Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimation de la variance pour un ratio en présence de données imputées

par David Haziza

Décembre 2007

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2007
Vol. 33, n° 2, pp. 179-186
Statistique Canada, n° 12-001-X au catalogue

Estimation de la variance pour un ratio en présence de données imputées

David Haziza [1]

Résumé

Dans le présent article, nous étudions le problème de l'estimation de la variance pour un ratio de deux totaux quand l'imputation hot deck aléatoire marginale est utilisée pour remplacer les données manquantes. Nous considérons deux approches d'inférence. Dans la première, l'établissement de la validité d'un modèle d'imputation est nécessaire. Dans la seconde, la validité d'un modèle d'imputation n'est pas nécessaire, mais il faut estimer les probabilités de réponse, auquel cas il est nécessaire d'établir la validité d'un modèle de non-réponse. Nous obtenons les estimateurs de la variance sous deux cadres distincts, à savoir le cadre à deux phases habituel et le cadre inversé.

Mots clés : Modèle d'imputation; modèle de non-réponse; imputation hot deck aléatoire marginale; cadre inversé; cadre à deux phases; estimation de la variance.

1. Introduction

L'estimation de la variance en présence de données imputées pour des paramètres univariés simples, tels que des totaux ou des moyennes de population, a fait l'objet de nombreuses études ces dernières années; voir, par exemple, Särndal (1992), Deville et Särndal (1994), Rao et Shao (1992), Rao (1996), et Shao et Steel (1999). Il est fréquent, en pratique, de devoir estimer le ratio de deux totaux de population, $R = Y / X$, où $(Y, X) = \sum_{i \in U} (y_i, x_i)$, y et x sont deux variables d'intérêt pour lesquelles des données peuvent manquer et U est la population finie (de taille N) étudiée. Alors que l'estimation de la variance dans le cas d'un ratio en présence de données imputées est un problème qui se pose souvent en pratique (surtout dans le cas des enquêtes auprès des entreprises), autant que nous sachions, la question n'a pas été étudiée à fond dans la littérature. Dans le présent article, nous considérons le cas de l'imputation hot deck aléatoire marginale (HDAM) effectuée dans le même ensemble de classes d'imputation pour les deux variables y et x. Autrement dit, pour corriger la non-réponse, on procède séparément à une imputation hot deck aléatoire pour chaque variable dans le même ensemble de classes d'imputation. Cette situation se présente fréquemment en pratique. Pour simplifier, nous considérons le cas d'une seule classe d'imputation. Les extensions aux classes d'imputation multiples sont relativement simples pour la plupart des calculs présentés ici.

Nous obtenons dans le présent article des estimateurs de la variance qui tiennent compte de l'échantillonnage, de la non-réponse et de l'imputation. Deux cadres distincts d'estimation de la variance ont été étudiés dans la littérature : i) le cadre à deux phases habituel (par exemple, Särndal (1992)) et ii) le cadre inversé (par exemple, Shao et Steel (1999)). Dans le cadre à deux phases, la non-réponse est considérée comme une deuxième phase de sélection. En d'autres termes, un échantillon aléatoire est sélectionné parmi la population selon le plan d'échantillonnage donné. Puis, étant donné l'échantillon sélectionné, l'ensemble de répondants est généré selon le mécanisme de non-réponse. Dans le cadre inversé, l'ordre de l'échantillonnage et de la réponse est inversé. Autrement dit, la population est d'abord répartie aléatoirement en une population de répondants et une population de non-répondants, selon le mécanisme de non-réponse. Puis, un échantillon aléatoire est sélectionné à partir de la population (contenant les répondants et les non-répondants), selon le plan d'échantillonnage. Comme nous le verrons à la section 4, le cadre inversé facilite le calcul des estimations de la variance, mais contrairement au cadre à deux phases, nécessite l'hypothèse supplémentaire que le mécanisme de non-réponse ne dépend pas de l'échantillon sélectionné. Cette hypothèse est satisfaite dans de nombreuses situations observées en pratique. Pour chaque cadre, l'inférence peut être fondée sur un modèle d'imputation (MI) ou sur un modèle de non-réponse (MN). L'approche MI requiert la validité d'un modèle d'imputation et l'approche MN requiert la validité d'un modèle de non-réponse.

À la section 2, nous introduisons la notation, les hypothèses et l'estimateur imputé d'un ratio dans le cas de l'imputation HDAM pondérée. Les approches MI et MN sont présentées aux sections 2.1 et 2.2. À la section 2.3, nous discutons du biais de l'estimateur imputé. À la section 3, nous obtenons les estimateurs de la variance sous le cadre à deux phases et l'approche MI en utilisant la méthode proposée par Särndal (1992). Nous montrons que, sous l'imputation HDAM, l'estimateur de la variance naïf (qui traite les valeurs imputées comme des valeurs observées) surestime généralement la variance d'échantillonnage quand y et x sont positivement corrélées. À la section 4, nous

1. David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, (Québec), H3C 3J7, Canada. Courriel : David.Haziza@umontreal.ca.

obtenons les estimateurs de la variance sous le cadre inversé et l'approche MI ainsi que MN, en utilisant la méthode proposée par Shao et Steel (1999). Enfin, nous présentons nos conclusions à la section 5.

