Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Imputation fractionnaire hot deck pour une inférence robuste sous un modèle de non-réponse partielle en échantillonnage

par Jae Kwang Kim et Shu Yang

Date de diffusion : 19 décembre 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Comment accéder à ce produit

Le produit no 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 235-256
Statistique Canada, nº 12-001-X au catalogue

Imputation fractionnaire hot deck pour une inférence robuste sous un modèle de non-réponse partielle en échantillonnage

Jae Kwang Kim et Shu Yang[1]

Résumé

L'imputation fractionnaire paramétrique (IFP) proposée par Kim (2011) est un outil d'estimation des paramètres à usage général en cas de données manquantes. Nous proposons une imputation fractionnaire hot deck (IFHD), qui est plus robuste que l'IFP ou l'imputation multiple. Selon la méthode proposée, les valeurs imputées sont choisies parmi l'ensemble des répondants, et des pondérations fractionnaires appropriées leur sont assignées. Les pondérations sont ensuite ajustées pour répondre à certaines conditions de calage, ce qui garantit l'efficacité de l'estimateur IFHD résultant. Deux études de simulation sont présentées afin de comparer la méthode proposée aux méthodes existantes.

Mots-clés : Algorithme EM; information de Kullback-Leibler; valeurs manquant au hasard; imputation multiple

1 Introduction

L'imputation est une méthode courante de compensation de la non-réponse partielle dans les enquêtes sur échantillon. Soit y la variable étudiée sujette à la non-réponse et $\mathbf{x}$ le vecteur des variables auxiliaires complètement observées. On utilise souvent un modèle de distribution conditionnelle $f(y|\mathbf{x})$ afin de générer des valeurs imputées pour la donnée y_i manquante. Cette méthode d'imputation fondée sur un modèle a fait l'objet de nombreuses études. L'imputation multiple de Rubin (1987) est une approche bayésienne d'imputation fondée sur un modèle. L'algorithme EM Monte Carlo de Wei et Tanner (1990) peut être traité comme une approche fréquentiste d'imputation fondée sur un modèle. Kim (2011) proposait une imputation fractionnaire paramétrique pour traiter les données manquantes multivariées.

Cependant, la méthode d'imputation fondée sur un modèle qui génère des valeurs imputées à partir de $f(y|\mathbf{x})$ n'est pas une imputation hot deck en ce sens que les valeurs artificielles sont construites après l'imputation. Une caractéristique souhaitable de l'imputation hot deck est que toutes les valeurs imputées sont des valeurs observées. Par exemple, les valeurs imputées pour des variables catégoriques seront elles aussi catégoriques et le nombre de catégories est le même que celui observé pour les répondants. Pour cette raison, l'imputation hot deck est la méthode d'imputation la plus populaire, particulièrement dans les enquêtes-ménages. L'imputation par la méthode du plus proche voisin est une autre imputation hot deck. Chen et Shao (2001), Beaumont et Bocci (2009), Kim, Fuller et Bell (2011) ont eux aussi examiné l'imputation par la méthode du plus proche voisin en contexte d'échantillonnage. Durrant (2009), Haziza (2009) et Andridge et Little (2010) ont donné des aperçus détaillés des méthodes d'imputation hot deck en échantillonnage.

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. Courriel : jkim@iastate.edu; Shu Yang, Department of Statistics, Iowa State University, Ames, IA 50011.

Kalton et Kish (1984) ont proposé une imputation fractionnaire hot deck afin d'assurer l'efficacité de l'imputation hot deck. Kim et Fuller (2004) et Fuller et Kim (2005) ont soumis l'imputation fractionnaire hot deck à un examen rigoureux et examiné l'estimation de la variance. Cependant, leur approche s'applique seulement lorsque $\mathbf{x}$ est catégorique. Pour les covariables continues, l'appariement d'après la moyenne prédictive peut être traité comme une méthode d'imputation par le plus proche voisin fondée sur la valeur prédite obtenue à partir de $f(y \mid \mathbf{x})$, mais ses propriétés statistiques ne sont pas traitées de façon approfondie dans la littérature.

Dans le présent article, nous proposons une nouvelle méthode d'imputation fractionnaire hot deck (IFHD) fondée sur un modèle paramétrique de $f(y \mid \mathbf{x})$ qui permet des covariables continues. La méthode proposée présente plusieurs avantages par rapport aux méthodes existantes. Premièrement, cette imputation hot deck préserve la structure de corrélation entre les éléments. Deuxièmement, elle est robuste en ce sens que l'estimateur résultant est moins sensible à l'échec du modèle théorique $f(y \mid \mathbf{x})$. Troisièmement, elle fournit des estimateurs de variance convergents pour différents paramètres sans exiger la condition de compatibilité de Meng (1994). L'imputation multiple exige toutefois la condition de compatibilité pour valider l'estimation de la variance. Lorsque la condition de compatibilité n'est pas satisfaite, l'imputation multiple donne souvent lieu à une inférence prudente qui, à son tour, réduit la puissance des tests. Voir la section 5.2 pour plus de détails.

La présentation de l'article suit. Dans la section 2, nous décrivons la configuration de base. La méthode proposée est présentée dans la section 3. La robustesse de l'IFHD est traitée dans la section 4. Dans la section 5, nous présentons les résultats de deux études par simulation et, dans la section 6, nous formulons nos conclusions.

2 Configuration de base

Considérons une population finie de N éléments identifiés par un ensemble d'indices $U = \{1, 2, \ldots, N\}$, où N est connu. À chaque unité i de la population sont associées les variables étudiées $\mathbf{x}_i$ et y_i, où $\mathbf{x}_i$ est toujours observée et y_i est sujette à la non-réponse. Soit A l'ensemble d'indices pour les éléments d'un échantillon sélectionné par échantillonnage probabiliste. Nous voulons estimer η, définie comme étant une solution (unique) à l'équation d'estimation de population $\sum_{i=1}^{N} U(\eta; \mathbf{x}_i, y_i) = 0$. Par exemple, la moyenne de population peut être obtenue en posant que $U(\eta; \mathbf{x}_i, y_i) = \eta - y_i$. Sous réponse complète, un estimateur convergent de η est obtenu en résolvant

$$\sum_{i \in A} w_i U(\eta; \mathbf{x}_i, y_i) = 0, \tag{2.1}$$

où $w_i = \{Pr(i \in A)\}^{-1}$ est l'inverse de la probabilité d'inclusion d'ordre un de l'unité i. Binder et Patak (1994) et Rao, Yung et Hidiroglou (2002) ont examiné les propriétés asymptotiques de l'estimateur obtenu au moyen de l'équation (2.1). Lorsqu'il manque des données, nous définissons

$$\delta_i = \begin{cases} 1 & \text{si } y_i \text{ est observée;} \\ 0 & \text{sinon.} \end{cases}$$

Nous obtenons alors un estimateur convergent de η en prenant l'espérance conditionnelle et en résolvant

$$\sum_{i\in A} w_i \left[\delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) E\left\{ U\left(\eta; \mathbf{x}_i, Y\right) \mid \mathbf{x}_i, \delta_i = 0 \right\} \right] = 0 \tag{2.2}$$

pour η. L'équation d'estimation (2.2) est parfois qualifiée d'équation d'estimation prévue (Wang et Pepe 2000).

