Y}(d)) \approx \sum_u \sum_{i \in u} \frac{(1 - p_{1i})}{p_{1i}} \left( y_i(d) - \frac{Y_u(d)}{N_u} \right)^2 + \sum_v \sum_{i \in v} \frac{(1 - p_{2i})}{p_{1i} p_{2i}} \left( y_i(d) - \frac{Y_v(d)}{N_v} \right)^2,$$

où $Y_u(d)$ et $Y_v(d)$ sont les totaux de la variable $y$ pour les parties du domaine $d$ contenues dans les strates a posteriori $u$ et $v$ respectivement.

L'estimateur de cette variance est défini par l'expression

$$\hat{V}(\hat{Y}(d)) = \sum_u \sum_v \left( \frac{N_u}{\check{N}_u} \right)^2 \left( \frac{\tilde{N}_v}{\dot{N}_v} \right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} \left( y_i(d) - \frac{\hat{Y}_u(d)}{\hat{N}_u} \right)^2 + \sum_u \sum_v \left( \frac{N_u}{\check{N}_u} \right)^2 \left( \frac{\tilde{N}_v}{\dot{N}_v} \right)^2 \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{2i})}{(p_{1i} p_{2i})^2} \left( y_i(d) - \frac{\hat{Y}_v(d)}{\hat{N}_v} \right)^2,$$

où les estimations $\hat{N}_u$ et $\hat{N}_v$ sont calculées à l'aide des poids finals.

La présence du facteur $(N_u/\check{N}_u)^2(\tilde{N}_v/\dot{N}_v)^2$ peut s'expliquer par l'amélioration des propriétés conditionnelles de l'estimateur (Royall et Eberhardt 1975). Wu (1982) a lui aussi étudié un estimateur de variance qui contient un facteur de correction analogue; il s'agit de l'estimateur de la variance d'un estimateur par quotient pour un échantillonnage à une phase. L'étude empirique de Wu et Deng (1983) montre que l'utilisation du facteur de correction contribue à améliorer les propriétés de couverture des intervalles de confiance basés sur l'approximation normale.

L'estimateur $\hat{Y}(d)$ est un cas particulier de $\hat{Y}_{\text{GREG}}(d)$, qui peut s'appliquer lorsque la pondération des échantillons de première et de seconde phase fait intervenir une seule variable auxiliaire qui a la valeur un pour tous les déclarants. Dans ce cas, nous avons $g_{1i} = N_u/\check{N}_u$ pour tous les déclarants de la strate a posteriori $u$ de la première phase et $g_{2i} = \tilde{N}_v/\dot{N}_v$ pour tous les déclarants de la strate a posteriori $v$ de la seconde phase. Notons que le choix de ces variables auxiliaires élimine la possibilité de poids $g$ négatifs. L'estimateur de variance $\hat{V}(\hat{Y}(d))$ diffère peu de l'estimateur $\hat{V}(\hat{Y}_{\text{GREG}}(d))$ pour ce cas particulier de $\hat{Y}_{\text{GREG}}(d)$. Le poids $g$ de seconde phase apparaît dans le premier terme de $\hat{V}(\hat{Y}(d))$ mais non dans $\hat{V}(\hat{Y}_{\text{GREG}}(d))$.

## 4. ÉTUDE EMPIRIQUE

Afin de comparer le rendement de $\hat{Y}_{H-T}(d)$, $\hat{Y}(d)$ et $\hat{Y}_{\text{GREG}}(d)$, nous avons effectué une étude empirique à l'aide de données de la province de Québec pour l'année d'imposition 1989. Comme l'estimateur $\hat{Y}(d)$ est un cas particulier de $\hat{Y}_{\text{GREG}}(d)$, on l'appellera $\hat{Y}_{\text{GREG-HDP}}(d)$ dans l'analyse qui va suivre. (HDP est le sigle utilisé pour "Hájek deux phases".) Deux autres estimateurs par régression généralisés ont été considérés. Dans les deux cas, $x$ et $z$ contiennent une variable qui prend la valeur un pour tous les déclarants. L'un des deux estimateurs implique un calage par rapport au revenu du déclarant dans la pondération de seconde phase. (Le revenu du déclarant est inclus dans $z$ comme deuxième variable auxiliaire.) L'autre estimateur implique un calage par rapport au revenu du déclarant dans les deux phases de pondération. (Le revenu du déclarant est inclus dans $x$ et $z$ comme deuxième variable auxiliaire.) Dans l'analyse qui suit, nous désignerons les estimations de totaux de domaine calculées à l'aide de ces deux estimateurs par $\hat{Y}_{\text{GREG-R2}}(d)$ et $\hat{Y}_{\text{GREG-R1R2}}(d)$ respectivement.