2. Notation et hypothèses

Notre but est d'estimer R. Nous sélectionnons un échantillon aléatoire, s, de taille n, selon un plan d'échantillonnage donné $p(s)$. Un estimateur basé sur des données complètes est donné par

$$\hat{R} = \frac{\hat{Y}_{\mathrm{HT}}}{\hat{X}_{\mathrm{HT}}}, \qquad (2.1)$$

où $(\hat{Y}_{\mathrm{HT}}, \hat{X}_{\mathrm{HT}}) = \sum_{i \in s} w_i (y_i, x_i)$ désignent les estimateurs Horvitz-Thompson de Y et X, respectivement et $w_i = 1/\pi_i$ représente le poids d'échantillonnage de l'unité i, où π_i est la probabilité d'inclusion dans l'échantillon. Dans (2.1), l'estimateur $\hat{R}$ est asymptotiquement p - sans biais pour R, c'est-à-dire $E_p(\hat{R}) \approx R$, où l'indice p désigne l'espérance et la variance par rapport au plan d'échantillonnage $p(s)$. Puisque $\hat{R}$ est une fonction non linéaire des totaux estimés, sa variance exacte sous le plan, $V_p(\hat{R})$, ne peut pas être calculée facilement. Pour contourner ce problème, on recourt souvent à la linéarisation de Taylor pour approximer la variance exacte. Un estimateur asymptotiquement p - sans biais de la variance approximative de $\hat{R}$ est donné par

$$\hat{V}_{\mathrm{SAM}} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} e_i e_j, \qquad (2.2)$$

où $e_i = 1/\hat{X}_{\mathrm{HT}} (y_i - \hat{R} x_i)$, $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j)/\pi_{ij} \pi_i \pi_j$ et π_{ij} est la probabilité de sélection conjointe des unités i et j. Notons que $\pi_{ii} = \pi_i$. Dans le cas de l'échantillonnage simple sans remise, l'estimateur de la variance (2.2) se réduit à

$$\hat{V}_{\mathrm{SAM}} = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}^2} \left(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}\right), \qquad (2.3)$$

où

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2, \; s_x^2 = \frac{1}{n-1} \sum_{i \in s} (x_i - \bar{x})^2$$

et

$$s_{xy} = \frac{1}{n-1} \sum_{i \in s} (x_i - \bar{x})(y_i - \bar{y})$$

avec

$$(\bar{y}, \bar{x}) = \frac{1}{n} \sum_{i \in s} (y_i, x_i).$$

Nous examinons maintenant le cas où des données pourraient manquer pour les variables x et y. Soit a_i l'indicateur de réponse de l'unité i, tel que $a_i = 1$ si l'unité i répond à la variable y et $a_i = 0$, autrement. De même, soit b_i l'indicateur de réponse de l'unité i, tel que $b_i = 1$ si l'unité i répond à la variable x et $b_i = 0$, autrement. Soit $s_r^{(y)}$ l'ensemble de répondants pour la variable y de taille r_y et $s_r^{(x)}$, l'ensemble de répondants pour la variable x de taille r_x. En outre, soit r_{xy} le nombre de répondants aux deux variables y et x. Enfin, soit y_i^* et x_i^*, les valeurs imputées pour remplacer les valeurs manquantes y_i et x_i, respectivement. Un estimateur imputé de R est donné par

$$\hat{R}_I = \frac{\sum_{i \in s} w_i \tilde{y}_i}{\sum_{i \in s} w_i \tilde{x}_i}, \qquad (2.4)$$

où $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$ et $\tilde{x}_i = b_i x_i + (1 - b_i) x_i^*$. Sous imputation HDAM pondérée, pour remplacer la valeur manquante y_i^*, on sélectionne un donneur j aléatoirement avec remise à partir de $s_r^{(y)}$ de sorte que

$$P(y_i^* = y_j) = \frac{w_j}{\sum_{l \in s} w_l a_l}.$$

De même, pour remplacer la valeur manquante x_i^*, on sélectionne un donneur k aléatoirement avec remise à partir de $s_r^{(x)}$, de sorte que

$$P(x_i^* = x_k) = \frac{w_k}{\sum_{l \in s} w_l b_l}.$$

Notons que, si y_i ainsi que x_i manquent, j n'est généralement pas égal à k sous l'imputation HDAM.

L'imputation hot deck aléatoire dans les classes est très répandue en pratique, parce que i) elle préserve la variabilité des données originales et que ii) elle produit des valeurs plausibles. Le deuxième point est particulièrement important dans le cas de variables d'intérêt catégoriques. Cependant, l'imputation hot deck aléatoire dans les classes souffre d'une composante de variance supplémentaire due à l'utilisation d'un mécanisme d'imputation aléatoire. La principale raison motivant l'utilisation de l'imputation HDAM pondérée est qu'elle mène à un estimateur asymptotiquement sans biais sous l'approche NM (voir la section 2.1), contrairement à l'imputation HDAM non pondérée.

Soit $E_I(. \mid s, s_r^{(y)}, s_r^{(x)})$, $V_I(. \mid s, s_r^{(y)}, s_r^{(x)})$ et $\mathrm{Cov}_I(.,. \mid s, s_r^{(y)}, s_r^{(x)})$, les opérateurs d'espérance, de variance et de covariance conditionnelles par rapport au mécanisme d'imputation aléatoire (ici, l'imputation HDAM pondérée). Par un développement en série de Taylor du premier ordre, on peut montrer que

$$E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) \approx \frac{\bar{y}_r}{\bar{x}_r} \circ \hat{R}_r, \qquad (2.5)$$

où

$$\bar{y}_r = \frac{\sum_{i \in s} w_i a_i y_i}{\sum_{i \in s} w_i a_i}$$

et

$$\bar{x}_r = \frac{\sum_{i \in s} w_i b_i y_i}{\sum_{i \in s} w_i b_i}$$

désignent les moyennes pondérées des répondants pour les variables y et x, respectivement. L'approximation (2.5) sera valide si la taille d'échantillon dans les classes est suffisamment grande, ce que nous supposons être le cas. Maintenant, soient

$$s_{yr}^2 = \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r)^2$$

et

$$s_{xr}^2 = \frac{1}{\sum_{i \in s} w_i b_i} \sum_{i \in s} w_i b_i (x_i - \bar{x}_r)^2$$

la variabilité des valeurs de y et des valeurs de x dans les ensembles de répondants $s_r^{(y)}$ et $s_r^{(x)}$, respectivement. En notant que, sous l'imputation HDAM pondérée,

$$V_I(y_i^*) = s_{yr}^2, \; V_I(x_i^*) = s_{xr}^2$$

et

$$\mathrm{Cov}_I(y_i^*, x_i^*) = 0,$$

nous pouvons approximer $V_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$ par

$V_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$

$$\approx \frac{1}{\bar{x}_r^2} \left[\sum_{i \in s} w_i^2 (1 - a_i) s_{yr}^2 + \hat{R}_r^2 \sum_{i \in s} w_i^2 (1 - b_i) s_{xr}^2 \right]. \qquad (2.6)$$

Les expressions (2.5) et (2.6) seront utilisées aux sections suivantes lors de la discussion du biais et de la variance de l'estimateur imputé $\hat{R}_I$. Comme nous le verrons aux sections 3 et 4, la variance conditionnelle (2.6) est une mesure de la variabilité due au mécanisme d'imputation.

