No 11-522-XIF au catalogue

**La série des symposiums internationaux de Statistique Canada - Recueil**


# Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques reliés aux besoins futurs d'information

# ESTIMATEURS DE VARIANCE PAR LINÉARISATION POUR LES PARAMÈTRES D'UN MODÈLE À PARTIR DE DONNÉES D'ENQUÊTE COMPLEXES

A. Demnati et J. N. K. Rao[1]

## RÉSUMÉ

En échantillonnage, on utilise souvent la linéarisation de Taylor pour obtenir des estimateurs de variance pour des estimateurs par calage de totaux et de paramètres non linéaires de population finie (ou recensement), comme des ratios, ou des coefficients de régression et de corrélation, qui peuvent être exprimés sous forme de fonctions lisses de totaux. La linéarisation de Taylor est généralement applicable à tout plan d'échantillonnage, mais elle peut produire de multiples estimateurs de variance qui sont asymptotiquement sans biais par rapport au plan de sondage sous un échantillonnage répété. Pour choisir lequel de ces estimateurs utiliser, il faut tenir compte d'autres critères, comme i) l'absence approximative de biais pour la variance par rapport au modèle de l'estimateur sous un modèle hypothétique, et ii) la validité sous l'échantillonnage répété conditionnel. Demnati et Rao (2004) ont proposé une nouvelle approche pour calculer les estimateurs de variance par linéarisation de Taylor, qui mène directement à un estimateur de variance unique qui satisfait aux critères susmentionnés pour des plans de sondage généraux. Ensuite, Demnati et Rao (2002) ont envisagé le cas de réponses manquantes lorsqu'on utilise un ajustement pour la non-réponse complète et l'imputation pour la non-réponse partielle fondées sur des fonctions lisses de valeurs observées, notamment l'imputation par le quotient. Lorsqu'on analyse des données d'enquête, on suppose souvent que les populations finies sont générées à partir de modèles de superpopulation et on s'intéresse aux inférences analytiques basées sur les paramètres du modèle. Si les fractions d'échantillonnage sont négligeables, la variance d'échantillonnage capte presque entièrement la variation générée par les deux processus aléatoires fondés sur le plan de sondage et sur le modèle. Toutefois, lorsque les fractions d'échantillonnage ne sont pas négligeables, on doit tenir compte de la variance par rapport au modèle afin de construire des inférences valides pour les paramètres du modèle selon les deux processus de randomisation. Dans le présent exposé, nous nous penchons sur l'estimation de la variance totale selon la méthode de Demnati-Rao en supposant que les caractéristiques d'intérêt sont des variables aléatoires générées à partir d'un modèle de superpopulation. Nous illustrons la méthode en utilisant des estimateurs par le quotient et des estimateurs définis comme solutions à des équations d'estimation pondérées par calage. Nous appliquons également la méthode à un modèle de Poisson avec excès de zéros.

MOTS CLÉS : Calage; équations d'estimation pondérées; estimateurs par le quotient; variance totale; modèle de Poisson avec excès de zéros.

## 1. INTRODUCTION

La linéarisation de Taylor est une méthode très répandue d'estimation de la variance pour des statistiques complexes, comme les estimateurs par le quotient et les estimateurs par régression, ainsi que les estimateurs des coefficients de régression logistique. Elle s'applique généralement à tout plan d'échantillonnage qui permet une estimation sans biais de la variance des estimateurs linéaires, contrairement à une méthode de rééchantillonnage comme celle du jackknife, et requiert des calculs plus simples que cette dernière méthode. Cependant, elle peut produire des estimateurs multiples de la variance qui sont asymptotiquement sans biais par rapport au plan de sondage sous échantillonnage répété. Par conséquent, pour déterminer lequel de ces estimateurs il convient d'utiliser, il faut tenir compte d'autres critères comme i) l'absence approximative de biais pour la variance de l'estimateur par rapport au modèle sous un modèle hypothétique et ii) la validité sous l'échantillonnage conditionnel répété. Par exemple, dans le contexte de l'échantillonnage aléatoire simple et de l'estimateur par le quotient, $\hat{Y}_R = (\bar{y}/\bar{x})X$ , du total de population $Y$ , Royall et Cumberland (1981) montrent qu'un estimateur de la variance par linéarisation utilisé couramment, $\vartheta_L = N^2(n^{-1} - N^{-1})s_z^2$ , ne capte pas la variance conditionnelle de $\hat{Y}_R$ sachant $\bar{x}$ ,