Des estimations ont été calculées pour deux variables en particulier: le revenu transcrit et les dépenses totales. Il y a quelques différences conceptuelles entre le revenu transcrit et le revenu du déclarant. Par exemple, dans de nombreuses industries les gains en capital et les produits exceptionnels sont inclus dans le revenu du déclarant mais exclus du revenu transcrit. En outre, le revenu du déclarant donne lieu plus souvent à des erreurs de saisie que le revenu transcrit étant donné qu'il n'est pas soumis aux mêmes normes de contrôle qualitatif.

L'univers étudié comprenait environ 140,000 déclarants T2 qui avaient indiqué des revenus de plus de 25,000$ pour l'année d'imposition 1989. Nous avons appliqué les probabilités d'échantillonnage de première et de seconde phase qui avaient servi à l'échantillonnage de déclarations produites pour l'année d'imposition 1989. L'échantillon de première phase comptait environ 31,000 déclarants tandis que l'échantillon de seconde phase comprenait environ 23,000 entreprises. La corrélation entre le revenu du déclarant et le revenu transcrit pour les entreprises de l'échantillon de seconde phase était de 0.969, tandis que la corrélation entre le revenu du déclarant et les dépenses totales était de 0.960.

Une forte proportion des unités des échantillons de première et de seconde phase ont été échantillonnées avec une probabilité égale à un. Toutes les unités ayant une probabilité de sélection égale à un dans la première phase ont été exclues de la pondération de première phase et la valeur des poids $g$ correspondants a été fixée à un. Les unités qui avaient une probabilité de sélection égale à un dans la seconde phase ont été traitées de la même manière dans la pondération de seconde phase. L'échantillon de première phase comptait 9,884 unités dont la probabilité de sélection en première phase était différente de un, tandis que l'échantillon de seconde phase comptait 910 unités dont la probabilité de sélection en seconde phase était elle aussi différente de un. Chaque strate a posteriori de

première phase était constituée d'une ou de plusieurs des strates d'échantillonnage de première phase qui avaient servi à l'échantillonnage des déclarations produites pour 1989. Ces strates étaient définies selon cinq catégories de revenu. Les strates contenues dans une strate a posteriori de première phase en particulier correspondaient toutes à la même catégorie de revenu. Chaque strate a posteriori de première phase devait contenir au moins vingt unités échantillonnées, cela à cause de la possibilité d'un biais dans $\hat{V}(\hat{Y}_{\mathrm{GREG}}(d))$ lorsque le nombre d'unités échantillonnées servant à l'estimation des coefficients de régression est très petit (Rao 1968b). Si une strate d'échantillonnage de première phase contenait moins de vingt unités échantillonnées, elle était combinée à d'autres strates correspondant aux mêmes codes CTI2 et à la même catégorie de revenu, jusqu'à ce qu'on obtienne une strate a posteriori contenant au moins vingt unités échantillonnées. Nous avons constitué ainsi 166 strates a posteriori de première phase. Les strates a posteriori de seconde phase ont été formées de la même façon, c'est-à-dire en combinant des strates d'échantillonnage ayant trait aux mêmes codes CTI4 jusqu'à ce que chaque strate a posteriori compte au moins vingt unités échantillonnées. Nous avons obtenu ainsi 30 strates a posteriori.