Nous allons maintenant décrire deux approches d'inférence qui seront utilisées pour obtenir les estimateurs de la variance aux sections 3 et 4, à savoir l'approche du modèle de non-réponse (MN) et l'approche du modèle d'imputation (MI).

2.1 Approche du modèle de non-réponse

Dans l'approche MN, l'inférence est faite par rapport à la distribution conjointe induite par le plan d'échantillonnage et le modèle de non-réponse. Ce dernier est un ensemble d'hypothèses au sujet de la loi inconnue des indicateurs de réponse $\mathbf{R}_s = \{(a_i, b_i); i \in s\}$. Cette loi inconnue est souvent appelée mécanisme de non-réponse. Soit $p_{yi} = P(a_i = 1 \mid s, \mathbf{Z}_s)$, la probabilité de réponse de l'unité i pour la variable y, où $\mathbf{Z}_s = \{\mathbf{z}_i; i \in s\}$ et $\mathbf{z}_i$ est un vecteur de variables auxiliaires disponibles pour toutes les unités de l'échantillon utilisées pour former les classes d'imputation. De même, soit $p_{xi} = P(b_i = 1 \mid s, \mathbf{Z}_s)$ la probabilité de réponse de l'unité i pour la variable x. Nous supposons que les unités répondent indépendamment les unes des autres; c'est-à-dire que $p_{yij} = P(a_i = 1, a_j = 1 \mid s, \mathbf{Z}_s) = p_{yi} p_{yj}$ pour $i \neq j$ et $p_{xij} = P(b_i = 1, b_j = 1 \mid s, \mathbf{Z}_s) = p_{xi} p_{xj}$ pour $i \neq j$. Cependant, nous ne supposons pas que, pour une unité donnée i, la réponse pour la variable y est indépendante de celle pour la variable x. Autrement dit, si nous posons que $p_{xyi} = P(a_i = 1, b_i = 1 \mid s, \mathbf{Z}_s)$, nous avons alors $p_{xyi} \neq p_{xi} p_{yi}$, en général. Dans une classe d'imputation, nous supposons que le mécanisme de réponse est uniforme, de sorte que $p_{yi} = p_y$, $p_{xi} = p_x$ et $p_{xyi} = p_{xy}$.

Nous supposons aussi que, après conditionnement sur s et $\mathbf{Z}_s$, le mécanisme de non-réponse est indépendant de toutes les autres variables qui interviennent dans l'estimateur imputé (2.4), ainsi que des probabilités de sélection conjointe. Autrement dit, la distribution de $\mathbf{R}_s$ ne dépend pas de $\mathbf{Y}_s = \{y_i; i \in s\}$, $\mathbf{W}_s = \{w_i; i \in s\}$ et $\mathbf{\Pi}_s = \{\pi_{ij}; i \in s, j \in s\}$, après conditionnement sur s et $\mathbf{Z}_s$. Par conséquent, sauf pour les indicateurs de réponse a_i et b_i, nous supposons que toutes les variables qui interviennent dans l'estimateur imputé (2.4) ainsi que les probabilités de sélection conjointe sont traitées comme étant fixes lorsque l'on prend en considération les espérances et les variances par rapport au modèle de non-réponse. À partir d'ici, nous utilisons l'indice q pour désigner l'espérance et la variance par rapport au mécanisme de non-réponse.

2.2 Approche du modèle d'imputation

Dans l'approche MI, l'inférence est faite par rapport à la distribution conjointe induite par le modèle d'imputation, le plan d'échantillonnage et le modèle de non-réponse. Le modèle d'imputation est un ensemble d'hypothèses au sujet de la loi inconnue de $(\mathbf{Y}_U, \mathbf{X}_U) = \{(y_i, x_i); i \in U\}$. Dans une classe d'imputation, sous imputation HDAM, le modèle d'imputation, m, est donné par

$$m: \begin{cases} y_i = \mu_y + \varepsilon_i \\ x_i = \mu_x + h_i \end{cases} \quad (2.7)$$

où ε_i est un terme d'erreur aléatoire tel que $E_m(\varepsilon_i) = 0$, $E_m(\varepsilon_i \varepsilon_j) = 0$, pour $i \neq j$, $V_m(\varepsilon_i) = \sigma_\varepsilon^2$ et h_i est un terme d'erreur aléatoire tel que $E_m(h_i) = 0$, $E_m(h_i h_j) = 0$, pour $i \neq j$, $V_m(\varepsilon_i) = \sigma_h^2$. En outre, nous supposons que $E_m(\varepsilon_i h_i) = \sigma_{\varepsilon h}$. Ici, $E_m(.)$, $V_m(.)$ et $\mathrm{Cov}_m(.)$ désignent respectivement les opérateurs d'espérance, de variance et de covariance par rapport au modèle m. La notation tient compte implicitement du fait que les espérances ou les variances par rapport au modèle m sont conditionnelles à $\mathbf{Z}_U = \{\mathbf{z}_i;\ i \in U\}$. Dans cette approche, nous supposons que la distribution des erreurs $(\boldsymbol{\varepsilon}_U, \boldsymbol{\eta}_U) = \{(\varepsilon_i, h_i);\ i \in U\}$ ne dépend pas de s, $s_r^{(y)}$, $s_r^{(x)}$, $\mathbf{W}_U = \{w_i;\ i \in U\}$ ni $\boldsymbol{\Pi}_U = \{\pi_{ij};\ i \in U,\ j \in U\}$, après conditionnement sur $\mathbf{Z}_U$. Par conséquent, sauf les variables d'intérêt y et x, toutes les variables qui interviennent dans l'estimateur imputé (2.4) sont traitées comme étant fixes lorsqu'on considère les espérances et les variances par rapport au modèle d'imputation.