Pour calculer l'espérance conditionnelle en (2.2), nous supposons que la population finie étudiée est réalisée à partir d'une population infinie appelée superpopulation. Dans le modèle de la superpopulation, nous postulons souvent une distribution conditionnelle paramétrique de y étant donnée $\mathbf{x}$, $f\left(y \middle| \mathbf{x}; \theta\right)$, qui est connue jusqu'au paramètre θ dans l'espace des paramètres Ω. Sous le modèle spécifié, nous pouvons calculer un estimateur convergent $\hat{\theta}$ de θ puis utiliser une méthode Monte Carlo pour évaluer l'espérance conditionnelle en (2.2) étant donné l'estimation $\hat{\theta}$. Si le mécanisme de réponse est manquant au hasard ou est ignorable au sens de Rubin (1976), nous pouvons approximer ainsi l'équation d'estimation prévue en (2.2)

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \frac{1}{m} \sum_{j=1}^{m} U\left(\eta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{2.3}$$

où

$$y_i^{*(1)}, \ldots, y_i^{*(m)} \overset{i.i.d.}{\sim} f\left(y_i \mid \mathbf{x}_i; \hat{\theta}\right).$$

Nous utilisons souvent l'estimateur du maximum de vraisemblance $\hat{\theta}$, ce qui résout

$$S\left(\theta\right) = \sum_{i\in A} w_i \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) = 0, \tag{2.4}$$

où $S\left(\theta; \mathbf{x}, y\right) = \partial \log f\left(y \mid \mathbf{x}; \theta\right) / \partial\theta$. Il est à noter que nous utilisons les poids d'échantillonnage w_i dans l'équation de score (2.4). Nous supposons implicitement que le modèle d'imputation, qui génère les valeurs imputées, est le modèle concernant les valeurs de la population finie $f\left(y_i \mid \mathbf{x}_i\right)$, plutôt que les valeurs d'échantillon. Nous convenons ainsi que le mécanisme d'échantillonnage peut être informatif dans le sens de Pfeffermann (2011). Par contraste, l'imputation multiple utilise le modèle d'échantillon, $f_s\left(y_i \mid \mathbf{x}_i\right) \equiv f\left(y_i \mid \mathbf{x}_i, i \in A\right)$, pour générer les valeurs imputées et suppose souvent que le mécanisme d'échantillonnage n'est pas informatif. Ainsi, l'imputation multiple suppose que les données sont manquantes au hasard dans l'échantillon étudié tandis que l'imputation fractionnaire suppose que les données sont manquantes au hasard dans la population. Sous un plan d'échantillonnage informatif, la génération de valeurs imputées à partir du modèle d'échantillon $f_s\left(y_i \mid \mathbf{x}\right)$ ne mène pas nécessairement à une inférence valide même quand la condition de données manquant au hasard dans l'échantillon est remplie. Voir la section 8.4 de Kim et Shao (2013) pour un examen plus approfondi des données manquant au hasard sous échantillonnage informatif.

L'imputation fractionnaire paramétrique (IFP) de Kim (2011) peut être utilisée pour calculer l'espérance conditionnelle en (2.2) de manière efficace. En IFP, les valeurs imputées sont générées à partir d'une distribution proposée acceptable $h\left(y \mid \mathbf{x}_i\right)$ puis l'équation d'estimation imputée (2.3) est remplacée par

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j=1}^{m} w_{ij}^* U\left(\eta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{2.5}$$

où

$$w_{ij}^* = \frac{f\left(y_i^{*(j)} \mid \mathbf{x}_i; \hat{\theta}\right) \Big/ h\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{\sum_{k=1}^{m} \left\{ f\left(y_i^{*(k)} \mid \mathbf{x}_i; \hat{\theta}\right) \Big/ h\left(y_i^{*(k)} \mid \mathbf{x}_i\right) \right\}}. \tag{2.6}$$

Le choix de distribution proposée $h(\cdot)$ est un peu arbitraire. Nous examinerons un choix particulier qui pourrait mener à une estimation robuste.

La convergence de l'estimateur $\hat{\eta}$ résultant de (2.3) ou (2.5) peut être établie en supposant que la distribution conditionnelle $f\left(y \mid \mathbf{x}; \theta\right)$ est correctement spécifiée (selon un argument semblable à celui utilisé dans la preuve du corollaire II.2 d'Andersen et Gill (1982), et la preuve n'est pas faite ici). Dans le présent article, nous examinons un différent type d'imputation fractionnaire qui est plus robuste en cas d'échec de l'hypothèse du modèle d'imputation.

3 Méthode proposée

Nous examinons d'abord une méthode d'imputation fractionnaire hot deck appelée **imputation fractionnaire complète**, où les valeurs imputées sont tirées de l'ensemble de répondants désigné par $A_R = \left\{ i \in A; \delta_i = 1 \right\}$. C'est-à-dire que la j-ième valeur imputée de la donnée manquante y_i, désignée par $y_i^{*(j)}$, est égale à la j-ième valeur de y dans l'ensemble A_R. Nous proposons une méthode d'imputation fractionnaire hot deck qui utilise l'hypothèse du modèle paramétrique $f\left(y \mid \mathbf{x}; \theta\right)$. Si tous les éléments de A_R sont choisis comme valeurs imputées de la donnée manquante y_i, nous pouvons traiter $\left\{ y_j; j \in A_R \right\}$ comme une réalisation de $f\left(y_j \mid \delta_j = 1\right)$ et, si $h\left(y_j \mid \mathbf{x}_i\right) = f\left(y_j \mid \delta_j = 1\right)$ est choisi en (2.6), le poids fractionnaire assigné au donneur y_j pour la donnée manquante y_i devient

$$\begin{aligned} w_{ij}^* &\propto f\left(y_j \mid \mathbf{x}_i, \delta_i = 0; \hat{\theta}\right) \Big/ f\left(y_j \mid \delta_j = 1\right) \\ &\propto f\left(y_j \mid \mathbf{x}_i, \hat{\theta}\right) \Big/ f\left(y_j \mid \delta_j = 1\right), \end{aligned} \tag{3.1}$$

où $\sum_{j;\delta_j=1} w_{ij}^* = 1$ et $\hat{\theta}$ est l'estimateur du maximum de vraisemblance (EMV) obtenu de l'équation (2.4). La deuxième ligne découle de l'hypothèse des données manquant au hasard. Nous pouvons aussi écrire

$$\begin{aligned} f\left(y_j \mid \delta_j = 1\right) &= \int f\left(y_j \mid \mathbf{x}, \delta_j = 1\right) f\left(\mathbf{x} \mid \delta_j = 1\right) d\mathbf{x} \\ &= \int f\left(y_j \mid \mathbf{x}\right) f\left(\mathbf{x} \mid \delta_j = 1\right) d\mathbf{x} \\ &\cong \frac{1}{N_R} \sum_{k=1}^{N} \delta_k f\left(y_j \mid \mathbf{x}_k\right), \end{aligned} \tag{3.2}$$

où la deuxième égalité découle de l'hypothèse des valeurs manquant au hasard, et la dernière égalité (approximative) est obtenue en approximant l'intégrale par distribution empirique de la population. N_R est le nombre de répondants dans la population. En utilisant les poids d'enquête, nous pouvons approximer

$$f\left(y_j \mid \delta_j = 1\right) \cong \frac{\sum_{k \in A_R} w_k f\left(y_j \mid \mathbf{x}_k\right)}{\sum_{k \in A_R} w_k}$$

et les poids fractionnaires en (3.1) sont calculés comme suit :

$$w_{ij}^* \propto \frac{f\left(y_j \mid \mathbf{x}_i; \hat{\theta}\right)}{\sum_{k \in A_R} w_k f\left(y_j \mid \mathbf{x}_k; \hat{\theta}\right)} \tag{3.3}$$

où $\sum_{j \in A_R} w_{ij}^* = 1$. En (3.3), la masse ponctuelle w_{ij}^* assignée au donneur y_j pour l'unité manquante i est exprimée par le ratio de la densité $f\left(y \mid \mathbf{x}\right)$. Ainsi, pour chaque unité manquante i, $n_R = \left|A_R\right|$, nous utilisons les observations comme donneurs pour l'imputation hot deck et w_{ij}^* comme poids fractionnaires. Cette méthode d'imputation fractionnaire peut être qualifiée d'imputation fractionnaire complète (IFC) en l'absence de caractère aléatoire attribuable au mécanisme d'imputation. L'estimateur IFC de η, défini par $\sum_{i=1}^{N} U\left(\eta; \mathbf{x}_i, y_i\right) = 0$ est alors calculé en résolvant

$$\sum_{i \in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1 - \delta_i\right) \sum_{j \in A_R} w_{ij}^* U\left(\eta; \mathbf{x}_i, y_j\right) \right\} = 0, \tag{3.4}$$

où w_{ij}^* est défini en (3.3). Il est à noter que l'équation d'estimation imputée (3.4) est une bonne approximation de l'équation d'estimation prévue en (2.2).

