ESTIMATEURS DE LA VARIANCE PAR LINÉARISATION POUR DES DONNÉES D'ENQUÊTE AVEC DES RÉPONSES MANQUANTES

Abdellatif DEMNATI (*), J. N. K. RAO (**)

(*) Statistique Canada, Division des méthodes d'enquêtes sociales
(**) Université Carlton, Étude des Mathématiques et des Statistiques

Introduction

La linéarisation de Taylor est une méthode d'estimation de la variance fréquemment utilisée pour des statistiques complexes, comme les estimateurs par quotient ou par régression, ainsi que les estimateurs des coefficients de régression logistique. Elle est généralement applicable à tout plan d'échantillonnage qui permet d'obtenir une estimation non biaisée de la variance d'estimateurs linéaires, et les calculs sont plus simples que ceux des méthodes de rééchantillonnage, comme le jackknife. Cependant, elle peut donner plusieurs estimateurs de la variance asymptotiquement non biaisés par rapport au plan d'échantillonnage en cas d'échantillonnage répété. Par conséquent, pour choisir l'estimateur de la variance, il faut tenir compte d'autres facteurs, dont (i) l'absence approximative du biais de la variance de l'estimateur sous le modèle, pour un modèle supposé et (ii) la validité dans des conditions d'échantillonnage répété conditionnel. Par exemple, dans le contexte de l'échantillonnage aléatoire simple et de l'estimateur par quotient, $\hat{Y}_R = (\bar{y} / \bar{x})X$, du total de la population Y, Royall et Cumberland [5] ont montré qu'un estimateur par linéarisation utilisé couramment, $v_L = N^2 (n^{-1} - N^{-1}) s_z^2$, ne permet pas de suivre la variance conditionnelle de $\hat{Y}_R$ étant donné $\bar{x}$, contrairement à l'estimateur jackknife de la variance, v_J. Ici, $\bar{y}$ et $\bar{x}$ sont les moyennes d'échantillon, X est le total connu de la population d'une variable auxiliaire x, s_z^2 est la variance dans l'échantillon des résidus $z_i = y_i - (\bar{y} / \bar{x}) x_i$ et (n, N) représente les tailles de l'échantillon et de la population. Si nous linéarisons l'estimateur jackknife de la variance, v_J, nous obtenons un estimateur par linéarisation de la variance différent, $v_{JL} = (\overline{X} / \bar{x})^2 v_L$, qui suit la variance conditionnelle ainsi que la variance non conditionnelle, où $\overline{X} = X/N$ est la moyenne de x. Par conséquent, on pourrait préférer v_{JL} ou v_J à v_L. Yung et Rao [9] ont considéré des estimateurs par régression généralisée et des estimateurs ajustés par quotient stratifiés a posteriori sous un plan stratifié à plusieurs degrés et ont obtenu un estimateur par linéarisation jackknife de la variance, v_{JL}, par la linéarisation de v_J. Valliant [8] a aussi obtenu v_{JL} pour l'estimateur stratifié a posteriori et a réalisé une étude de simulation pour démontrer que v_J et v_{JL} possèdent tous deux de bonnes propriétés

conditionnelles, étant donné les estimations des tailles des strates. Särndal, Swensson et Wretman [6] ont montré que v_{JL} est à la fois asymptotiquement non biaisé par rapport au plan de sondage et asymptotiquement non biaisé par rapport au modèle au sens où $E_m(v_{JL}) = Var_m(\hat{Y}_R)$, où E_m représente l'espérance par rapport au modèle et $Var_m(\hat{Y}_R)$ est la variance de $\hat{Y}_R$ par rapport au modèle sous un modèle par quotient: $E_m(y_i) = b\, x_i$; $i = 1,\ldots,N$ et les y_i sont indépendants avec comme variance du modèle $Var_m(y_i) = s^2 x_i$, $s^2 > 0$. Donc, v_{JL} est un bon choix tant du point de vue du plan d'échantillonnage que du point de vue du modèle. Demnati et Rao [2] ont proposé une nouvelle méthode d'estimation de la variance qui est justifiable théoriquement et qui donne directement un estimateur de la variance de type v_{JL} pour des plans d'échantillonnage généraux. Cette méthode est présentée à la deuxième section.