2.3 Biais de l'estimateur imputé

Pour étudier le biais de l'estimateur imputé (2.4), nous utilisons la décomposition standard de l'erreur totale de $\hat{R}_I$:

$$\hat{R}_I - R = (\hat{R} - R) + (E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R}) + (\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})). \quad (2.8)$$

Le premier terme $\hat{R} - R$ du deuxième membre de (2.8) est appelé erreur d'échantillonnage, le deuxième terme $E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R}$ est appelé erreur de non-réponse, tandis que le troisième terme $\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$ est appelé erreur d'imputation.

À l'aide d'un développement en série de Taylor du premier ordre, il est facile de montrer que, sous l'approche MN, l'estimateur imputé (2.4) est asymptotiquement pqI - sans biais, c'est-à-dire que $E_{pqI}(\hat{R}_I - R) \approx 0$. De plus, sous l'approche MI et le modèle (2.7), nous pouvons montrer que l'estimateur imputé (2.4) est asymptotiquement $mpqI$ - sans biais, c'est-à-dire que $E_{mpqI}(\hat{R}_I - R) \approx 0$. Donc, l'estimateur imputé est robuste en ce sens qu'il est valide sous l'approche MN ou sous l'approche MI. Notons que, pour que le biais asymptotique soit nul sous les deux approches, nous imposons que la taille d'échantillon soit suffisamment grande dans chaque classe d'imputation. Dans ce qui suit, nous supposons donc que le biais de $\hat{R}_I$ est négligeable.

3. Estimation de la variance : cadre à deux phases

À la présente section, nous obtenons des estimateurs de la variance sous le cadre à deux phases et l'approche MI, suivant la méthode proposée par Särndal (1992), ainsi que par Deville et Särndal (1994). En utilisant la décomposition (2.8), la variance totale de $\hat{R}_I$ peut être approximée par

$$V_{mpqI}(\hat{R}_I - R) \approx E_{mpqI}(\hat{R}_I - R)^2$$

$$= V_{\mathrm{SAM}} + V_{\mathrm{NR}} + \tilde{V}_I + 2V_{\mathrm{MIX}}, \quad (3.1)$$

où $V_{\mathrm{SAM}} = E_m V_p(\hat{R}) = E_m(V_{\mathrm{SAM}})$ est la variance d'échantillonnage de l'estimateur sous données complètes $\hat{R}$, $V_{\mathrm{NR}} = E_{pq} V_m(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ est la variance de non-réponse de l'estimateur imputé $\hat{R}_I$, $\tilde{V}_I = E_{mpq} V_I(\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) \mid s, s_r^{(y)}, s_r^{(x)})$ est la variance due à l'imputation de l'estimateur imputé $\hat{R}_I$, et $V_{\mathrm{MIX}} = E_{pqm}[(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s, s_r^{(y)}, s_r^{(x)}]$ une composante mixte. Notons que l'expression (3.1) ne contient qu'un seul terme de produit croisé, $2V_{\mathrm{MIX}}$, parce que tous les autres sont asymptotiquement nuls.

3.1 Estimation de la variance d'échantillonnage V_{SAM}

Soit $\hat{V}_{\mathrm{ORD}}$ l'estimateur de variance naïf de $\hat{R}_I$, c'est-à-dire l'estimateur de variance obtenu en traitant les valeurs imputées comme s'il s'agissait de valeurs observées. Nous parvenons à cet estimateur en remplaçant e_i par $\tilde{e}_i = 1/\hat{X}_I(\tilde{y}_i - \hat{R}_I \tilde{x}_i)$ dans (2.2), ce qui donne

$$\hat{V}_{\mathrm{ORD}} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \tilde{e}_i \tilde{e}_j. \quad (3.2)$$

Comme nous le montrons maintenant dans le cas de l'échantillonnage simple sans remise, $\hat{V}_{\mathrm{ORD}}$ surestime V_{SAM} sous l'imputation HDAM quand $\sigma_{\varepsilon h} > 0$ (comme cela est habituellement le cas en pratique). Après certains calculs algébriques, nous obtenons

$$E_{ml}(\hat{V}_{\mathrm{ORD}} - \hat{V}_{\mathrm{SAM}} \mid s, s_r^{(y)}, s_r^{(x)}) \approx \frac{2}{\mu_x^2}\left(\frac{\mu_y}{\mu_x}\right)\left(1 - \frac{n}{N}\right)\left(1 - \frac{r_{xy}}{n}\right)\frac{\sigma_{\varepsilon h}}{n}. \quad (3.3)$$

L'expression (3.3) montre que $\hat{V}_{\mathrm{ORD}}$ est $mpqI$ - biaisé pour $\hat{V}_{\mathrm{SAM}}$, à moins que $\sigma_{\varepsilon h} = 0$, $r_{xy} = n$ (ce qui est le cas des données complètes) ou que $n = N$ (ce qui est le cas d'un recensement). Le fait que $\hat{V}_{\mathrm{ORD}}$ ne soit pas un estimateur valide de V_{SAM} s'explique facilement en notant que, si l'imputation HDAM préserve la variabilité, c'est-à-dire s_x^2 et s_y^2, correspondant aux variables x et y, elle ne