En échantillonnage, un ensemble de données imputées où la quantité d'imputation est importante n'est pas toujours souhaitable. Au lieu d'utiliser toutes les observations en A_R comme donneurs pour chaque donnée manquante, nous pouvons sélectionner un sous-ensemble de A_R afin de réduire la taille de l'ensemble donneur de la donnée manquante y_i. Ainsi, la sélection des donneurs est considérée comme un problème d'échantillonnage et nous utilisons un plan d'échantillonnage et des techniques de pondération efficaces pour obtenir des estimateurs par imputation efficaces. Des plans d'échantillonnage efficaces, comme un échantillonnage stratifié ou un échantillonnage systématique avec probabilité proportionnelle à la taille (PPT), peuvent être utilisés pour sélectionner des donneurs de taille m. Un échantillonnage PPT systématique pour l'imputation fractionnaire hot deck peut être décrit comme suit :

1. Dans chaque i où $\delta_i = 0$, trier les donneurs de l'ensemble complet de répondants $\left\{y_j; \delta_j = 1\right\}$ par ordre croissant où $y_{(1)} \le \cdots \le y_{(r)}$ et utiliser $w_{i(j)}^*$ pour désigner le poids fractionnaire associé à $y_{(j)}$, c'est-à-dire $w_{i(j)}^* = w_{ik}^*$ pour $y_{(j)} = y_k$.

2. Partitionner $[0,1]$ par $\left\{ I_j \equiv \left[\sum_{k=0}^{j} w_{i(j)}^*, \sum_{k=0}^{j+1} w_{i(j)}^* \right), j = 1, \ldots, r-1 \right\}$, où $w_{i(0)}^* = 0$.

3. Générer $u \sim \text{uniforme}(0,1/m)$ et poser $u_k = u + k/m,\ k = 0,\ldots,m-1$. Pour $k = 0,\ldots,m-1,$ si $u_k \in I_j$ pour certains $0 \le j \le r-1,$ inclure j dans l'échantillon D_i.

Après avoir sélectionné D_i dans l'ensemble complet de répondants, nous assignons les poids fractionnaires initiaux $w_{ij0}^* = 1/m$ aux donneurs choisis en D_i. D'autres ajustements sont apportés aux poids fractionnaires afin de satisfaire

$$\sum_{i\in A} w_i \left\{ (1-\delta_i) \sum_{j\in D_i} w_{ij,c}^* \mathbf{q}(\mathbf{x}_i, y_j) \right\} = \sum_{i\in A} w_i \left\{ (1-\delta_i) \sum_{j\in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j) \right\}, \tag{3.5}$$

pour certains $\mathbf{q}(\mathbf{x}_i, y_j)$, et $\sum_{j\in D_i} w_{ij,c}^* = 1$ pour tous les i où $\delta_i = 0,$ w_{ij}^* étant les poids fractionnaires pour la méthode d'IFC définie en (3.3). En ce qui concerne le choix de la fonction de contrôle $\mathbf{q}(\mathbf{x}, y)$ en (3.5), nous pouvons utiliser $\mathbf{q}(\mathbf{x}, y) = (y, y^2)'$, ce qui rapproche le plus possible les distributions empiriques de y pour D_i et A_R en ce sens que les premier et second moments de y sont les mêmes. D'autres choix peuvent être envisagés. Voir Fuller et Kim (2005).

Le problème d'ajustement des poids initiaux afin de respecter certaines contraintes est souvent qualifié de calage et les poids fractionnaires résultants peuvent être qualifiés de poids fractionnaires calés. En utilisant la pondération par régression, nous pouvons calculer des poids fractionnaires finaux de calage qui satisfont à (3.5) et $\sum_j w_{ij,c}^* = 1$ comme suit :

$$w_{ij,c}^* = w_{ij0}^* + w_{ij0}^* \Delta \left(\mathbf{q}_{ij}^* - \overline{\mathbf{q}}_{i\cdot}^* \right), \tag{3.6}$$

où $\mathbf{q}_{ij}^* = \mathbf{q}(\mathbf{x}_i, y_j),$ $\overline{\mathbf{q}}_{i\cdot}^* = \sum_{j\in A_R} w_{ij0}^* \mathbf{q}_{ij}^*,$

$$\Delta = \left\{ C_q - \sum_{i\in A} w_i (1-\delta_i) \sum_{j\in A_R} w_{ij0}^* \mathbf{q}_{ij}^* \right\}^T \left\{ \sum_{i\in A} w_i (1-\delta_i) \sum_{j\in A_R} w_{ij0}^* \left(\mathbf{q}_{ij}^* - \overline{\mathbf{q}}_{i\cdot}^* \right)^{\otimes 2} \right\}^{-1}$$

et $C_q = \sum_{i\in A} w_i \left\{ (1-\delta_i) \sum_{j\in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j) \right\}$. Ici, $B^{\otimes 2}$ désigne BB^T. Certains des poids fractionnaires calculés en (3.6) peuvent prendre des valeurs négatives. Le cas échéant, il faut utiliser des algorithmes remplaçant la pondération par régression. Par exemple, considérons la pondération par l'entropie, où les poids fractionnaires de la forme

$$w_{ij,c}^* = \frac{w_{ij}^* \exp\left(\Delta \mathbf{q}_{ij}^* \right)}{\sum_{k\in A_R} w_{ik}^* \exp\left(\Delta \mathbf{q}_{ik}^* \right)} \tag{3.7}$$

sont à peu près égaux aux poids fractionnaires par régression en (3.6) et sont toujours positifs. Après avoir obtenu les poids fractionnaires de calage, nous pouvons calculer l'estimateur IFHD de η en résolvant

$$\sum_{i\in A} w_i \left\{ \delta_i U(\eta; \mathbf{x}_i, y_i) + (1-\delta_i) \sum_{j\in D_i} w_{ij,c}^* U(\eta; \mathbf{x}_i, y_j) \right\} = 0. \tag{3.8}$$

Une méthode par rééchantillonnage peut être utilisée pour estimer la variance. L'annexe A.1 contient une brève discussion de l'estimateur de variance par rééchantillonnage pour la méthode proposée.

La méthode proposée peut aussi traiter la non-réponse non ignorable sous spécification correcte du modèle de réponse. Voir l'annexe A.3 pour l'extension à un cas de non-réponse non ignorable.

4 Robustesse

Nous examinons maintenant la robustesse de la méthode proposée pour tenir compte d'un léger écart par rapport au modèle paramétrique présumé. La robustesse de l'estimateur proposé protège contre les erreurs de spécification du modèle d'imputation, une légère inclinaison exponentielle du modèle vrai. Pour simplifier la présentation, supposons que le plan d'échantillonnage est un échantillonnage aléatoire simple et que l'échantillon réalisé est un échantillon aléatoire tiré du modèle de superpopulation.

Nous supposons que le modèle vrai $g(y|x)$ n'appartient pas à $\{f(y|x;\theta);\theta\in\Omega\}$. Nous pouvons quand même spécifier un modèle de travail $f(y|x;\theta)$ et calculer l'EMV de θ. Il est bien connu (White 1982) que l'EMV converge vers θ^*, le minimiseur de l'information de Kullback-Leibler

$$K(\theta)=E_g\left[\log\left\{\frac{g(Y|x)}{f(Y|x;\theta)}\right\}\right]$$

pour $\theta\in\Omega$. Sung et Geyer (2007) ont examiné les propriétés asymptotiques de l'EMV Monte Carlo de θ sous données manquantes.

Pour discuter formellement de la robustesse, supposons que la distribution véritable $g(y|x)$ appartient au voisinage

$$\mathcal{N}_\varepsilon=\left\{g;D(g,f)<\frac{1}{2}\varepsilon^2\right\} \tag{4.1}$$

pour un rayon $\varepsilon>0$, où

$$D(g,f)=\int log\left(\frac{g}{f}\right)g\ dy, \tag{4.2}$$

est la mesure de la distance de Kullback-Leibler. Le voisinage (4.1) peut être décrit de la façon suivante. Posons que $z(x,y,\theta)$ est une fonction de x, y et θ, normalisée pour satisfaire $E_{Y|x}(z)=0$ et $Var_{Y|x}(z)=1$, et définissons

$$g(y|x)=f(y|x;\theta)\exp\{\varepsilon z(x,y,\theta)-\kappa(x,\theta)\}, \tag{4.3}$$

où

$$\kappa=\log\left(E_{Y|x}\left[\exp\{\varepsilon z(x,Y,\theta)\}\right]\right).$$

Pour un petit $\varepsilon>0$, il peut être démontré que

$$\kappa\cong D(g,f)\cong\frac{1}{2}\varepsilon^2. \tag{4.4}$$

L'équation (4.3) représente un vaste ensemble de distributions proches de $f(y|x;\theta)$ créées en variant $z(x,y,\theta)$ sur différentes fonctions normalisées, où z et ε contiennent des interprétations géométriques qui représentent respectivement la direction et la grandeur des erreurs de spécification. Pour le paramètre θ, de dimension p, nous pouvons spécifier les directions des erreurs de spécification comme suit :

$$\left(z_1, z_2, \ldots, z_p\right)^T = I_\theta^{-1/2} s(x,y,\theta),$$

où $s(x,y,\theta) = \partial \log f(y|x;\theta)/\partial\theta$ et I_θ est la matrice d'information pour θ. Représentons $z(x,y,\theta)$ comme

$$z(x,y,\theta) = \lambda^T I_\theta^{-1/2} s(x,y,\theta),$$

où $\sum_{i=1}^{p} \lambda_i^2 = 1$. $z(x,y,\theta)$ satisfait alors aux critères de normalisation $E_{Y|x}(z) = 0$ et $Var_{Y|x}(z) = 1$. Voir Copas et Eguchi (2001) pour une discussion plus approfondie de cette expression.