Nous avons calculé les poids de première et de seconde phase pour $\hat{Y}_{\mathrm{GREG\text{-}HDP}}(d)$, $\hat{Y}_{\mathrm{GREG\text{-}R2}}(d)$ et $\hat{Y}_{\mathrm{GREG\text{-}R1R2}}(d)$ à l'aide d'une version modifiée de la macro SAS CALMAR (Sautory, 1991). L'ensemble de poids d'échantillonnage de première phase calculés pour l'estimateur GREG-R1R2 comprenait douze poids négatifs. En revanche, nous n'avons calculé aucun poids négatif en seconde phase pour GREG-R2 ou GREG-R1R2. (On ne peut avoir de poids négatifs avec l'estimateur GREG-HDP.) En nous servant des trois estimateurs GREG et de $\hat{Y}_{\mathrm{H\text{-}T}}(d)$, nous avons calculé des estimations du revenu transcrit et des dépenses totales pour 77 domaines définis par un code CTI2, 256 domaines définis par un code CTI3 et 587 domaines définis par un code CTI4. Comme l'estimateur GREG-R1R2 n'a produit aucune valeur négative, nous n'avons pas tenté de modifier les poids négatifs rattachés à cet estimateur.

Les tableaux 1 et 2 donnent les résultats de la comparaison de l'estimateur GREG-HDP et de l'estimateur de Horvitz-Thompson (H-T). La moyenne indiquée pour les gains et les pertes est une moyenne de rapports de coefficients de variation. On constate que l'estimateur GREG-HDP est plus efficace que l'estimateur H-T pour la majorité des domaines. Sa supériorité est notable particulièrement dans le cas des domaines CTI2. En ce qui concerne les domaines CTI4, le coefficient de variation (CV) estimé de l'estimation GREG-HDP des dépenses totales est inférieur au CV estimé de l'estimation H-T pour 60.5% des domaines, et l'écart entre les deux CV estimés est en moyenne de 5.5%. Lorsque le CV estimé de l'estimation GREG-HDP est supérieur au CV estimé de l'estimation H-T, l'écart est de 7.9% en moyenne. Outre l'information contenue dans les tableaux 1 et 2, il y a une autre raison de préférer GREG-HDP à l'estimateur H-T: chaque année, il arrive que les dossiers fiscaux de certains déclarants de l'échantillon ne parviennent pas à Statistique Canada ou qu'il ne soient pas utilisables parce qu'ils ne contiennent pas les états financiers voulus. À supposer que l'on n'ait pas à tenir compte de ces cas de non-réponse, l'estimateur GREG-HDP s'impose comme un moyen de compensation de la non-réponse.

**Tableau 1**

Comparaison de l'estimateur GREG-HDP et de l'estimateur H-T pour le revenu transcrit, coefficients de variation estimés

| Type de domaine | Gains engendrés par l'utilisation de GREG-HDP | | Pertes engendrées par l'utilisation de GREG-HDP | |
|---|---|---|---|---|
| | Nombre | Moyenne | Nombre | Moyenne |
| CTI2 | 57 | 0.768 | 20 | 1.113 |
| CTI3 | 175 | 0.909 | 81 | 1.082 |
| CTI4 | 359 | 0.945 | 228 | 1.079 |

**Tableau 2**

Comparaison de l'estimateur GREG-HDP et de l'estimateur H-T pour les dépenses totales, coefficients de variation estimés

| Type de domaine | Gains engendrés par l'utilisation de GREG-HDP | | Pertes engendrées par l'utilisation de GREG-HDP | |
|---|---|---|---|---|
| | Nombre | Moyenne | Nombre | Moyenne |
| CTI2 | 57 | 0.773 | 20 | 1.100 |
| CTI3 | 175 | 0.910 | 81 | 1.082 |
| CTI4 | 355 | 0.945 | 232 | 1.079 |

Les résultats des tableaux 1 et 2 montrent que l'efficacité relative des estimateurs GREG-HDP et H-T est semblable pour l'une et l'autre variables étudiées. De même, les résultats des autres comparaisons faites dans le cadre de cette étude empirique varient peu selon la variable étudiée. Par conséquent, le reste des tableaux ne donnent que les résultats relatifs aux dépenses totales.

