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

Techniques d'enquête

Appariement statistique par imputation fractionnaire

par Jae Kwang Kim, Emily Berg et Taesung Park

Date de diffusion : le 22 juin 2016

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 STATCAN.infostats-infostats.STATCAN@canada.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

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 21-43
Statistique Canada, nº 12-001-X au catalogue

Appariement statistique par imputation fractionnaire

Jae Kwang Kim, Emily Berg et Taesung Park[1]

Résumé

L'appariement statistique est une technique permettant d'intégrer deux ou plusieurs ensembles de données lorsque les renseignements nécessaires pour apparier les enregistrements des participants individuels dans les ensembles de données sont incomplets. On peut considérer l'appariement statistique comme un problème de données manquantes en vertu duquel on souhaite effectuer une analyse conjointe de variables qui ne sont jamais observées ensemble. On utilise souvent une hypothèse d'indépendance conditionnelle pour créer des données imputées aux fins d'appariement statistique. Nous examinons une approche générale de l'appariement statistique faisant appel à l'imputation fractionnaire paramétrique de Kim (2011) pour créer des données imputées en vertu de l'hypothèse que le modèle spécifié est entièrement identifié. La méthode proposée ne produit pas une séquence espérance-maximisation (EM) convergente si le modèle n'est pas identifié. Nous présentons aussi des estimateurs de variance convenant à la procédure d'imputation. Nous expliquons comment la méthode s'applique directement à l'analyse des données obtenues à partir de plans de sondage à questionnaire scindé et aux modèles d'erreur de mesure.

Mots-clés : Combinaison de données; fusion de données; imputation hot deck; plan de sondage à questionnaire scindé; modèle d'erreur de mesure.

1 Introduction

L'échantillonnage d'enquête est un outil scientifique permettant de faire des inférences à propos de la population cible. Toutefois, il arrive souvent que toutes les données nécessaires ne soient pas recueillies dans le cadre d'une même enquête, à cause de contraintes de temps et de coût. Dans ce cas, on souhaite exploiter le plus possible les données existantes provenant d'autres sources portant sur la même population cible. L'appariement statistique, que l'on appelle parfois « fusion de données » (Baker, Harris et O'Brien 1989) ou « combinaison de données » (Ridder et Moffit 2007), vise à intégrer deux ou plusieurs ensembles de données lorsque les renseignements nécessaires pour apparier les enregistrements des participants individuels dans les ensembles de données sont incomplets. D'Orazio, Zio et Scanu (2006) ainsi que Leulescu et Agafitei (2013) présentent un bon aperçu des techniques d'appariement statistique dans l'échantillonnage d'enquête.

L'appariement statistique peut être considéré comme un problème de données manquantes en vertu duquel on souhaite effectuer une analyse conjointe de variables qui ne sont jamais observées ensemble. Moriarity et Scheuren (2001) proposent un cadre théorique pour l'appariement statistique en vertu d'une hypothèse de normalité multivariée. Rässler (2002) a mis au point des techniques d'imputation multiple pour l'appariement statistique à l'aide de valeurs prédéterminées pour les paramètres non identifiables. Lahiri et Larsen (2005) traitent de l'analyse par régression à l'aide de données couplées. Ridder et Moffit (2007) présentent un traitement rigoureux des hypothèses et des approches pour l'appariement statistique dans le domaine de l'économétrie.

L'appariement statistique vise à construire des fichiers de données entièrement augmentées pour effectuer des analyses conjointes statistiquement valides. Pour simplifier la mise en situation, supposons

1. Jae Kwang Kim, Département de statistique, Iowa State University, Ames, IA 50011, États-Unis. Courriel : jkim@iastate.edu; Emily Berg, Département de statistique, Iowa State University, Ames, Iowa, États-Unis. Courriel : emilyb@iastate.edu; Taesung Park, Département de statistique, Université nationale de Séoul, Séoul, Corée. Courriel : taesungp@gmail.com.

que deux enquêtes, l'enquête A et l'enquête B, offrent des données partielles à propos de la population, et que l'on observe x et y_1 dans l'échantillon de l'enquête A et x et y_2 dans l'échantillon de l'enquête B. Le tableau 1.1 illustre une structure de données simple pour l'appariement. Si l'échantillon de l'enquête B (échantillon B) est un sous-ensemble de l'échantillon de l'enquête A (échantillon A), on peut employer les techniques de couplage d'enregistrements (Herzog, Scheuren et Winkler 2007) pour obtenir les valeurs de y_1 pour l'échantillon de l'enquête B. Toutefois, dans de nombreux cas, un tel appariement parfait n'est pas possible (par exemple, parce que les échantillons peuvent contenir des sous-ensembles non chevauchants); on dépend alors d'une méthode probabiliste d'identification des « jumeaux statistiques » de l'autre échantillon, c'est-à-dire que l'on doit créer y_1 pour chaque élément de l'échantillon B en trouvant son plus proche voisin dans l'échantillon A. L'imputation par la méthode du plus proche voisin a été examinée par de nombreux auteurs, dont Chen et Shao (2001) et Beaumont et Bocci (2009), dans le contexte des réponses manquantes.

Tableau 1.1
Structure de données simple pour l'appariement

	$\boldsymbol{X}$	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$
Échantillon A	o	o	
Échantillon B	o		o

La détermination du plus proche voisin repose souvent sur la « proximité » en fonction de la valeur de x seulement. Ainsi, dans de nombreux cas, l'appariement statistique est fondé sur l'hypothèse que y_1 et y_2 sont indépendants, conditionnellement à x, c'est-à-dire

$$y_1 \perp y_2 \,|\, x. \tag{1.1}$$

L'hypothèse (1.1) est souvent appelée « hypothèse d'indépendance conditionnelle (IC) » et est très utilisée dans la pratique.

Dans le présent article, nous examinons une autre approche, qui ne repose pas sur l'hypothèse d'IC. Nous présentons les hypothèses à la section 2, puis les méthodes proposées à la section 3. Nous examinons en outre deux extensions de l'approche, l'une aux plans de sondage à questionnaire scindé (section 4) et l'autre aux modèles d'erreur de mesure (section 5). Les résultats de deux études par simulation sont présentés à la section 6. La section 7 conclut l'article.

2 Scénario de base

Pour simplifier la présentation, nous considérons deux enquêtes indépendantes réalisées auprès de la même population cible consistant en N éléments. Comme il est précisé à la section 1, supposons que l'échantillon A comporte des données uniquement à propos de x et de y_1 et que l'échantillon B comporte des données uniquement à propos de x et de y_2.

Pour illustrer cette idée, supposons pour l'instant que les variables (x, y_1, y_2) sont générées à partir d'une distribution normale comme suit :

$$\begin{pmatrix} x \\ y_1 \\ y_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{1x} & \sigma_{2x} \\ & \sigma_{11} & \sigma_{12} \\ & & \sigma_{22} \end{pmatrix}\right].$$

Selon la structure de données présentée dans le tableau 1.1, il est clair que le paramètre σ_{12} ne peut pas être estimé à partir des échantillons. Il découle de l'hypothèse d'indépendance conditionnelle énoncée en (1.1) que $\sigma_{12} = \sigma_{1x}\sigma_{2x} / \sigma_{xx}$ et que $\rho_{12} = \rho_{1x}\rho_{2x}$, c'est-à-dire que σ_{12} est entièrement déterminé à partir d'autres paramètres, plutôt qu'estimé directement à partir des échantillons réalisés.

Dans ce cas, l'imputation de données synthétiques en vertu de l'hypothèse d'indépendance conditionnelle peut se faire en deux étapes :

[Étape 1] Estimer $f(y_1 \mid x)$ à partir de l'échantillon A, et désigner l'estimation $\hat{f}_a(y_1 \mid x)$.

[Étape 2] Pour chaque élément i de l'échantillon B, utiliser la valeur de x_i pour générer des valeurs imputées de y_1 à partir de $\hat{f}_a(y_1 \mid x_i)$.

Comme les valeurs de y_1 ne sont jamais observées dans l'échantillon B, des valeurs synthétiques de y_1 sont créées pour tous les éléments de l'échantillon B, ce qui donne lieu à une imputation synthétique. Haziza (2009) présente un bon examen des publications relatives à la méthodologie d'imputation. Kim et Rao (2012) présentent une approche assistée par modèle pour l'imputation synthétique lorsque seul x est disponible dans l'échantillon B. Une telle imputation synthétique ne tient absolument pas compte des données observées pour y_2 dans l'échantillon B.

L'appariement statistique fondé sur l'indépendance conditionnelle suppose que $\mathrm{Cov}(y_1, y_2 \mid x) = 0$. Ainsi, la régression de y_2 sur x et y_1 à partir des données imputées issues de l'imputation synthétique ci-dessus estimera un coefficient de régression nul pour y_1. Autrement dit, l'estimation $\hat{\beta}_2$ pour

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y_1$$

donnera une valeur nulle. De telles analyses peuvent être trompeuses si l'IC n'est pas vérifiée. Pour comprendre pourquoi, posons un problème de régression à variable omise :

$$\begin{aligned} y_1 &= \beta_0^{(1)} + \beta_1^{(1)} x + \beta_2^{(1)} z + e_1 \\ y_2 &= \beta_0^{(2)} + \beta_1^{(2)} x + \beta_2^{(2)} z + e_2 \end{aligned}$$

où les variables z, e_1, e_2 sont indépendantes et non observées. Sauf si $\beta_2^{(1)} = \beta_2^{(2)} = 0$, la variable latente z est un facteur de confusion non observable qui explique pourquoi $\mathrm{Cov}(y_1, y_2 \mid x) \neq 0$. Ainsi, le coefficient de y_1 dans la régression de la population de y_2 sur x et y_1 n'est pas nul.