Soit $w_{ij,g}^*$ le poids fractionnaire de la forme (3.3) selon la vraie densité g où $w_{ij,f}^*$ est le poids fractionnaire correspondant selon la « densité de travail » f. La construction spéciale du poids nous permet d'établir

$$w_{ij,g}^* \cong w_{ij,f}^* + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta}\left(w_{ij,f}^*\right). \tag{4.5}$$

La preuve de (4.5) est donnée à l'annexe A.2. Ainsi

$$\begin{aligned}\sum_i w_i \sum_j w_{ij,g}^* U(\eta; x_i, y_j) \cong & \sum_i w_i \sum_j w_{ij,f}^* U(\eta; x_i, y_j) \\ & + \varepsilon \lambda^T I_\theta^{-1/2} \sum_i w_i \sum_j \frac{\partial}{\partial\theta}\left(w_{ij,f}^*\right) U(\eta; x_i, y_j).\end{aligned} \tag{4.6}$$

Pour un petit ε, nous avons

$$\sum_i w_i \sum_j w_{ij,g}^* U(\eta; x_i, y_j) \cong \sum_i w_i \sum_j w_{ij,f}^* U(\eta; x_i, y_j),$$

et l'estimateur η résultant de $\sum_i w_i \sum_j w_{ij,f}^* U(\eta; x_i, y_j) = 0$ approchera donc la valeur réelle η_0.

5 Étude par simulation

Nous avons effectué deux études par simulation. Dans la section 5.1, nous avons comparé la performance de la méthode proposée avec celle d'autres méthodes d'imputation dans un modèle correctement spécifié et un modèle mal spécifié, respectivement, avec données manquantes ignorables. Dans la section 5.2, nous avons comparé la puissance statistique d'un test fondé sur l'IFHD plutôt que sur l'imputation multiple (IM).

5.1 Première simulation

La première étude de simulation testait la performance de la méthode proposée sous l'hypothèse de données manquantes ignorables. Nous avons utilisé deux ensembles de modèles pour générer les observations. Dans le modèle A, $y_i = 0,5x_i + e_i$, où $x_i \sim \exp(1)$, $e_i \sim N(0,1)$, x_i et e_i étant indépendants. Dans le modèle B, $y_i = 0,5x_i + e_i$, où $x_i \sim \exp(1)$, $e_i \sim \{\chi^2(2) - 2\}/2$, x_i et e_i étant indépendants. Les échantillons aléatoires de taille $n = 200$ ont été générés séparément à partir des deux modèles. Outre (x_i, y_i), nous avons généré δ_i à partir d'une Bernoulli(π_i), où $\pi_i = \{1 + \exp(-0,2 - x_i)\}^{-1}$. La variable x_i était toujours observée, mais la variable y_i l'était si et seulement si $\delta_i = 1$. Les taux de réponse globaux étaient d'environ 65% dans les deux cas. Nous avons utilisé $B = 2\ 000$ échantillons Monte Carlo lors de la simulation.

À partir de chacun des échantillons Monte Carlo, dont l'un avait été généré à l'aide du modèle A et l'autre à l'aide du modèle B, nous avons calculé les huit estimateurs suivants :

1. L'estimateur d'échantillon complet qui est calculé sur l'échantillon complet.

2. L'appariement d'après la moyenne prédictive (AMP) est une méthode d'imputation semi-paramétrique, qui attribue une valeur de manière aléatoire à partir des observations les plus proches de la valeur prédite tirée de $f(y \mid \mathbf{x})$. L'AMP a été mis en œuvre au moyen de la fonction « mice.impute.pmm » de R.

3. L'estimateur par imputation multiple (IM) où la taille du groupe d'imputation est $m = 10$, et où les valeurs imputées sont générées à partir du modèle de régression fondé sur la théorie normale, tel que l'envisagent Schenker et Welsh (1988).

4. L'estimateur par imputation fractionnaire paramétrique sans calage (IFP) où la taille du groupe d'imputation est $m = 10$.

5. L'estimateur par imputation fractionnaire paramétrique avec calage (IFP_cal) où la taille du groupe d'imputation est $m = 10$. Les poids fractionnaires sont calculés selon la méthode de calage en (3.6) où $\mathbf{q} = (y, y^2)$.

6. L'estimateur par imputation fractionnaire complète (IFC) utilisant l'ensemble complet de répondants comme valeurs d'imputation, c.-à-d. que la taille du groupe d'imputation est $m = n_R$, où n_R est la taille de A_R.

7. L'estimateur par imputation fractionnaire hot deck sans calage (IFHD) utilisant un petit sous-ensemble de répondants de taille $m = 10$ comme valeurs d'imputation.

8. L'estimateur par imputation fractionnaire hot deck avec calage (IFHD_cal) utilisant un petit sous-ensemble de répondants de taille $m = 10$ comme valeurs d'imputation. Les poids fractionnaires sont calculés selon la méthode de calage en (3.6) où $\mathbf{q} = (y, y^2)$.

L'imputation multiple est une façon de générer des valeurs imputées avec estimation simplifiée de la variance. Cette procédure envisage des méthodes bayésiennes de génération des valeurs imputées, où $m>1$ valeurs imputées sont générées à partir d'une loi prédictive a posteriori. À partir des valeurs imputées $\mathbf{y}^{*(1)},\ldots,\mathbf{y}^{*(m)}$, l'estimateur par imputation multiple η, désigné par $\hat{\eta}_{IM}$ est

$$\hat{\eta}_{IM}=\frac{1}{m}\sum_{k=1}^{m}\hat{\eta}^{(k)}$$

où $\hat{\eta}^{(k)}$ est l'estimateur de réponse complète appliqué au k-ième ensemble de données imputées. La formule de Rubin peut être utilisée pour estimer la variance par IM,

$$\hat{V}_{IM}\left(\hat{\eta}_{IM}\right)=W_m+\left(1+\frac{1}{m}\right)B_m, \tag{5.1}$$

où $W_m=m^{-1}\sum_{k=1}^{m}\hat{V}^{(k)}$, $B_m=\left(m-1\right)^{-1}\sum_{k=1}^{m}\left(\hat{\eta}^{(k)}-\hat{\eta}_{IM}\right)^2$, et $\hat{V}^{(k)}$ est l'estimateur de la variance de $\hat{\eta}^{(k)}$ sous réponse complète appliqué au k-ième ensemble de données imputées.

Dans les deux modèles, nous avons utilisé la densité normale de moyenne $\beta_0+\beta_1 x$ et de variance σ^2 comme modèle de travail pour l'imputation. Le modèle de travail est donc le vrai modèle dans le modèle A, mais pas dans le modèle B.

Nous avons considéré trois paramètres : $\theta_1=E\left(Y\right)$, la moyenne de population de y, $\theta_2=Pr\left(Y<1\right)$, la proportion de Y inférieure à un, et θ_3, le quantile 0,5 de Y. Pour estimer θ_2 sous échantillon complet, nous avons utilisé $\hat{\theta}_{2,n}=n^{-1}\sum_{i=1}^{n}I\left(y_i<1\right)$. Pour estimer θ_3 sous échantillon complet, nous avons utilisé $\hat{\theta}_{3,n}=\hat{F}^{-1}\left(p\right)=\inf\left\{y:\hat{F}\left(y\right)>p\right\}$, où $\hat{F}\left(y\right)=n^{-1}\sum_{i=1}^{n}I\left(y_i<y\right)$ et $p=0,5$.