En cas de réponses manquantes, on procède souvent à l'ajustement des poids pour compenser la non-réponse complète, tandis que l'on a recourt couramment à l'imputation pour compenser la non-réponse partielle afin d'obtenir des données complètes pour calculer des estimations. Cependant, traiter les poids ajustés comme des poids de sondage et les valeurs imputées comme des valeurs réelles, et appliquer des formules standard d'estimation de la variance peut produire une sous-estimation considérable si le taux de non-réponse est important. Ces dernières années, on a proposé plusieurs méthodes pour estimer correctement la variance d'un estimateur en cas d'imputation. Rao [4], Shao et Steel [7] et d'autres ont étudié l'estimation de la variance sous imputation par quotient, cependant le problème d'estimation de la variance en cas d'ajustement de la pondération et d'imputation reste ouvert.

L'objectif principal du présent article est d'étendre la méthode de Demnati et Rao [2] au cas des réponses manquantes, lorsqu'on procède à la correction pour la non-réponse totale et à l'imputation basée sur des fonctions lisses des valeurs observées, en particulier l'imputation par quotient. À la deuxième section, nous décrivons brièvement la méthode dans le cas de la réponse complète. À la troisième section, nous décrivons la méthode de Shao et Steel [7], tandis qu'à la quatrième section, nous présentons l'extension de la méthode de Demnati-Rao.

1. Réponse complète

Pour motiver l'application de la méthode de Demnati-Rao [2] en cas de réponse complète, supposons qu'un estimateur $\hat{q}$ d'un paramètre q s'exprime sous forme d'une fonction dérivable $g(\hat{\underset{\sim}{Y}})$ des totaux estimés $\hat{\underset{\sim}{Y}} = (\hat{Y}_1,\ldots,\hat{Y}_m)^T$, où $\hat{Y}_j = \sum_{i\in U} d_i(s)\, y_{ij}$ est un estimateur du total de la population Y_j, $j = 1,\ldots,m$ et $d_i(s) = 0$ si l'unité i ne fait pas partie de l'échantillon s, U est l'ensemble des unités de la population et $q = g(\underset{\sim}{Y})$, avec $\underset{\sim}{Y} = (Y_1,\ldots,Y_m)^T$. Nous pouvons écrire $\hat{q}$ sous la forme $\hat{q} = f(\underset{\sim}{d}(s), \underset{\sim}{A}_y)$ et $q = f(\underset{\sim}{1}, \underset{\sim}{A}_y)$, où $\underset{\sim}{A}_y$ est une matrice $m \times N$ dont la j^e colonne est $\underset{\sim}{y}_j = (y_{1j},\ldots,y_{mj})^T$, $j = 1,\ldots,N$, $\underset{\sim}{d}(s) = (d_1(s),\ldots,d_N(s))^T$ et $\underset{\sim}{1}$ est le vecteur de taille N dont toutes les coordonnées valent 1. Par exemple, si $\hat{q}$ représente l'estimateur par quotient $\hat{Y}_R = \left[\sum_{i\in U} d_i(s) y_i / \sum_{i\in U} d_i(s) x_i\right] X$, alors $m = 2$, $y_{1i} = y_i$, $y_{2i} = x_i$ et $f(\underset{\sim}{1}, \underset{\sim}{A}_y)$ se réduit au total Y, en remarquant que $(Y/X)\, X = Y$. Notons que $\hat{Y}_R$ est une fonction de $\underset{\sim}{d}(s)$, $\underset{\sim}{y}$ et $\underset{\sim}{x}$, et du total connu X, mais que, par souci de simplicité, nous laissons tomber X et écrivons $\hat{Y}_R = f(\underset{\sim}{d}(s), \underset{\sim}{y}, \underset{\sim}{x})$. Si nous utilisons les coefficients de pondération d'Horvitz-Thompson,

alors $d_i(s) = 1/p_i$ pour tout $i \in s$, où p_i est la probabilité de sélectionner l'unité i dans l'échantillon s.

Soit $\hat{Y} = \sum b_i y_i$ pour des nombres réels arbitraires $b = (b_1,...,b_N)^T$, et $f(b, A_y) = f(b)$. Demnati et Rao [2] ont montré que la linéarisation de Taylor de $\hat{q} - q$, c'est-à-dire

$$\hat{q} - q = g(\hat{Y}) - g(Y) \approx \left(\partial g(a)/\partial a\right)^T \Big|_{a=y} (\hat{Y} - Y),$$

est équivalente à

$$\begin{aligned} \hat{q} - q &\approx \sum_{k=1}^{N} \left(\partial f(b)/\partial b_k\right)\Big|_{b=1} (d_k(s) - 1) \\ &= z^T (d(s) - 1) \end{aligned} \tag{1.1}$$