préserve pas la covariance, s_{xy}, dans (2.3). En effet, l'imputation a tendance à sous-estimer la relation entre les variables qui sont positivement corrélées. Donc, $\hat{V}_{\text{ORD}}$ surestime V_{SAM}, à cause de la présence du signe moins devant s_{xy} dans (2.3). Pour surmonter cette difficulté, Särndal (1992) a proposé d'estimer $V_{\text{DIF}} = E_{mI}(\hat{V}_{\text{SAM}} - \hat{V}_{\text{ORD}} \mid s, s_r^{(y)}, s_r^{(x)})$ au moyen d'un estimateur mI - sans biais, $\hat{V}_{\text{DIF}}$; c'est-à-dire $E_{mI}(\hat{V}_{\text{DIF}} \mid s, s_r^{(y)}, s_r^{(x)}) = V_{\text{DIF}}$. Toutefois, le calcul de cette composante pour un plan arbitraire comporte des opérations algébriques fastidieuses dans le cas d'un ratio. Par conséquent, nous proposons une solution de rechange qui ne nécessite aucun calcul, mais comprend la construction d'un nouvel ensemble de valeurs imputées. Elle peut être décrite comme suit : chaque fois que $a_i = 0$ et (ou) $b_i = 0$, on choisit un donneur j aléatoirement avec remise à partir de l'ensemble de répondants aux deux variables y *et* x (c'est-à-dire l'ensemble d'unités échantillonnées pour lesquelles $a_i = 1$ et $b_i = 1$) avec la probabilité $w_j / \sum_{i \in s} w_i a_i b_i$ et on impute le vecteur (x_j, y_j). Autrement dit, chaque fois qu'une réponse manque pour une variable, la valeur observée est écartée et dite manquante; les valeurs manquantes sont ensuite remplacées par les valeurs d'un donneur sélectionné au hasard parmi l'ensemble de répondants aux deux variables x et y (souvent appelé ensemble de donneurs communs). De même, lorsque les valeurs des deux variables manquent, le vecteur (x_j, y_j) d'un donneur j est imputé. Puis, on applique l'estimateur de variance standard (2.2) valide sous réponse complète en utilisant ces valeurs imputées. Soit $\hat{V}_{\text{ORD}}^{*}$ l'estimateur de variance résultant. Notons que ce nouvel ensemble de valeurs imputées est utilisé pour obtenir un estimateur valide de la variance d'échantillonnage, mais ne l'est pas pour estimer le paramètre d'intérêt R. On peut montrer que $\hat{V}_{\text{ORD}}^{*}$ est un estimateur de V_{SAM} asymptotiquement $mpqI$ - sans biais. En pratique, on pourrait, par exemple, créer un fichier d'estimation de la variance contenant le nouvel ensemble de valeurs imputées et utiliser l'un des systèmes standard d'estimation de la variance (employés dans le cas de données complètes) pour obtenir l'estimation de la variance d'échantillonnage.

3.2 Estimation de la variance de non-réponse V_{NR}

Un estimateur $\hat{V}_{\text{NR}}$ de $V_{\text{NR}} = E_{pq} V_m (E_I (\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ peut s'obtenir en estimant simplement $V_m (E_I (\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$. Par un développement en série de Taylor de premier ordre, nous arrivons à

$$V_{\text{NR}} \approx \frac{1}{\bar{x}_I^2}\left[\left(\sum_{i \in s} w_i^2 a_i / \hat{N}_a^2 + \sum_{i \in s} w_i^2 / \hat{N}^2 - 2\sum_{i \in s} w_i^2 a_i / \hat{N}\hat{N}_a\right)\sigma_e^2 + \left(\sum_{i \in s} w_i^2 b_i / \hat{N}_b^2 + \sum_{i \in s} w_i^2 / \hat{N}^2 - 2\sum_{i \in s} w_i^2 b_i / \hat{N}\hat{N}_b\right)\left(\frac{\bar{y}_U}{\bar{x}_U}\right)^2 \sigma_h^2 - 2\left(\sum_{i \in s} w_i^2 a_i b_i / \hat{N}_a \hat{N}_b + \sum_{i \in s} w_i^2 / \hat{N}^2 - \sum_{i \in s} w_i^2 a_i / \hat{N}\hat{N}_a - \sum_{i \in s} w_i^2 b_i / \hat{N}\hat{N}_b\right)\left(\frac{\bar{y}_U}{\bar{x}_U}\right)\sigma_{eh}\right], \quad (3.4)$$

où $(\hat{N}, \hat{N}_a, \hat{N}_b) = \sum_{i \in s} w_i (1, a_i, b_i)$. Maintenant, posons que $s_{Ix}^2 = 1/\hat{N} \sum_{i \in s} w_i (\tilde{x}_i - \bar{x}_I)^2$ et $s_{Iy}^2 = 1/\hat{N} \sum_{i \in s} w_i (\tilde{y}_i - \bar{y}_I)^2$ avec $(\bar{x}_I, \bar{y}_I) = 1/\hat{N} \sum_{i \in s} w_i (\tilde{x}_i, \tilde{y}_i)$. Notons que s_{Ix}^2 et s_{Iy}^2 désignent respectivement la variabilité d'échantillon des valeurs de x et des valeurs de y après l'imputation. Nous pouvons montrer que s_{Ix}^2 et s_{Iy}^2 sont, respectivement, asymptotiquement mI - sans biais pour les variances sous le modèle σ_h^2 et σ_e^2. En outre, posons que $s_{xyr} = 1/\hat{N}_{ab} \sum_{i \in s} w_i a_i b_i (x_i - \bar{x}_{rr})(y_i - \bar{y}_{rr})$, où $\hat{N}_{ab} = \sum_{i \in s} w_i a_i b_i$ et $(\bar{x}_{rr}, \bar{y}_{rr}) = \hat{N}_{ab}^{-1} \sum_{i \in s} w_i a_i b_i (x_i, y_i)$. Notons que s_{xyr} est m - sans biais pour la covariance sous le modèle σ_{eh}. Il s'ensuit que $\hat{V}_{\text{NR}}$ s'obtient par estimation des quantités inconnues dans (3.4), ce qui mène à

$$\hat{V}_{\text{NR}} = \frac{1}{\bar{x}_I^2}\left[\left(\sum_{i \in s} w_i^2 a_i / \hat{N}_a^2 + \sum_{i \in s} w_i^2 / \hat{N}^2 - 2\sum_{i \in s} w_i^2 a_i / \hat{N}\hat{N}_a\right)s_{Iy}^2 + \left(\sum_{i \in s} w_i^2 b_i / \hat{N}_b^2 + \sum_{i \in s} w_i^2 / \hat{N}^2 - 2\sum_{i \in s} w_i^2 b_i / \hat{N}\hat{N}_b\right)\hat{R}_I^2 s_{Ix}^2 - 2\left(\sum_{i \in s} w_i^2 a_i b_i / \hat{N}_a \hat{N}_b + \sum_{i \in s} w_i^2 / \hat{N}^2 - \sum_{i \in s} w_i^2 a_i / \hat{N}\hat{N}_a - \sum_{i \in s} w_i^2 b_i / \hat{N}\hat{N}_b\right)\hat{R}_I s_{xyr}\right]. \quad (3.5)$$