Les tableaux 5.1 et 5.2 montrent les moyennes Monte Carlo, la variance normalisée et les erreurs quadratiques moyennes normalisées des huit estimateurs sous les modèles A et B, respectivement. La variance normalisée (erreur quadratique moyenne) est le ratio de la variance (erreur quadratique moyenne) et la variance (erreur quadratique moyenne) de l'estimateur d'échantillon complet multiplié par 100, ce qui mesure la variance (erreur quadratique moyenne) accrue attribuable à l'imputation par rapport à l'estimateur d'échantillon complet. En ce qui concerne les moyennes Monte Carlo (4e colonne), les estimateurs par imputation sont tous sans biais pour l'estimation de θ_1, θ_2, et θ_3 sous le modèle A. Sous le modèle B, l'AMP, l'IM, l'IFP et l'IFP_cal pour l'estimation de θ_3 sont beaucoup plus biaisés en valeurs absolues que l'IFC, l'IFHD et l'IFHD_cal lorsque le modèle est mal spécifié dans cette simulation. Pour ce qui est de la variance normalisée et de l'erreur quadratique moyenne normalisée (5e et 6e colonnes), l'IFP est plus efficace que l'IFHD parce qu'en IFP, les valeurs imputées sont générées directement en fonction de la distribution conditionnelle $f\left(y\mid x\right)$ tandis qu'en IFHD, les valeurs imputées peuvent être tirées des répondants dont les poids fractionnaires sont dominants. La taille effective des données imputées est déterminée par les observations imputées qui ont des poids fractionnaires importants, ce qui contribue aussi à la perte d'efficacité. L'IFHD perd en efficacité, mais gagne en robustesse. Enfin, l'IFHD où $m=10$ a une variance normalisée un peu plus grande que l'IFC pour θ_2, en raison de la variabilité additionnelle due à la procédure d'échantillonnage. En comparant l'IFP à l'IFP_cal et l'IFHD à l'IFHD_cal, l'étape du calage améliore légèrement l'efficacité. L'AMP affiche la plus grande variance quel que soit le scénario.

Tableau 5.1
Moyenne, variance normalisée (VN) et erreur quadratique moyenne normalisée (EQMN) Monte Carlo des estimateurs ponctuels dans le modèle A de la première simulation

Modèle	Paramètre	Méthode	Moyenne	VN	EQMN
A	μ_y	Complet	0,50	100	100
		AMP	0,50	175	175
		IM ($m = 10$)	0,50	135	135
		IFP ($m = 10$)	0,50	130	130
		IFP cal ($m = 10$)	0,50	130	130
		IFC ($m = n_R$)	0,50	130	130
		IFHD ($m = 10$)	0,50	156	156
		IFHD cal ($m = 10$)	0,50	130	130
	$Pr(Y < 1)$	Complet	0,68	100	100
		AMP	0,68	168	167
		IM ($m = 10$)	0,68	112	112
		IFP ($m = 10$)	0,68	110	110
		IFP cal ($m = 10$)	0,68	109	109
		IFC ($m = n_R$)	0,68	130	130
		IFHD ($m = 10$)	0,68	137	136
		IFHD cal ($m = 10$)	0,68	132	132
	Quantile	Complet	0,47	100	100
		AMP	0,47	184	184
		IM ($m = 10$)	0,47	111	111
		IFP ($m = 10$)	0,47	111	111
		IFP cal ($m = 10$)	0,47	111	111
		IFC ($m = n_R$)	0,47	135	135
		IFHD ($m = 10$)	0,47	142	142
		IFHD cal ($m = 10$)	0,47	141	141

Tableau 5.2
Moyenne, variance normalisée (VN) et erreur quadratique moyenne normalisée (EQMN) Monte Carlo des estimateurs ponctuels dans le modèle B de la première simulation

Modèle	Paramètre	Méthode	Moyenne	VN	EQMN
B	μ_y	Complet	0,50	100	100
		AMP	0,50	172	172
		IM ($m = 10$)	0,50	131	131
		IFP ($m = 10$)	0,50	131	131
		IFP cal ($m = 10$)	0,50	128	128
		IFC ($m = n_R$)	0,50	127	127
		IFHD ($m = 10$)	0,50	147	147
		IFHD cal ($m = 10$)	0,50	127	127
	$Pr(Y < 1)$	Complet	0,75	100	100
		AMP	0,75	166	166
		IM ($m = 10$)	0,73	140	170
		IFP ($m = 10$)	0,73	138	168
		IFP cal ($m = 10$)	0,73	137	169
		IFC ($m = n_R$)	0,75	137	137
		IFHD ($m = 10$)	0,75	145	145
		IFHD cal ($m = 10$)	0,75	140	141
	Quantile	Complet	0,26	100	100
		AMP	0,24	191	198
		IM ($m = 10$)	0,31	122	159
		IFP ($m = 10$)	0,31	123	160
		IFP cal ($m = 10$)	0,31	122	159
		IFC ($m = n_R$)	0,26	135	135
		IFHD ($m = 10$)	0,26	144	144
		IFHD cal ($m = 10$)	0,26	139	139

Nous avons examiné l'estimation de la variance par rééchantillonnage pour l'IFC et l'IFHD, particulièrement l'estimation de la variance jackknife avec suppression d'un groupe, qui est décrite à l'annexe A.1. Nous avons aussi envisagé l'estimation de la variance en IM, qui utilise la formule de Rubin (5.1).

Le tableau 5.3 montre les biais relatifs Monte Carlo des estimateurs de variance, qui sont calculés comme $\left[E_{MC}\left\{\hat{V}\right\}-V_{MC}\left\{\hat{\theta}\right\}\right]\Big/V_{MC}\left\{\hat{\theta}\right\}$, où $E_{MC}\left\{\hat{V}\right\}$ est la moyenne Monte Carlo des estimations de la variance $\hat{V}$, et $V_{MC}\left\{\hat{\theta}\right\}$ est la variance Monte Carlo des estimations ponctuelles $\hat{\theta}$. Le biais relatif de l'estimateur de la variance en IFC et IFHD est raisonnablement faible pour tous les paramètres envisagés dans les deux modèles, ce qui suggère que l'estimateur de la variance par rééchantillonnage est valide. Le biais relatif et la statistique t de l'estimateur de la variance en IM sont faibles pour θ_1 mais assez importants pour θ_2 même quand le modèle de travail est vrai (modèle A). La formule de Rubin repose sur la décomposition suivante :

$$V\left(\hat{\theta}_{IM}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{IM}-\hat{\theta}_n\right), \tag{5.2}$$

où $\hat{\theta}_n$ est l'estimateur d'échantillon complet de η. Essentiellement, le terme W_m en (5.1) estime $V\left(\hat{\theta}_n\right)$ et le terme $\left(1+m^{-1}\right)B_m$ en (5.1) estime $V\left(\hat{\theta}_{IM}-\hat{\theta}_n\right)$. La décomposition (5.2) est vérifiée lorsque $\hat{\theta}_n$ est l'EMV de θ, ce qui est la condition de compatibilité de $\hat{\theta}_n$ (Meng 1994). Dans les cas généraux, nous avons

$$V\left(\hat{\theta}_{IM}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{IM}-\hat{\theta}_n\right)+2Cov\left(\hat{\theta}_{IM}-\hat{\theta}_n,\hat{\theta}_n\right) \tag{5.3}$$

et l'estimateur de la variance de Rubin peut être biaisé si $Cov\left(\hat{\theta}_{IM}-\hat{\theta}_n,\hat{\theta}_n\right)\neq 0$. La condition de compatibilité est vérifiée pour l'estimation de la moyenne de population; elle ne l'est toutefois pas pour l'estimateur de $Pr\left(Y<1\right)$ par la méthode des moments. Il est à noter que l'estimateur imputé de $\theta_2=Pr\left(Y<1\right)$ peut s'exprimer comme suit :

$$\hat{\theta}_{2,I}=n^{-1}\sum_{i=1}^{n}\left[\delta_i I\left(y_i<1\right)+\left(1-\delta_i\right)E\left\{I\left(y_i<1\right)\mid x_i;\hat{\mu},\hat{\sigma}\right\}\right]. \tag{5.4}$$

Ainsi, les estimateurs imputés de θ_2 « empruntent de l'information » en utilisant des données additionnelles associées à $f\left(y\mid x\right)$, c'est-à-dire que la normalité de $f\left(y\mid x\right)$ est utilisée pour calculer l'espérance conditionnelle en (5.4), ce qui améliore l'efficacité de l'estimateur imputé pour θ_2. Le même phénomène s'applique à θ_3. Au tableau 5.1, l'augmentation de la variance due à l'imputation pour l'IM où $m=10$ est d'environ 35 % pour θ_1 mais de seulement 12 % et 11 % pour θ_2 et θ_3, respectivement, ce qui illustre le phénomène de l'« emprunt d'information » pour l'estimation de θ_2 et θ_3 grâce à l'utilisation de données additionnelles à l'étape de l'imputation. Ainsi, lorsque les conditions de compatibilité ne sont pas remplies, l'estimateur imputé améliore l'efficacité, mais l'estimateur de la variance de Rubin ne reconnaît pas cette amélioration.

