Techniques d'enquête, juin 2001
Vol. 27, N° 1, pp. 81-90
Statistique Canada, N° 12-001 au catalogue

Estimation de la variance après imputation

Jae-Kwang Kim [1]

Résumé

On recourt fréquemment à l'imputation pour compenser la non-réponse partielle. L'estimation de la variance après imputation a suscité de nombreux débats et plusieurs estimateurs ont été proposés. Nous proposons un estimateur de la variance fondé sur un ensemble de pseudo-données créé uniquement pour estimer la variance. L'application des estimateurs type de la variance de données complètes à l'ensemble de pseudo-données produit des estimateurs cohérents dans le cas des estimateurs linéaires pour diverses méthodes d'imputation, y compris l'imputation par la méthode hot deck sans remise et avec remise. Nous illustrons l'équivalence asymptotique de la méthode proposée et de la méthode corrigée du jackknife de Rao et Sitter (1995). La méthode proposée s'applique directement à l'estimation de la variance en cas d'échantillonnage à deux phases.

Mots clés : Échantillonnage à deux phases; non-réponse partielle; imputation déterministe; imputation aléatoire.

1. Introduction

L'imputation, c'est-à-dire l'insertion de valeurs pour remplacer les réponses manquantes, est utilisée fréquemment lorsque des données d'enquête manquent. L'un des avantages de l'imputation est sa commodité. Autrement dit, nous pouvons appliquer aux ensembles contenant des données imputées des programmes types conçus pour des données complètes en vue de calculer des estimations ponctuelles. Rubin (1996), Fay (1996) et Rao (1996) ont passé en revue divers problèmes posés par l'imputation.

Toute méthode d'imputation s'appuie sur un modèle. Après avoir précisé ce dernier, nous pouvons recourir à l'imputation déterministe ou à l'imputation aléatoire selon le modèle. Dans le cas de l'imputation aléatoire, on procède à une forme d'échantillonnage probabiliste en vue d'imputer des données pour remplacer les valeurs manquantes. Nous donnons à ce mécanisme aléatoire supplémentaire le nom de mécanisme d'imputation. L'imputation déterministe, quant à elle, ne nécessite pas l'ajout d'un mécanisme aléatoire supplémentaire. Si nous considérons l'ensemble de répondants comme un échantillon aléatoire de l'échantillon original, nous donnons au mécanisme de sélection des répondants le nom de mécanisme de réponse. Ce dernier est souvent considéré comme étant la deuxième phase d'échantillonnage. Pour plus de précisions, consulter Särndal et Swensson (1987).

Le choix d'un modèle et d'une méthode d'imputation appropriés permet de réduire considérablement le biais dû à la non-réponse qui entache les résultats lorsque l'on se sert uniquement des données observées. Cependant, il est reconnu qu'un estimateur de la variance qui traite les données imputées comme s'il s'agissait de données observées est incohérent.

Diverses méthodes ont été proposées pour estimer la variance après l'imputation. Rubin et Schenker (1986) et Rubin (1987) recommandent l'imputation multiple qui crée plusieurs ensembles de données et produit des statistiques complètes pour chaque ensemble contenant des données imputées. L'estimateur de la variance est calculé par combinaison de deux termes, à savoir la variance dans les ensembles de données et la variance entre les ensembles de données. La méthode d'imputation multiple consiste à appliquer un estimateur type de la variance à chaque ensemble de données pour calculer les termes de variance à l'intérieur de l'ensemble de données et des estimateurs ponctuels type pour calculer la variance entre ensembles contenant des données imputées. L'application de cette méthode dépend du choix d'une méthode d'imputation appropriée. Autrement dit, l'imputation devrait remplir les conditions 1 à 3 décrites par Rubin (1987, pages 118 et 119). Ces conditions ne sont pas toujours faciles à satisfaire. (Par exemple, voir Fay 1992.) Même dans Schafer (1997), il n'est pas montré que les méthodes d'imputation multiples décrites sont appropriées au sens de Rubin. Comme le fait remarquer Rao (1996), certaines méthodes d'imputation utilisées couramment, y compris l'imputation par la méthode hot deck et l'imputation par régression, ne conviennent pas.

Rao et Shao (1992) et Rao et Sitter (1995) ont proposé un estimateur jackknife corrigé de la variance. La méthode proposée est applicable à plusieurs méthodes d'imputation et à plusieurs plans d'échantillonnage. Le calcul réel, au moyen d'un logiciel standard applicable à des données complètes n'est pas facile, car des calculs spéciaux sont exécutés pour corriger les valeurs imputées pour chaque pseudo-répétition. En outre, Särndal (1992) a proposé une méthode d'estimation de la variance qui tient compte explicitement du modèle utilisé pour l'imputation.

Essentiellement, la méthode de Rubin produit plusieurs ensembles de pseudo-données pour l'estimation de la variance et consiste à appliquer un estimateur type de la

1. Jae-Kwang Kim, Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, États-Unis.

variance à chaque ensemble de données pour calculer le terme de variance à l'intérieur de l'ensemble de données, tandis que la méthode de Rao et celle de Särndal consiste à appliquer un estimateur particulier de la variance à l'ensemble contenant des données imputées. Ici, nous proposons une méthode pour créer un ensemble de pseudo-données unique pour l'estimation de la variance. À la section 2, nous présentons la nouvelle méthode dans le cadre de l'échantillonnage à deux phases. À la section 3, nous illustrons l'extension de la méthode proposée à la méthode d'imputation aléatoire. À la section 4, nous étendons la méthode proposée aux plans d'échantillonnage complexes. A la section 5, nous comparons les résultats à ceux de l'estimateur jackknife corrigé de la variance. À la section 6, nous présentons une étude en simulation limitée. Enfin, nous présentons certaines conclusions à la section 7. L'exposé de certaines preuves figure en annexe.

