Approche par modèle de non-réponse pour l'inférence en présence de données imputées

David HAZIZA et J.N.K. RAO

Statistique Canada, DMEE & Université Carleton

1. Introduction

L'imputation est fréquemment utilisée en pratique pour traiter le problème de la non-réponse partielle. L'imputation par la régression, qui comprend l'imputation par la moyenne ou par le ratio comme cas particuliers, est une méthode.

En présence de données imputées, deux approches sont utilisées pour mener des inférences : (i) l'approche par Modèle de non-réponse Uniforme (MU); (ii) l'approche par Modèle d'Imputation (MI). L'approche MU repose sur l'hypothèse suivante :

Hypothèse MU : La probabilité de réponse à une variable d'intérêt donnée est constante pour toutes les unités et les unités répondent indépendamment les unes des autres.

L'approche MU a été considérée par Rao (1990, 1996), Rao et Sitter (1995) et Shao et Steel (1999). Cette approche consiste à décrire complètement le mécanisme de non-réponse. L'approche MI, quant à elle, repose sur l'hypothèse suivante :

Hypothèse MI : Le mécanisme de non-réponse est non confondu en ce sens que la probabilité de réponse peut dépendre de variables auxiliaires mais pas de la variable d'intérêt (celle que l'on impute). Nous faisons alors appel à un modèle d'imputation qui, dans le cas de l'imputation par la régression, est de la forme

$$E_m\left(y_i\right)=\mathbf{z}_i'\boldsymbol{\beta},V_m\left(y_i\right)=\sigma_i^2=\sigma^2\boldsymbol{\lambda}'\mathbf{z}_i\equiv c_i^{-1},Cov_m\left(y_i,y_j\right)=0 \text{ si } i\neq j, \qquad (1)$$

où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus de dimensions $q\times 1$, $\mathbf{z}_i$ est un vecteur de variables auxiliaires disponibles pour toutes les unités échantillonnées (répondants et non-répondants) de dimension $q\times 1$, σ^2 est un paramètre inconnu et $E_m(.),V_m(.)\text{et}\,Cov_m(.)$ désignent respectivement l'espérance, la variance et la covariance par rapport au modèle d'imputation. Notons que la restriction $\sigma_i^2=c_i^{-1}$ ne restreint pas trop sévèrement l'éventail des modèles d'imputation. L'approche AMI a été considérée par Särndal (1992), Deville et Särndal (1994) et Shao-Steel (1999). Contrairement à l'approche MU, on ne cherche pas ici à décrire le mécanisme de non-réponse. On fait plutôt appel à un modèle d'imputation qui décrit le lien entre la variable d'intérêt et un ensemble de variables auxiliaires.

Nous proposons une troisième approche, appelée approche par Modèle de non-réponse Non-Confondu (MNC). L'AMNC repose sur l'hypothèse suivante :

Hypothèse MNC : Le mécanisme de non-réponse est non-confondu et les unités répondent indépendamment les unes des autres. Nous supposons que la probabilité de réponse, p_i, pour l'unité i est liée à un vecteur de variables auxiliaires $\mathbf{u}_i$ de dimension $l \times 1$ selon le modèle de régression logistique

$$p_i = \frac{e^{\mathbf{u}_i'\boldsymbol{\eta}}}{1+e^{\mathbf{u}_i'\boldsymbol{\eta}}}, \tag{2}$$

où $\boldsymbol{\eta}$ est un vecteur de paramètres inconnus de dimensions $l \times 1$. Le modèle (2) est appelé modèle de non-réponse. Notons que l'approche MU est un cas particulier de l'approche MNC.

En pratique, on a fréquemment recours à l'imputation par la régression (déterministe ou aléatoire) pondérée qui utilise les poids de sondage lors de la construction des valeurs imputées. L'estimateur imputé pour un total est alors approximativement sans biais sous les deux approches MU et MI. Cet estimateur est donc robuste en ce sens qu'il est valide sous les deux approches. Cependant, l'estimateur imputé est généralement biaisé sous l'approche MNC. Dans cet article, nous proposons une nouvelle méthode d'imputation qui mène à un estimateur imputé approximativement sans biais sous les approches MI et MNC.