Soulignons que l'hypothèse d'IC concerne l'identification du modèle. L'hypothèse de variable instrumentale (VI) constitue une autre hypothèse d'identification, décrite ci-dessous.

Remarque 2.1 *Nous présentons une description formelle de l'hypothèse de VI. D'abord, présumons que* x *se décompose en* $x = (x_1, x_2)$ *de sorte que*

(*i*) $f(y_2 | x_1, x_2, y_1) = f(y_2 | x_2, y_1)$

(*ii*) $f(y_1 | x_2, x_1 = a) \neq f(y_1 | x_2, x_1 = b)$

pour $a \neq b$. *Ainsi,* x_1 *est conditionnellement indépendante de* y_2 *sachant* x_2 *et* y_1, *mais* x_1 *est corrélée avec* y_1 *sachant* x_2. *Soulignons que* x_2 *peut être nulle ou avoir une distribution dégénérée, par exemple une ordonnée à l'origine. La variable* x_1 *satisfaisant aux deux conditions ci-dessus est souvent appelée une variable instrumentale (VI) pour* y_1. *Le graphe acyclique orienté de la figure 2.1 illustre la structure de dépendance d'un modèle assorti d'une variable instrumentale. Ridder et Moffit (2007) ont utilisé des « contraintes d'exclusion » pour décrire l'hypothèse de variable instrumentale. Les enquêtes répétées sont un exemple de cas où il est raisonnable de poser une hypothèse de variable instrumentale. Supposons une enquête répétée où* y_t *est la variable étudiée à l'année* t *et vérifie la propriété de Markov*

$$P(y_{t+1} | y_1, \ldots, y_t) = P(y_{t+1} | y_t),$$

où $P(y_t)$ *désigne une fonction de distribution cumulative. Dans ce cas,* y_{t-1} *est une variable instrumentale pour* y_t. *En fait, la dernière observation de* y_s $(s \leq t)$, *quelle qu'elle soit, est la variable instrumentale pour* y_t.

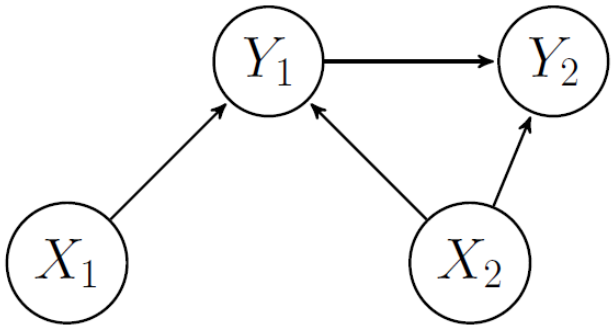

Figure 2.1 Structure de dépendance pour un modèle où x_1 est une variable instrumentale pour y_1 et où x_2 est une covariable supplémentaire dans les modèles pour y_2 et y_1.

En vertu de l'hypothèse de variable instrumentale, on peut utiliser une régression en deux étapes pour estimer les paramètres de régression d'un modèle linéaire. L'exemple suivant présente les concepts de base.

Exemple 2.1 *Prenons la structure des données de deux échantillons présentée dans le tableau 1.1. On présume le modèle de régression linéaire suivant :*

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \beta_2 x_{2i} + e_i, \tag{2.1}$$

où $e_i \sim (0, \sigma_e^2)$ *et* e_i *est indépendante de* (x_{1j}, x_{2j}, y_{1j}) *pour toutes les valeurs de* i, j. *Dans ce cas, on peut obtenir un estimateur convergent de* $\beta = (\beta_0, \beta_1, \beta_2)'$ *à l'aide de la méthode des moindres carrés en deux étapes (MC2E) comme suit :*

1. *À partir de l'échantillon A, on ajuste le « modèle de travail » suivant pour* y_1

$$y_{1i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i, \quad u_i \sim (0, \sigma_u^2) \tag{2.2}$$

pour obtenir un estimateur convergent de $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ *défini par*

$$\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)' = (X'X)^{-1} X'Y_1$$

où $X = [X_0, X_1, X_2]$ *est une matrice dont la* i^e *ligne est* $(1, x_{1i}, x_{2i})$ *et* Y_1 *est un vecteur dont* y_{1i} *est la* i^e *composante.*

2. *On obtient un estimateur convergent de* $\beta = (\beta_0, \beta_1, \beta_2)'$ *à l'aide de la méthode des moindres carrés pour la régression de* y_{2i} *sur* $(1, \hat{y}_{1i}, x_{2i})$ *où* $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$.

La question de l'absence asymptotique de biais de l'estimateur par les MC2E en vertu de l'hypothèse de variable instrumentale est abordée à l'annexe A. La méthode des MC2E n'est pas applicable directement si le modèle de régression (2.1) n'est pas linéaire. En outre, bien que la méthode des MC2E permette d'estimer les paramètres de régression, elle ne fournit pas des estimateurs convergents pour les paramètres plus généraux comme $\theta = \Pr(y_2 < 1 | y_1 < 3)$. L'imputation stochastique peut constituer une solution pour estimer une classe plus générale de paramètres. Nous expliquons comment modifier l'imputation fractionnaire paramétrique de Kim (2011) pour effectuer une estimation générale dans le contexte d'un problème d'appariement statistique.

3 Imputation fractionnaire

Nous allons maintenant décrire les méthodes d'imputation fractionnaire aux fins d'appariement statistique sans avoir recours à l'hypothèse d'IC. L'utilisation de l'imputation fractionnaire pour l'appariement statistique a été présentée pour la première fois dans le chapitre 9 de Kim et Shao (2013) en vertu de l'hypothèse de VI. Dans le présent article, nous présentons la méthodologie sans recourir à l'hypothèse de VI. Nous présumons seulement que le modèle spécifié est entièrement identifié. L'identifiabilité du modèle spécifié peut facilement être vérifiée dans le calcul de la procédure proposée.

Pour expliquer l'idée, rappelons que la variable y_1 est absente de l'échantillon B et que le but est de générer y_1 à partir de la distribution conditionnelle de y_1 sachant les observations. Autrement dit, nous voulons générer y_1 à partir de

$$f(y_1 | x, y_2) \propto f(y_2 | x, y_1) f(y_1 | x). \tag{3.1}$$

Pour ce faire, on peut utiliser la stratégie d'imputation en deux étapes suivante :

1. Générer y_1^* à partir de $\hat{f}_a(y_1 | x)$.

2. Accepter y_1^* si $f\left(y_2 \mid x, y_1^*\right)$ est suffisamment grande.

Soulignons que la première étape est la méthode habituelle en vertu de l'hypothèse d'IC. La deuxième étape intègre l'information dans y_2. Pour déterminer si $f\left(y_2 \mid x, y_1^*\right)$ est suffisamment grande à l'étape 2, on applique souvent une méthode Monte Carlo par chaîne de Markov (MCMC), par exemple l'algorithme de Metropolis-Hastings (Chib et Greenberg 1995). Soit $y_1^{(t-1)}$ la valeur courante de y_1 dans la chaîne de Markov; on accepte alors y_1^* selon la probabilité

$$R\left(y_1^*, y_1^{(t-1)}\right) = \min\left\{1, \frac{f\left(y_2 \mid x, y_1^*\right)}{f\left(y_2 \mid x, y_1^{(t-1)}\right)}\right\}.$$

De tels algorithmes peuvent devenir fastidieux à calculer, à cause de la convergence lente de l'algorithme MCMC.

L'imputation fractionnaire paramétrique de Kim (2011) permet de générer les valeurs imputées en (3.1) sans recourir à la méthode MCMC. On peut utiliser l'algorithme espérance-maximisation (EM) par imputation fractionnaire suivant :

1. Pour tout $i \in B$, générer m valeurs imputées de y_{1i}, désignées par $y_{1i}^{*(1)}, \ldots, y_{1i}^{*(m)}$, à partir de $\hat{f}_a(y_1 \mid x_i)$, où $\hat{f}_a(y_1 \mid x)$ correspond à la densité estimée pour la distribution conditionnelle de y_1 sachant x obtenue à partir de l'échantillon A.
2. Soit $\hat{\theta}_t$ la valeur courante du paramètre θ dans $f(y_2 \mid x, y_1)$. Pour la j^{e} valeur imputée $y_{1i}^{*(j)}$, affecter le poids fractionnaire

   $$w_{ij(t)}^* \propto f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right)$$

   de sorte que $\sum_{j=1}^m w_{ij}^* = 1$.
3. Résoudre l'équation de score obtenue par imputation fractionnaire pour θ

   $$\sum_{i \in B} w_{ib} \sum_{j=1}^m w_{ij(t)}^* S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0 \tag{3.2}$$

   pour obtenir $\hat{\theta}_{t+1}$, où $S(\theta; x, y_1, y_2) = \partial \log f(y_2 \mid x, y_1; \theta) / \partial \theta$, et w_{ib} correspond au poids d'échantillonnage de l'unité i dans l'échantillon B.
4. Reprendre à l'étape 2 et poursuivre jusqu'à la convergence.