[1] *A. Demnati, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6 (Abdellatif.Demnati@statcan.ca); J. N. K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6 (JRao@math.carleton.ca).*

contrairement à l'estimateur de la variance par le jackknife $\vartheta_J$. Ici, $\bar{y}$ et $\bar{x}$ sont les moyennes d'échantillon, $X$ est le total connu de population d'une variable auxiliaire $x$, $s_z^2$ est la variance d'échantillon des résidus $z_k = y_k - (\bar{y}/\bar{x})x_k$ et $(n, N)$ représente les tailles d'échantillon et de population. Par linéarisation de l'estimateur jackknife de la variance, $\vartheta_J$, on obtient un estimateur de la variance par linéarisation différent, $\vartheta_{JL} = (\overline{X}/\bar{x})^2 \vartheta_L$ ; ce dernier capte la variance conditionnelle, ainsi que la variance inconditionnelle, où $\overline{X} = X/N$ est la moyenne de $x$. Par conséquent, on pourrait préférer utiliser $\vartheta_{JL}$ ou $\vartheta_J$ plutôt que $\vartheta_L$. Valliant (1993) obtient $\vartheta_{JL}$ pour l'estimateur poststratifié et réalise une étude en simulation pour démontrer que $\vartheta_J$ et $\vartheta_{JL}$ possèdent tous deux de bonnes propriétés conditionnelles étant donné les estimations des comptes des strates a posteriori. Särndal, Swensson et Wretman (1989) montrent que $\vartheta_{JL}$ est à la fois asymptotiquement sans biais par rapport au plan de sondage et approximativement sans biais par rapport au modèle au sens de $E_m(\vartheta_{JL}) = V_m(\hat{Y}_R)$, où $E_m$ représente l'espérance fondée sur le modèle et $V_m(\hat{Y}_R)$ représente la variance fondée sur le modèle de $\hat{Y}_R$ sous un « modèle du quotient » : $E_m(y_k) = \beta x_k$ ; $k = 1,..., N$ et les $y_k$ sont indépendants avec une variance fondée sur le modèle $V_m(y_k) = \sigma^2 x_k$, $\sigma^2 > 0$. Donc, $\vartheta_{JL}$ est un bon choix aussi bien du point de vue des propriétés par rapport au plan de sondage que par rapport au modèle. Demnati et Rao (2004) ont proposé une nouvelle méthode d'estimation de la variance qui est justifiable théoriquement et qui, par la même occasion, mène directement à un estimateur de la variance de type $\vartheta_{JL}$ pour des plans de sondage généraux. Ils ont appliqué la méthode selon l'approche fondée sur le plan de sondage à divers problèmes, dont les estimateurs par calage d'un total $Y$ et d'autres estimateurs définis explicitement ou implicitement comme étant des solutions d'équations d'estimation. Ils ont obtenu un nouvel estimateur de la variance pour une classe générale d'estimateurs par calage qui inclut l'estimateur par la méthode itérative du quotient (raking ratio) généralisée et des estimateurs par régression généralisée. Ils ont aussi étendu la méthode à l'échantillonnage à deux phases pour obtenir un estimateur de la variance qui utilise plus complètement les données de l'échantillon de première phase que les estimateurs de la variance par linéarisation classiques. Demnati et Rao (2002) ont étendu leur méthode au cas des réponses manquantes lorsqu'on utilise des ajustements pour la non-réponse complète et l'imputation pour la non-réponse partielle fondée sur des fonctions lisses de valeurs observées, notamment l'imputation par le quotient.

Lorsqu'on analyse des données d'enquête, on suppose souvent que les valeurs de la population finie $\boldsymbol{y} = (y_1,..., y_N)^T$ sont générées à partir d'un modèle de superpopulation et on s'intéresse aux inférences analytiques sur les paramètres du modèle. Si les fractions d'échantillonnage sont négligeables, la variance d'échantillonnage capte presque entièrement la variation générée par les processus aléatoires fondés sur le plan de sondage et sur le modèle. Toutefois, lorsque les fractions d'échantillonnage ne sont pas négligeables, la variance par rapport au modèle ne l'est pas non plus par rapport à la variance totale. Dans ce cas, on doit aussi tenir compte de la variance par rapport au modèle afin de construire des inférences valides pour les paramètres du modèle selon les deux processus aléatoires.