où $\partial g(a)/\partial a = (\partial g(a)/\partial a_1,...,\partial g(a)/\partial a_m)^T$ et $z = (z_1,...,z_N)^T$ avec $z_k = \partial f(b)/\partial b_k\Big|_{b=1}$. Il découle de (1.1) qu'un estimateur de la variance de $\hat{q}$ est donné approximativement par l'estimateur de la variance du total estimé $\sum d_i(s) z_i = \hat{Y}(z)$; c'est-à-dire, $var(\hat{q}) \approx v(z)$, où $v(y)$ représente l'estimateur de la variance de $\hat{Y} = \hat{Y}(y)$ en notation opérationnelle. Maintenant, nous remplaçons z_k par $z_k = \partial f(b)/\partial b_k\Big|_{b=d(s)}$, puisque les z_k sont inconnus, pour obtenir un estimateur par linéarisation de la variance

$$v_L(\hat{q}) = v(z). \tag{1.2}$$

Notons que $v_L(\hat{q})$ donné par (1.2) s'obtient simplement à partir de la formule de $v(y)$ pour $\hat{Y} = \hat{Y}(y)$ en remplaçant y_i par z_i pour $i \in s$. Soulignons que nous ne commençons pas par évaluer les dérivées partielles $\partial f(b)/\partial b_k\Big|$ en $b = 1$ afin d'obtenir z pour ensuite substituer les estimations aux composantes inconnues de z. Par conséquent, notre méthode est similaire dans l'esprit à l'approche de Binder [1]. L'estimateur de la variance $v_L(\hat{q})$ est valide, car z_i est un estimateur convergent de z_i.

Supposons que $\hat{q}$ soit l'estimateur par quotient $\hat{Y}_R = X\left[(\sum d_i(s) y_i)/\sum d_i(s) x_i)\right]$, où $\sum$ représente la sommation sur $i \in U$. Alors $f(b) = X\left[(\sum b_i y_i)/(\sum b_i x_i)\right] = X\left[\hat{Y}(b)/\hat{X}(b)\right]$ et

$$z_k = \partial f(b)/b_k\Big|_{b=d(s)} = \frac{X}{\hat{X}}(y_k - \hat{R}x_k).$$

Dans le cas de l'échantillonnage aléatoire simple, $v_L(\hat{Y}_R) = v(z)$ coïncide avec $v_{JL} = (\bar{X}/\bar{x})^2 v_L$.

Demnati et Rao [2] ont appliqué la méthode à divers problèmes, qui englobent les estimateurs par régression et 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 les estimateurs par la méthode itérative du quotient (raking ratio) généralisée et les estimateurs par régression généralisée. Ils ont étendu également la méthode à l'échantillonnage à deux degrés et ont obtenu un estimateur de la variance qui utilise mieux les données du premier degré d'échantillonnage que les estimateurs de la variance par linéarisation classique.

2. Non-réponse partielle

En s'inspirant de Fay [3], Shao et Steel [7] ont proposé une méthode de calcul des estimateurs de la variance de l'estimateur du total de type Horvitz-Thompson, $\hat{Y}^{\bullet}$, lorsque des valeurs sont imputées pour tenir compte de la non-réponse partielle. Ils ont supposé que l'on peut exprimer le total estimé $\hat{Y}^{\bullet}$ sous forme d'une fonction lisse des totaux $\hat{Y}^{\bullet} = y(\hat{\underset{\sim}{T}}_o)$, où $\hat{\underset{\sim}{T}}_o = \sum d_i(s) diag(\underset{\sim}{o}_i)\underset{\sim}{t}_i$, t_{ki} est la valeur de y_i ou celle d'une autre variable utilisée pour imputer une valeur à y_i, et $\underset{\sim}{o}_i = (o_{i1},...,o_{ip})^T$ est le vecteur des variables indicatrices de réponse. Par exemple, considérons l'imputation par quotient lorsqu'on dispose des valeurs de la variable auxiliaire x_i pour tous les i de s. Une valeur de y_i manquante est alors imputée par $\hat{y}_i = \hat{R}_o x_i$, où $\hat{R}_o = (\sum d_i(s) o_i y_i)/(\sum d_i(s) o_i x_i)$ et o_i est l'indicateur de réponse pour y_i, c.-à-d. $o_i = 1$ si la valeur de y_i est observée et $o_i = 0$ si elle est manquante. L'estimateur imputé $\hat{Y}^{\bullet}$ est donné par