Une fois le modèle identifié, la séquence EM obtenue à partir de la méthode d'imputation fractionnaire paramétrique ci-dessus converge. Si le modèle spécifié n'est pas identifiable, c'est qu'il n'y a pas de solution unique pour maximiser la vraisemblance observée et la séquence EM ci-dessus ne converge pas. Soulignons qu'en (3.2), pour une valeur suffisamment grande de m,

$$\begin{aligned}\sum_{j=1}^m w_{ij(t)}^* S\left(\theta; x_i, y_{1i}^{*(j)}, y_{2i}\right) &\cong \frac{\int S(\theta; x_i, y_1, y_{2i})\, f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 \mid x_i)\, dy_1}{\int f\left(y_{2i} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a(y_1 \mid x_i)\, dy_1} \\ &= E\left\{S(\theta; x_i, Y_1, y_{2i}) \mid x_i, y_{2i}; \hat{\theta}_t\right\}.\end{aligned}$$

Si y_{i1} est catégorique, le poids fractionnaire peut être établi par la probabilité conditionnelle correspondant à la valeur imputée réalisée (Ibrahim 1990). On a recours à l'étape 2 pour intégrer les données observées de y_{i2} dans l'échantillon B. Précisons que l'étape 1 n'est pas répétée pour chaque itération. Seules les étapes 2 et 3 sont reprises jusqu'à ce qu'il y ait convergence. Comme l'étape 1 n'est pas répétée, la convergence est garantie et la vraisemblance observée augmente, à condition que le modèle soit identifiable (voir le théorème 2 de Kim [2011]).

Remarque 3.1 *À la section 2, il est question de VI uniquement parce que c'est la façon la plus répandue d'assurer l'identifiabilité. La méthode proposée ici ne dépend pas de cette hypothèse. Pour illustrer une situation où il est possible d'identifier le modèle sans l'hypothèse de VI, supposons le modèle suivant :*

$$\begin{aligned} y_2 &= \beta_0 + \beta_1 x + \beta_2 y_1 + e_2 \\ y_1 &= \alpha_0 + \alpha_1 x + e_1 \end{aligned}$$

où $e_1 \sim N\left(0, x^2\sigma_1^2\right)$ *et* $e_2 \,|\, e_1 \sim N\left(0, \sigma_2^2\right)$. *Alors*

$$f\left(y_2 \,|\, x\right) = \int f\left(y_2 \,|\, x, y_1\right) f\left(y_1 \,|\, x\right) dy_1$$

est aussi une distribution normale de moyenne $\left(\beta_0 + \beta_2\alpha_0\right) + \left(\beta_1 + \beta_2\alpha_1\right)x$ *et de variance* $\sigma_2^2 + \beta_2^2\sigma_1^2x^2$. *En vertu de la structure de données présentée dans le tableau 1.1, on peut identifier ce modèle sans poser l'hypothèse de VI. L'hypothèse d'absence d'interaction entre* y_1 *et* x *dans le modèle pour* y_2 *est essentielle pour s'assurer que le modèle est identifiable.*

Au lieu de générer $y_{1i}^{*(j)}$ à partir de $\hat{f}_a\left(y_1 \,|\, x_i\right)$, on peut utiliser une méthode d'imputation fractionnaire hot deck (IFHD), en vertu de laquelle toutes les valeurs observées de y_{1i} dans l'échantillon A sont utilisées comme valeurs imputées. Dans ce cas, les poids fractionnaires de l'étape 2 sont donnés par

$$w_{ij}^*\left(\hat{\theta}_t\right) \propto w_{ij0}^* f\left(y_{2i} \,|\, x_i, y_{1i}^{*(j)}; \hat{\theta}_t\right),$$

où

$$w_{ij0}^* = \frac{\hat{f}_a\left(y_{1j} \,|\, x_i\right)}{\sum_{k\in A} w_{ka}\,\hat{f}_a\left(y_{1j} \,|\, x_k\right)}. \tag{3.3}$$

Le poids fractionnaire initial w_{ij0}^* en (3.3) est calculé par l'application d'une pondération préférentielle à l'aide de

$$\hat{f}_a\left(y_{1j}\right) = \int \hat{f}_a\left(y_{1j} \,|\, x\right) \hat{f}_a\left(x\right) dx \propto \sum_{i\in A} w_{ia}\,\hat{f}_a\left(y_{1j} \,|\, x_i\right)$$

comme densité proposée pour y_{1j}. L'étape de maximisation est la même que pour l'imputation fractionnaire paramétrique. Pour en savoir davantage à propos de l'IFHD, voir Kim et Yang (2014). Dans la pratique, on peut utiliser une seule valeur imputée pour chaque unité. Dans ce cas, les poids fractionnaires peuvent être utilisés comme probabilité de sélection dans l'échantillonnage avec probabilité proportionnelle à la taille (PPT) de taille $m = 1$.

Pour estimer la variance, on peut utiliser une méthode de linéarisation ou une méthode de ré-échantillonnage. On examine d'abord l'estimation de la variance pour l'estimateur du maximum de vraisemblance (EMV) de θ. Si on a recours à un modèle paramétrique $f(y_1|x) = f(y_1|x;\theta_1)$ et $f(y_2|x,y_1;\theta_2)$, on obtient l'EMV de $\theta = (\theta_1,\theta_2)$ en résolvant

$$\left[S_1(\theta_1), \bar{S}_2(\theta_1,\theta_2)\right] = (0,0), \tag{3.4}$$

où $S_1(\theta_1) = \sum_{i\in A} w_{ia} S_{i1}(\theta_1)$, $S_{i1}(\theta_1) = \partial \log f(y_{1i}|x_i;\theta_1)/\partial\theta_1$ est la fonction de score de θ_1,

$$\bar{S}_2(\theta_1,\theta_2) = E\{S_2(\theta_2)|X,Y_2;\theta_1,\theta_2\},$$

$S_2(\theta_2) = \sum_{i\in B} w_{ib} S_{i2}(\theta_2)$, et $S_{i2}(\theta_2) = \partial \log f(y_{2i}|x_i,y_{1i};\theta_2)/\partial\theta_2$ est la fonction de score de θ_2. Soulignons qu'on peut écrire $\bar{S}_2(\theta_1,\theta_2) = \sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2)|x_i,y_{2i};\theta\}$. Ainsi,

$$\begin{aligned}
\frac{\partial}{\partial\theta_1'}\bar{S}_2(\theta) &= \sum_{i\in B} w_{ib}\frac{\partial}{\partial\theta_1'}\left[\frac{\int S_{i2}(\theta_2) f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}{\int f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}\right] \\
&= \sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2) S_{i1}(\theta_1)'|x_i,y_{2i};\theta\} \\
&\quad -\sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2)|x_i,y_{2i};\theta\} E\{S_{i1}(\theta_1)'|x_i,y_{2i};\theta\}
\end{aligned}$$

et

$$\begin{aligned}
\frac{\partial}{\partial\theta_2'}\bar{S}_2(\theta) &= \sum_{i\in B} w_{ib}\frac{\partial}{\partial\theta_2'}\left[\frac{\int S_{i2}(\theta_2) f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}{\int f(y_1|x_i;\theta_1) f(y_{2i}|x_i,y_1;\theta_2)\,dy_1}\right] \\
&= \sum_{i\in B} w_{ib} E\left\{\frac{\partial}{\partial\theta_2'} S_{i2}(\theta_2)|x_i,y_{2i};\theta\right\} \\
&\quad +\sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2) S_{i2}(\theta_2)'|x_i,y_{2i};\theta\} \\
&\quad -\sum_{i\in B} w_{ib} E\{S_{i2}(\theta_2)|x_i,y_{2i};\theta\} E\{S_{2i}(\theta_2)'|x_i,y_{2i};\theta\}.
\end{aligned}$$

Maintenant, on peut estimer de manière convergente $\partial\bar{S}_2(\theta)/\partial\theta_1'$ par

$$\hat{B}_{21} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* S_{2ij}^*\left(\hat{\theta}_2\right)\left\{S_{1ij}^*\left(\hat{\theta}_1\right) - \bar{S}_{1i}^*\left(\hat{\theta}_1\right)\right\}', \tag{3.5}$$

où $S_{1ij}^*\left(\hat{\theta}_1\right) = S_1\left(\hat{\theta}_1; x_i, y_{1i}^{*(j)}\right)$, $S_{2ij}^*\left(\hat{\theta}_2\right) = S_2\left(\hat{\theta}_2; x_i, y_{1i}^{*(j)}, y_{2i}\right)$, et $\bar{S}_{1i}^*\left(\hat{\theta}_1\right) = \sum_{j=1}^{m} w_{ij}^* S_1\left(\hat{\theta}_1; x_i, y_{1i}^{*(j)}\right)$. De plus, on peut estimer $\partial\bar{S}_2(\theta)/\partial\theta_2'$ de manière convergente par

$$-\hat{I}_{22} = \sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^* \dot{S}_{2ij}^*\left(\hat{\theta}_2\right) - \hat{B}_{22} \tag{3.6}$$

où

$$\hat{B}_{22} = \sum_{i\in B} w_{ib} \sum_{j=1}^{m} w_{ij}^{*} S_{2ij}^{*}\left(\hat{\theta}_2\right)\left\{S_{2ij}^{*}\left(\hat{\theta}_2\right) - \overline{S}_{2i}^{*}\left(\hat{\theta}_2\right)\right\}',$$

$\dot{S}_{2ij}^{*}(\theta_2) = \partial S_2\left(\theta_2; x_i, y_{1i}^{*(j)}, y_{2i}\right)\big/\partial\theta_2'$ et $\overline{S}_{2i}^{*}(\theta_2) = \sum_{j=1}^{m} w_{ij}^{*} S_{2ij}^{*}(\theta_2)$.