2. Une méthode d'estimation de la variance

Nous décrivons une méthode d'estimation de la variance applicable aux échantillons à deux phases et aux échantillons contenant des données imputées. En plus de l'ensemble de données totalisées, la méthode nécessite un ensemble distinct de données pour l'estimation de la variance. Pour présenter la méthode et illustrer les concepts, considérons un échantillon à deux phases. Supposons que la deuxième phase consiste en un échantillon aléatoire simple de taille r sélectionné à partir de l'échantillon de première phase, qui est un échantillon aléatoire simple de taille n sélectionné à partir d'une population infinie. Supposons que l'estimateur par régression de la moyenne d'une caractéristique y est

$$\hat{\mu}_y = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)\hat{\beta}, \tag{1}$$

où

$$(\bar{y}_2, \bar{x}_2) = r^{-1}\sum_{i=1}^{r}(y_i, x_i),$$

$$\bar{x}_1 = n^{-1}\sum_{i=1}^{n} x_i,$$

$$\hat{\beta} = \left[\sum_{i=1}^{r}(x_i - \bar{x}_2)^2\right]^{-1}\sum_{i=1}^{r}(x_i - \bar{x}_2)(y_i - \bar{y}_2)$$

et que les unités de deuxième phase sont marquées d'un indice dont la valeur varie de un à r. Il est bien connu (par exemple, Cochran 1977, équation 12.51) que la variance de l'estimateur par régression est, approximativement,

$$V\{\hat{\mu}_y\} = [n^{-1}\rho^2 + r^{-1}(1-\rho^2)]\sigma_y^2, \tag{2}$$

où ρ est la corrélation de population entre y et x et où σ_y^2 est la variance de population de y. Selon la théorie classique de la régression, un estimateur de la variance prend la forme

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2 + r^{-1}(r-2)^{-1}\sum_{i=1}^{r}(y_i - \hat{y}_i)^2 \tag{3}$$

où $\hat{y}_i = \bar{y}_2 + (x_i - \bar{x}_2)\hat{\beta}$ pour $i = 1, 2, \ldots, n$, et $\bar{y}_I = n^{-1}\sum_{i=1}^{n}\hat{y}_i$. Notons que $\bar{y}_I$ est une autre façon d'écrire $\hat{\mu}_y$ dans (1).

Supposons que

$$c_r = [n(n-1)r^{-1}(r-2)^{-1}]^{1/2} \tag{4}$$

et

$$y_i^* = \begin{cases} \hat{y}_i, & i = r+1, r+2, \ldots, n \\ \hat{y}_i + c_r(y_i - \hat{y}_i), & i = 1, 2, \ldots, r. \end{cases} \tag{5}$$

Alors,

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(y_i^* - \bar{y}_I)^2 \tag{6}$$

où $\bar{y}_I$ représente la moyenne des y_i^*, ainsi que la moyenne des $\hat{y}_i$, parce que la somme des termes $y_i - \hat{y}_i$ est nulle. L'équation (6) est la forme opérationnelle de l'estimateur proposé. L'ensemble de données pour l'estimation de la variance contient la pseudo-observation y_i^*.

Dans la mesure où le modèle utilisé pour l'imputation correspond au modèle d'échantillonnage à deux phases, l'équation (6) peut être appliquée à un ensemble contenant des données imputées. Par exemple, si nous supposons que les données manquantes le sont au hasard et que nous utilisons la régression pour imputer les valeurs manquantes de $\hat{y}_i$, alors l'équation (6) est applicable immédiatement. Naturellement, l'imputation par régression ou l'échantillonnage à deux phases peuvent comprendre l'utilisation d'un vecteur x.

3. Extensions à l'imputation aléatoire

Une extension modérée de la méthode décrite à la section 2 nous permet d'estimer la variance d'une moyenne d'échantillon par imputation aléatoire. En fait, d'autres méthodes sont possibles.

Par exemple, supposons que le modèle d'imputation est le modèle de regression

$$y_i = \mathbf{x}_i\beta + e_i \tag{7}$$

où le premier élément de chaque $\mathbf{x}_i$ est égal à 1 et où les e_i sont des variables aléatoires non corrélées $(0, \sigma_e^2)$.

Supposons que le modèle est estimé et que les valeurs imputées sont données par

$$\ddot{y}_i = \hat{y}_i + \ddot{e}_i, \quad i = r+1, r+2, ..., n \tag{8}$$

où $\hat{y}_i = \mathbf{x}_i \hat{\beta}$ avec $\hat{\beta} = (\sum_{i=1}^{r} \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_{i=1}^{r} \mathbf{x}_i' y_i$ et où $\ddot{e}_i$ est choisi au hasard dans l'ensemble $\hat{\mathbf{e}}_r = \{\hat{e}_i = y_i - \hat{y}_i; i = 1, 2, ..., r\}$. L'estimateur de la moyenne y est

$$\hat{\mu}_y = n^{-1} \sum_{i=1}^{n} \ddot{y}_i \tag{9}$$

où $\ddot{y}_i = y_i$ si $i = 1, 2, ..., r$.

Si nous choisissons les $\ddot{e}_i$ avec remise et avec probabilité égale à partir de l'ensemble $\hat{\mathbf{e}}_r$, alors la variance $\hat{\mu}_y$ est donnée, approximativement, par

$$V\{\hat{\mu}_y\} = [n^{-1}R^2 + (r^{-1} + n^{-2}m)(1 - R^2)]\sigma_y^2, \tag{10}$$

où $m = n - r$ et où R^2 représente le carré du coefficient de corrélation multiple entre y et $\mathbf{x}$. L'augmentation de la variance due à l'utilisation de l'imputation aléatoire avec $\ddot{e}_i$, plutôt que l'utilisation de $\ddot{e}_i \equiv 0$, est $n^{-2}m(1 - R^2)\sigma_y^2$.

Par conséquent, un estimateur de la variance de la moyenne d'échantillon imputée est donné par (6) où le c_r de (4) est donné par

$$c_I = [n(n-1)(r^{-1} + n^{-2}m)(r-p)^{-1}]^{1/2}, \tag{11}$$

et p représente la dimension de β. Nous avons

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_I)^2 + (r^{-1} + n^{-2}m)(r-p)^{-1} \sum_{i=1}^{r} (y_i - \hat{y}_i)^2 \tag{12}$$

où $\bar{y}_I = \sum_{i=1}^{n} \hat{y}_i$. L'estimateur de la variance calculé en utilisant c_I de l'équation (11) est un estimateur de la variance non conditionnelle, c'est-à-dire la moyenne sur tous les échantillons possibles contenant des données imputées. Les calculs de (10) et (12) figurent à l'annexe A.

Pour considérer une autre méthode d'estimation de la variance, nous supposons que l'on applique une méthode de sélection aléatoire pour l'imputation, mais n'imposons aucune contrainte, si ce n'est que les probabilités de sélection sont inversement proportionnelles à la probabilité que la valeur y réponde. En outre, nous enregistrons le nombre de fois qu'une valeur $\hat{e}$ est utilisée comme donneur dans l'imputation.