L'estimateur (3.5) est asymptotiquement $mpqI$ - sans biais pour V_{NR}. Dans le cas particulier de l'échantillonnage aléatoire simple sans remise, l'expression (3.5) se réduit à

$$\hat{V}_{\mathrm{NR}} = \frac{1}{\bar{x}_I^2}\left[\left(1 - \frac{r_y}{n}\right)\frac{s_{Iy}^2}{r_y} + \hat{R}_I^2\left(1 - \frac{r_x}{n}\right)\frac{s_{Ix}^2}{r_x} - 2\hat{R}_I\left(1 - \frac{r_x r_y}{n r_{xy}}\right)\frac{r_{xy}}{r_x}\frac{r_{xy}}{r_y}\frac{s_{xyr}}{r_{xy}}\right].$$

3.3 Estimation de la variance d'imputation $\tilde{V}_I$

Un estimateur $\hat{V}_I$ de $\tilde{V}_I = E_{mpq} V_I(\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ peut s'obtenir en estimant simplement $V_I(\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ donnée par (2.6). Un estimateur asymptotiquement I - sans biais de $V_I(\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ est alors donné par

$$\hat{V}_I = \frac{1}{\bar{x}_I^2}\left[\sum_{i\hat{I}s} w_i^2 (1 - a_i) s_{Iy}^2 + \hat{R}_I^2 \sum_{i\hat{I}s} w_i^2 (1 - b_i) s_{Ix}^2\right]. \quad (3.6)$$

Il en découle que $\hat{V}_I$ dans (3.6) est asymptotiquement $mpqI$ - sans biais pour $\tilde{V}_I$. Dans le cas particulier de l'échantillonnage aléatoire simple sans remise, l'expression (3.6) se réduit à

$$\hat{V}_I = \frac{N^2}{\bar{x}_I^2}\left[\left(1 - \frac{r_y}{n}\right)\frac{s_{Iy}^2}{n} + \hat{R}_I^2\left(1 - \frac{r_x}{n}\right)\frac{s_{Ix}^2}{n}\right].$$

3.4 Estimation de la composante mixte V_{MIX}

Enfin, nous obtenons un estimateur $\hat{V}_{\mathrm{MIX}}$ de V_{MIX} en estimant

$$E_m[(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s, s_r^{(y)}, s_r^{(x)}].$$

Grâce à un développement en série de Taylor de premier ordre, nous obtenons

$$E_m[(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s, s_r^{(y)}, s_r^{(x)}]$$

$$\approx \frac{1}{m_x^2}\left[\left(\sum_{i\hat{I}s} \frac{w_i^2 a_i}{\hat{N}\hat{N}_a} - \sum_{i\hat{I}s} \frac{w_i^2}{\hat{N}^2}\right) s_e^2 + \left(\sum_{i\hat{I}s} \frac{w_i^2 b_i}{\hat{N}\hat{N}_b} - \sum_{i\hat{I}s} \frac{w_i^2}{\hat{N}^2}\right)\left(\frac{m_y}{m_x}\right)^2 s_h^2 - 2\left(\sum_{i\hat{I}s} \frac{w_i^2 a_i}{\hat{N}\hat{N}_a} + \sum_{i\hat{I}s} \frac{w_i^2 b_i}{\hat{N}\hat{N}_b} - 2\sum_{i\hat{I}s} \frac{w_i^2}{\hat{N}^2}\right)\frac{m_y}{m_x} s_{eh}\right]. \quad (3.7)$$

Un estimateur de (3.7) est donc donné par

$$\hat{V}_{\mathrm{MIX}} = \frac{1}{\bar{x}_I^2}\left[\left(\sum_{i\hat{I}s} \frac{w_i^2 a_i}{\hat{N}\hat{N}_a} - \sum_{i\hat{I}s} \frac{w_i^2}{\hat{N}^2}\right) s_{Iy}^2 + \left(\sum_{i\hat{I}s} \frac{w_i^2 b_i}{\hat{N}\hat{N}_b} - \sum_{i\hat{I}s} \frac{w_i^2}{\hat{N}^2}\right)\hat{R}_r^2 s_{Ix}^2 - 2\left(\sum_{i\hat{I}s} \frac{w_i^2 a_i}{\hat{N}\hat{N}_a} + \sum_{i\hat{I}s} \frac{w_i^2 b_i}{\hat{N}\hat{N}_b} - 2\sum_{i\hat{I}s} \frac{w_i^2}{\hat{N}^2}\right)\hat{R}_I s_{xyr}\right]. \quad (3.8)$$

L'estimateur (3.8) est asymptotiquement $mpqI$ - sans biais pour V_{MIX}. Dans le cas de l'échantillonnage aléatoire simple sans remise, la composante $\hat{V}_{\mathrm{MIX}}$ est nulle. De façon plus générale, la composante $\hat{V}_{\mathrm{MIX}}$ est nulle pour tout plan de sondage autopondéré à une étape (c'est-à-dire un plan d'échantillonnage pour lequel tous les poids d'échantillonnage sont égaux). Dans le cas de plans d'échantillonnage avec probabilités inégales, il est important d'inclure la composante $\hat{V}_{\mathrm{MIX}}$, parce que sa contribution (positive ou négative) à la variance globale peut être importante (Brick, Kalton et Kim (2004)).