Tableau 5.3
Biais relatif Monte Carlo de l'estimateur de la variance par rééchantillonnage dans la première simulation

Modèle	Paramètre	Méthode	B.R. (%)
*A	$V(\hat{\theta}_1)$	IM ($m=10$)	-2,33
		IFC ($m=n_R$)	-0,80
		IFHD_cal ($m=10$)	-0,80
	$V(\hat{\theta}_2)$	IM ($m=10$)	8,20
		IFC ($m=n_R$)	-5,01
		IFHD_cal ($m=10$)	-5,12
	$V(\hat{\theta}_3)$	IM ($m=10$)	19,84
		IFC ($m=n_R$)	4,50
		IFHD_cal ($m=10$)	3,78
*B	$V(\hat{\theta}_1)$	IM ($m=10$)	2,60
		IFC ($m=n_R$)	-0,56
		IFHD_cal ($m=10$)	-0,56
	$V(\hat{\theta}_2)$	IM ($m=10$)	-3,33
		IFC ($m=n_R$)	-1,89
		IFHD_cal ($m=10$)	-3,25
	$V(\hat{\theta}_3)$	IM ($m=10$)	-8,99
		IFC ($m=n_R$)	3,50
		IFHD_cal ($m=10$)	3,80

5.2 Deuxième simulation

La deuxième simulation testait la puissance de la méthode proposée dans un test d'hypothèse utilisant le modèle nul comme modèle d'imputation. Les échantillons de données bivariées (x_i, y_i) de taille $n=100$ ont été générés à partir de

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(x_i^2 - 1\right) + e_i \tag{5.5}$$

où $(\beta_0;\beta_1;\beta_2) = (0;0{,}9;0{,}06)$, $x_i \sim N(0;1)$, $e_i \sim N(0;0{,}16)$, x_i et e_i étant indépendants. La variable x_i est toujours observée, mais la probabilité que y_i réponde est de $0{,}5$. Les échantillons Monte Carlo ont été générés indépendamment $B=10\ 000$ fois. Nous voulons tester $H_0: \beta_2 = 0$ pour les répondants. Nous avons comparé l'IFHD à l'IM en utilisant la même taille de groupe d'imputation, soit $m=30$. Le modèle d'imputation est le modèle nul,

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

C'est-à-dire que le modèle d'imputation utilise des informations supplémentaires de $\beta_2 = 0$. À partir des données imputées, nous avons ajusté le modèle (5.5) et calculé la puissance d'un test $H_0: \beta_2 = 0$ au niveau significatif de $0{,}05$. Nous avons également envisagé la méthode des cas complets (MCC), qui utilise les répondants seulement pour la régression.

Le tableau 5.4 montre la moyenne et la variance Monte Carlo des estimateurs ponctuels, le biais relatif de l'estimateur de la variance et la puissance Monte Carlo des tests $H_0: \beta_2 = 0$. Dans chacun des échantillons Monte Carlo, nous avons construit un intervalle de confiance à 95% de Wald de β_2 en utilisant la formule $\left(\hat{\beta}_2 - 1{,}96\hat{V}^{1/2}; \hat{\beta}_2 + 1{,}96\hat{V}^{1/2}\right)$ et nous rejetons l'hypothèse nulle si $\beta_2 = 0$ ne tombe pas dans l'intervalle de confiance de Wald. La puissance Monte Carlo est calculée comme étant la fréquence relative du rejet de l'hypothèse nulle dans les échantillons Monte Carlo. Dans la deuxième colonne, les estimateurs IFHD et IM sont biaisés pour β_2, comme prévu, car le modèle d'imputation est le modèle nul et il est légèrement différent du modèle vrai qui a généré l'échantillon. Le biais de l'IFHD est plus petit que celui de l'IM en raison de la robustesse de l'IFHD examinée à la section 4. En IM, 50% des données imputées viennent du modèle nul et les 50% qui restent sont tirées du modèle vrai, de sorte que la pente β_2 est ramenée à zéro par la moitié de la vraie pente. En IFHD, même si nous avons utilisé le modèle nul pour calculer les poids fractionnaires, les données imputées viennent du modèle vrai, ce qui réduit le biais. L'IM fournit des estimateurs ponctuels plus efficaces que la MCC, mais l'estimation de la variance est très prudente (surestimation d'environ 180%). Étant donné le biais positif marqué de l'estimateur de variance IM, la puissance statistique des tests fondés sur l'IM est plus faible que celle des tests fondés sur la MCC. Par ailleurs, l'IFHD fournit des estimateurs ponctuels plus efficaces que la MCC, l'estimation de la variance est essentiellement sans biais, et la puissance statistique des tests fondés sur l'IFHD est plus élevée que celle des tests fondés sur la MCC.

Tableau 5.4
Résultats fondés sur 10 000 échantillons Monte Carlo de la deuxième simulation

Méthode	$E\left(\hat{\beta}_2\right)$	$V\left(\hat{\beta}_2\right)$	B.R.$\left(\hat{V}\right)$	Puissance
IFHD	0,046	0,00146	0,02	0,314
IM	0,028	0,00056	1,81	0,044
MCC	0,060	0,00234	-0,01	0,285

6 Conclusion

Nous avons proposé une méthode d'imputation fractionnaire hot deck qui utilise un modèle paramétrique pour $f(y|\mathbf{x})$ quand $\mathbf{x}$ contient des composantes continues. La méthode proposée fournit une estimation robuste pour les paramètres en ce sens que le modèle d'imputation n'est pas nécessairement égal au modèle générateur de données. Le prix que nous payons dans l'IFHD est la perte d'efficacité dans l'estimation ponctuelle. Dans notre première simulation, l'estimateur IFHD pour $P(Y<1)$ affiche la deuxième variance en importance, mais la plus petite erreur quadratique moyenne lorsque le modèle de travail n'est pas vrai, comparativement à d'autres estimateurs.

La perte d'efficacité tient principalement au fait que les poids fractionnaires sont plus variables que ceux obtenus selon la méthode de l'IFP parce que certains des $\mathbf{x}_j$ n'aident pas à imputer y_i, c'est-à-dire que la valeur de $f\left(y_i|\mathbf{x}_j;\hat{\theta}\right)$ peut être très faible. Lorsque le groupe d'imputation est très petit (p. ex.

$m = 10$), l'imputation fractionnaire hot deck ne fait pas augmenter la variance de façon significative, comme nous pouvons le voir au tableau 5.1 sous le modèle A.

En fait, la méthode d'imputation fractionnaire peut être utilisée pour élaborer une méthode d'imputation unique en appliquant l'IFHD où $m = 1,$ ce qui sélectionne une valeur imputée ayant une probabilité proportionnelle au poids fractionnaire pour chaque unité manquante. En l'occurrence, l'IFHD peut être utilisée pour élaborer une méthode d'imputation unique qui reste robuste aux erreurs de spécification du modèle. Le calage de pondération est toutefois incompatible avec une imputation unique. Nous pouvons quand même respecter les contraintes de calage en employant la méthode d'imputation équilibrée examinée par Chauvet, Deville et Haziza (2011), ou l'échantillonnage réjectif de Poisson de Fuller (2009). Un examen plus approfondi suivant cette piste fera l'objet d'une prochaine étude.

Remerciements

Nous remercions deux examinateurs anonymes et le rédacteur associé de leurs commentaires très utiles. Les travaux de recherche ont été financés en partie par une subvention du Conseil de recherches en sciences naturelles et en génie (MMS-121339) et par l'entente de coopération conclue entre le *Natural Resources Conservation Service* de l'USDA et le *Center for Survey Statistics and Methodology* de l'*Iowa State University*.