Supposons que

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, ..., n \\ \hat{y}_i + c_r(y_i - \hat{y}_i) & i = 1, 2, ..., r \end{cases} \tag{13}$$

et que

$$c_r = [n^{-1}(n-1)r(r-p)^{-1}]^{1/2}(1 + d_i) \tag{14}$$

où d_i est le nombre de fois que $\hat{e}_i$ est utilisé comme donneur. Le terme $[n^{-1}(n-1)r(r-p)^{-1}]^{1/2}$ est utilisé pour tenir compte de l'effet de l'estimation de p paramètres de régression. Alors, l'estimateur de la variance (6) peut être représenté par

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_I)^2 + n^{-2}r(r-p)^{-1} \sum_{i=1}^{r} (1 + d_i)^2 (y_i - \hat{y}_i)^2. \tag{15}$$

Si la méthode d'imputation est une méthode d'échantillonnage aléatoire simple avec remise, alors, subordonnée à l'échantillon et aux répondants,

$$E_I\{(1 + d_i)^2\} = \left(\frac{n}{r}\right)^2 + \frac{m}{r}\left(1 - \frac{1}{r}\right) \tag{16}$$

où nous utilisons la notation I pour représenter l'espérance en rapport avec le mécanisme d'imputation produit par imputation aléatoire. L'égalité dans (16) établit l'équivalence de (12) et (15) en cas de sélection avec remise. Nous montrons à l'annexe B que $\hat{V}\{\hat{\mu}_y\}$ donné par l'équation (15) est aussi un estimateur valide si les donneurs sont sélectionnés sans remise. Puisque la méthode proposée d'estimation de la variance produit la variance conditionnelle, étant donné l'échantillon imputé réalisé, son applicabilité est très grande.

4. Plans de sondage complexes

4.1 Imputation déterministe

La méthode proposée est applicable à des plans de sondage complexes, ainsi qu'à l'échantillonnage aléatoire simple. Supposons que l'estimateur pour l'échantillon complet de la moyenne de y peut être représenté par $\bar{y} = \sum_{i=1}^{n} w_i y_i$, où w_i est le poids d'échantillonnage de l'unité i dans l'échantillon. Supposons aussi que $\sum_{i=1}^{n} w_i = 1$.

Si les r premiers éléments sont observés et que les $n - r$ autres éléments manquent, alors l'estimateur de la moyenne de y dans des conditions d'imputation par régression est

$$\bar{y}_I = \sum_{i=1}^{r} w_i y_i + \sum_{i=r+1}^{n} w_i \hat{y}_i \tag{17}$$

où

$$\hat{y}_i = \mathbf{x}_i \hat{\beta},$$

$$\hat{\beta} = \left[\sum_{i=1}^{r} w_i^* \mathbf{x}_i' \mathbf{x}_i\right]^{-1} \sum_{i=1}^{r} w_i^* \mathbf{x}_i' y_i.$$

Ici, w_i^* représente le poids d'échantillonnage de l'unité i dans l'échantillon de deuxième phase et est défini par

$$w_i^* = \begin{bmatrix} \Pr(i \text{ soit dans l'échantillon de deuxième phase} \\ |i \text{ soit dans l'échantillon de première phase}) \end{bmatrix}^{-1} w_i.$$

En outre, $\sum_{i=1}^r w_i^* = 1$. Si nous supposons que l'échantillon de deuxième phase est un échantillon aléatoire de taille r sélectionné à partir de l'échantillon de première phase n, alors $w_i^* = nr^{-1}w_i$. Dans certaines conditions, nous pouvons écrire l'estimateur donné par (17) sous la forme

$$\bar{y}_I = \sum_{i=1}^n w_i \hat{y}_i. \tag{18}$$

La représentation (18) est vérifiée si $(w_i^*)^{-1}w_i$ se trouve dans l'espace colonne de la matrice $\mathbf{X} = (\mathbf{x}_1', ..., \mathbf{x}_r')'$ parce qu'alors, nous avons $\sum_{i=1}^r w_i(y_i - \hat{y}_i) = 0$ provenant de $\sum_{i=1}^r w_i^* \mathbf{x}_i'(y_i - \hat{y}_i) = \mathbf{0}$.

Nous supposons une série d'échantillons et de populations finies telle que celle décrite dans Fuller (1998). Définissons $\bar{\mathbf{x}}_1 = \sum_{i=1}^n w_i \mathbf{x}_i$ et $(\bar{\mathbf{x}}_2, \bar{y}_2) = \sum_{i=1}^r w_i^*(\mathbf{x}_i, y_i)$. Nous adoptons aussi les mêmes hypothèses que dans Fuller (1998). C'est-à-dire

$$E(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2) = (\mu_x, \mu_x, \mu_y), \tag{19}$$

et

$$V\{(\hat{\beta} - \beta)', \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2\} = O(n^{-1}), \tag{20}$$

où $(\mu_{\mathbf{x}}, \mu_y) = N^{-1}\sum_{i=1}^N (\mathbf{x}_i, y_i)$ et $\beta = (\sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_{i=1}^N \mathbf{x}_i' y_i$.

Pour $i = 1, 2, ..., N$, définissons

$$a_i = \begin{cases} 1 & \text{si l'unité } i \text{ répond à l'échantillonnage} \\ 0 & \text{autrement,} \end{cases}$$

et $\mathbf{a} = (a_1, a_2, ..., a_N)$. La définition étendue de a_i est dictée par Fay (1991) et utilisée par Shao et Steel (1999). Maintenant, supposons que

$$\bar{y}_{II} = \sum_{i=1}^n w_i \tilde{y}_i^* \tag{21}$$

où

$$\tilde{y}_i^* = \tilde{y}_i + a_i w_i^{-1} w_i^*(y_i - \tilde{y}_i) \tag{22}$$

avec $\tilde{y}_i = \mathbf{x}_i\beta$. Alors, nous avons $\bar{y}_I = \bar{y}_{II} + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\hat{\beta} - \beta)$. Avec (19) et (20), nous avons $\bar{y}_I = \bar{y}_{II} + O_p(n^{-1})$ et $V(\bar{y}_I - \bar{Y}_N) = V(\bar{y}_{II} - \bar{Y}_N) + o(n^{-1})$. Maintenant,

$$V(\bar{y}_{II} - \bar{Y}_N) = V[E(\bar{y}_{II} - \bar{Y}_N|\mathbf{a})] + E[V(\bar{y}_{II} - \bar{Y}_N)|\mathbf{a}]. \tag{23}$$