Enfin, un estimateur asymptotiquement $mpqI$ - sans biais de la variance totale $V_{\mathrm{TOT}} = V_{mpqI}(\hat{R}_I - R)$ est donc donné par

$$\hat{V}_{\mathrm{TOT}}^{(\mathrm{TP})} = \hat{V}_{\mathrm{ORD}}^{*} + \hat{V}_{\mathrm{NR}} + \hat{V}_I + 2\hat{V}_{\mathrm{MIX}}.$$

4. Estimation de la variance : cadre inversé

Nous allons maintenant obtenir des estimateurs de la variance sous le cadre inversé et l'approche MN ainsi que MI, selon la méthode proposée par Shao et Steel (1999). Rappelons que, sous ce cadre, nous devons formuler l'hypothèse supplémentaire que les probabilités de réponse ne dépendent pas de l'échantillon s. Sous l'approche MN, la variance totale de $\hat{R}_I$ peut être approximée par

$$V(\hat{R}_I - R) \approx E_q V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + E_{pq} V_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}), \quad (4.1)$$

où $\mathbf{a} = (a_1, \ldots, a_N)'$ et $\mathbf{b} = (b_1, \ldots, b_N)'$ désignent les vecteurs des indicateurs de réponse aux variables y et x, respectivement.

Sous l'approche MI, la variance totale de $\hat{R}_I$ peut être approximée par

$$V(\hat{R}_I - R) \approx E_{mq} V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + E_{mpq} V_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}). \quad (4.2)$$

Sous l'approche MN ainsi que sous l'approche MI, nous pouvons obtenir un estimateur du premier terme du deuxième membre de (4.1) et de (4.2) en trouvant un estimateur asymptotiquement pI - sans biais de $V_p E_I(\hat{R}_I | \mathbf{a}, \mathbf{b})$. En outre, nous pouvons estimer le deuxième terme du deuxième membre de (4.1) et de (4.2) par $\hat{V}_I$ donné par (3.6). Sous l'approche MN, nous pouvons obtenir un estimateur du dernier terme du deuxième membre de (4.1) en estimant $V_q E_{pI}(\hat{R}_I \mid \mathbf{a}, \mathbf{b})$, tandis que sous l'approche MI, nous pouvons obtenir un estimateur du dernier terme du deuxième membre de (4.2) en estimant $V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$. Par conséquent, les estimateurs des deux premiers termes de (4.1) et de (4.2) sont identiques et donc valides, quelle que soit l'approche (MN ou MI) utilisée pour l'inférence. Seul le troisième terme du deuxième membre de (4.1) et de (4.2) dépendra de l'approche choisie. Dans le cas de l'approche MI, la spécification et la validation du modèle d'imputation sont essentielles à l'obtention de l'absence de biais de la troisième composante, tandis que dans l'approche MN, l'absence de biais de la troisième composante dépend de la spécification correcte du modèle de non-réponse.

4.1 Estimation de $V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$

Partant d'un développement en série de Taylor du premier ordre et de l'expression (2.5), un estimateur de $V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$, dénoté par $\hat{V}_1$, est donné par

$$\hat{V}_1 = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \kappa_i \kappa_j, \tag{4.3}$$

où

$$\kappa_i = \frac{1}{\bar{x}_r}\left[\frac{1}{\hat{N}_a} a_i (y_i - \bar{y}_r) - \left(\frac{\bar{y}_r}{\bar{x}_r}\right)\frac{1}{\hat{N}_b} b_i (x_i - \bar{x}_r)\right].$$

Autrement dit, l'estimateur $\hat{V}_1$ est obtenu à partir de l'estimateur de variance sous données complètes (2.2) en remplaçant e_i par κ_i. Dans le cas de l'échantillonnage aléatoire simple sans remise, l'estimateur (4.3) se réduit à

$$\hat{V}_1 = \left(1 - \frac{n}{N}\right)\frac{1}{\bar{x}_r^2}\left[s_{yr}^2 + \hat{R}_r^2 \frac{s_{xr}^2}{r_x} - 2\hat{R}_r\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right]. \tag{4.4}$$

4.2 Estimation de $V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ sous l'approche MN

Pour commencer, notons

$$E_{pI}(\hat{R}_I) \approx \frac{Y_a N_b}{N_a X_b},$$

où $(Y_a, N_a) = \sum_{i \in U} a_i (y_i, 1)$ et $(X_b, N_b) = \sum_{i \in U} b_i (x_i, 1)$.

Par un développement en série de Taylor de premier ordre, nous pouvons montrer que $V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ peut être approximé par

$$V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) \approx \frac{1}{N\bar{X}^2}\left[\left(\frac{1}{p_y} - 1\right) S_y^2 + R^2\left(\frac{1}{p_x} - 1\right) S_x^2 - 2R\left(\frac{p_{xy}}{p_x p_y} - 1\right) S_{xy}\right], \tag{4.5}$$

où

$$S_y^2 = \frac{1}{N-1}\sum_{i \in U} (y_i - \bar{Y})^2,$$

$$S_x^2 = \frac{1}{N-1}\sum_{i \in U} (x_i - \bar{X})^2$$

et

$$S_{xy} = \frac{1}{N-1}\sum_{i \in U} (x_i - \bar{X})(y_i - \bar{Y})$$

avec

$$(\bar{Y}, \bar{X}) = \frac{1}{N}\sum_{i \in U} (y_i, x_i).$$

Nous obtenons un estimateur de $V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ en estimant les quantités inconnues dans (4.5), ce qui donne

$$\begin{aligned}\hat{V}_2^{(\mathrm{NM})} &= \frac{1}{\hat{N}\bar{x}_I^2}\left[\left(\frac{1}{\hat{p}_y} - 1\right) s_{Iy}^2 + \hat{R}_I^2\left(\frac{1}{\hat{p}_x} - 1\right) s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{p}_{xy}}{\hat{p}_x\hat{p}_y} - 1\right) s_{xyr}\right] \\ &= \frac{1}{\bar{x}_I^2}\left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}}\right) s_{Iy}^2 + \hat{R}_I^2\left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}}\right) s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{N}_{ab}}{\hat{N}_b\hat{N}_b} - \frac{1}{N}\right) s_{xyr}\right],\end{aligned} \tag{4.6}$$

où $\hat{p}_y = \dfrac{\hat{N}_a}{\hat{N}}$, $\hat{p}_x = \dfrac{\hat{N}_b}{\hat{N}}$ et $\hat{p}_{xy} = \dfrac{\hat{N}_{ab}}{\hat{N}}$.