En effectuant un développement en série de Taylor par rappport à θ_1,

$$\begin{aligned}\overline{S}_2\left(\hat{\theta}_1, \theta_2\right) &\cong \overline{S}_2(\theta_1, \theta_2) - E\left\{\frac{\partial}{\partial\theta_1'}\overline{S}_2(\theta)\right\}\left[E\left\{\frac{\partial}{\partial\theta_1'}S_1(\theta_1)\right\}\right]^{-1} S_1(\theta_1)\\ &= \overline{S}_2(\theta) + KS_1(\theta_1),\end{aligned}$$

on peut écrire

$$V\left(\hat{\theta}_2\right) \doteq \left\{E\left(\frac{\partial}{\partial\theta_2'}\overline{S}_2\right)\right\}^{-1} V\left\{\overline{S}_2(\theta) + KS_1(\theta_1)\right\}\left\{E\left(\frac{\partial}{\partial\theta_2'}\overline{S}_2\right)\right\}^{-1'}.$$

En écrivant

$$\overline{S}_2(\theta) = \sum_{i\in B} w_{ib}\overline{s}_{2i}(\theta),$$

où $\overline{s}_{2i}(\theta) = E\left\{S_{i2}(\theta_2)\,\middle|\, x_i, y_{2i}; \theta\right\}$, on peut obtenir un estimateur convergent de $V\left\{\overline{S}_2(\theta)\right\}$ en appliquant un estimateur de la variance convergent par rapport au plan à $\sum_{i\in B} w_{ib}\hat{s}_{2i}$ où $\hat{s}_{2i} = \sum_{j=1}^{m} w_{ij}^{*} S_{2ij}^{*}\left(\hat{\theta}_2\right)$. En vertu d'un échantillonnage aléatoire simple pour l'échantillon B, on obtient

$$\hat{V}\left\{\overline{S}_2(\theta)\right\} = n_B^{-2}\sum_{i\in B}\hat{s}_{2i}\hat{s}_{2i}'.$$

De plus, on peut estimer de manière convergente $V\left\{KS_1(\theta_1)\right\}$ par

$$\hat{V}_2 = \hat{K}\hat{V}(S_1)\hat{K}',$$

où $\hat{K} = \hat{B}_{21}\hat{I}_{11}^{-1}$, $\hat{B}_{21}$ est défini selon l'équation (3.5), et $\hat{I}_{11} = -\partial S_1(\theta_1)\big/\partial\theta_1'$ est évalué à $\theta_1 = \hat{\theta}_1$. Comme les deux termes $\overline{S}_2(\theta)$ et $S_1(\theta_1)$ sont indépendants, on peut estimer la variance par

$$\hat{V}\left(\hat{\theta}\right) \doteq \hat{I}_{22}^{-1}\left[\hat{V}\left\{\overline{S}_2(\theta)\right\} + \hat{V}_2\right]\hat{I}_{22}^{-1'},$$

où $\hat{I}_{22}$ est défini selon l'équation (3.6).

De façon plus générale, on pourrait considérer l'estimation d'un paramètre η défini comme une racine de l'équation d'estimation de recensement $\sum_{i=1}^{N} U(\eta; x_i, y_{1i}, y_{2i}) = 0$. L'estimation de la variance de l'estimateur par imputation fractionnaire de η calculée à partir de $\sum_{i\in B} w_{ib}\sum_{j=1}^{m} w_{ij}^{*} U\left(\eta; x_i, y_{1i}^{*(j)}, y_{2i}\right) = 0$ est présentée à l'annexe B.

4 Plan de sondage à questionnaire scindé

À la section 3, on examine le cas où l'échantillon A et l'échantillon B sont deux échantillons indépendants de la même population cible. Nous allons maintenant examiner un autre cas, celui d'un plan de sondage à questionnaire scindé en vertu duquel l'échantillon initial S est sélectionné à partir d'une population cible, puis l'échantillon A et l'échantillon B sont sélectionnés au hasard de sorte que $A \cup B = S$ et $A \cap B = \phi$. On observe (x, y_1) dans l'échantillon A et (x, y_2) dans l'échantillon B. On souhaite créer des données entièrement augmentées avec observation de (x, y_1, y_2) dans S.

De tels plans de sondage à questionnaire scindé gagnent en popularité parce qu'ils réduisent le fardeau de réponse (Raghunathan et Grizzle 1995; Chipperfield et Steel 2009). Des plans de sondage à questionnaire scindé ont notamment été explorés dans le cadre de la *Consumer Expenditure Survey* (Gonzalez et Eltinge 2008) et de la *National Assessment of Educational Progress (NAEP) Survey* aux États-Unis. Les analystes qui utilisent les résultats des enquêtes à questionnaire scindé peuvent s'intéresser à des paramètres multiples, comme la moyenne pour y_1 et la moyenne pour y_2, en plus du coefficient de la régression de y_2 sur y_1.

Nous avons examiné un plan de sondage où l'échantillon initial S est divisé en deux sous-échantillons : A et B. On suppose que x_i est observé pour $i \in S$, que y_{1i} est recueilli pour $i \in A$ et que y_{2i} est recueilli pour $i \in B$. La probabilité de sélection dans A ou B peut dépendre de x_i mais ne dépend pas de y_{1i} ni de y_{2i}. En conséquence, le plan de sondage utilisé pour sélectionner les sous-échantillons A et B est non informatif pour le modèle spécifié (Fuller 2009, chapitre 6). Soit w_i le poids d'échantillonnage associé à l'échantillon complet S. On suppose qu'il existe une procédure pour estimer la variance d'un estimateur de la forme $\hat{Y} = \sum_{i \in S} w_i y_i$, et on désigne l'estimateur de la variance par $\hat{V}_s \left(\sum_{i \in S} w_i y_i \right)$.

Décrivons maintenant une procédure pour obtenir un ensemble de données entièrement imputées. D'abord, on utilise la méthode décrite à la section 3 pour obtenir les valeurs imputées $\{y_{1i}^{*(j)} : i \in B, j = 1, \ldots, m\}$ et une estimation $\hat{\theta}$ du paramètre de la distribution $f(y_2 \mid y_1, x; \theta)$. On obtient l'estimation $\hat{\theta}$ en résolvant

$$\sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^* S_2 \left(\theta; x_i, y_{1i}^{*(j)}, y_{2i} \right) = 0, \tag{4.1}$$

où $S_2(\theta; x, y_1, y_2) = \partial \log f(y_2 \mid y_1, x; \theta) / \partial \theta$. Sachant $\hat{\theta}$, on génère les valeurs imputées $y_{2i}^{*(j)} \sim f \left(y_2 \mid y_{1i}, x_i; \hat{\theta} \right)$, pour $i \in A$ et $j = 1, \ldots, m$.

Si l'on suppose que le modèle est identifié, l'estimateur de paramètre $\hat{\theta}$ généré par la résolution de (4.1) est entièrement efficace au sens où la valeur imputée de y_{2i} pour l'échantillon A ne donne lieu à aucun gain d'efficacité. Pour le voir, notons que l'équation de score utilisant la valeur imputée de y_{2i} se calcule comme suit :

$$\sum_{i \in A} w_i m^{-1} \sum_{j=1}^{m} S_2 \left(\theta; x_i, y_{1i}, y_{2i}^{*(j)} \right) + \sum_{i \in B} w_i \sum_{j=1}^{m} w_{ij}^* S_2 \left(\theta; x_i, y_{1i}^{*(j)}, y_{2i} \right) = 0. \tag{4.2}$$

Comme $y_{2i}^{*(1)},\ldots,y_{2i}^{*(m)}$ sont générées à partir de $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$,

$$p\lim_{m\to\infty} \sum_{i\in A} w_i m^{-1} \sum_{j=1}^{m} S_2\left(\theta; x_i, y_{1i}, y_{2i}^{*(j)}\right) = \sum_{i\in A} w_i E\left\{S_2\left(\theta; x_i, y_{1i}, Y_2\right) \mid y_{1i}, x_i; \hat{\theta}\right\}.$$

Ainsi, en vertu de la propriété de la fonction de score, le premier terme de (4.2) évalué à $\theta = \hat{\theta}$ est proche de zéro et la solution de l'équation (4.2) est essentiellement la même que celle de l'équation (4.1), c'est-à-dire qu'on ne gagne pas en efficience en utilisant la valeur imputée de y_{2i} pour calculer l'EMV pour θ dans $f(y_2 \mid y_1, x; \theta)$.