Le premier terme du deuxième membre de (23) est nul, car $E(\bar{y}_{II} - \bar{Y}_N|\mathbf{a}) = 0$ en vertu du modèle (7). Pour estimer le deuxième terme de (23), notons que, subordonné à $\mathbf{a}$, $\bar{y}_{II}$ est un estimateur linéaire. Donc, la méthode standard d'estimation de la variance appliquée à l'ensemble de pseudo-données $\tilde{\mathbf{Y}}^* \equiv \{\tilde{y}_i^*; i = 1, 2, ..., n\}$ produira une estimation non biaisée de la variance de $\bar{y}_{II} = \sum_{i=1}^n w_i \tilde{y}_i^*$. Puisque l'ensemble $\tilde{\mathbf{Y}}^*$ ne peut être observé, nous pouvons utiliser l'ensemble $\mathbf{Y}^* \equiv \{y_i^*; i = 1, 2, ..., n\}$, où

$$y_i^* = \hat{y}_i + a_i w_i^{-1} w_i^*(y_i - \hat{y}_i) \tag{24}$$

pour obtenir un estimateur de la variance cohérent.

Pour montrer que nous pouvons utiliser l'ensemble $\mathbf{Y}^*$ pour obtenir un estimateur approximatif de la variance, supposons que l'estimateur de la variance de l'échantillon complet de $\bar{y}$ peut être représenté par

$$\hat{V} = \sum_{i=1}^L c_i(\bar{y}^{(i)} - \bar{y})^2$$

où L est le nombre de répétitions, c_i est le i^{e} facteur de répétition et $\bar{y}^{(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j$ est la i^{e} répétition de $\bar{y}$. Le terme $M_j^{(i)}$ est le multiplicateur de répétition appliqué au poids de l'unité j lors de la i^{e} répétition. Par exemple, en cas d'échantillonnage aléatoire simple, le multiplicateur jackknife est

$$M_j^{(i)} = \begin{cases} (n-1)^{-1}n & \text{si } i \neq j \\ 0 & \text{si } i = j. \end{cases}$$

Supposons que nous appliquons l'estimateur répété de la variance $\hat{V}$ à l'ensemble $\mathbf{Y}^*$ pour obtenir

$$\hat{V}^* = \sum_{i=1}^L c_i(\bar{y}_I^{*(i)} - \bar{y}_I)^2$$

où $\bar{y}_I^{*(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j^*$ avec y_j^* définit dans (24). Alors, nous obtenons

$$\bar{y}_I^{*(i)} - \bar{y}_I = \bar{y}_{II}^{*(i)} - \bar{y}_{II} + (\bar{\mathbf{x}}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)} - \bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)(\hat{\beta} - \beta) \tag{25}$$

où

$$(\bar{\mathbf{x}}_1^{(i)}, \bar{\mathbf{x}}_2^{(i)}) = \sum_{j=1}^n w_j M_j^{(i)}(\mathbf{x}_j, a_j w_j^{-1} w_j^* \mathbf{x}_j).$$

Nous montrons à l'annexe C que,

$$\hat{V}^* = \sum_{i=1}^L c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})^2 + o_p(n^{-1}). \tag{26}$$

Par conséquent, nous pouvons utiliser l'estimateur jackknife type de la variance appliqué à l'ensemble de pseudo-données $\mathbf{Y}^*$ comme approximation de l'estimateur jackknife type de la variance appliqué à l'ensemble de pseudo-données $\tilde{\mathbf{Y}}^*$.

4.2 Imputation aléatoire

Les arguments concernant l'estimation de la variance en cas d'imputation aléatoire sont fort semblables à ceux

décrits pour l'imputation déterministe à la sous-section qui précède. En premier lieu, définissons la fonction indicatrice d'imputation

$$d_{ij} = \begin{cases} 1 & \text{si l'unité } i \text{ est utilisée comme donneur pour l'unité } j \\ 0 & \text{autrement.} \end{cases} \tag{27}$$

Alors, l'estimateur de la moyenne de y en recourant à l'imputation aléatoire est

$$\bar{y}_I = \sum_{i=1}^{n} w_i y_i^* \tag{28}$$

où

$$\bar{y}_i^* = \hat{y}_i + a_i(1+d_i)(y_i - \hat{y}_i) \tag{29}$$

et

$$d_i = \sum_{j=1}^{n} (1-a_j) d_{ij} w_i^{-1} w_j. \tag{30}$$

Si les poids d'échantillonnage originaux sont les mêmes, alors d_i représente le nombre de fois que l'unité i est utilisée comme donneur. Nous supposons que

$$E[a_i(1+d_i)|F_1] = 1 \tag{31}$$

où $F_1 = \{(i, \mathbf{x}_i, y_i); i = 1, 2, ..., n\}$. Dans (31), l'espérance a trait à la distribution combinée du mécanisme de réponse et du mécanisme d'imputation. Alors, nous avons

$$E(\bar{y}_I | F_1) \doteq \bar{y}.$$

Si nous supposons que les probabilités de réponse sont égales, alors, en vertu de l'hypothèse (31), la probabilité de sélection des donneurs devrait être proportionnelle aux poids, ce qui correspond aux conditions établies par Rao et Shao (1992) pour l'imputation aléatoire.