Molina, Smith et Sugden (2001) ont obtenu des expressions générales de la moyenne et de la covariance des données-échantillons $diag(\boldsymbol{a}(s))\boldsymbol{y}$ ainsi que des totaux d'échantillon selon les processus conjoints, où $\boldsymbol{a}(s) = (a_1(s),..., a_N(s))^T$, $a_k(s) = 1$ si l'élément $k$ appartient à l'échantillon $s$ et $a_k(s) = 0$ dans le cas contraire. Il ne fait aucun doute que le processus combiné de sélection de l'échantillon et de génération de la population finie doit être à la base des inférences analytiques, comme le soutiennent Molina, Smith et Sugden (2001). Toutefois, une méthode d'application générale est nécessaire pour l'estimation de la variance totale. Dans la section 2, nous étendons la méthode de Demnati-Rao à l'estimation de la variance totale en supposant que les caractéristiques d'intérêt sont des variables aléatoires générées à partir d'un modèle de superpopulation. La méthode est théoriquement justifiable, tout en menant directement à un estimateur unique de la variance totale possédant des propriétés souhaitables. Toujours dans la section 2, nous appliquons la méthode à des estimateurs par le quotient des paramètres du modèle. Dans la section 3, nous étendons la méthode à des estimateurs définis comme des solutions à des équations d'estimation pondérées, en utilisant des poids de calage par la régression généralisée (GREG).

## 2. ESTIMATEUR PAR LE QUOTIENT LORSQUE $y_k$ EST ALÉATOIRE

Supposons que $\hat{\theta}$ est l'estimateur par le quotient $\hat{\theta} = X(\sum y_k d_k(s))/(\sum x_k d_k(s)) \equiv X\hat{R}$ et que le paramètre du modèle est $\theta = E_m(Y) = \sum E_m(y_k)$, où la somme est sur tous les $N$ éléments de la population, $d_k(s) = 0$ si l'élément $k$ n'appartient pas à l'échantillon $s$ et $X$ est le total connu de $x$. Soit $\boldsymbol{d}_k = (d_{1k}, d_{2k})^T$, où $d_{1k} = d_k(s)$, $d_{2k} = d_k(s)y_k$, et $s$ est supprimé dans $\boldsymbol{d}_k(s)$ pour simplifier. Nous pouvons alors écrire $\hat{\theta}$ sous la forme $\hat{\theta} = f(\boldsymbol{A}_d) = X(\sum d_{2k})/(\sum x_k d_{1k})$, où $\boldsymbol{A}_d$ est une matrice de dimension $2 \times N$ dont la $k^e$ colonne est $\boldsymbol{d}_k$. Nous utilisons les poids de Horvitz-Thompson (HT) $d_k(s) = a_k(s)/\pi_k$, où $a_k(s)$ est la variable indicatrice d'appartenance à l'échantillon et $\pi_k$ est la probabilité d'inclusion. Dans ce cas, si $E = E_m E_p$ représente l'espérance totale, nous avons $E(d_{1k}) = E_m E_p(d_k(s)) = E_m(1) = 1 \equiv \mu_{1k}$ et $E(d_{2k}) = E_m(y_k E_p(d_k(s))) = E_m(y_k) \equiv \mu_{2k}$, où $E_p$ représente l'espérance à l'égard du plan. Donc, $E\hat{\theta} \approx f(\boldsymbol{A}_\mu) = \theta$, où $\boldsymbol{A}_\mu$ est une matrice de dimension $2 \times N$ dont la $k^e$ colonne est $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k})^T$.

Demnati et Rao (2004) ont montré qu'on pouvait formuler l'expansion de Taylor de $\hat{\theta} - \theta$ sous la forme

$$\hat{\theta} - \theta \approx \sum \tilde{\boldsymbol{z}}_k^T (\boldsymbol{d}_k - \boldsymbol{\mu}_k), \tag{2.1}$$

où $\tilde{\boldsymbol{z}}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_\mu}$ et $\boldsymbol{A}_b$ est une matrice de dimension $2 \times N$ de nombres réels arbitraires dont la $k^e$ colonne est $\boldsymbol{b}_k$. Ce résultat se vérifie pour n'importe quelle estimateur $\hat{\theta}$, qu'on peut exprimer sous forme de fonction lisse d'estimateurs de totaux. En utilisant la notation de l'opérateur, soit $\mathcal{G}(\boldsymbol{u})$ qui représente l'estimateur de la variance totale d'un estimateur linéaire $\hat{U} = \sum \boldsymbol{u}_k^T \boldsymbol{d}_k$. L'estimateur de Demnati-Rao (DR) de la variance par linéarisation de $\hat{\theta}$ est alors donné simplement par