Les tableaux 3 et 4 servent à comparer l'estimateur GREG-HDP aux estimateurs GREG-R2 et GREG-R1R2 respectivement. D'après les coefficients de variation estimés, GREG-R2 est un peu plus efficace que GREG-HDP. Comme une forte proportion des unités de l'échantillon de seconde phase ont une probabilité de sélection égale à un et que les deux estimateurs comparés, GREG-R2 et GREG-HDP, utilisent les mêmes variables auxiliaires dans la pondération de première phase, il n'est pas étonnant de constater le peu de différence entre les deux estimateurs. Le CV estimé de l'estimation GREG-R1R2 est généralement moins élevé que celui de l'estimation GREG-HDP et l'efficacité relative de GREG-R1R2 s'accroît à mesure qu'augmente la taille du domaine. Néanmoins, GREG-R1R2 n'est supérieur à GREG-HDP que pour 64% des domaines CTI4, et l'augmentation moyenne du CV estimé pour les domaines où GREG-R1R2 est moins efficace que GREG-HDP est plus forte que la diminution moyenne du CV estimé pour les domaines où GREG-R1R2 est plus efficace.

**Tableau 3**

Comparaison de l'estimateur GREG-R2 et de l'estimateur GREG-HDP pour les dépenses totales, coefficients de variation estimés

| Type de domaine | Gains engendrés par l'utilisation de GREG-R2 | | Aucune différence | Pertes engendrées par l'utilisation de GREG-R2 | |
|---|---|---|---|---|---|
| | Nombre | Moyenne | Nombre | Nombre | Moyenne |
| CTI2 | 38 | 0.993 | 26 | 13 | 1.001 |
| CTI3 | 58 | 0.991 | 158 | 40 | 1.002 |
| CTI4 | 88 | 0.988 | 439 | 60 | 1.009 |

**Tableau 4**

Comparaison de l'estimateur GREG-R1R2 et de l'estimateur GREG-HDP pour les dépenses totales, coefficients de variation estimés

| Type de domaine | Gains engendrés par l'utilisation de GREG-R1R2 | | Pertes engendrées par l'utilisation de GREG-R1R2 | |
|---|---|---|---|---|
| | Nombre | Moyenne | Nombre | Moyenne |
| CTI2 | 51 | 0.867 | 26 | 1.170 |
| CTI3 | 160 | 0.934 | 96 | 1.093 |
| CTI4 | 377 | 0.954 | 210 | 1.074 |

Les résultats des tableaux 3 et 4 indiquent que nous serions mal avisés de remplacer l'estimateur GREG-HDP en usage actuellement par l'estimateur GREG-R1R2, malgré les perspectives intéressantes que celui-ci laisse entrevoir. Les améliorations que l'on observe avec GREG-R1R2 sont relativement négligeables, compte tenu de la forte corrélation entre le revenu du déclarant et les dépenses totales. Ces améliorations pourraient être plus notables 1) s'il y avait toujours une concordance entre les codes CTI utilisés pour la stratification de première et de seconde phase et les codes CTI qui servent à déterminer l'effectif de domaine pour des unités échantillonnées et 2) s'il n'était pas nécessaire de regrouper des strates d'échantillonnage pour faire en sorte que chaque strate a posteriori de première et de seconde phase contienne un nombre minimum d'unités.