Maintenant, supposons que

$$\bar{y}_{II} = \sum_{i=1}^{n} w_i [\tilde{y}_i + a_i(1+d_i)(y_i - \tilde{y}_i)] \tag{32}$$

où $\tilde{y}_i = \mathbf{x}_i \beta$. Alors, nous avons aussi $\bar{y}_I = \bar{y}_{II} + (\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1)(\hat{\beta} - \beta)$ où $\bar{\mathbf{x}}_d = \sum_{i=1}^{n} w_i a_i (1+d_i) \mathbf{x}_i$. En vertu de l'hypothèse (31), nous avons $E(\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 | F_1) = 0$. Si les contraintes sont peu sévères, $\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 = O_p(n^{-1/2})$ et $\bar{y}_I = \bar{y}_{II} + O_p(n^{-1})$. Maintenant,

$$V(\bar{y}_I - \bar{Y}_N) = V[E(\bar{y}_I - \bar{Y}_N | \mathbf{a}, \mathbf{d})] + E[V(\bar{y}_I - \bar{Y}_N | \mathbf{a}, \mathbf{d})]$$

où $\mathbf{d} = (d_1, d_2, ..., d_N)$. Subordonné à $\mathbf{a}$ et à $\mathbf{d}$, l'estimateur $\bar{y}_{II}$ est linéaire. Donc, nous pouvons utiliser les pseudo-données

$$y_i^* = \hat{y}_i + a_i(1+d_i)(y_i - \hat{y}_i) \tag{33}$$

pour estimer la variance de $\bar{y}_I$.

5. Comparaison à la méthode corrigée du jackknife

Rao et Sitter (1995) ont proposé un estimateur jackknife corrigé de la variance pour résoudre le problème de l'imputation par la méthode du quotient. Dans les conditions décrites à la section 4, l'estimateur imputé par quotient de μ_y est

$$\hat{\mu}_I = \sum_{i=1}^{n} w_i [a_i y_i + (1-a_i)\hat{y}_i]$$

où $\hat{y}_i = x_i \hat{R}$ et $\hat{R} = (\sum_{i=1}^{n} w_i a_i x_i)^{-1} \sum_{i=1}^{n} w_i a_i y_i$. L'estimateur de la variance de Rao et Sitter (1995) est donné par

$$V_a = \sum_{i=1}^{L} c_i (\hat{\mu}_I^{(i)} - \hat{\mu}_I)^2, \tag{34}$$

où la répétition jackknife corrigée obtenue à la i^{e} itération est donnée par

$$\hat{\mu}_I^{(i)} = \sum_{j=1}^{n} w_j M_j^{(i)} y_j^{*(i)} \tag{35}$$

où

$$y_j^{*(i)} = \begin{cases} x_j \hat{R}^{(i)} & \text{si } a_i = 1 \\ x_j \hat{R} & \text{si } a_i = 0 \end{cases} \tag{36}$$

avec $\hat{R}^{(i)} = (\sum_{j=1}^{n} w_j M_j^{(i)} a_j x_j)^{-1} \sum_{j=1}^{n} w_j M_j^{(i)} a_j y_j$. Les valeurs corrigées (36) de la méthode de Rao et Sitter (1995) peuvent aussi être considérées comme des pseudo-données pour l'estimation de la variance. Notons que le calcul des pseudo-données (36) oblige à recalculer $\hat{R}^{(i)}$ pour chaque i, avec $a_i = 1$.

Nous modifions le calcul des pseudo-valeurs y_i^* dans (5) pour obtenir

$$y_i^* = \begin{cases} \hat{y}_i & \text{si } a_i = 0 \\ \hat{y}_i + c_r \left(\dfrac{\bar{x}_1}{\bar{x}_2} \right)(y_i - \hat{y}_i) & \text{si } a_i = 1, \end{cases} \tag{37}$$

où $\bar{x}_2 = \sum_{i=1}^{n} w_i r^{-1} n a_i x_i, \bar{x}_1 = n^{-1} \sum_{i=1}^{n} w_i x_i$ et $c_r \doteq r^{-1} n$. Nous insérons le terme $(\bar{x}_1 / \bar{x}_2)$ pour améliorer les propriétés conditionnelles de V_J, étant donné l'échantillon de première phase. L'estimateur résultant de la variance est approximativement équivalent à l'estimateur jackknife corrigé de la variance donné par (34). Pour le confirmer, notons que les valeurs corrigées (35) peuvent s'écrire sous la forme

$$\hat{\mu}_I^{(i)} = \left(\sum_{j=1}^{n} w_j M_j^{(i)} x_j \right) \frac{\sum_{j=1}^{n} w_j M_j^{(i)} a_j y_j}{\sum_{j=1}^{n} w_j M_j^{(i)} a_j x_j} =: \hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}},$$

où $A =: B$ indique que nous définissons B pour qu'il soit équivalent à A. En outre, définissons $\hat{Z} = \sum_{j=1}^{n} w_j x_j$, $\hat{S} = \sum_{j=1}^{n} w_j a_j y_j$, et $\hat{T} = \sum_{j=1}^{n} w_j a_j x_j$. Alors, en vertu du développement par série de Taylor de premier ordre,

$$\begin{aligned}\hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}} &\doteq \hat{Z} \frac{\hat{S}}{\hat{T}} + (\hat{Z}^{(i)} - \hat{Z}) \frac{\hat{S}}{\hat{T}} \\ &\quad + (\hat{S}^{(i)} - \hat{S}) \frac{\hat{Z}}{\hat{T}} - (\hat{T}^{(i)} - \hat{T}) \frac{\hat{Z}\hat{S}}{\hat{T}^2} \\ &= \left[\hat{Z}^{(i)} \frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} \left(\hat{S}^{(i)} - \hat{T}^{(i)} \frac{\hat{S}}{\hat{T}} \right) \right]. \qquad (38)\end{aligned}$$

Notons que le deuxième membre de l'équation (38) est exactement égal à

$$\sum_{j=1}^{n} w_j M^{(i)} \left[\frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_j \left(y_j - \frac{\hat{S}}{\hat{T}} \right) \right].$$

Donc, les pseudo-données produites pour estimer la variance peuvent s'écrire sous la forme

$$y_i^* = \frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_i \left(y_i - \frac{\hat{S}}{\hat{T}} \right),$$

que l'on peut réduire à (37). Par conséquent, la méthode proposée représente exactement une linéarisation par série de Taylor de premier ordre de la méthode de Rao et Sitter dans le cas de l'imputation par la méthode du quotient. Par conséquent, en principe, la méthode que nous proposons doit avoir les mêmes propriétés asymptotiques que la méthode de Rao et Sitter jusqu'à l'ordre n^{-1}.

La méthode d'estimation de la variance basée sur l'ensemble de pseudo-données calculées au moyen de l'équation (37) est facile à appliquer parce que nous pouvons nous servir directement d'un logiciel existant, ce qui est plus difficile dans le cas de la méthode de Rao et Shao (1992) ou de Rao et Sitter (1995). En outre, si nous calculons les pseudo-données au moyen de l'équation (13), alors l'ensemble de données convient pour l'imputation par la méthode hot deck sans remise ainsi que pour l'imputation par la méthode hot deck avec remise.