$$\begin{aligned} \hat{Y}^{\bullet} &= \sum d_i(s) o_i y_i + \sum d_i(s)(1-o_i)\hat{R}_o x_i \\ &= \sum d_i(s) o_{1i} y_i (1 + \sum d_i(s) o_{2i} x_i / \sum d_i(s) o_{1i} x_i) \end{aligned} \tag{2.1}$$

où $o_{i1} = o_i$ et $o_{i2} = 1 - o_i$. Il découle de (2.1) que $\hat{Y}^{\bullet}$ est de la forme $y(\hat{T}_o)$ avec $\underset{\sim}{o}_i = (o_{i1}, o_{i1}, o_{i2})^T$ et . $t_{\sim i} = (y_i, x_i, x_i)^T$

Nous supposons que l'imputation est déterministe. Nous avons $Var(\hat{Y}^{\bullet} - Y) = V_1 + V_2$, où $V_1 = E_o(Var_s(\hat{Y}^{\bullet} - Y)), V_2 = Var_o(E_s(\hat{Y}^{\bullet} - Y))$, E_o et Var_o représentent l'espérance et la variance sous le mécanisme de réponse, et E_s et Var_s représentent l'espérance et la variance sous le mécanisme d'échantillonnage selon un plan d'échantillonnage donné. Shao et Steel [7] ont obtenu un estimateur de la variance v_1 de V_1 en se servant d'un estimateur par linéarisation de la variance de $y(\hat{T}_o)$ étant donné les $\underset{\sim}{o}_i$.

Ils ont obtenu aussi un estimateur, v_2 de $V_2 = \left[\nabla j(E_o \underset{\sim}{T}_o)\right]^T C \left[\nabla j(E_o \underset{\sim}{T}_o)\right]$, où $j(T_o) = y(E_s \hat{\underset{\sim}{T}}_o) - Y$, en calculant $\underset{\sim}{C}$ dont le kl^e élément $c_{kl} = cov_o(\sum o_{ki} t_{ki}, \sum o_{li} t_{li})$ et en remplaçant les quantités inconnues par des estimateurs. Pour l'échantillonnage aléatoire simple et l'imputation par quotient, Shao et Steel [7] ont obtenu v_1 sous la forme

$$v_1 = N^2 \frac{(1-n/N)}{n(n-1)} \left[\left(\frac{\bar{x}}{\bar{x}_o} \right)^2 \frac{s_d^2}{n_o} + 2 \frac{\bar{x}}{\bar{x}_o} \frac{\hat{R}_o s_{dx}}{n} + \frac{\hat{R}_o^2 s_x^2}{n} \right], \tag{2.2}$$

où $\bar{x}$ et s_x^2 sont la moyenne et la variance d'échantillon des x_i, $\bar{x}_o = \sum_{i \in s} o_i x_i / n_o$ est la moyenne des x_i pour les répondants, n_o est le nombre de répondants, $s_d^2 = \sum_{i \in s} (y_i - \hat{R}_o x_i)^2 / (n_o - 1)$, et $s_{dx} = \sum_{i \in s} o_i x_i (y_i - \hat{R}_o x_i)^2 / (n_o - 1)$.

Aussi, sous l'hypothèse de réponse uniforme (c.-à-d. si les o_i sont indépendamment et identiquement distribués en ayant une moyenne p_y et une variance $p_y(1-p_y)$), Shao et Steel [7] ont obtenu v_2 sous la forme

$$v_2 = (X / X_o)^2 \hat{p}_y (1-\hat{p}_y) N s_d^2, \quad (2.3)$$

où $\hat{p}_y = \sum o_i d_i(s) / \sum d_i(s)$. La somme de (2.2) et (2.3) donne l'estimateur de la variance de $\hat{Y}^\bullet$.

La méthode de Shao et Steel [7] est fondée sur la méthode de linéarisation classique qui consiste à i) exprimer l'estimateur en fonction de composantes élémentaires, ii) calculer les dérivées partielles au niveau de la population et iii) estimer les paramètres inconnus dans la formule. Par conséquent, l'estimateur correspondant de la variance n'est pas nécessairement unique. Notre méthode permet d'éviter d'exprimer l'estimateur en fonction de composantes élémentaires, donc produit directement un estimateur unique de la variance ayant les propriétés souhaitées. Nous présentons notre méthode pour l'imputation par quotient à la sous-section 3.1, et nous examinons le cas de l'estimation de la variance sous ajustement de la pondération pour tenir compte de la non-réponse complète et imputation par quotient pour tenir compte de la non-réponse partielle à la sous-section 3.2.