**Tableau 5**

Comparaison de l'estimateur GREG-R1R2 et de l'estimateur GREG-HDP pour les dépenses totales, coefficients de variation estimés, aucune erreur de classement

| Type de domaine | Gains engendrés par l'utilisation de GREG-R1R2 | | Pertes engendrées par l'utilisation de GREG-R1R2 | |
|---|---|---|---|---|
| | Nombre | Moyenne | Nombre | Moyenne |
| CTI2 | 66 | 0.778 | 11 | 1.057 |
| CTI3 | 184 | 0.916 | 72 | 1.047 |
| CTI4 | 402 | 0.944 | 185 | 1.034 |

Les résultats du tableau 5 ont été obtenus après que l'on eut modifié, pour certaines unités échantillonnées, le code CTI attribué par Revenu Canada ainsi que le code CTI utilisé pour la stratification de l'échantillon de seconde phase, cela afin d'éliminer toute incohérence entre ces codes et ceux qui servent à déterminer l'effectif des domaines. Une comparaison des tableaux 4 et 5 montre que l'efficacité relative de GREG-R1R2 est beaucoup plus grande lorsqu'il n'y a pas d'erreur de classement. Avec l'estimateur GREG-R1R2, le CV estimé diminue de plus de 22% (en moyenne) pour plus de 85% des domaines CTI2.

Les résultats empiriques présentés dans cette section montrent invariablement que les gains d'efficacité réalisés grâce à l'utilisation d'information supplémentaire s'accroissent à mesure qu'augmente la taille du domaine. Cette constatation rejoint les observations faites dans la section 2 au sujet des conditions dans lesquelles la corrélation entre $y(d)$ et les vecteurs de variables auxiliaires, $x$ et $z$, sera élevée. Pourvu que la variable étudiée soit fortement corrélée avec les variables auxiliaires, les corrélations qui impliquent $y(d)$ seront le plus élevées si chaque strate a posteriori qui contient au moins une unité échantillonnée du domaine $d$ ne contient pas trop d'unités ne faisant pas partie du domaine $d$.

## 5. CONCLUSIONS

L'estimation par régression généralisée offre un cadre propice à l'utilisation d'information supplémentaire. Dans cet article, nous avons déterminé un estimateur par régression généralisé pour un plan d'échantillonnage à deux phases avec échantillonnage de Poisson à chaque phase. Nous avons étudié l'efficacité de cet estimateur en l'appliquant à l'échantillon à deux phases de dossiers fiscaux prélevé par Statistique Canada en vue d'établir des estimations annuelles de la production économique des petites entreprises. La méthode d'estimation actuellement en usage dans ce programme consiste notamment à appliquer des coefficients de correction aux estimateurs de strate formée a posteriori durant la pondération des échantillons de première et de seconde phase pour tenir compte de la différence entre la taille réelle et la taille prévue de l'échantillon. Cet estimateur de stratification a posteriori est un cas particulier de l'estimateur par régression généralisé.

Selon notre étude empirique, l'estimateur par régression généralisé en usage actuellement à Statistique Canada (GREG-HDP) est beaucoup plus efficace que l'estimateur de Horvitz-Thompson. Notre étude compare aussi GREG-HDP à deux autres estimateurs par régression généralisés. L'utilisation de ces deux estimateurs entraîne des gains d'efficacité pour les grands domaines. Toutefois, en ce qui a trait aux petits domaines, qui intéressent particulièrement les utilisateurs d'estimations basées sur l'échantillon à deux phases de dossiers fiscaux, le rendement de ces estimateurs ne justifie pas qu'on les substitue entièrement à l'estimateur actuellement en usage.

## REMERCIEMENTS

Les auteurs tiennent à remercier René Boyer, qui leur a fourni une version de la macro SAS CALMAR adaptée à leur étude empirique, ainsi que K.P. Srinath et Michael Hidiroglou, qui ont participé à des échanges fructueux avec eux. Ils remercient également Michael Bankier et Jean Leduc pour leurs utiles commentaires sur une version antérieure de cet article.