6. Étude par simulation

Pour éprouver la théorie qui précède, nous avons réalisé une étude en simulation portant sur une population artificielle, finie, à partir de laquelle nous avons sélectionné des échantillons répétés. La population compte $L = 32$ strates, N_h grappes dans la strate h et 20 unités finales dans chaque grappe. Les valeurs des paramètres de population ont été choisies de façon à ce qu'elles correspondent aux valeurs de population réelles observées dans le cadre de la U.S. *National Assessment of Educational Progress Study* (Hansen et Tepping 1985) et sont énumérées au tableau 1. Les unités de la population finie sont représentées par

$$y_{hij} = y_{hi} + e_{hij},$$

où

$$y_{hi} \overset{\text{iid}}{\sim} N(\mu_h, \sigma_h^2),\ h = 1, 2, ..., L,\ i = 1, 2, ..., N_h,$$

et

$$e_{hij} \overset{\text{iid}}{\sim} N\left(0, \frac{1-\rho}{\rho}\sigma_h^2\right),\ j = 1, 2, ..., 20.$$

Shao, Chen et Chen (1998) se sont aussi servi de la même population pour réaliser leur étude en simulation. La valeur de la corrélation intra-grappe ρ considérée dans la simulation est $\rho = 0,3$. Les simulations réalisées en prenant d'autres valeurs de ρ ont donné des résultats similaires et ne sont donc pas présentés ici par souci de concision.

Tableau 1
Paramètres de la population finie utilisés pour la simulation

h	N_h	μ_h	σ_h	h	N_h	μ_h	σ_h
1	13	100,0	20,0	2	16	95,0	19,0
3	20	90,0	18,0	4	25	98,0	19,6
5	25	93,0	18,6	6	25	98,0	19,6
7	25	96,0	19,2	8	28	94,0	18,8
9	28	92,0	18,4	10	28	96,0	19,2
11	31	94,0	18,8	12	31	92,0	18,4
13	31	90,0	18,0	14	31	96,0	19,2
15	31	94,0	18,8	16	31	92,0	18,4
17	31	90,0	18,0	18	31	88,0	17,6
19	31	86,0	17,2	20	34	84,0	16,8
21	34	82,0	16,4	22	34	80,0	16,0
23	34	90,0	18,0	24	37	85,0	17,0
25	37	80,0	16,0	26	37	90,0	18,0
27	37	85,0	17,0	28	39	80,0	16,0
29	39	75,0	15,0	30	42	75,0	15,0
31	42	75,0	15,0	32	42	75,0	15,0

Nous considérons un plan d'échantillonnage par grappe stratifié où on sélectionne $n_h = 2$ grappes avec remise dans la strate h avec probabilité égale et où toutes les unités finales qui figurent dans les grappes sélectionnées sont dans l'échantillon. La fraction d'échantillonnage est de 6,4 %. Pour chaque unité échantillonnée y_{hij}, nous produisons une variable indicatrice de réponse a_{hij} à partir de

$$a_{hij} \overset{\text{iid}}{\sim} \text{Bernoulli}\,(p),$$

et cet a_{hij} est indépendant de y_{hij}. Les valeurs de p considérées pour la simulation sont p = 0,9, 0,8, 0,7, 0,6 et 0,5.

Nous sélectionnons un ensemble de 5 000 échantillons selon le même plan d'échantillonnage. Pour chaque échantillon sélectionné, nous considérons trois méthodes d'imputation.

[M1] Imputation par la méthode hot deck pondérée avec remise considérée par Rao et Shao (1992), où une valeur manquante est remplacée par imputation d'une valeur sélectionnée au hasard parmi les répondants avec remise et probabilité proportionnelle au poids de sondage.

[M2] Imputation par la méthode hot deck pondérée sans remise, qui est identique à la méthode [M1], hormis le fait que la sélection est effectuée par échantillonnage sans remise. La sélection de donneurs sans remise est exécutée systématiquement selon la méthode décrite par Hansen, Hurwitz et Madow (1953, page 343) pour les répondants triés par ordre aléatoire.

[M3] Imputation globale de la moyenne, où l'on impute la moyenne pondérée calculée pour les répondants qui figurent dans l'échantillon.

Par conséquent, toutes les méthodes d'imputation comportent une cellule d'imputation unique qui regroupe toutes les strates.

Pour chaque ensemble contenant de données imputées, nous calculons trois estimateurs de la variance, à savoir $\hat{V}_n$, qui est l'estimateur naïf de la variance obtenu lorsque l'on traite les données imputées comme s'il s'agissait de données observées, $\hat{V}_a$, qui est l'estimateur jackknife corrigé de la variance proposé par Rao et Shao (1992) pour [M1] et [M2] et par Rao et Sitter (1995) pour [M3], et $\hat{V}^*$, qui est l'estimateur jackknife de la variance fondé sur les pseudo-données. Nous produisons l'ensemble de pseudo-données au moyen de l'équation (29) pour [M1] et [M2] et au moyen de l'équation (24) pour [M3]. L'estimateur complet de la variance d'échantillon est un estimateur jackknife type pour l'échantillonnage par grappe stratifié, selon lequel une grappe est supprimée à chaque itération. Notons que l'estimateur jackknife type est un estimateur cohérent de la variance dans les conditions du modèle avec corrélation intra-grappe non nulle. Donc, nous pouvons appliquer la méthode type du jackknife fondée sur les pseudo-données à l'ensemble de données considéré. Les estimateurs ponctuels de la moyenne de population sont non biaisés dans le cas des trois scénarios distincts d'imputation et nous ne les présentons pas ici.

Le tableau 2 donne le biais relatif des trois estimateurs de la variance, l'écart-type du biais relatif des estimateurs de la variance et le coefficient de corrélation d'échantillon entre l'estimateur jackknife corrigé de la variance de Rao et le nouvel estimateur de la variance fondés sur les 5 000 échantillons. Nous calculons le biais relatif de $\hat{V}$, en tant qu'estimateur de la variance de $\bar{y}_I$, au moyen de $[\mathrm{Var}_B(\bar{y}_I)]^{-1}[E_B(\hat{V}) - \mathrm{Var}_B(\bar{y}_I)]$, où l'indice B représente la distribution produite par la simulation de Monte Carlo. Enfin, nous calculons les coefficients de corrélation des deux estimateurs de la variance pour donner une mesure de la linéarité relative de ces estimateurs.