3. Nouvelle méthode : réponses manquantes

Après ajustement de la pondération pour tenir compte de la non-réponse complète et imputation pour tenir compte de la non-réponse partielle, nous estimons le total de la population Y au moyen d'un total pondéré à partir de l'échantillon

$$\hat{Y}^\bullet = \sum \tilde{w}_i(s) o_i y_i + \sum \tilde{w}_i(s)(1-o_i)\hat{y}_i^\bullet, \quad (3.1)$$

où $\tilde{w}_i(s)$ est le poids ajusté et $\hat{y}_i^\bullet$ représente la valeur imputée pour l'unité i. Nous pouvons réécrire l'estimateur (3.1) sous la forme

$$\hat{Y}^\bullet = \sum \tilde{w}_i(s)\hat{y}_i = \underset{\sim}{\hat{y}}^T \underset{\sim}{\tilde{w}}(s), \quad (3.2)$$

où $\underset{\sim}{\hat{y}} = (\hat{y}_1, ..., \hat{y}_N)^T$ et $\hat{y}_i = o_i y_i + (1-o_i)\hat{y}_i^\bullet$. À la sous-section 3.1, nous étudions le cas de la non-réponse partielle uniquement (c.-à-d. $\tilde{w}_i(s) = d_i(s)$) en supposant que nous pouvons exprimer $\hat{Y}^\bullet$ sous la forme d'une fonction lisse de totaux $\sum d_i(s) diag(\underset{\sim}{o}_i)\underset{\sim}{t}_i$, où t_{ki} est la valeur de y_i ou celle d'une autre variable utilisée pour imputer une valeur à y_i. À la sous-section 3.2, nous traitons le cas plus général de l'ajustement de la pondération pour tenir compte de la non-réponse complète et de l'imputation pour tenir compte de la non-réponse partielle.

3.1 Imputation pour tenir compte de la non-réponse partielle

Nous supposons que l'estimateur avec données imputées $\hat{Y}^\bullet$ est une fonction lisse de totaux, $\sum d_i(s) diag(\underset{\sim}{o}_i)\underset{\sim}{t}_i$, à l'instar de Shao et Steel [7]. Dans ce cas, nous pouvons exprimer $\hat{Y}^\bullet$ sous la forme $f(\underset{\sim}{A}_w, \underset{\sim}{A}_y)$, où $\underset{\sim}{A}_w = diag(\underset{\sim}{d}(s))\underset{\sim}{A}_o$ est une matrice $m \times N$ dont la j^e colonne est $\underset{\sim}{w}_j(s) = (w_{1j}(s), ..., w_{mj}(s))^T$, $j = 1, ..., N$. Le vecteur $\underset{\sim}{w}_j(s)$ est défini comme étant

$$\underset{\sim}{w}_j(s) = (w_{1j}(s),..., w_{mj}(s))^T$$
$$= (o_{1j} d_j(s),..., o_{mj} d_j(s))^T = \underset{\sim}{o}_j \; d_j(s),$$

où $\underset{\sim}{o}_j = (o_{1j},..., o_{mj})^T$ est le vecteur des variables indicatrices correspondant au vecteur $\underset{\sim}{y}_j = (y_{1j},..., y_{mj})^T$.

Par souci de simplicité, nous laissons tomber $\underset{\sim}{A}_y$ et écrivons $\hat{Y}^\bullet = f(\underset{\sim}{A}_w)$. En cas d'imputation par quotient, nous avons $m = 2$, $y_{1j} = x_j$, $y_{2j} = y_j$, $o_{1j} = 1$, $o_{2j} = o_j$, $\underset{\sim}{w}_j(s) = (w_{1j}(s), w_{2j}(s))^T = (d_j(s), o_j d_j(s))^T$ et

$$\hat{Y}^\bullet = \sum w_{2i}(s)(y_i - \hat{R}_o x_i) + \sum w_{1i}(s)\hat{R}_o x_i,$$

où $\hat{R}_o = (\sum w_{2i}(s) y_i) / (\sum w_{2i}(s) x_i)$.
Comme l'estimateur $\hat{Y}^\bullet = f(\underset{\sim}{A}_w)$ est une fonction de totaux, nous pouvons appliquer la méthode de Demnati et Rao [2] pour approximer sa variance au moyen de la variance d'une fonction linéaire