$$\mathcal{G}_{DR}(\hat{\theta}) = \mathcal{G}(\boldsymbol{z}), \tag{2.2}$$

qu'on obtient à partir de $\mathcal{G}(\boldsymbol{u})$ en remplaçant $\boldsymbol{u}_k$ par $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$. Notons que $\boldsymbol{z}_k$ est un estimateur cohérent de $\tilde{\boldsymbol{z}}_k$. Pour l'estimateur par le quotient $\hat{\theta}$, nous avons

$$\boldsymbol{z}_k = (z_{1k}, z_{2k})^T = (X/\hat{X})(-\hat{R}\, x_k, 1)^T. \tag{2.3}$$

Il reste à évaluer $\mathcal{G}(\boldsymbol{u})$. Nous avons

$$\mathcal{G}(\boldsymbol{u}) = \sum\sum \boldsymbol{u}_k^T \operatorname{cov}(\boldsymbol{d}_k, \boldsymbol{d}_t)\boldsymbol{u}_t, \tag{2.4}$$

où

$$\operatorname{cov}(\boldsymbol{d}_k, \boldsymbol{d}_t) = d_{kt}(s)\begin{bmatrix} 0 & 0 \\ 0 & \operatorname{cov}_m(y_k, y_t) \end{bmatrix} + d_{kt}(s)\frac{(1-\omega_{kt})}{\omega_{kt}} \boldsymbol{v}_k \boldsymbol{v}_t^T. \tag{2.5}$$

Dans l'équation (2.5), $\boldsymbol{v}_k = (1, y_k)^T$,

$d_{kt}(s) = a_k(s)a_t(s)/\pi_{kt}$, $d_{kk}(s) = d_k(s)$,

$\omega_{kt} = \pi_k \pi_t / \pi_{kt}$, $(1-\omega_{kt})/\omega_{kt} = (\pi_{kt} - \pi_k \pi_t)/(\pi_k \pi_t)$,

et $\operatorname{cov}_m(y_k, y_t)$ est un estimateur de la covariance de $y_k$ et $y_t$ selon le modèle, où $\pi_{kt}$ est la probabilité d'inclusion conjointe pour $k \neq t$, et $\pi_{kk} = \pi_k$. Lorsque la covariance fondée sur le modèle de $y_k$ et $y_t$ est nulle alors $\operatorname{cov}_m(y_k, y_t)$ est mise à zéro.

En remplaçant $\boldsymbol{z}_k$ dans (2.3) par $\boldsymbol{u}_k$ dans (2.4), on obtient

$$\begin{aligned}\vartheta_{DR}(\hat{\theta}) &= \sum\sum d_{kt}(s) z_{k;m} z_{t;m} \operatorname{cov}_m(y_k, y_t) + \sum\sum d_{kt}(s) z_{k;s} z_{t;s} (1-\omega_{kt})/\omega_{kt} \\ &\equiv \vartheta_m + \vartheta_s\end{aligned} . \tag{2.6}$$

où $z_{k;m} = z_{2k} = X/\hat{X}$ et $z_{k;s} = \boldsymbol{z}_k^T \boldsymbol{v}_k = z_{1k} + z_{2k} y_k = (X/\hat{X})(y_k - \hat{R} x_k)$. Notons que le premier terme, $\vartheta_m$, du côté droit de l'équation (2.6), correspond au modèle et que le deuxième terme, $\vartheta_s$, correspond au plan d'échantillonnage.

Selon un échantillon aléatoire simple,

$$\vartheta_s = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{X}{\hat{X}}\right)^2 s_e^2 , \tag{2.7}$$

où $s_e^2 = \sum a_k(s)(y_k - \hat{R} x_k)^2/(n-1)$ et $\hat{R} = \bar{y}/\bar{x}$. En outre, selon le modèle du quotient

$$E_m(y_k) = \beta x_k \ , \ Cov_m(y_k, y_t) = 0 \ , \ k \neq t \ , \tag{2.8}$$

$\theta = \beta X$, la variance par rapport au modèle de $y_k$, $V_m(y_k) = E_m(y_k - \beta x_k)^2$ est estimée de manière robuste par $(y_k - \hat{R} x_k)^2$, et

$$\vartheta_m = \frac{N}{n}\left(\frac{X}{\hat{X}}\right)^2 (n-1) s_e^2 . \tag{2.9}$$