## ANNEXE A: DÉFINITION DE LA VARIANCE DE $\hat{Y}_{\mathrm{GREG}}(d)$ ET DE L'ESTIMATEUR DE LA VARIANCE

On peut déterminer la variance de $\hat{Y}_{\mathrm{GREG}}(d)$ au moyen de l'identité

$$V(\hat{Y}_{\mathrm{GREG}}(d)) = E_1 V_2(\hat{Y}_{\mathrm{GREG}}(d)) + V_1 E_2(\hat{Y}_{\mathrm{GREG}}(d)).$$

Premièrement, considérons la variance de l'estimateur par rapport à la seconde phase d'échantillonnage, étant donné les résultats du calage de première phase. Si le vecteur de variables auxiliaires pour la pondération de seconde phase, $z$, contient une variable qui prend la valeur un pour tous les déclarants (ou si l'on peut construire une combinaison linéaire de variables auxiliaires qui soit égale à un pour tous les déclarants), l'estimateur par régression généralisé peut être défini par l'expression

$$\hat{Y}_{\mathrm{GREG}}(d) = \sum_{i \in s2} w_{1i} w_{2i} y_i(d)$$

$$= \sum_{v} \sum_{i \in s2 \cap v} w_{1i}(y_i(d) - z_i' \hat{B}_v)/p_{2i} + \sum_{v} \tilde{Z}_v \hat{B}_v .$$

Si l'on fait abstraction de la variabilité due à l'estimation des coefficients de régression dans la pondération de seconde phase, on a

$$E_1 V_2(\hat{Y}_{\mathrm{GREG}}) \approx E_1 V_2\left( \sum_{i \in s2} w_{1i} Q_{2i}/p_{2i} \right)$$

$$= E_1\left( \sum_{i \in s1} \frac{(1 - p_{2i})}{p_{2i}} w_{1i}^2 Q_{2i}^2 \right).$$

D'après l'estimateur de la variance des estimateurs de calage proposés par Deville et Särndal (1992, p. 380), on peut formuler l'estimateur de $E_1 V_2(\hat{Y}_{\mathrm{GREG}}(d))$ de la façon suivante:

$$\hat{S}_1 = \sum_{i \in s2} \frac{(1 - p_{2i})}{(p_{1i} p_{2i})^2} (g_{1i} g_{2i} q_{2i})^2.$$

Si l'on fait abstraction de la variabilité due à l'estimation des coefficients de régression dans la pondération de première phase, le deuxième terme de l'expression de la variance peut s'écrire

$$V_1 E_2(\hat{Y}_{\mathrm{GREG}}(d)) = V_1\left( \sum_{i \in s1} w_{1i} y_i(d) \right)$$

$$= \sum_{i} \frac{(1 - p_{1i})}{p_{1i}} Q_{1i}^2,$$

et son estimateur est

$$\hat{S}_2 = \sum_{i \in s2} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2.$$

## BIBLIOGRAPHIE


ARMSTRONG, J., BLOCK, C., et SRINATH, K.P. (1993). Two-phase sampling of tax records for business surveys. *Journal of Business and Economic Statistics*, 11, 407-416.

BANKIER, M., RATHWELL, S., et MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census of Population. *Statistics Sweden, Workshop on the Uses of Auxiliary Information in Surveys.*

BREWER, K.R.W., EARLY, L.J., et JOYCE, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.

CHOUDHRY, G.H., LAVALLÉE, P., et HIDIROGLOU, M. (1989). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.

DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 34, 911-934.

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

HIDIROGLOU, M.A., SÄRNDAL, C.-E., et BINDER, D.A. (1993). Pondération et estimation dans les enquêtes-entreprises. Présenté en décembre 1993, à l'Institut national de la statistique et des études économiques, Paris. Journées de méthodologie statistique.

LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

RAO, J.N.K. (1968a). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

RAO, J.N.K. (1968b). Some small sample results in ratio and regression estimation. *Journal of the Indian Statistical Association*, 6, 160-168.

ROYALL, R.M., et EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, Série C, 37, 43-52.

SÄRNDAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SAUTORY, O. (1991). La macro SAS: CALMAR. Document non publié, Institut national de la statistique et des études économiques, Paris.

STATISTIQUE CANADA (1980). *Classification type des industries*. N° 12-501F au catalogue, Statistique Canada.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: A useful technique. *Journal of Official Statistics*, 2, 161-168.

WU, C.F.J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183-189.

WU, C.F.J., et DENG, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. Dans Box, G.E.P. et coll. (Éds.), *Scientific Inference, Data Analysis and Robustness*, New York: Academic Press, 245-277.