Toutefois, les valeurs imputées de y_{2i} peuvent améliorer l'efficience des inférences pour les paramètres de la distribution conjointe de (y_{1i}, y_{2i}). À titre d'exemple simple, prenons l'estimation de μ_2, la moyenne marginale de y_{2i}. En vertu d'un échantillonnage aléatoire simple, l'estimateur imputé de $\mu = E(Y_2)$ est

$$\hat{\mu}_{I,m} = \frac{1}{n}\left\{\sum_{i\in A}\left(m^{-1}\sum_{j=1}^{m} y_{2i}^{*(j)}\right) + \sum_{i\in B} y_{2i}\right\}, \tag{4.3}$$

où les valeurs de $y_{2i}^{*(1)},\ldots,y_{2i}^{*(m)}$ sont générées à partir de $f\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right)$. Pour des valeurs de m, suffisamment grandes, on peut écrire

$$\begin{aligned}\hat{\mu}_{I,\infty} &= \frac{1}{n}\left\{\sum_{i\in A} \hat{y}_{2i} + \sum_{i\in B} y_{2i}\right\} \\ &= \frac{1}{n}\left\{\sum_{i\in A} E\left(y_2 \mid y_{1i}, x_i; \hat{\theta}\right) + \sum_{i\in B} y_{2i}\right\}.\end{aligned}$$

En vertu du scénario de l'exemple 2.1, on peut écrire $\hat{y}_{2i} = \hat{\beta}_0 + \hat{\beta}_1 y_{1i} + \hat{\beta}_2 x_{2i}$ où $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ satisfont

$$\sum_{i\in B}\left(y_{2i} - \hat{\beta}_0 - \hat{\beta}_{1i}\hat{y}_{1i} - \hat{\beta}_2 x_{2i}\right) = 0$$

et $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$ où $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ satisfont $\sum_{i\in A}(y_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i}) = 0$. Ainsi, en ignorant les termes d'ordre plus faible, on obtient

$$V\left(\hat{\mu}_{I,\infty}\right) = \frac{1}{n}V(y_2) + \left(\frac{1}{n_b} - \frac{1}{n}\right)V(y_2 - \hat{y}_2),$$

qui est inférieure à la variance de l'estimateur direct $\hat{\mu}_b = n_b^{-1}\sum_{i\in B} y_{2i}$.

5 Modèles d'erreur de mesure

Examinons maintenant l'application d'un appariement statistique au problème des modèles d'erreur de mesure. Supposons que l'on s'intéresse au paramètre θ de la distribution conditionnelle $f(y_2 \mid y_1; \theta)$. Dans l'échantillon initial, au lieu d'observer (y_{1i}, y_{2i}), on observe (x_i, y_{2i}), où x_i est une version contaminée

de y_{1i}. Comme il est possible que l'inférence pour θ fondée sur (x_i, y_{2i}) soit biaisée, d'autres renseignements sont nécessaires. L'une des façons courantes d'obtenir ces renseignements supplémentaires est de recueillir (x_i, y_{1i}) dans le cadre d'une étude de calage externe. Dans ce cas, on observe (x_i, y_{1i}) dans l'échantillon A et (x_i, y_{2i}) dans l'échantillon B, l'échantillon A étant l'échantillon de calage et l'échantillon B, l'échantillon principal. Guo et Little (2011) présentent une application d'un calage externe.

Le cadre de calage externe peut s'exprimer sous forme de problème d'appariement statistique. Le tableau 5.1 établit de façon explicite le lien entre l'appariement statistique et le calage externe. Une hypothèse de variable instrumentale permet l'inférence pour θ en fonction de données selon la structure présentée dans le tableau 1.1. Dans la notation du modèle d'erreur de mesure, l'hypothèse de variable instrumentale est

$$f(y_{2i} \mid y_{1i}, x_i) = f(y_{2i} \mid y_{1i}) \quad \text{et} \quad f(y_{1i} \mid x_i = a) \neq f(y_{1i} \mid x_i = b), \tag{5.1}$$

pour certains $a \neq b$. L'hypothèse de variable instrumentale peut être considérée raisonnable dans les applications relatives à l'erreur dans les covariables parce que le modèle d'intérêt en question est $f(y_{2i} \mid y_{1i})$, et x_i est une version contaminée de y_{1i} ne contenant aucun renseignement supplémentaire à propos de y_{2i} sachant y_{1i}.

Tableau 5.1
Structure de données pour le modèle d'erreur de mesure

	x_i	y_{1i}	y_{2i}
Enquête A (étude de calage)	o	o	
Enquête B (étude principale)	o		o

Dans le cas où $f(y_{2i} \mid y_{1i})$ et $f(y_{1i} \mid x_i)$ sont entièrement paramétriques, on peut utiliser l'imputation fractionnaire paramétrique pour exécuter l'algorithme EM. Cette méthode exige une évaluation de l'espérance conditionnelle de la fonction de score des données complètes sachant les valeurs observées. Pour évaluer l'espérance conditionnelle par imputation fractionnaire, on écrit d'abord la distribution conditionnelle de y_1 sachant (x, y_2) comme suit :

$$f(y_1 \mid x, y_2) \propto f(y_1 \mid x) f(y_2 \mid y_1). \tag{5.2}$$

Soit un estimateur $\hat{f}_a(y_{1i} \mid x_i)$ de $f(y_{1i} \mid x_i)$ provenant de l'échantillon de calage (échantillon A). La mise en œuvre de l'algorithme EM par imputation fractionnaire se déroule comme suit :

1. Pour chaque $i \in B$, générer $y_{1i}^{*(j)}$ à partir de $\hat{f}_a(y_1 \mid x_i)$, pour $j = 1, \ldots, m$.
2. Calculer les poids fractionnaires

$$w_{ij(t)}^{*} \propto f\left(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_t\right)$$

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

3. Mettre à jour θ en résolvant

$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0,$$

où $S(\theta; y_1, y_2) = \partial \log f(y_2 \mid y_1; \theta) / \partial \theta$.

4. Reprendre à l'étape 2 jusqu'à la convergence.

Cette méthode exige que l'on génère des données à partir de $f(y_1 \mid x)$. Dans le cas de certains modèles non linéaires ou de modèles assortis de variances non constantes, la simulation à partir de la distribution conditionnelle de y_1 sachant x peut exiger le recours à des méthodes Monte Carlo comme l'acceptation-rejet ou l'algorithme de Metropolis-Hastings. La simulation présentée à la section 6.2 est un bon exemple d'une simulation dans laquelle la distribution conditionnelle de $y_1 \mid x$ n'a pas d'expression de forme explicite. Dans ce cas, on peut envisager une autre solution plus simple à calculer. Pour décrire cette solution, posons $h(y_1 \mid x)$ comme distribution conditionnelle « de travail », par exemple la distribution normale, à partir de laquelle les échantillons peuvent être facilement générés. On présume que les estimations $\hat{f}_a(y_1 \mid x)$ et $\hat{h}_a(y_1 \mid x)$ de $f(y_1 \mid x)$ et $h(y_1 \mid x)$, respectivement, peuvent être obtenues à partir de l'échantillon A. La mise en œuvre de l'algorithme EM par imputation fractionnaire s'effectue ensuite comme suit :

1. Pour chaque $i \in B$, générer $x_i^{*(j)}$ à partir de $\hat{h}_a(y_1 \mid x_i)$, pour $j = 1, \ldots, m$.
2. Calculer les poids fractionnaires

$$w_{ij(t)}^{*} \propto f\left(y_{2i} \mid y_{1i}^{*(j)}; \hat{\theta}_t\right) \hat{f}_a\left(y_{1i}^{*(j)} \mid x_i\right) \Big/ \hat{h}_a\left(y_{1i}^{*(j)} \mid x_i\right) \tag{5.3}$$

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

3. Mettre à jour θ en résolvant

$$\sum_{i \in B} w_{ib} \sum_{j=1}^{m} w_{ij(t)}^{*} S\left(\theta; y_{1i}^{*(j)}, y_{2i}\right) = 0.$$

4. Reprendre à l'étape 2 jusqu'à la convergence.

L'estimation de la variance est une application directe de la méthode de linéarisation présentée à la section 3. La méthode d'imputation fractionnaire hot deck décrite à la section 3 assortie des poids fractionnaires définis en (3.3) s'applique aussi directement au contexte de l'erreur de mesure.

6 Études par simulation

Pour mettre à l'essai notre théorie, nous présentons deux études par simulation limitées. La première porte sur la combinaison de deux enquêtes indépendantes avec observation partielle afin d'effectuer une analyse conjointe. La deuxième porte sur la définition de modèles d'erreur de mesure avec calage externe.

6.1 Première simulation

Pour comparer les méthodes proposées avec les méthodes actuelles, nous avons généré 5 000 échantillons Monte Carlo comprenant (x_i, y_{1i}, y_{2i}) et de taille $n = 400$, où

$$\begin{pmatrix} y_{1i} \\ x_i \end{pmatrix} \sim N\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0{,}7 \\ 0{,}7 & 1 \end{bmatrix} \right),$$

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + e_i, \tag{6.1}$$

$e_i \sim N(0, \sigma^2)$, et $\beta = (\beta_0, \beta_1, \sigma^2)' = (1,1,1)'$. Soulignons que dans ce scénario, on trouve $f(y_2 \mid x, y_1) = f(y_2 \mid y_1)$; la variable x joue donc le rôle de variable instrumentale pour y_1.

Au lieu d'observer (x_i, y_{1i}, y_{2i}) conjointement, on présume que seuls (y_1, x) sont observés dans l'échantillon A, et que seuls (y_2, x) sont observés dans l'échantillon B; l'échantillon A est obtenu par sélection des $n_a = 400$ premiers éléments de l'échantillon initial, et l'échantillon B, par sélection des $n_b = 400$ éléments qui restent. On veut estimer quatre paramètres : trois paramètres de régression $\beta_0, \beta_1, \sigma^2$ et $\pi = P(y_1 < 2, y_2 < 3)$, la proportion de $y_1 < 2$ et de $y_2 < 3$. Quatre méthodes sont envisagées pour estimer ces paramètres :

1. Estimation de l'échantillon complet (EEC) : Utiliser toutes les observations de (y_{1i}, y_{2i}) de l'échantillon B.
2. Imputation par régression stochastique (IRS) : Utiliser la régression de y_1 sur x dans l'échantillon A pour obtenir $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\sigma}_1^2)$, où le modèle de régression est $y_1 = \alpha_0 + \alpha_1 x + e_1$ avec $e_1 \sim (0, \sigma_1^2)$. Pour chaque $i \in B$, $m = 10$ valeurs imputées sont générées par $y_{1i}^{*(j)} = \hat{\alpha}_0 + \hat{\alpha}_1 x_i + e_i^{*(j)}$ où $e_i^{*(j)} \sim N(0, \hat{\sigma}_1^2)$.
3. Imputation fractionnaire paramétrique (IFP) avec $m = 10$ selon l'hypothèse de variable instrumentale.
4. Imputation fractionnaire hot deck (IFHD) avec $m = 10$ selon l'hypothèse de variable instrumentale.