Notons que $\vartheta_m$ reste valide malgré une erreur de spécification de la variance par rapport au modèle de $y_k$.
En combinant maintenant (2.7) avec (2.9), on obtient un estimateur de la variance totale de $\hat{\theta}$ de la forme

$$\vartheta_{DR}(\hat{\theta}) = \frac{N^2}{n}\left(\frac{X}{\hat{X}}\right)^2 \frac{N-1}{N} s_e^2 . \tag{2.10}$$

Il est intéressant de noter que le « poids g » $X/\hat{X}$ figure automatiquement dans $\vartheta_{DR}(\hat{\theta})$ et que la correction de la population finie $1-n/N$ est absente de $\vartheta_{DR}(\hat{\theta})$. La méthode de Demnati-Rao mène à un choix unique d'estimateur de la variance qui préserve automatiquement les facteurs g.

Selon la méthode traditionnelle d'estimation de la variance totale, $V(\hat{\theta})$ est écrit sous la forme $E_m V_p(\hat{\theta}) + V_m E_p(\hat{\theta}) \approx E_m V_p(\hat{\theta}) + V_m(Y) = E_m V_p(\hat{\theta}) + \sum E_m(y_k - \beta x_k)^2$ selon le modèle du quotient. Le premier terme est estimé par un estimateur de $V_p(\hat{\theta})$, qui utilise habituellement $\vartheta_s$ sans le facteur g. Le deuxième terme est estimé de manière robuste par $\sum d_k(s)(y_k - \hat{R} x_k)^2 = (N/n)(n-1)s_e^2$. La somme des deux termes estimés égale (2.10) sans le facteur g. Le choix de l'estimateur de la variance selon la méthode traditionnelle n'est pas unique.

Si le paramètre d'intérêt est $\beta = \theta/X$, alors $\hat{\beta} = \hat{\theta}/X = \hat{R}$ et

$$\vartheta_{DR}(\hat{\beta}) = \vartheta_{DR}(\hat{\theta}/X) = \frac{N^2}{n}\frac{1}{\hat{X}^2}\frac{N-1}{N} s_e^2 , \tag{2.11}$$

selon le modèle du quotient. La méthode traditionnelle utilise habituellement le même estimateur de la variance de $\hat{\beta}$.

## 3. ÉQUATIONS D'ESTIMATION PONDÉRÉES

Supposons que le modèle de superpopulation des réponses $y_k$ est spécifié par un modèle linéaire généralisé dont la moyenne est $E_m(y_k) = \mu_k(\boldsymbol{\theta}) = h(\boldsymbol{u}_k^T \boldsymbol{\theta})$, où $\boldsymbol{u}_k$ est un vecteur de dimension $p \times 1$ de variables explicatives et $h(.)$ est une fonction « lien ». Le paramètre de modèle d'intérêt est $\boldsymbol{\theta}$. Par exemple, le choix $h(a) = a$ donne un modèle

de régression linéaire et $h(a) = e^a/(1+e^a)$ donne un modèle de régression logistique pour les réponses binaires $y_k$.

Nous définissons des équations d'estimation de recensement de forme $\boldsymbol{l}(\boldsymbol{\theta}) = \sum \boldsymbol{l}_k(\boldsymbol{\theta}) = \mathbf{0}$ avec $E_m \boldsymbol{l}_k(\boldsymbol{\theta}) = \mathbf{0}$, dont la solution donne le paramètre de recensement $\hat{\boldsymbol{\theta}}_N$. Par exemple, $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{u}_k(y_k - \boldsymbol{\mu}_k(\boldsymbol{\theta}))$ pour les modèles de régression linéaire et de régression logistique. Nous utilisons des poids de calage par la régression généralisée (GREG) $w_k(s) = d_k(s) g_k(\boldsymbol{d}(s))$, où les « poids g » sont donnés par

$$g_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T [\sum d_k(s) c_k \boldsymbol{x}_k \boldsymbol{x}_k^T]^{-1} c_k \boldsymbol{x}_k , \tag{3.1}$$

pour des $c_k$ spécifiés, $\hat{\boldsymbol{X}} = \sum d_k(s) \boldsymbol{x}_k$ est l'estimateur HT du total connu $\boldsymbol{X}$ d'un vecteur de dimension $q \times 1$ de variables de calage $\boldsymbol{x}_k$ et $\boldsymbol{d}(s)$ est le vecteur de dimension $N \times 1$ de poids HT $d_k(s)$. L'estimateur GREG résultant du total $Y$, soit $\hat{Y} = \sum w_k(s) y_k$, possède la propriété de calage $\sum w_k(s) \boldsymbol{x}_k = \boldsymbol{X}$ (Särndal *et coll.*, 1989).