Annexe

A.1 Estimation de la variance par rééchantillonnage

Des méthodes de répliques peuvent être utilisées pour estimer la variance. Soit $w_i^{[k]}$ les k-ième poids de rééchantillonnage de sorte que

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k \left(\hat{Y}^{[k]} - \hat{Y} \right)^2$$

est convergent pour la variance de $\hat{Y} = \sum_{i \in A} w_i y_i,$ où L est la taille des répliques, c_k est le k-ième facteur de réplication qui dépend de la méthode des répliques et du mécanisme d'échantillonnage, et $\hat{Y}^{[k]} = \sum_{i \in A} w_i^{[k]} y_i$ est la k-ième répétition de $\hat{Y}$. En estimation de la variance jackknife avec suppression d'un groupe, $L = n$ et $c_k = (n-1)/n$.

Pour appliquer la méthode des répliques en IFC, nous devons d'abord appliquer les poids de rééchantillonnage $w_i^{[k]}$ en (2.4) afin de calculer $\hat{\theta}^{[k]}$. Après avoir obtenu $\hat{\theta}^{[k]}$, nous utilisons les mêmes valeurs imputées pour calculer les poids de rééchantillonnage fractionnaires initiaux

$$w_{ij}^{*[k]} \propto w_j^{[k]} w_j^{-1} f\left(y_j \mid x_i; \hat{\theta}^{[k]} \right) \Bigg/ \left\{ \sum_{l \in A_R} w_l^{[k]} f\left(y_j \mid x_l; \hat{\theta}^{[k]} \right) \right\}, \tag{A.1}$$

où $\sum_{j \in A_R} w_{ij}^{*[k]} = 1.$ La variance de $\hat{\eta}_{IFC},$ calculé en (3.4), est ensuite calculée comme suit :

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k \left(\hat{\eta}_{IFC}^{[k]} - \hat{\eta}_{IFC} \right)^2,$$

où $\hat{\eta}_{IFC}^{[k]}$ est obtenu en résolvant

$$\sum_{i \in A} w_i^{[k]} \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j \in A_R} w_{ij}^{*[k]} U\left(\eta; \mathbf{x}_i, y_j\right) \right\} = 0,$$

et $w_{ij}^{*[k]}$ est défini en (A.1).

Examinons maintenant l'estimation de la variance par rééchantillonnage de l'estimateur IFHD $\hat{\eta}_{IFHD}$ calculé en (3.8). Définissons $d_{ij} = 1$ si $j \in D_i$ et $d_{ij} = 0$ autrement. Il est à noter que $\hat{\eta}_{IFHD}$ est calculé en deux étapes. Dans la première étape, nous utilisons un échantillonnage PPT systématique où la probabilité de sélection est proportionnelle aux poids fractionnaires de la méthode IFC. Dans la deuxième étape, nous utilisons la méthode de pondération par calage en appliquant la contrainte (3.5) où $\sum_{j \in A_R} d_{ij} w_{ij,c}^{*} = 1$. Ainsi, les poids de rééchantillonnage fractionnaires sont eux aussi calculés en deux étapes. Premièrement, le poids de rééchantillonnage fractionnaire initial pour $w_{ij0}^{*} = 1/m$ est alors donné par

$$w_{ij0}^{*[k]} = \frac{d_{ij}\left(w_{ij}^{*[k]} / w_{ij}^{*}\right)}{\sum_{l \in A_R} d_{il}\left(w_{il}^{*[k]} / w_{il}^{*}\right)}, \tag{A.2}$$

où w_{ij}^{*} est le poids fractionnaire pour l'IFC défini en (2.6) et $w_{ij}^{*[k]}$ est le k-ième poids de rééchantillonnage fractionnaire pour l'IFC défini en (A.1). Deuxièmement, les poids de rééchantillonnage fractionnaires sont ajustés afin de respecter les contraintes de calage. L'équation de calage pour les poids de rééchantillonnage fractionnaires correspondant à (3.5) est alors

$$\sum_{i \in A} w_i^{[k]} \left\{ \left(1-\delta_i\right) \sum_{j \in D_i} w_{ij,c}^{*[k]} \mathbf{q}\left(\mathbf{x}_i, y_j\right) \right\} = \sum_{i \in A} w_i^{[k]} \left\{ \left(1-\delta_i\right) \sum_{j \in A_R} w_{ij}^{*[k]} \mathbf{q}\left(\mathbf{x}_i, y_j\right) \right\} \tag{A.3}$$

et $\sum_{j \in D_i} w_{ij,c}^{*[k]} = 1$. Nous pouvons utiliser la pondération par régression ou la pondération par entropie pour obtenir des poids de rééchantillonnage fractionnaires respectant les contraintes. Après avoir obtenu les poids de rééchantillonnage fractionnaires, nous calculons l'estimation par répliques $\hat{\eta}^{[k]}$ en résolvant

$$\sum_{i \in A} w_i^{[k]} \left\{ \delta_i U\left(\eta; x_i, y_i\right) + \left(1-\delta_i\right) \sum_{j \in A_R} w_{ijc}^{*[k]} U\left(\eta; x_i, y_j\right) \right\} = 0.$$

L'estimateur de la variance par rééchantillonnage de $\hat{\eta}$, calculé en (3.8) est donné par

$$\hat{V}_{rep}\left(\hat{\eta}\right) = \sum_{k=1}^{L} c_k \left(\hat{\eta}^{[k]} - \hat{\eta} \right)^2.$$

Comme $\hat{\eta}$ est une fonction lisse de $\hat{\theta}$, la convergence de $\hat{V}_{rep}(\hat{\eta})$ découle directement de l'argument standard de l'estimation de la variance par rééchantillonnage (Shao et Tu 1995).

A.2 Preuve de la formule (4.5)

Utilisons

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} = \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)} \exp\left(\varepsilon \Delta_{ik|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)$$

où $\Delta_{ik|j} = z\left(x_i, y_j; \theta\right) - z\left(x_k, y_j; \theta\right)$. En nous fondant sur la linéarisation de Taylor et sur (4.4), nous avons

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} \cong \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)} \left\{1 + \varepsilon \Delta_{ik|j}\right\}.$$

Si nous connaissons la véritable densité, les poids fractionnaires corrects en (3.3) peuvent être exprimés comme suit :

$$\begin{aligned}
w_{ij,g}^{*} &\propto \frac{g\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k g\left(y_j \mid x_k\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{g(y_j \mid x_k)}{g(y_j \mid x_i)}\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \exp\left(\varepsilon \Delta_{ki|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)\right\}} \\
&\cong \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \left(1 + \varepsilon \Delta_{ki|j}\right)\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \sum_{k;\delta_k=1} w_k \left[\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \left\{z\left(x_k, y_j\right) - z\left(x_i, y_j\right)\right\}\right]} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \left(\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \left\{s\left(x_k, y_j\right) - s\left(x_i, y_j\right)\right\}\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta} \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}
\end{aligned}$$

$$\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{ \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta} \left\{ \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}}{\sum_{k;\delta_k=1} w_k \left\{ \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}} \right]$$

$$\propto \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\frac{\partial}{\partial \theta} \left\{ \sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \right\}}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}} \right]$$

$$= \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \left\{ \frac{1}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}} \right\}$$

$$= a_{ij} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij},$$

où $a_{ij} = f\left(y_j \mid x_i\right) \Big/ \sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)$ et $a_{i+} = \sum_{j;\delta_j=1} a_{ij}$. Ainsi, $w_{ij,f}^* = a_{ij} / a_{i+}$ et

$$w_{ij,g}^* \cong \frac{a_{ij} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij}}{a_{i+} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{i+}}$$

$$= \frac{a_{ij}}{a_{i+}} \left(1 + \frac{\varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij}}{a_{ij}} \right) \left(\frac{a_{i+}}{a_{i+} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{i+}} \right)$$

$$\cong \frac{a_{ij}}{a_{i+}} \left(1 + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \log a_{ij} \right) \left(1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \log a_{i+} \right)$$

$$\cong \frac{a_{ij}}{a_{i+}} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{a_{ij}}{a_{i+}} \left(\frac{\partial}{\partial \theta} \log a_{ij} - \frac{\partial}{\partial \theta} \log a_{i+} \right)$$

$$= \frac{a_{ij}}{a_{i+}} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \left(\frac{a_{ij}}{a_{i+}} \right)$$

$$= w_{ij,f}^* + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} \left(w_{ij,f}^* \right),$$

ce qui prouve (4.5).