Le tableau 6.1 présente les moyennes et les variances Monte Carlo des estimateurs ponctuels des quatre paramètres d'intérêt. La méthode IRS comporte des biais importants pour tous les paramètres étudiés, parce qu'elle repose sur l'hypothèse d'indépendance conditionnelle. Les méthodes IFP et IFHD fournissent des estimateurs presque sans biais pour tous les paramètres. Les estimateurs obtenus à l'aide de la méthode IFHD sont légèrement plus efficients que ceux obtenus par la méthode IFP, parce que la démarche en deux étapes de la méthode IFHD utilise l'ensemble complet des répondants à la première étape. La variance asymptotique théorique de $\hat{\beta}_1$ calculée à partir de la méthode IFP est

$$V(\hat{\beta}_1) \doteq \frac{1}{(0{,}7)^2}\frac{1}{400} 2\left(1 - \frac{0{,}7^2}{2}\right) + \frac{1}{(0{,}7)^2}\frac{1}{400}(1 - 0{,}7^2) \doteq 0{,}0103$$

ce qui correspond au résultat de simulation présenté dans le tableau 6.1. En plus de l'estimation ponctuelle, on calcule aussi les estimateurs de variance pour les méthodes IFP et IFHD. Les estimateurs de variance

montrent de faibles biais relatifs (moins de 5 % en valeur absolue) pour tous les paramètres. Les résultats de l'estimation de la variance ne sont pas présentés ici par souci de concision.

Tableau 6.1
Moyennes et variances Monte Carlo des estimateurs ponctuels pour la première simulation. (EEC : estimation de l'échantillon complet; IRS : imputation par régression stochastique; IFP : imputation fractionnaire paramétrique; IFHD : imputation fractionnaire hot deck)

Paramètre	Méthode	Moyenne	Variance
β_0	EEC	1,00	0,0123
	IRS	1,90	0,0869
	IFP	1,00	0,0472
	IFHD	1,00	0,0465
β_1	EEC	1,00	0,00249
	IRS	0,54	0,01648
	IFP	1,00	0,01031
	IFHD	1,00	0,01026
σ^2	EEC	1,00	0,00482
	IRS	1,73	0,01657
	IFP	0,99	0,02411
	IFHD	0,99	0,02270
π	EEC	0,374	0,00058
	IRS	0,305	0,00255
	IFP	0,375	0,00059
	IFHD	0,375	0,00057

La méthode proposée repose sur l'hypothèse de variable instrumentale. Pour déterminer la sensibilité de la méthode d'imputation fractionnaire proposée aux violations de l'hypothèse de variable instrumentale, nous avons réalisé une étude par simulation supplémentaire. Au lieu de générer y_{2i} à partir de (6.1), on utilise

$$y_{2i} = 0{,}5 + y_{1i} + \rho(x_i - 3) + e_i, \tag{6.2}$$

où $e_i \sim N(0,1)$ et ρ peuvent prendre des valeurs non nulles. Nous avons utilisé trois valeurs de ρ, $\rho \in \{0; 0{,}1; 0{,}2\}$, pour l'analyse de sensibilité et nous avons employé la même procédure d'IFP et d'IFHD fondée sur l'hypothèse que x est une variable instrumentale pour y_1. Cette hypothèse est satisfaite pour $\rho = 0$, mais elle est légèrement enfreinte pour $\rho = 0{,}1$ ou $\rho = 0{,}2$. À l'aide des données obtenues par imputation fractionnaire dans l'échantillon B, nous avons estimé trois paramètres, $\theta_1 = E(Y_1)$, θ_2 la pente de la régression simple de y_2 sur y_1, et $\theta_3 = P(y_1 < 2, y_2 < 3)$, la proportion de $y_1 < 2$ et $y_2 < 3$. Le tableau 6.2 présente les moyennes et les variances Monte Carlo des estimateurs ponctuels pour les trois paramètres en vertu des trois différents modèles. Dans le tableau 6.2, on constate que les valeurs absolues de l'écart entre l'estimateur obtenu par imputation fractionnaire et l'estimateur obtenu à partir de l'échantillon complet augmente avec la valeur de ρ, ce qui est normal puisque l'hypothèse de variable instrumentale est plus gravement enfreinte pour les valeurs élevées de ρ, mais les écarts sont relativement faibles dans tous les cas. Plus particulièrement, l'estimateur de θ_1 n'est pas touché par la dérogation à

l'hypothèse de variable instrumentale, parce que l'estimateur par imputation obtenu en vertu du modèle d'imputation inexact fournit quand même un estimateur non biaisé pour la moyenne de population, à condition que le modèle d'imputation par régression contienne un terme d'ordonnée à l'origine (Kim et Rao 2012). Ainsi, cette analyse de sensibilité limitée indique que la méthode proposée semble produire des estimations comparables lorsque l'hypothèse de variable instrumentale est légèrement enfreinte.

Tableau 6.2
Moyennes et variances Monte Carlo des deux estimateurs ponctuels pour l'analyse de sensibilité de la première simulation (EEC : estimation de l'échantillon complet; IFP : imputation fractionnaire paramétrique; IFHD : imputation fractionnaire hot deck)

Modèle	Paramètre	Méthode	Moyenne	Variance
$\rho = 0$	θ_1	EEC	2,00	0,00235
		IFP	2,00	0,00352
		IFHD	2,00	0,00249
	θ_2	EEC	1,00	0,00249
		IFP	1,00	0,01031
		IFHD	1,00	0,01026
	θ_3	EEC	0,43	0,00061
		IFP	0,43	0,00059
		IFHD	0,43	0,00057
$\rho = 0,1$	θ_1	EEC	2,00	0,00235
		IFP	2,00	0,00353
		IFHD	2,00	0,00250
	θ_2	EEC	1,07	0,00248
		IFP	1,14	0,01091
		IFHD	1,14	0,01081
	θ_3	EEC	0,44	0,00061
		IFP	0,45	0,00062
		IFHD	0,45	0,00059
$\rho = 0,2$	θ_1	EEC	2,00	0,00235
		IFP	2,00	0,00353
		IFHD	2,00	0,00250
	θ_2	EEC	1,14	0,00250
		IFP	1,28	0,01115
		IFHD	1,28	0,01102
	θ_3	EEC	0,44	0,00061
		IFP	0,46	0,00066
		IFHD	0,46	0,00062

6.2 Deuxième simulation

Dans la deuxième étude par simulation, on examine une variable de réponse binaire y_{2i}, où

$$y_{2i} \sim \text{Bernoulli}(p_i), \tag{6.3}$$

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 y_{1i},$$

et $y_{1i} \sim N\left(\mu_1, \sigma_1^2\right)$. Dans l'échantillon principal, désigné par la lettre B, au lieu d'observer (y_{1i}, y_{2i}), on observe (x_i, y_{2i}), où

$$x_i = \beta_0 + \beta_1 y_{1i} + u_i, \tag{6.4}$$

et $u_i \sim N\left(0, \sigma^2 \left| y_{1i} \right|^{2\alpha}\right)$. On observe (x_i, y_{1i}), $i = 1, \ldots, n_A$ dans un échantillon de calage, désigné par la lettre A. Aux fins de la simulation, $n_A = n_B = 800$, $\gamma_0 = 1$, $\gamma_1 = 1$, $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0{,}25$, $\alpha = 0{,}4$, $\mu_1 = 0$, et $\sigma_1^2 = 1$. Le but principal est d'estimer γ_1 et de mettre à l'essai l'hypothèse nulle que $\gamma_1 = 1$. La taille de l'échantillon Monte Carlo (MC) est 1 000.

On compare les estimateurs IFP de γ_1 aux trois autres estimateurs. Comme la distribution conditionnelle de y_{1i} sachant x_i n'est pas standard, on utilise les poids de (5.3) pour réaliser l'IFP, où la distribution proposée $\hat{h}_a(y_{1i} \mid x_i)$ est une estimation de la distribution marginale de y_{1i} fondée sur les données de l'échantillon A. On considère les quatre estimateurs suivants :

1. *IFP :* Pour l'IFP, la distribution proposée pour générer $y_{1i}^{*(j)}$ est une distribution normale de moyenne $\hat{\mu}_1$ et de variance $\hat{\sigma}_1^2$, où $\hat{\mu}_1$ et $\hat{\sigma}_1^2$ sont les estimations du maximum de vraisemblance de μ_1 et σ_1^2, respectivement, en fonction de l'échantillon A. Les poids fractionnaires définis en (5.3) se présentent sous la forme

$$w_{ij}^* \propto \hat{p}_{ij}^{\,y_{2i}} \left(1 - \hat{p}_{ij}\right)^{1-y_{2i}} \hat{f}_a\left(y_{1i}^{*(j)} \mid x_i\right), \tag{6.5}$$

   où $\hat{p}_{ij} = \left\{1 + \exp\left(-\hat{\gamma}_0 - \hat{\gamma}_1 y_{1i}^{*(j)}\right)\right\}^{-1}$ et $\hat{f}_a(y_{1i} \mid x_i)$ est l'estimation de $f(y_{1i} \mid x_i)$ fondée sur l'estimation du maximum de vraisemblance à partir des données de l'échantillon A. La taille de la classe d'imputation est $m = 800$.