Nous utilisons les poids de calage pour estimer l'équation d'estimation de recensement. Les équations d'estimation pondérées par calage sont données par

$$\hat{\boldsymbol{l}}(\boldsymbol{\theta}) = \sum d_k(s) g_k(\boldsymbol{d}(s)) \boldsymbol{l}_k(\boldsymbol{\theta}) = \mathbf{0} . \tag{3.2}$$

La solution à l'équation (3.2) donne l'estimateur pondéré de calage $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$, et $\hat{\boldsymbol{\theta}}$ est approximativement sans biais par rapport au plan et au modèle pour $\boldsymbol{\theta}$, c.-à-d. $E_m E_p(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. Il découle de (3.2) que $\hat{\boldsymbol{\theta}}$ est de la forme $\boldsymbol{f}(\boldsymbol{A}_d)$ avec $\boldsymbol{d}_k = (d_k(s), d_k(s) \boldsymbol{l}_k^T(\boldsymbol{\theta}))^T$, où $\boldsymbol{f}(\boldsymbol{A}_d)$ est un vecteur de dimension $p \times 1$ et $\boldsymbol{A}_d$ est une matrice de dimension $(p+1) \times N$ dont la $k^e$ colonne est $\boldsymbol{d}_k$.

En suivant la méthode de différenciation implicite de Demnati et Rao (2004), on évalue $\boldsymbol{Z}_k = \partial \boldsymbol{f}(\boldsymbol{A}_b) / \partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$ sous la forme

$$\boldsymbol{Z}_k^T = [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1} g_k(\boldsymbol{d}(s)) (-\hat{\boldsymbol{B}}_l^T \boldsymbol{x}_k, \boldsymbol{I}_p) , \tag{3.3}$$

avec

$$\hat{\boldsymbol{B}}_l = [\sum d_k(s) c_k \boldsymbol{x}_k \boldsymbol{x}_k^T]^{-1} \sum d_k(s) c_k \boldsymbol{x}_k \boldsymbol{l}_k^T(\hat{\boldsymbol{\theta}}) , \tag{3.4}$$

$$\hat{\boldsymbol{J}}(\boldsymbol{\theta}) = \sum d_k(s) g_k(\boldsymbol{d}(s)) (\partial \boldsymbol{l}_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T) , \tag{3.5}$$

et $\boldsymbol{I}_p$ est la matrice d'identité. On obtient l'estimateur DR de la variance par linéarisation de $\hat{\boldsymbol{\theta}}$ à partir de (2.4) et de (2.5) en remplaçant $\boldsymbol{u}_k^T$ par la matrice de dimension $p \times (p+1)$ $\boldsymbol{Z}_k^T$, $\boldsymbol{v}_k$ par $(1, \boldsymbol{l}_k^T(\boldsymbol{\theta}))^T$ et $\mathrm{cov}_m(y_k, y_t)$ par un estimateur de la matrice de covariance de dimension $p \times p$ de $\boldsymbol{l}_k(\boldsymbol{\theta})$ selon le modèle. Après simplification, on obtient

$$\vartheta_{DR}(\hat{\boldsymbol{\theta}}) = \vartheta_m + \vartheta_s , \tag{3.6}$$

où $\vartheta_s$ est l'estimateur de la matrice de covariance d'échantillonnage donnée par

$$\vartheta_s = [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1} \sum\sum d_{kt}(s) g_k(\boldsymbol{d}(s)) g_t(\boldsymbol{d}(s)) (1 - \omega_{kt}) / \omega_{kt} \boldsymbol{e}_k^*(\hat{\boldsymbol{\theta}}) \boldsymbol{e}_t^*(\hat{\boldsymbol{\theta}}) [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1\mathrm{T}} , \tag{3.7}$$

avec

$$\boldsymbol{e}_k^*(\hat{\boldsymbol{\theta}}) = \boldsymbol{l}_k(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{B}}_l^T \boldsymbol{x}_k . \tag{3.8}$$