$$Var(\hat{Y}^\bullet) \approx Var(\hat{Y}_L^\bullet)$$

avec

$$\hat{Y}_L^\bullet = \sum (\underset{\sim}{o}_i \; d_i(s))^T \; \underset{\sim}{\tilde{z}}_i = \sum w_i^T(s) \, \underset{\sim}{\tilde{z}}_i,$$

où $\underset{\sim}{\tilde{z}}_i$ est le vecteur des dérivées de $f(\underset{\sim}{A}_b)$ par rapport à $\underset{\sim}{b}_k$ évaluées à $\underset{\sim}{A}_b = E(\underset{\sim}{A}_w)$, où $\underset{\sim}{A}_b$ est une matrice $m \times N$ de nombres réels arbitraires, $f(\underset{\sim}{A}_b)$ est obtenue par remplacement de $\underset{\sim}{A}_w$ par $\underset{\sim}{A}_b$ dans la formule de $\hat{Y}^\bullet$ et $\underset{\sim}{b}_k$ est un vecteur colonne de $\underset{\sim}{A}_b$. Nous pouvons estimer la variance totale de $\hat{Y}_L^\bullet$ selon

$$v(\hat{Y}_L^\bullet) = v_s(\underset{\sim}{o}^T \; \underset{\sim}{z}) + v_o(\underset{\sim}{z}), \tag{3.4}$$

où $\underset{\sim}{z}_k$ est le vecteur des dérivées de l'estimateur $f(\underset{\sim}{A}_b)$ par rapport à $\underset{\sim}{b}_k$ évaluées à $\underset{\sim}{A}_b = \underset{\sim}{A}_w$, et $v_o(\underset{\sim}{z})$ est un estimateur de $Var(\sum diag(\underset{\sim}{o}_i) \, \underset{\sim}{z}_i)$. Sous un mécanisme de réponse indépendante,

$$v_o(\underset{\sim}{z}) = \sum \underset{\sim}{z}_i^T \; \mathrm{cov}_o(\underset{\sim}{o}_i) \, \underset{\sim}{z}_i, \tag{3.5}$$

où $\mathrm{cov}_o(\underset{\sim}{o}_i)$ est un estimateur (approximativement) non biaisé de $E(\underset{\sim}{o}_i \, \underset{\sim}{o}_i^T) - E(\underset{\sim}{o}_i) E(\underset{\sim}{o}_i^T)$.

Sous imputation par quotient, nous avons

$$\begin{aligned} \underset{\sim}{z}_k &= \frac{\partial}{\partial \underset{\sim}{b}_k} \left(\sum b_{2i} (y_i - \hat{R}_o(\underset{\sim}{A}_b) x_i + \sum b_{1i} \hat{R}_o(\underset{\sim}{A}_b) x_i \right)_{\underset{\sim}{A}_b = \underset{\sim}{A}_w} \\ &= \left(\hat{R}_o x_k, (\hat{X} / \hat{X}_o)(y_k - \hat{R}_o x_k) \right)^T \end{aligned} \tag{3.6}$$

Il découle de (3.6) que

$$z_k = o_k (\hat{X} / \hat{X}_o)(y_k - \hat{R}_o x_k) + \hat{R}_o x_k. \tag{3.7}$$

Par conséquent, $v_s(\underset{\sim}{o}^T \underset{\sim}{z})$ est égal à $v_1 = v(z)$. Dans les conditions d'échantillonnage aléatoire simple, *v(z)* avec z_k donné par (3.7) concorde avec l'expression (2.2) de Shao et Steel [7]. Aussi,

$$\mathrm{cov}_o(\underset{\sim}{o}_i) = d_i(s)\begin{pmatrix} 0 & 0 \\ 0 & o_i(1-\hat{x}_{io}) \end{pmatrix},$$

où $\hat{x}_{io}$ est un estimateur de la probabilité de réponse de l'unité i. Par conséquent, $v_o(\underset{\sim}{z})$, donné par (3.5), se réduit à

$$v_o(\underset{\sim}{z}) = (\hat{X} / \hat{X}_o)^2 \sum d_i(s) o_i (1-\hat{x}_{io})(y_i - \hat{R}_o x_i)^2 . \qquad (3.8)$$

Dans le cas d'un échantillonnage aléatoire simple et d'un mécanisme de réponse uniforme, (3.8) se réduit à

$$v_o(\underset{\sim}{z}) = \frac{N}{n}(\hat{X} / \hat{X}_o)^2 (1-n_o/n) \sum o_i (y_i - \hat{R}_o x_i)^2 \qquad (3.9)$$

qui est l'estimateur v_2 de Shao et Steel [7] donné par (2.3).