A.3 Extension à un cas de données manquantes non ignorables

Nous considérons une extension de la méthode proposée à un cas de données manquantes non ignorables. Selon l'hypothèse des données manquantes non ignorables, le modèle conditionnel $f\left(y \mid x\right)$ et

le modèle de probabilité de réponse $P(\delta=1|\mathbf{x},y)$ sont nécessaires pour évaluer la fonction d'estimation prévue en (4.6). Soit le modèle de probabilité de réponse donné par $Pr(\delta_i=1|\mathbf{x}_i,y_i)=\pi(\mathbf{x}_i,y_i;\phi)$ pour certains ϕ ayant une fonction $\pi(\cdot)$ connue. Nous supposons que les paramètres sont identifiables, comme en ont discuté Wang, Shao et Kim (2013).

En IFP, selon Kim et Kim (2012), l'EMV $\left(\hat{\theta},\hat{\phi}\right)$ peut être obtenu en résolvant

$$\sum_{i\in A} w_i\left\{\delta_i S(\theta;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*} S\left(\theta;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0, \tag{A.4}$$

et

$$\sum_{i\in A} w_i\left\{\delta_i S(\phi;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*} S\left(\phi;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0, \tag{A.5}$$

où $S(\theta;\mathbf{x},y)=\partial\log f(y|\mathbf{x};\theta)/\partial\theta$, $S(\phi;\mathbf{x},y)=\partial\log\pi(\mathbf{x},y;\phi)/\partial\phi$, et les poids fractionnaires sont donnés par

$$w_{ij}^{*}(\theta,\phi)=\frac{f\left(y_i^{*(j)}|\mathbf{x}_i;\theta\right)\left\{1-\pi\left(\mathbf{x}_i,y_i^{*(j)},\phi\right)\right\}/h\left(y_i^{*(j)}|\mathbf{x}_i\right)}{\sum_{k=1}^{m}\left[f\left(y_i^{*(k)}|\mathbf{x}_i;\theta\right)\left\{1-\pi\left(\mathbf{x}_i,y_i^{*(k)},\phi\right)\right\}/h\left(y_i^{*(k)}|\mathbf{x}_i\right)\right]}. \tag{A.6}$$

La solution de (A.4) et (A.5) peut être obtenue au moyen de l'algorithme EM. Dans l'algorithme EM, l'étape E calcule les poids fractionnaires en (A.6) en utilisant les valeurs paramétriques actuelles et l'étape M met à jour les valeurs paramétriques $\hat{\theta}^{(t+1)}$ et $\hat{\phi}^{(t+1)}$ en résolvant

$$\sum_{i\in A} w_i\left\{\delta_i S(\theta;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*}\left(\hat{\theta}^{(t)},\hat{\phi}^{(t)}\right) S\left(\theta;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0,$$

et

$$\sum_{i\in A} w_i\left\{\delta_i S(\phi;\mathbf{x}_i,y_i)+(1-\delta_i)\sum_{j=1}^{m} w_{ij}^{*}\left(\hat{\theta}^{(t)},\hat{\phi}^{(t)}\right) S\left(\phi;\mathbf{x}_i,y_i^{*(j)}\right)\right\}=0.$$

Dans la méthode IFC proposée, les poids fractionnaires sont donnés par

$$\begin{aligned} w_{ij}^{*} &\propto f\left(y_j|\mathbf{x}_i,\delta_i=0;\theta,\phi\right)/f\left(y_j|\delta_j=1\right) \\ &\propto f\left(y_j|\mathbf{x}_i,\theta\right)\left\{1-\pi\left(\mathbf{x}_i,y_j;\phi\right)\right\}/f\left(y_j|\delta_j=1\right), \end{aligned}$$

où $\sum_{j;\delta_j=1} w_{ij}^{*}=1$. Parce que

$$\begin{aligned} f\left(y_j|\delta_j=1\right) &= \int\pi\left(\mathbf{x},y_j\right) f\left(y_j|\mathbf{x}\right) f(\mathbf{x})d\mathbf{x} \\ &\cong \sum_{k\in A} w_k\pi\left(\mathbf{x}_k,y_j\right) f\left(y_j|\mathbf{x}_k\right). \end{aligned} \tag{A.7}$$

Les poids fractionnaires peuvent être calculés à partir de

$$w_{ij}^{*} \propto \frac{f\left(y_j \mid \mathbf{x}_i;\theta\right)\left\{1-\pi\left(\mathbf{x}_i, y_j;\phi\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j;\phi\right) f\left(y_j \mid \mathbf{x}_k;\theta\right)}. \tag{A.8}$$

où $\sum_{j\in A_R} w_{ij}^{*} = 1.$

Nous pouvons donc utiliser l'algorithme EM suivant pour obtenir les estimations paramétriques désirées.

(Étape I) Pour chaque unité manquante $i \in A_M = \left\{i \in A; \delta_i = 0\right\}$, prendre m valeurs imputées comme $y_i^{(1)},\ldots,y_i^{(m)}$ à partir de A_R, où $m = r.$

(Étape E) Les poids fractionnaires sont donnés par

$$w_{ij}^{*(t)} \propto \frac{f\left(y_j \mid \mathbf{x}_i, \hat{\theta}^{(t)}\right)\left\{1-\pi\left(\mathbf{x}_i, y_j;\hat{\phi}^{(t)}\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j;\hat{\phi}^{(t)}\right) f\left(y_j \mid \mathbf{x}_k;\hat{\theta}^{(t)}\right)}$$

et $\sum_{j=1}^{m} w_{ij}^{*(t)} = 1.$

(Étape M) Mettre à jour les paramètres $\hat{\theta}^{(t+1)}$ et $\hat{\phi}^{(t+1)}$ en résolvant les équations de score imputées suivantes :

$$\sum_{i\in A} w_i \left\{\delta_i S\left(\theta;\mathbf{x}_i, y_i\right) + \left(1-\delta_i\right)\sum_{j\in A_R} w_{ij}^{*(t)} S\left(\theta;\mathbf{x}_i, y_j\right)\right\} = 0,$$

et

$$\sum_{i\in A} w_i \left\{\delta_i S\left(\phi;\mathbf{x}_i, y_i\right) + \left(1-\delta_i\right)\sum_{j\in A_R} w_{ij}^{*(t)} S\left(\phi;\mathbf{x}_i, y_j\right)\right\} = 0.$$

Il est à noter que l'étape I n'a pas besoin d'être répétée dans l'algorithme EM. Une fois les estimations paramétriques finales obtenues, les poids fractionnaires sont calculés selon la formule en (A.8) et ils servent de probabilités de sélection pour l'IFHD lorsque le groupe d'imputation est de petite taille m. La méthode d'échantillonnage PPT systématique examinée à la section 3 peut aussi être utilisée pour obtenir l'IFHD.

Bibliographie

Andersen, P.K. et Gill, R.D. (1982). Cox's regression model for counting process: a large sample study. *The Annals of Statistics*, 10, 1100-1120.

Andridge, R.R. et Little, R.J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78, 40-64.

Beaumont, J.F. et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Binder, D. et Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Chauvet, G., Deville, J.-C. et Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98, 459-471.

Chen, J. et Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Copas, J.B. et Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Series B*, 63, 871-895.

Durrant, G.B. (2009). Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. *International Journal of Social Research Methodology*, 12, 293-304.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Fuller, W.A. et Kim, J.K. (2005). Imputation hot deck pour le modèle de réponse. *Techniques d'enquête*, 31, 153-164.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics*, 29, 215-246.

Kalton, G. et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, Series A*, 13, 1919-1939.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

Kim, J.K. et Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Fuller, W.A. et Bell, W.R. (2011). Variance estimation for nearest neighbor imputation for U.S. census long form data. *Annals of Applied Statistics*, 5, 824-842.

Kim, J.K. et Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman and Hall/CRC.

Kim, J.Y. et Kim, J.K. (2012). Parametric fractional imputation for nonignorable missing data. *Journal of the Korean Statistical Society*, 41, 291-303.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.

Pfeffermann, D. (2011). Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? *Techniques d'enquête*, 37, 123-146.

Rao, J.N.K., Yung, W. et Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *The Indian Journal of Statistics, Series A*, 64, 364-378.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

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

Schenker, N. et Welsh, A.H. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, 16, 1550-1566.

Shao, J. et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer.

Sung, Y.J. et Geyer, C.J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, 35, 990-1011.

Wang, C.-Y. et Pepe, M.S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B*, 62, 509-24.

Wang, S., Shao J. et Kim, J.K. (2013). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*. In press.

Wei, G.C. et Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699-704.

White, H. (1982). Maximum likelihood estimation of misspecifed models. *Econometrica*, 50, 1-25.