L'estimateur de la matrice de covariance fondée sur le modèle dépend de la structure de covariance du modèle hypothétique. Si $Cov_m(\boldsymbol{l}_k(\boldsymbol{\theta}), \boldsymbol{l}_t^T(\boldsymbol{\theta})) = \mathbf{0}$ pour $k \neq t$, et $\boldsymbol{V}_m(\boldsymbol{l}_k(\boldsymbol{\theta})) = E_m(\boldsymbol{l}_k(\boldsymbol{\theta}) \boldsymbol{l}_k^T(\boldsymbol{\theta}))$ est estimé de manière robuste par $\boldsymbol{l}_k(\hat{\boldsymbol{\theta}}) \boldsymbol{l}_k^T(\hat{\boldsymbol{\theta}})$, l'estimateur de la matrice de covariance fondée sur le modèle, $\vartheta_m$, se réduit à

$$\vartheta_m = [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1} \sum d_k(s) g_k^2(\boldsymbol{d}(s)) \boldsymbol{l}_k(\hat{\boldsymbol{\theta}}) \boldsymbol{l}_k^T(\hat{\boldsymbol{\theta}}) [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1\mathrm{T}} . \tag{3.9}$$

Notons que pour les modèles de régression linéaire et de régression logistique, $Cov_m(\boldsymbol{l}_k(\boldsymbol{\theta}), \boldsymbol{l}_t^T(\boldsymbol{\theta})) = \mathbf{0}$ pour $k \neq t$ si

les $y_k$ sont non corrélés selon le modèle, et que $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{u}_k(y_k - \boldsymbol{\mu}_k(\boldsymbol{\theta}))$.

***Exemple*** : Modèle de régression de Poisson avec excès de zéros

Nous présentons maintenant les résultats d'une étude de simulation sur le biais du nouvel estimateur de la variance $\vartheta_{DR}(\hat{\boldsymbol{\theta}})$ donné par l'équation (3.6) à partir d'échantillons de taille finie. Nous considérons un modèle de régression de Poisson avec excès de zéros qui est souvent utilisé pour des données de comptes contenant trop de zéros. Le modèle suppose qu'avec la probabilité $1-p_k$, la valeur du $k^e$ élément, $y_k$, est toujours nulle et qu'avec la probabilité $p_k$, elle est de $a_k\ (\geq 0)$ d'après une distribution de Poisson ($\lambda_k$) (Lambert, 1992). Nous supposons que $p_k = p = e^{\alpha}/(1+e^{\alpha}) \approx 0.62$ et $\lambda_k = \lambda = e^{\beta} \approx 2.7$ avec $\alpha = 0.5$ et $\beta = 1$, de sorte que les paramètres du modèle sont $\alpha$ et $\beta$. À partir du modèle de Poisson avec excès de zéros susmentionné, nous avons généré 300 populations, chacune de taille $N = 1\,000$, et de chacune d'elles nous avons choisi 300 échantillons selon la méthode d'échantillonnage de Bernoulli avec probabilité $\pi = 0{,}1$. Pour procéder au calage, nous avons généré des constantes $x_k$ $(k = 1,...,N)$ d'après l'échantillonnage de Bernoulli (0,6) et les avons fixées tout au long de la simulation. À l'aide de ces $x_k$ et des poids déterminés par le plan d'échantillonnage $d_k(s) = a_k(s)/\pi$, nous avons calculé les poids GREG avec $c_k = 1$.

La fonction d'estimation $\boldsymbol{l}_k(\boldsymbol{\theta})$ selon le modèle ci-dessus comprend deux composantes, où $\boldsymbol{\theta} = (\alpha, \beta)^T$ :

$$l_{1k}(\boldsymbol{\theta}) = \frac{I_k(1-p_k) + p_k f(y_k)}{1-p_k + p_k f(y_k)} \partial \log f(y_k)/\partial\beta \tag{3.10}$$

$$l_{2k}(\boldsymbol{\theta}) = \frac{I_k - p_k + p_k f(y_k)}{1-p_k + p_k f(y_k)} \partial \log p_k/\partial\alpha, \tag{3.11}$$