4.2 Ajustement de la pondération et imputation pour tenir compte de la non-réponse partielle

Soit r_i, la variable indicatrice de réponse partielle pour la i^e unité, c.-à-d. $r_i = 0$ si la non-réponse est complète et $r_i = 1$ si la réponse est partielle. La variable indicatrice de réponse partielle r_i est reliée aux variables indicatrices de réponse à une question o_p, $p = 1..m$ par

$$r_i = 1 - \prod_{p=1}^{m} (1-o_{ip}) . \qquad (3.10)$$

Nous avons

$$Cov(r_i, o_{ip}) = E(r_i o_{ip}) - E(r_i)E(o_{ip}) ,$$

pour toute variable indicatrice de réponse o_{ip}. Notant que $r_i o_{ip} = o_{ip}$ pour tout o_{ip},

$$r_i o_{ip} = \left[1-(1-o_{ip})\prod_{q\neq p}(1-o_{iq})\right] o_{ip} = o_{ip} .$$

Donc,

$$Cov(r_i, o_{ip}) = E(o_{ip}) - E(r_i)E(o_{ip}) = E(o_{ip})(1-E(r_i)) .$$

Nous pouvons donner un estimateur de $Cov(r_i, o_{ip})$ sous la forme

$$\mathrm{cov}(r_i, o_{ip}) = o_{ip}(1-\hat{x}_{ir})$$

où $\hat{x}_{ir} = \hat{E}(r_i)$ et $\hat{E}(.)$ représente un estimateur de $E(.)$.

Une méthode très répandue de correction de la non-réponse complète consiste à employer un nouvel ensemble de poids, $\tilde{w}(s)$, dont le i^e élément est égal à

$$\tilde{w}_i(s) = d_i(s)\ r_i\ g_i(\underset{\sim}{d}(s), \underset{\sim}{r}, \underset{\sim}{A}_c)\,, \tag{3.11}$$

où les $g_i(\underset{\sim}{d}(s), \underset{\sim}{r}, \underset{\sim}{A}_c)$ sont connus sont le nom de poids g dans le contexte de l'estimateur par régression et $\underset{\sim}{A}_c$ est une matrice de variables auxiliaires dont les valeurs sont connues pour toutes les unités figurant dans l'échantillon. L'estimateur par quotient est un cas particulier de (4.11) pour lequel les poids g se réduisent à

$$g_i(\underset{\sim}{d}(s), \underset{\sim}{r}, \underset{\sim}{A}_c) = \frac{\sum d_i(s)\, c_i}{\sum d_i(s)\, r_i\, c_i} = \frac{\hat{c}}{\hat{c}_r}\,, \tag{3.12}$$

où $\hat{c} = \sum d_i(s)\, c_i$ et $\hat{c}_r = \sum d_i(s)\, r_i\, c_i$. L'ajustement des poids au moyen du quotient (3.12) est un cas particulier de la classe de poids de calage obtenus en utilisant l'estimateur par régression. On utilise également les poids obtenus par ajustement itératif du quotient généralisé pour compenser pour la non-réponse complète. Un autre moyen de tenir compte de la non-réponse complète consiste à pondérer chaque observation par la probabilité inverse de réponse, auquel cas

$$g_i(\underset{\sim}{d}(s), \underset{\sim}{r}, \underset{\sim}{A}_c) = \hat{x}_{ir}^{-1}\,,$$

où

$$\hat{x}_{ir} = x_{ir}(\underset{\sim}{r}, \underset{\sim}{d}(s)) = \Pr(r_i = 1 | \underset{\sim}{d}(s), \underset{\sim}{A}_c)\,,$$

est l'estimateur de la probabilité de réponse défini comme étant la solution d'une équation d'estimation de la forme

$$\hat{U}(\underset{\sim}{\hat{x}}_r) = \sum d_i(s) u_i(r_i, \underset{\sim}{c}_i, \hat{x}_{ir}) = 0\,.$$

Dans le cas logistique, nous avons

$$\hat{U}(\underset{\sim}{\hat{x}}_r) = \sum d_i(s)(r_i - \hat{x}_{ir})\, \underset{\sim}{c}_i = 0\,,$$

où

$$u_i(r_i, \underset{\sim}{c}_i, \hat{x}_{ir}) = (r_i - \hat{x}_{ir})\, \underset{\sim}{c}_i\,,$$

$$x_{ir} = \exp(\underset{\sim}{c}_i^T\ \underset{\sim}{b})/(1 + \exp(\underset{\sim}{c}_i^T\ \underset{\sim}{b})) = \Pr(r_i = 1 | \underset{\sim}{c}_i, \underset{\sim}{b})\,,$$

et $\underset{\sim}{c}_i$ est le vecteur des variables explicatives.