2. Estimation d'un total

Soit U une population de taille N. Le but est d'estimer le total $Y = \sum_{i \in U} y_i$ lorsque l'imputation a été utilisée pour traiter la non-réponse à la variable d'intérêt y. Supposons qu'un échantillon aléatoire, s, de taille n est tiré de la population selon un plan de sondage $p(s)$. En l'absence de non-réponse à la variable y, un estimateur de Y est l'estimateur de Horvitz-Thompson défini selon

$$\hat{Y} = \sum_{i \in s} w_i y_i, \tag{3}$$

où $w_i = \frac{1}{\pi_i}$ désigne le poids de sondage de l'unité i et $\pi_i = P(i \in s)$ désigne la probabilité d'inclusion d'ordre 1 pour l'unité i. Il est bien connu que

$$E_p(\hat{Y}) = Y,$$

où $E_p(.)$ désigne l'espérance par rapport au plan de sondage $p(s)$. Autrement dit, l'estimateur $\hat{Y}$ est un estimateur sans biais sous le plan de sondage de Y. En présence de non-réponse à la variable y, il n'est plus possible de calculer $\hat{Y}$ en (3) puisque certaines valeurs de y sont manquantes. On définit plutôt un estimateur imputé selon

$$\hat{Y}_I = \sum_{i \in s} w_i a_i y_i + \sum_{i \in s} w_i (1-a_i) y_i^* = \sum_{i \in s} w_i \tilde{y}_i, \tag{4}$$

où a_i désigne la variable indicatrice de réponse pour l'unité i telle que $a_i = 1$ si l'unité i a répondu à la variable y et $a_i = 0$ sinon, y_i^* désigne la valeur imputée utilisée pour remplacer la valeur manquante y_i et $\tilde{y}_i = y_i$ si l'unité i a répondu à la variable y et $\tilde{y}_i = y_i^*$ sinon. L'estimateur $\hat{Y}_I$ peut être calculé à partir du fichier de données qui contient les poids de sondage w_i ainsi que les valeurs de la variable $\tilde{y}_i$; dans ce cas, la présence des variables indicatrices de réponse a_i dans le fichier n'est pas requise.

L'erreur totale $\hat{Y}_I - Y$ peut être décomposée comme suit :

$$\hat{Y}_I - Y = \left(\hat{Y} - Y\right) + \left(\hat{Y}_I - \hat{Y}\right). \tag{5}$$

Le terme $\hat{Y} - Y$ en (5) est l'erreur due à l'échantillonnage alors que le terme $\hat{Y}_I - \hat{Y}$ est l'erreur due à la non-réponse et à l'imputation. Notons que dans le cas d'une imputation déterministe, il n'y a pas d'erreur due à l'imputation. L'erreur due à l'échantillonnage ne dépendant pas du mécanisme de non-réponse ou de la méthode d'imputation, nous évaluons les propriétés de l'estimateur imputé $\hat{Y}_I$ (biais et variance) étant donné l'échantillon s.

Par souci de simplicité, nous considérons le cas d'une seule classe d'imputation mais la généralisation au cas de classes multiples est relativement aisée.

2.1 Imputation par la régression déterministe

L'imputation par la régression déterministe utilise les valeurs imputées

$$y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r, \tag{6}$$

où $\hat{\mathbf{B}}_r$ est l'estimateur des moindres carrés pondérés calculé à partir des unités répondantes et donné par

$$\hat{\mathbf{B}}_r = \left(\sum_{i \in s} w_i a_i c_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in s} w_i a_i c_i \mathbf{z}_i y_i. \tag{7}$$

Dans le cas de l'imputation par la régression déterministe (6), l'estimateur imputé (4) s'écrit comme

$$\hat{Y}_I = \hat{Y}_r + \left(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r\right)' \hat{\mathbf{B}}_r, \tag{8}$$

où $\hat{Y}_r = \sum_{i \in s} w_i a_i y_i, \hat{\mathbf{Z}} = \sum_{i \in s} w_i \mathbf{z}_i$ et $\hat{\mathbf{Z}}_r = \sum_{i \in s} w_i a_i \mathbf{z}_i$. Notons que l'estimateur imputé en (8) est similaire à un estimateur par la régression dans le cas l'échantillonnage à deux phases. Sous l'approche MU, on a $E_r\left(a_i | s\right) = p$. Il s'ensuit que le biais conditionnel de $\hat{Y}_I$, $Biais\left(\hat{Y}_I | s\right) = E_r\left(\hat{Y}_I - \hat{Y} | s\right)$ est approximativement égal à 0, où $E_r(.)$ désigne l'espérance par rapport au mécanisme de non-réponse. De plus, on peut facilement montrer que sous l'approche MI et le modèle de régression (1), le biais conditionnel, $Biais\left(\hat{Y}_I | s\right) = E_r E_m\left(\hat{Y}_I - \hat{Y} | s\right)$ est égal à 0.