où $I_k = 0$ si $y_k = 0$ et $I_k = 1$ si $y_k > 0$, et $f(y_k)$ est la densité de Poisson ($\lambda_k$). En utilisant (3.10) et (3.11) dans (3.2), nous avons calculé $\hat{\boldsymbol{\theta}}$ pour chaque échantillon simulé. En utilisant $\hat{\boldsymbol{\theta}}$, nous avons ensuite calculé l'estimation de la variance totale $\vartheta_{DR}(\hat{\boldsymbol{\theta}})$ et ses composantes $\vartheta_m$ et $\vartheta_s$ pour chaque échantillon. À l'aide de ces valeurs, nous avons évalué la matrice de covariance simulée de $\hat{\boldsymbol{\theta}}$ et les valeurs moyennes de $\vartheta_m$, $\vartheta_s$ et $\vartheta_{DR}(\hat{\boldsymbol{\theta}})$ et $\hat{\boldsymbol{\theta}}$ pour les simulations. Le biais de l'estimation de $\boldsymbol{\theta}$ est ici négligeable : $\alpha = 0{,}5$ contre $\overline{\hat{\alpha}} = 0{,}5169$ et $\beta = 1$ contre $\overline{\hat{\beta}} = 0{,}9936$, où $\overline{\hat{\alpha}}$ et $\overline{\hat{\beta}}$ représentent les valeurs moyennes de $\hat{\alpha}$ et $\hat{\beta}$. En comparant les valeurs moyennes $\overline{\vartheta}_m$, $\overline{\vartheta}_s$ et $\overline{\vartheta}_{DR}$ aux valeurs simulées $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$, nous obtenons les résultats suivants :

| | | | |
|---|---|---|---|
| $V(\hat{\alpha}) = 0{,}561$, | $\overline{\vartheta}_{DR}(\hat{\alpha}) = 0{,}583$, | $\overline{\vartheta}_s(\hat{\alpha}) = 0{,}525$, | $\overline{\vartheta}_m(\hat{\alpha}) = 0{,}0058$, |
| $V(\hat{\beta}) = 0{,}0076$, | $\overline{\vartheta}_{DR}(\hat{\beta}) = 0{,}0076$, | $\overline{\vartheta}_s(\hat{\beta}) = 0{,}0069$, | $\overline{\vartheta}_m(\hat{\beta}) = 0{,}0007$, |
| $Cov(\hat{\alpha}, \hat{\beta}) = -0{,}0041$, | $\overline{\vartheta}_{DR}(\hat{\alpha}, \hat{\beta}) = -0{,}0042$, | $\overline{\vartheta}_s(\hat{\alpha}, \hat{\beta}) = -0{,}0038$, | $\overline{\vartheta}_m(\hat{\alpha}, \hat{\beta}) = 0{,}0004$, |

Il est clair que les estimateurs DR de la variance et de la covariance captent les valeurs correspondantes de la population, alors que l'utilisation de $\vartheta_s$ ne mène qu'à une sous-estimation de $V(\hat{\alpha})$ et $V(\hat{\beta})$.

## CONCLUSION

Pour les estimateurs de paramètres de modèle définis comme solutions aux équations d'estimation GREG pondérées par calage, nous avons étudié l'estimation de la variance totale en supposant que les caractéristiques $y_k$ de la population finie étaient générées à partir d'un modèle de superpopulation. Nous avons obtenu un estimateur de la variance par linéarisation en utilisant la méthode de Demnati et Rao (2004). L'estimateur proposé de la variance préserve automatiquement les « poids g ». De plus, il reste valide malgré une mauvaise spécification de la variance

par rapport au modèle de $y_k$, étant supposé que la covariance fondée sur le modèle de $y_k$ et $y_t$ est nulle pour $k \neq t$. Nous prévoyons étendre nos résultats à des données d'enquêtes longitudinales en tenant compte, dans le temps, d'une mauvaise spécification de la covariance fondée sur le modèle. D'autres extensions à l'étude comprennent l'échantillonnage à deux phases et les réponses manquantes.

## RÉFÉRENCES


Demnati, A. et Rao, J. N. K. (2002), "Linearization Variance Estimators for Survey Data With Missing Responses", *Proceeding of the Section Survey Research Methods, American Statistical Association,* pp. 736-740.

Demnati, A. et Rao, J. N. K. (2004), « Estimateurs de variance par linéarisation pour des données d'enquête » (avec commentaires), *Techniques d'enquête*, vol. 30, p. 17 à 34.

Lambert, D. (1992), "Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing", *Technometrics*, 34, pp.1-14.

Molina, E. A., Smith, T. M. F. et Sugden, R. A. (2001), "Modeling Overdispersion for Complex Survey Data", *International Statistical Review, 69, pp. 373-384.*

Royall, R. M., et Cumberland, W. G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance", *Journal of the American Statistical Association*, 76, pp. 66-77.

Särndal, C.-E., Swensson, B., et Wretman, J.H. (1989), "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total", *Biometrika*, 76, pp. 527-537.

Valliant, R. (1993), "Poststratification and Conditional Variance Estimation", *Journal of the American Statistical Association*, 88, pp. 89-96.