Tableau 2
Biais relatif de l'estimateur de la variance, écart-type du biais relatif et coefficient de corrélation de l'échantillon entre l'estimateur de la variance de Rao et le nouvel estimateur de la variance fondés sur 5 000 échantillons

Taux de réponse (p)	Méthode d'imputation	Biais rel. × 100 (É. -T. × 100) Naïf	Rao	Nouveau	Coeff. de corr. r
0,9	M1	-17,40 (2,02)	1,61 (2,03)	1,70 (2,04)	0,967
	M2	-17,50 (2,00)	1,41 (2,01)	0,81 (2,03)	0,974
	M3	-18,03 (2,03)	1,16 (2,05)	1,15 (2,04)	1,000
0,8	M1	-34,45 (2,01)	0,65 (2,03)	0,49 (2,05)	0,939
	M2	-32,89 (2,01)	2,49 (2,04)	0,19 (2,03)	0,947
	M3	-34,96 (2,01)	1,59 (2,03)	1,59 (2,03)	1,000
0,7	M1	-48,96 (2,01)	0,21 (1,99)	0,41 (2,04)	0,912
	M2	-44,76 (2,02)	5,31 (2,05)	0,76 (2,05)	0,920
	M3	-50,21 (2,02)	1,53 (2,05)	1,52 (2,04)	1,000
0,6	M1	-59,80 (2,02)	1,58 (2,05)	1,27 (2,06)	0,892
	M2	-54,86 (2,03)	7,10 (2,07)	-0,75 (2,07)	0,899
	M3	-64,11 (2,00)	-0,35 (2,04)	-0,35 (2,01)	1,000
0,5	M1	-69,75 (1,99)	0,84 (2,03)	1,12 (2,03)	0,873
	M2	-59,90 (2,01)	15,07 (2,07)	2,27 (2,06)	0,872
	M3	-74,44 (1,97)	1,99 (2,00)	1,98 (2,00)	1,000

Le tableau 2 confirme notre théorie de la façon suivante.

1. Il est bien connu que l'estimateur naïf de la variance sous-estime fortement la variance réelle. L'estimateur jackknife corrigé de la variance donne de bons résultants pour [M1] et [M3], mais pas pour [M2]. La théorie qui sous-tend la méthode corrigée du jackknife suppose que les imputations par la méthode hot deck s'appuient sur l'échantillonnage avec remise, lequel n'est pas utilisé dans [M2]. À mesure que le taux de réponse diminue dans le tableau 2, le biais relatif de l'estimateur jackknife corrigé augmente.

2. La nouvelle méthode basée sur les pseudo-données donne de bons résultats même pour l'imputation avec échantillonnage sans remise [M2]. Comme nous en avons discuté à la fin de la section 3, la formule (29) suffit, à elle seule, à produire les pseudo-données pour une grande classe de méthodes d'imputation.

3. Comme le montrent les coefficients de corrélation, les comportements de l'estimateur jackknife corrigé et de l'estimateur proposé de la variance sont très semblables pour l'imputation de la moyenne [M3], parce que les deux estimateurs sont asymptotiquement équivalents, comme nous l'avons exposé à la section 5.

7. Conclusions

Nous avons décrit certaines méthodes qui permettent de produire des pseudo-données pour estimer la variance.

Généralement parlant, les pseudo-données peuvent être décrites par la formule

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, ..., n \\ \hat{y}_i + c_i g_i (y_i - \hat{y}_i) & i = 1, 2, ..., r, \end{cases} \quad (39)$$

où $\hat{y}_i$ est la valeur prévue de y_i dans les conditions du modèle utilisé pour l'imputation. Si $c_i g_i = 1$, alors l'estimateur de la variance traite les valeurs imputées comme des observations. En choisissant convenablement $c_i g_i > 1$, on obtient un estimateur de la variance cohérent. Si la méthode d'imputation est déterministe et que les répondants utilisés sont considérés comme étant un échantillon aléatoire de l'échantillon original, alors $c_i \doteq r^{-1} n > 1$. Pour un échantillonnage à deux phases avec plan de sondage complexe, $c_i = w_i^{-1} w_i^*$, où w_i représente le poids d'échantillonnage de l'unité i pour l'échantillon de première phase et w_i^* représente le poids d'échantillonnage de l'unité i pour l'échantillon de seconde phase.

Le g_i de l'équation (39) représente la correction faite pour améliorer les propriétés conditionnelles, étant donné la variable auxiliaire x. Pour l'imputation par la méthode du quotient,

$$g_i = (\bar{x}_2)^{-1} \bar{x}_1$$

où $\bar{x}_2 = \sum_{i=1}^r w_i^* x_i$ et $\bar{x}_1 = \sum_{i=1}^n w_i x_i$. Pour l'imputation par régression avec la grandeur scalaire x,

$$g_i = 1 + (\bar{x}_1 - \bar{x}_2) \left\{ \sum_{k=1}^r w_k^* (x_k - \bar{x}_2)^2 \right\}^{-1} (x_i - \bar{x}_2).$$

Dans l'un et l'autre cas, nous avons

$$\sum_{i=1}^r w_i^* g_i x_i = \bar{x}_1.$$

Pendant l'évaluation du présent article, Shao et Steel (1999) ont proposé des méthodes similaires dans le cas de l'imputation déterministe. Notre méthode est plus générale en ce sens que nous considérons aussi l'imputation aléatoire et que nous introduisons le terme c_i pour améliorer les propriétés de l'échantillon fini.

Remerciements

L'auteur remercie son conseiller de thèse, Wayne A. Fuller, pour les discussions fructueuses qu'ils ont eues. Il remercie aussi Pamela Abbitt, F. Jay Breidt, Lou Rizzo, Richard Valliant et les examinateurs pour leurs commentaires utiles, qui lui ont permis d'améliorer considérablement l'article. La plupart des travaux présentés ici ont été réalisés pendant que l'auteur poursuivait des études de cycles supérieures à la Iowa State University et financés en partie par l'entente de coopération 68-3A75-43 conclue par le USDA Natural Resources Conservation Service et la Iowa State University et par l'entente de coopération 43-3AEU-3-80088 conclue par la Iowa State University, le National Agricultural Statistics Service et le U.S. Bureau of Census.