Avec les méthodes d'ajustement de la pondération susmentionnées, nous pouvons obtenir la variance en suivant la méthode de Demnati et Rao [2] en exprimant $\hat{Y}^{\bullet}$ sous la forme $f(\underset{\sim}{A}_w)$, puis en dérivant $f(\underset{\sim}{A}_b)$ par rapport à $\underset{\sim}{b}_k$.

Nous omettons les détails par souci de simplicité, mais nous illustrons le calcul pour l'estimateur (3.1) sous ajustement de la pondération par quotient (3.12) et imputation par quotient, c'est-à-dire

$$\hat{Y}^{\bullet} = \sum_i o_i \tilde{w}_i(s) y_i + \sum_j (1 - o_j) \tilde{w}_j(s) \hat{R}_o x_j\,,$$

avec

$$\hat{R}_o = \frac{\sum \tilde{w}_i(s) o_i y_i}{\sum \tilde{w}_i(s) o_i x_i},$$

et

$$\tilde{w}_i(s) = d_i(s) r_i \hat{c} / \hat{c}_r .$$

Nous avons

$$w_i(s) = (1, o_i, r_i)^T d_i(s),$$

$$\underset{\sim}{z}_k = \left(x_k \hat{R}_o (\hat{X}_r / \hat{c}_r), (\hat{c} / \hat{c}_r)(\hat{X}_r / \hat{X}_o)(y_k - \hat{R}_o x_k), (\hat{c} / \hat{c}_r) \hat{R}_o (x_k - (\hat{X}_r / \hat{c}_r) c_k) \right)^T,$$

et

$$cov_o(\underset{\sim}{o}_i^T) = d_i(s) \begin{pmatrix} 0 & 0 & 0 \\ 0 & o_i(1 - \hat{x}_{io}) & o_i(1 - \hat{x}_{ir}) \\ 0 & o_i(1 - \hat{x}_{ir}) & r_i(1 - \hat{x}_{ir}) \end{pmatrix}.$$

Conclusion

Nous avons présenté une nouvelle méthode d'estimation de la variance en cas de réponses manquantes. Nous donnons un estimateur valide de la variance pour diverses méthodes d'ajustement de la pondération utilisées fréquemment pour tenir compte de la non-réponse complète, ainsi que pour l'imputation basée sur des fonctions lisses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'extension à l'imputation par le plus proche voisin et aux enquêtes par panel est à l'étude.

Bibliographie

[1] Binder, D., « Linearization methods for single phase and two-phase samples: a cookbook approach», *Survey Methodology, 22*, pp 17-22, 1996.

[2] Demnati, A. and Rao, J. N. K., Linearization variance estimators for survey data, Methodology Branch Working Paper, SSMD-2001-010E. Statistics Canada, 2001.

[3] Fay, R. E., « A design-based perspective on missing data variance», *in Proceeding of the 1991 Annual Research Conference, US Bureau of the census*, pp 429-440, 1991.

[4] Rao, J. N. K., « On variance estimation with imputed survey data (with discussion) », *Journal of the American Statistical Association, 91*, pp 499-520, 1996.

[5] Royall, R. M., and Cumberland, W. G., « An empirical study of the ratio estimator and estimators of its variance», *Journal of the American Statistical Association*, pp 76, 66-77, 1981.

[6] Särndal, C.-E., Swensson, B., and Wretman, J.H., «The Weighted residual technique for estimating the variance of the general regression estimator of the finite population total», *Biometrika, 76*, pp 527-537,1989.

[7] Shao, J. and Steel, P., « Variance estimation for survey data with composite imputation and nonnegligible sampling fractions», *Journal of the American Statistical Association, 94*, pp 254-265, 1999.

[8] Valliant, R.,«Postsratification and conditional variance estimation», *Journal of the American Statistical Association, 88*, pp 89-96, 1993.

[9] Yung, W. and Rao, J. N. K., « Jackknife linearization variance estimators under stratified multi-Stage sampling», *Survey Methodology, 22*, pp 23-31, 1996.