L'estimateur (4.6) est asymptotiquement pqI - sans biais pour la variance approximative (4.5), en notant que s_{Iy}^2, s_{Ix}^2 et s_{xyr} sont asymptotiquement pqI - sans biais pour S_y^2, S_x^2 et S_{xy}, respectivement. Dans le cas de l'échantillonnage aléatoire simple sans remise, l'estimateur (4.6) se réduit à

$$\hat{V}_2^{(\mathrm{NM})} = \frac{1}{\bar{x}_I^2}\left[\left(\frac{n}{N} - \frac{r_y}{N}\right)\frac{s_{Iy}^2}{r_y} + \hat{R}_I^2\left(\frac{n}{N} - \frac{r_x}{N}\right)\frac{s_{Ix}^2}{r_x}\right.$$

$$\left. - 2\hat{R}_I\left(\frac{n}{N} - \frac{r_x r_y}{N r_{xy}}\right)\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right]. \tag{4.7}$$

4.3 Estimation de $E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ sous l'approche MI

Par un développement en série de Taylor de premier ordre, nous pouvons montrer que $E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ peut être approximé par

$$E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) \approx \frac{1}{\bar{x}_I^2}\left[\left(\frac{1}{E_q(N_a)} - \frac{1}{N}\right)s_e^2 + \left(\frac{\bar{y}}{\bar{x}_I}\right)^2\left(\frac{1}{E_q(N_b)} - \frac{1}{N}\right)s_h^2 - 2\left(\frac{\bar{y}}{\bar{x}_I}\right)\left(\frac{N_{ab}}{E_q N_b N_b} - \frac{1}{N}\right)s_{eh}\right], \quad (4.8)$$

où $N_{ab} = \sum_{i \in U} a_i b_i$. Nous obtenons un estimateur de $E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ en estimant les quantités inconnues dans (4.8), ce qui mène à

$$\hat{V}_2^{(\mathrm{IM})} = \frac{1}{\bar{x}_I^2}\left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}}\right)s_{Iy}^2 + \hat{R}_I^2\left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}}\right)s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{N}_{ab}}{\hat{N}_b\hat{N}_b} - \frac{1}{N}\right)s_{xyr}\right]. \quad (4.9)$$

L'estimateur (4.9) est asymptotiquement *mpqI* - sans biais pour la variance approximative (4.8). Il est intéressant de souligner que, sous l'imputation HDAM pondérée, l'estimateur $\hat{V}_2^{(\mathrm{NM})}$ dans (4.6) obtenu sous l'approche MN est identique à l'estimateur $\hat{V}_2^{(\mathrm{IM})}$ dans (4.9) obtenu sous l'approche MI. Toutefois, il pourrait ne pas en être ainsi sous une méthode d'imputation différente. En outre, la composante $\hat{V}_2$ est négligeable par rapport à $\hat{V}_1$ quand la fraction d'échantillonnage n/N est négligeable, où $\hat{V}_2$ représente $\hat{V}_2^{(\mathrm{NM})}$ ou $\hat{V}_2^{(\mathrm{IM})}$. Dans ce cas, la composante $\hat{V}_2$ peut être omise des calculs.

Enfin, un estimateur de la variance totale sous le cadre inversé est donné par

$$\hat{V}_{\mathrm{TOT}}^{(\mathrm{RE})} = \hat{V}_1 + \hat{V}_I + \hat{V}_2.$$

Sous le cadre inversé, les approches MN et MI mènent toutes deux au même estimateur de la variance totale. Donc, l'estimateur de variance $\hat{V}_{\mathrm{TOT}}^{(\mathrm{RE})}$ est robuste en ce sens qu'il est valide sous l'approche MN ou MI.

5. Résumé et conclusion

Dans le présent article, nous avons obtenu des estimateurs de variance pour l'estimateur imputé d'un ratio sous deux cadres différents. Le cadre inversé facilite la détermination des expressions de la variance (comparativement au cadre à deux phases habituel), particulièrement si la fraction d'échantillonnage est faible, auquel cas nous pouvons omettre la composante $\hat{V}_2$. Toutefois, contrairement au cadre à deux phases, il nécessite l'hypothèse supplémentaire voulant que les probabilités de réponse ne dépendent pas de l'échantillon réalisé s. De plus, le cadre à deux phases s'appuie sur une décomposition naturelle de l'erreur totale qui aboutit à une décomposition naturelle de la variance totale. Par conséquent, cette dernière peut être exprimée comme la somme de la variance d'échantillonnage, de la variance de non-réponse et de la variance d'imputation, ce qui permet au statisticien d'enquête de se faire une idée de l'importance relative de chaque composante. Sous l'approche inverse, il n'existe aucune interprétation simple des composantes de la variance (sauf la variance d'imputation).

Nous avons considéré le cas de l'imputation HDAM pondérée dans les classes. Une autre version de l'imputation hot deck aléatoire pondérée, que nous appelons imputation hot deck aléatoire conjointe (HDAC) pondérée est identique à l'imputation HDAM pondérée, excepté que, lorsque les données manquent pour les deux variables, un donneur j est sélectionné au hasard parmi l'ensemble de donneurs communs (c'est-à-dire l'ensemble de répondants aux deux variables y et x) avec la probabilité $w_j / \sum_{l \in s} w_l a_l b_l$ et le vecteur (x_j, y_j) est imputé. Cette version de la méthode aide à préserver les relations entre les variables étudiées, contrairement à l'imputation indépendante de chaque variable. Les résultats de l'imputation HDAC peuvent être obtenus en utilisant des techniques comparables à celles présentées ici. Enfin, les résultats exposés dans le présent article peuvent être généralisés facilement à l'imputation par la régression tant déterministe qu'aléatoire dans les classes d'imputation.

Remerciements

L'auteur remercie un Éditeur Associé de ses suggestions et commentaires constructifs qui ont permis d'améliorer la qualité de l'article. Les travaux de recherché de David Haziza ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

Bibliographie

Brick, M.J., Kalton, G. et Kim, J.K. (2004). Estimation de variance pour l'imputation hot deck à l'aide d'un modèle. *Techniques d'enquête*, 30, 63-72.

Deville, J.-C., et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.