2. Estimateur *naïf :* Un estimateur *naïf* est l'estimateur de la pente dans la régression logistique de y_{2i} sur x_i pour $i \in B$.

3. Estimateur *bayésien :* On utilise l'approche de Guo et Little (2011) pour définir un estimateur bayésien. Le modèle de notre simulation diffère de celui de Guo et Little (2011) par le fait que la réponse d'intérêt est binaire. On établit un échantillonnage de Gibbs à l'aide du programme JAGS (Plummer 2003), en précisant les distributions a priori diffuses appropriées pour les paramètres du modèle. Soit

$$\theta_1 = \left(\log\left(\sigma^2\right), \log\left(\sigma_1^2\right), \mu_1, \beta_0, \beta_1, \gamma_0, \gamma_1\right);$$

   on suppose a priori que $\theta_1 \sim N\left(0, 10^6 I_7\right)$, où I_7 est une matrice identité de dimensions 7×7 et où la notation $N(0, V)$ désigne une distribution normale de moyenne 0 et de matrice de covariances V. La distribution a priori pour la puissance α est uniforme sur l'intervalle $[-5, 5]$.

   Pour évaluer la convergence, on examine les tracés des courbes et les facteurs de réduction d'échelle possibles définis par Gelman, Carlin, Stern et Rubin (2003) pour 10 ensembles de données simulées préliminaires. On produit trois chaînes MCMC, chacune d'une longueur de 1 500, à partir de valeurs initiales aléatoires, et on rejette les 500 premières itérations, considérées comme faisant partie du rodage. Les facteurs de réduction d'échelle possibles dans les

10 ensembles de données simulées vont de 1,0 à 1,1, et les tracés des courbes indiquent que les chaînes se combinent bien. Pour réduire le temps de calcul, on utilise 1 000 itérations d'une seule chaîne pour la simulation principale, après rejet des 500 premières itérations de rodage.

4. Un estimateur par *calage par régression pondérée (CRP)*. Cet estimateur par CRP est une modification de l'estimateur par calage par régression pondérée défini par Guo et Little (2011) pour une variable de réponse binaire. On calcule l'estimateur par calage par régression pondérée comme suit :

   (i) À l'aide des moindres carrés ordinaires (MCO), effectuer une régression de y_{1i} sur x_i pour l'échantillon de calage.

   (ii) Effectuer une régression du logarithme des résidus quadratiques obtenus à l'étape (i) sur le logarithme de x_i^2 pour l'échantillon de calage. Soit $\hat{\lambda}$ la pente estimée de la régression.

   (iii) À l'aide des moindres carrés pondérés (MCP) avec le poids $|x_i|^{2\hat{\lambda}}$, effectuer la régression de y_{1i} sur x_i pour l'échantillon de calage. Soient $\hat{\eta}_0$ et $\hat{\eta}_1$ l'ordonnée à l'origine et la pente estimées, respectivement, de la régression des MCP.

   (iv) Pour chaque unité i de l'échantillon principal, posons $\hat{y}_{1i} = \hat{\eta}_0 + \hat{\eta}_1 x_i$.

   (v) L'estimation de (γ_0, γ_1) est obtenue à partir de la régression logistique de y_{2i} sur $\hat{y}_{1i}$ dans l'échantillon principal.

Le tableau 6.3 indique le biais, la variance et l'EQM Monte Carlo des quatre estimateurs de γ_1. L'estimateur naïf a un biais négatif parce que x_i est une version contaminée de y_{1i}. L'estimateur par IFP est supérieur à l'estimateur bayésien et à l'estimateur par CRP.

On calcule une estimation de la variance des estimateurs par IFP de γ_1 à l'aide de l'expression de la variance fondée sur l'approximation linéaire. On définit le biais relatif MC comme étant le ratio de la différence entre la moyenne MC de l'estimateur de variance et la variance MC de l'estimateur à la variance MC de l'estimateur. Le biais relatif MC des estimateurs de variance pour l'IFP est négligeable (moins de 2 % en valeur absolue).

Tableau 6.3
Moyennes, variances et erreurs quadratiques moyennes Monte Carlo des estimateurs ponctuels de γ_1 pour la deuxième simulation. (IFP : imputation fractionnaire paramétrique; CRP : calage par régression pondérée; MC : Monte Carlo; EQM : erreur quadratique moyenne)

Méthode	Biais MC	Variance MC	EQM MC
IFP	0,0239	0,0386	0,0392
Estimateur naïf	-0,2241	0,0239	0,0742
Estimateur bayésien	0,0406	0,0415	0,0432
Estimateur par CRP	0,112	0,0499	0,0625

7 Conclusion

Nous considérons l'appariement statistique comme un problème de données manquantes et proposons la méthode d'IFP pour obtenir des estimateurs convergents et des estimateurs de variance correspondants. En vertu de l'hypothèse que le modèle spécifié est entièrement identifié, la méthode proposée permet d'obtenir les estimateurs du pseudo maximum de vraisemblance des paramètres du modèle.

Pour qu'un modèle puisse être considéré comme identifiable, il suffit qu'il comporte une variable instrumentale. Le cadre d'erreur de mesure énoncé aux sections 5 et 6.2, en vertu duquel un calage externe fournit une mesure indépendante de la vraie covariable d'intérêt, représente une situation dans laquelle le plan de sondage peut être considéré comme soutenant l'hypothèse de variable instrumentale. La méthodologie proposée peut être appliquée sans hypothèse de variable instrumentale, à condition que le modèle soit identifié. Si le modèle n'est pas identifiable, l'algorithme EM pour la méthode d'IFP proposée ne converge pas nécessairement. Dans la pratique, on peut considérer le modèle spécifié comme étant identifié si la séquence EM converge. Autrement dit, tant que la séquence EM converge, l'analyse connexe est convergente sous le modèle spécifié. C'est là l'un des nombreux avantages du recours à la méthode axée sur les fréquences par rapport à la méthode bayésienne. Avec la méthode bayésienne, il est possible d'obtenir les valeurs a posteriori même avec des modèles non identifiés et dans ce cas, l'analyse qui en découle peut être trompeuse.

En vertu de la structure de données présentée dans le tableau 1.1, il est beaucoup plus difficile, voire impossible, de déterminer si l'hypothèse de VI se vérifie dans les données disponibles. Compte tenu du modèle spécifié, on ne peut que vérifier si les paramètres du modèle peuvent être entièrement estimés. En revanche, il en va autrement de la détermination du caractère approprié du modèle spécifié par rapport aux données disponibles. Le diagnostic de modèles et la sélection d'un modèle parmi les différents modèles identifiables constituent d'importants sujets qu'il conviendrait d'approfondir dans le cadre de recherches futures.

L'appariement statistique peut aussi servir à évaluer les effets de traitements multiples dans les études d'observation. En utilisant adéquatement les techniques d'appariement statistique, on peut créer un fichier de données augmentées sur les résultats possibles afin d'étudier l'inférence causale (Morgan et Winship 2007). De telles utilisations seront présentées ailleurs.

Remerciements

Nous remercions le professeur Yanyuan Ma, un examinateur anonyme et le rédacteur adjoint pour leurs commentaires très constructifs. Les travaux de recherche de Jae Kwang Kim ont été en partie financés par le programme *Brain Pool* (131S-1-3-0476) de la *Korean Federation of Science and Technology Society* et par une subvention de la NSF (MMS-121339). Les travaux de recherche d'Emily Berg ont été financés en vertu d'une entente de coopération entre le *US Department of Agriculture Natural Resources Conservation Service* et la *Iowa State University*. Les travaux de recherche de Taesung Park ont été financés dans le cadre du *Bio-Synergy Research Project* (2013M3A9C4078158) du *Ministry of Science, ICT and Future Planning*, par l'entremise de la *National Research Foundation* de Corée.

Annexe

A. Absence asymptotique de biais de l'estimateur par les MC2E

Supposons que l'on observe (y_1, x) dans l'échantillon A et (y_2, x) dans l'échantillon B. Par souci de rigueur, on peut écrire (y_{1a}, x_a) pour désigner l'observation de (y_1, x) dans l'échantillon A, et (y_{2b}, x_b) pour désigner les observations dans l'échantillon B. Dans ce cas, le modèle peut s'écrire

$$y_{1a} = \phi_0 1_a + \phi_1 x_{1a} + \phi_2 x_{2a} + e_{1a}$$
$$y_{2b} = \beta_0 1_b + \beta_1 y_{1b} + \beta_2 x_{2b} + e_{2b}$$

avec $E(e_{1a} \mid x_a) = 0$ et $E(e_{2b} \mid x_b, y_{1b}) = 0$. Soulignons que y_{1b} n'est pas observée dans l'échantillon. On utilise plutôt $\hat{y}_{1b}$ produite grâce à l'estimation par les MCO obtenue à partir de l'échantillon A.