Annexe

A. Preuve des équations (10) et (12)

L'estimateur $\hat{\mu}_y$ qui figure dans l'équation (9) peut être écrit sous la forme

$$\hat{\mu}_y = n^{-1} \sum_{i=1}^n \hat{y}_i + n^{-1} \sum_{i=1}^r (1 + d_i) \hat{e}_i \quad (A.1)$$

où d_i est le nombre de fois que l'unité i est utilisée comme un donneur. En vertu du mécanisme d'imputation avec échantillonnage à probabilité égale et avec remise, nous avons

$$E_I(d_i) = r^{-1} m$$

et

$$\mathrm{Cov}_I(d_i, d_j) = \begin{cases} r^{-1} m (1 - r^{-1}) & \text{si } i = j \\ -r^{-2} m & \text{si } i \neq j \end{cases}$$

où l'indice I représente la variation due au mécanisme d'imputation. Il s'ensuit que $E_I(\hat{\mu}_y) = n^{-1} \sum_{i=1}^n \hat{y}_i$ et $V_I(\hat{\mu}_y) = n^{-2} r^{-1} m \sum_{i=1}^r \hat{e}_i^2$. Par conséquent,

$$V(\hat{\mu}_y) \doteq V\left(n^{-1} \sum_{i=1}^n \hat{y}_i \right) + E\left(n^{-2} r^{-1} m \sum_{i=1}^r \hat{e}_i^2 \right). \quad (A.2)$$

Maintenant, par un argument similaire à celui qui nous a mené à (2), nous obtenons

$$\mathrm{Var}\left(n^{-1} \sum_{i=1}^n \hat{y}_i \right) = [n^{-1} R^2 + r^{-1} (1 - R^2)] \, \sigma_y^2. \quad (A.3)$$

Puisque $\hat{y}_i - \bar{y}_I = (\mathbf{x}_i - \bar{\mathbf{x}}_i) \beta + o_p(1)$, nous appliquons la théorie classique de régression pour obtenir

$$E\left[(r - p)^{-1} \sum_{i=1}^r \hat{e}_i^2 \right] = (1 - R^2) \, \sigma_y^2, \quad (A.4)$$

et

$$E\left[(n - 1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_I)^2 \right] = R^2 \sigma_y^2. \quad (A.5)$$

Par conséquent, l'équation (10) est prouvée et l'estimateur donné par (12) est cohérent pour la variance donnée par l'équation (10).

B. Validité de l'équation (15) dans les conditions du mécanisme d'imputation sans remise

Nous supposons que $m = kr + t$ où k et t sont des nombres entiers non négatifs et que $t < r$. Supposons que l'estimateur de la moyenne de y a la forme (A.1). Supposons que l'imputation est exécutée de façon telle que t répondants sont utilisés $k+1$ fois pour l'imputation et que $r-t$ unités sont utilisés k fois pour l'imputation. Les t répondants qui sont utilisés $k+1$ fois sont choisis par échantillonnage aléatoire simple sans remise. Alors,

$$E_I(d_i) = k + r^{-1}t = r^{-1}m$$

et

$$\mathrm{Cov}_I(d_i, d_j) = \begin{cases} r^{-1}t(1-r^{-1}t) & \text{si } i = j \\ -r^{-2}t & \text{si } i \neq j. \end{cases}$$

Donc, par des arguments similaires à ceux utilisés dans la preuve de (A.2), nous obtenons

$$V(\hat{\mu}_y) \doteq V(\bar{y}_I) + E\left(n^{-2}r^{-1}t\sum_{i=1}^{r}\hat{e}_i^2\right). \qquad \text{(B.1)}$$

Par conséquent, en utilisant (A.3) et (A.4), nous obtenons

$$V\{\hat{\mu}_y\} = [n^{-1}R^2 + (r^{-1} + n^{-2}t)(1-R^2)]\,\sigma_y^2. \qquad \text{(B.2)}$$

Maintenant, subordonnée à l'échantillon réalisé et aux répondants, nous obtenons

$$E_I\{(1+d_i)^2\} = \left(\frac{n}{r}\right)^2 + \frac{t}{r}\left(1 - \frac{t}{r}\right)$$

de sorte que $\hat{V}\{\mu_y\}$ dans l'équation (15) satisfait

$$E_I(\hat{V}\{\mu_y\}) \doteq n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2 + [r^{-1} + n^{-2}t(1-r^{-1}t)](r-p)^{-1}\sum_{i=1}^{r}(y_i - \hat{y}_i)^2.$$

Par conséquent, en utilisant (A.4) et (A.5), nous obtenons l'absence de biais approximative de $\hat{V}\{\mu_y\}$ dans le cas du mécanisme d'imputation sans remise.

C. Preuve de l'équation (26)

Premièrement, définissons $R_n^{(i)} = (\mathbf{x}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)})(\hat{\beta} - \beta)$ et $R_n = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\hat{\beta} - \beta)$. D'après l'égalité (25),

$$\hat{V}^* = \sum_{i=1}^{L} c_i(\bar{y}_I^{*(i)} - \bar{y}_I)^2 = A_n + B_n + 2C_n$$

où $A_n = \sum_{i=1}^{L} c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})^2$, $B_n = \sum_{i=1}^{L} c_i(R_n^{(i)} - R_n)^2$, et $C_n = \sum_{i=1}^{L} c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})(R_n^{(i)} - R_n)$. Par conséquent, en vertu de l'hypothèse (20), il s'ensuit l'équation (26), car $A_n = O_p(n^{-1})$, $B_n = o_p(n^{-1})$, et $C_n = o_p(n^{-1})$. La dernière propriété provient de l'inégalité de Cauchy-Schwartz, $C_n^2 \leq A_n B_n$.

Bibliographie

Cochran, W.G. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the Bureau of the Census Annual Research conference*, 429-440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 227-232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.

Hansen, M., Hurwitz, W.N. et Madows, W.G. (1953). *Sample Survey Methods and Theory*. Vol. I, New York : John Wiley & Sons, Inc.

Hansen, M., et Tepping, B.J. (1985). Estimation for Variance in NAEP. Note de service inédite, Westat, Washington, D.C.

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.

Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley & Sons, Inc.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., et Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

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.

Särndal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de la Statistique*, 55, 279-294.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Shao, J., Chen, Y. et Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

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