En écrivant $X_a = [1_a, x_a]$ et $X_b = [1_b, x_b]$, on obtient $\hat{y}_{1b} = X_b (X_a' X_a)^{-1} X_a' y_{1a} = X_b \hat{\phi}_a$. L'estimateur par les MC2E de $\beta = (\beta_0, \beta_1, \beta_2)'$ est donc

$$\hat{\beta}_{\text{MC2E}} = (Z_b' Z_b)^{-1} Z_b' y_{2b}$$

où $Z_b = [1_b, \hat{y}_{1b}, x_{2b}]$. Ainsi, on obtient

$$\begin{aligned} \hat{\beta}_{\text{MC2E}} - \beta &= (Z_b' Z_b)^{-1} Z_b' (y_{2b} - Z_b \beta) \\ &= (Z_b' Z_b)^{-1} Z_b' \{\beta_1 (y_{1b} - \hat{y}_{1b}) + e_{2b}\}. \end{aligned} \tag{A.1}$$

On peut écrire

$$y_{1b} = \phi_0 1_b + \phi_1 x_b + e_{1b} = X_b \phi + e_{1b}$$

où $E(e_{1b} \mid x_b) = 0$. Comme

$$\begin{aligned} \hat{y}_{1b} &= X_b (X_a' X_a)^{-1} X_a' y_{1a} \\ &= X_b (X_a' X_a)^{-1} X_a' (X_a \phi + e_{1a}) \\ &= X_b \phi + X_b (X_a' X_a)^{-1} X_a' e_{1a}, \end{aligned}$$

on obtient

$$y_{1b} - \hat{y}_{1b} = e_{1b} - X_b (X_a' X_a)^{-1} X_a' e_{1a}$$

et (A.1) devient

$$\hat{\beta}_{\text{MC2E}} - \beta = (Z_b' Z_b)^{-1} Z_b' \left\{\beta_1 e_{1b} - \beta_1 X_b (X_a' X_a)^{-1} X_a' e_{1a} + e_{2b}\right\}. \tag{A.2}$$

Supposons que les deux échantillons sont indépendants. Ainsi, $E(e_{1b} \mid x_a, x_b, y_{1a}) = 0$. De plus, $E\left\{(Z_b' Z_b)^{-1} Z_b' e_{2b} \,\middle|\, x_a, x_b, y_{1a}, y_{1b}\right\} = 0$. Ainsi,

$$E\left\{\hat{\beta}_{\text{MC2E}} - \beta \,\middle|\, x_a, x_b, y_{1a}\right\} = E\left\{-\beta_1 (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} \,\middle|\, x_a, x_b, y_{1a}\right\}$$

et

$$\begin{aligned} (Z_b' Z_b)^{-1} Z_b' X_b (X_a' X_a)^{-1} X_a' e_{1a} &= (Z_b' Z_b)^{-1} Z_b' \left\{X_b (X_a' X_a)^{-1} X_a' (y_{1a} - X_a \phi)\right\} \\ &= (Z_b' Z_b)^{-1} Z_b' X_b \left(\hat{\phi}_a - \phi\right). \end{aligned}$$

Ce terme a une espérance nulle asymptotiquement parce que $n_b^{-1} Z_b' Z_b$ et $n_b^{-1} Z_b' X_b$ sont bornés en probabilité et que $\left(\hat{\phi}_a - \phi\right)$ converge vers zéro.

B. Estimation de la variance

Soit le paramètre d'intérêt défini par la solution de $U_N(\eta) = \sum_{i=1}^{N} U(\eta; y_{1i}, y_{2i}) = 0$. On suppose que $\partial U_N(\eta)/\partial\theta = 0$. Ainsi, le paramètre η est indépendant a priori de θ, qui est le paramètre de la distribution de production de données de (x, y_1, y_2).

En vertu du scénario de la section 3, on pose $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ comme l'EMV de $\theta = (\theta_1, \theta_2)$ obtenu par la résolution de (3.4). De plus, posons $\hat{\eta}$ comme la solution de $\bar{U}(\eta|\hat{\theta}) = 0$ où

$$\bar{U}(\eta|\theta) = \sum_{i\in B}\sum_{j=1}^{m} w_{ib} w_{ij}^{*} U\left(\eta; y_{1i}^{*(j)}, y_{2i}\right),$$

et

$$w_{ij}^{*} \propto f\left(y_{1i}^{*(j)} \middle| x_i; \hat{\theta}_1\right) f\left(y_{2i} \middle| y_{1i}^{*(j)}; \hat{\theta}_2\right) \Big/ h\left(y_{1i}^{*(j)} \middle| x_i\right)$$

avec $\sum_{j=1}^{m} w_{ij}^{*} = 1$. Ici, $h(y_1|x)$ est la distribution proposée pour la génération des valeurs imputées de y_1 dans l'imputation fractionnaire paramétrique. En introduisant la distribution proposée h, on peut sans risque ignorer la dépendance des valeurs imputées $y_{1i}^{*(j)}$ à la valeur de paramètre estimée $\hat{\theta}_1$.

En effectuant une linéarisation par série de Taylor,

$$\bar{U}(\eta|\hat{\theta}) \cong \bar{U}(\eta|\theta) + (\partial\bar{U}/\partial\theta_1')(\hat{\theta}_1 - \theta_1) + (\partial\bar{U}/\partial\theta_2')(\hat{\theta}_2 - \theta_2)$$

on constate que

$$\hat{\theta}_1 - \theta_1 \cong \{I_1(\theta_1)\}^{-1} S_1(\theta_1)$$

où $I_1(\theta_1) = -\partial S_1(\theta_1)/\partial\theta_1'$. De plus,

$$\hat{\theta}_2 - \theta_2 \cong \left\{-\frac{\partial}{\partial\theta_2'}\bar{S}_2(\theta)\right\}^{-1} \bar{S}_2(\theta)$$

où

$$\bar{S}_2(\theta) = \sum_{i\in B}\sum_{j=1}^{m} w_i w_{ij}^{*}(\theta) S_2\left(\theta_2; y_{1i}^{*(j)}, y_{2i}\right).$$

Ainsi, on peut établir

$$\bar{U}(\eta|\hat{\theta}) \cong \bar{U}(\eta|\theta) + K_1 S_1(\theta_1) + K_2 \bar{S}_2(\theta),$$

où $K_1 = D_{21} I_{11}^{-1}$ et $K_2 = D_{22} I_{22}^{-1}$ avec $I_{11} = -E(\partial S_1/\partial\theta_1')$, $I_{22} = -E(\partial\bar{S}_2/\partial\theta_2')$, $D_{21} = E\{U(\eta) S_1(\theta_1)'\}$ et $D_{22} = E\{U(\eta) S_2(\theta_2)'\}$, on obtient

$$V\left\{\bar{U}(\eta|\hat{\theta})\right\} = \tau^{-1}\{V_1 + V_2\}\tau^{-1'}$$

où $\tau = -E\{\partial \bar{U}(\eta|\theta)/\partial \eta'\}$,

$$V_1 = V\left\{\sum_{i\in B} w_i\left(\bar{u}_i^* + K_2 S_{2i}^*\right)\right\},$$

$\bar{u}_i^* = E\left[U(\hat{\eta}; y_{1i}, y_{2i})\,|\,y_{2i}; \hat{\theta}\right]$, et $V_2 = V\left\{K_1 \sum_{i\in A} w_i S_{1i}\right\}$. Un estimateur convergent de chaque composante peut être élaboré selon la technique décrite à la section 3.

Bibliographie

Baker, K.H., Harris, P. et O'Brien, J. (1989). Data fusion: An appraisal and experimental evaluation. *Journal of the Market Research Society*, 31, 152-212.

Beaumont, J.-F., et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 37, 3, 400-416.

Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chib, S., et Greenberg, E. (1995). Jackknife variance estimation for nearest neighbor imputation. *The American Statistician*, 46, 327-333.

Chipperfield, J.O., et Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 2, 227-244.

D'Orazio, M., Zio, M.D. et Scanu, M. (2006). *Statistical Matching: Theory and Practice.* Chichester, R.U.: Wiley.

Fuller, W.A. (2009). *Sampling Statistics*, Hoboken, NJ: John Wiley & Sons, Inc.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapman and Hall Texts in Statistical Science. Chapman and Hall/CRC, deuxième édition.

Gonzalez, J., et Eltinge, J. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2081-2088.

Guo, Y., et Little, R.J. (2011). Regression analysis with covariates that have heteroskedastic measurement error. *Statistics Medicine*, 30, 18, 2278-2294.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Dans *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*, (Éds., C.R. Rao et D. Pfeffermann), 215-246.

Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques.* New York: Springer.

Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85, 765-769.

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

Kim, J.K., et Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

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

Kim, J.K., et Yang, S. (2014). Imputation fractionnaire hot deck pour une inférence robuste sous un modèle de non-réponse partielle en échantillonnage. *Techniques d'enquête*, 40, 2, 235-256.

Lahiri, P., et Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 1265-1275.

Leulescu, A., et Agafitei, M. (2013). Statistical matching: A model based approach for data integration. *Eurostat Methodologies and Working Papers.*

Morgan, S.L., et Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, USA: Cambridge University Press.

Moriarity, C., et Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17, 407-422.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Dans *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Raghunathan, T.E., et Grizzle, J.E. (1995). A split questionnaire design. *Journal of the American Statistical Association*, 90, 54-63.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer-Verlag.

Ridder, S., et Moffit, R. (2007). The econometrics of data combination. *Handbook of Econometrics*, 5470-5544.