Cependant, l'estimateur imputé (8) est conditionnellement biaisé sous l'approche MNC. L'expression du biais est donnée par

$$Biais\left(\hat{Y}_I \middle| s\right) = E_r\left(\hat{Y}_I - \hat{Y} \middle| s\right) \approx \sum_{i \in s} w_i (1 - p_i)\left(y_i - \mathbf{z}_i' \hat{\mathbf{B}}_p\right), \tag{9}$$

où

$$\hat{\mathbf{B}}_p = \left(\sum_{i \in s} w_i p_i c_i \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \sum_{i \in s} w_i p_i c_i \mathbf{z}_i y_i. \tag{10}$$

L'expression du biais en (9) est obtenue en remarquant que, sous l'approche MNC, on a $E_r\left(a_i \middle| s\right) = p_i$. Notons que dans le cas particulier de réponse uniforme, $p_i = p$ (approche MU), le biais en (9) est égal à 0. L'expression du biais en (9) suggère donc que le choix des valeurs imputées (6) n'est, en général, pas adéquat sous l'approche de MNC.

2.2 Un estimateur ajusté

Supposons pour le moment que les probabilités de réponse sont connues. Une approche naturelle permettant d'éliminer le biais de $\hat{Y}_I$ sous l'approche MNC est de considérer un estimateur ajusté pour le biais de la forme

$$\hat{Y}_I^a = \hat{Y}_I - \hat{B}\left(\hat{Y}_I \middle| s\right), \tag{11}$$

où $\hat{B}\left(\hat{Y}_I \middle| s\right)$ est un estimateur du biais conditionnel (9) donné par

$$\hat{B}\left(\hat{Y}_I \middle| s\right) = \sum_{i \in s} w_i a_i \frac{(1 - p_i)}{p_i}\left(y_i - \mathbf{z}_i' \hat{\mathbf{B}}_r\right). \tag{12}$$

Notons que $E_r\left[\hat{B}\left(\hat{Y}_I \middle| s\right) s\right] \approx Biais\left(\hat{Y}_I \middle| s\right)$ sous l'approche MNC. L'estimateur ajusté $\hat{Y}_I^a$ s'écrit comme

$$\hat{Y}_I^a = \sum_{i \in s} \frac{w_i}{p_i} a_i \left(y_i - \mathbf{z}_i' \hat{\mathbf{B}}_r\right) + \sum_{i \in s} w_i \mathbf{z}_i' \hat{\mathbf{B}}_r. \tag{13}$$

Notons que l'estimateur ajusté $\hat{Y}_I^a$ en (13) possède la forme d'un estimateur par la régression. En pratique, les probabilités de réponse ne sont pas connues. Supposons que l'on puisse obtenir des estimateurs valides $\hat{p}_i$ de p_i en modélisant p_i selon le modèle de non-réponse (2). Un estimateur ajusté est alors obtenu en remplaçant p_i en (13) par $\hat{p}_i$. Cet estimateur est également approximativement sans biais sous l'approche MI. Par conséquent, l'estimateur ajusté est robuste en ce sens qu'il est valide sous les approches MNC et MI. Cependant, ce dernier présente deux désavantages pratiques. Premièrement, les indicateurs de réponse a_i doivent être fournis dans le fichier de données, ce qui n'est pas toujours satisfait en pratique. Deuxièmement, les probabilités de réponse estimées $\hat{p}_i$ doivent également être fournies dans le fichier. Ces deux désavantages peuvent être contournées en utilisant la nouvelle méthode d'imputation, présentée en section 2.3, et qui mène à un estimateur imputé approximativement sans biais sous les approches MI et MNC.

2.3 Imputation par la régression modifiée

Nous supposons, pour le moment, que les probabilités de réponse p_i sont connues. Nous proposons d'utiliser les valeurs imputées

$$y_i^* = \mathbf{z}_i'\boldsymbol{\beta} \tag{14}$$

pour remplacer la valeur manquante y_i et d'obtenir la forme de $\boldsymbol{\beta}$ qui mène à un estimateur imputé approximativement sans biais sous l'approche MNC.

2.3.1 Étude du biais

Le lemme suivant exhibe la forme de $\boldsymbol{\beta}$ qui mène à un estimateur approximativement sans biais sous l'approche MNC.

Lemme 1: Le choix de $\boldsymbol{\beta}$ qui mène à un estimateur sans biais sous l'approche MNC est donné par

$$\tilde{\mathbf{B}} = \left[\sum_{i\in s} w_i(1-p_i)c_i\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\sum_{i\in s} w_i(1-p_i)c_i z_i y_i. \tag{15}$$

Notons que $\tilde{\mathbf{B}}$ en (15) ne peut être calculé puisque les valeurs de y ne sont observées que pour les unités répondantes et que les probabilités de réponse p_i ne sont pas connues. Un estimateur de $\tilde{\mathbf{B}}$ calculé à partir des unités répondantes, est donné par

$$\tilde{\mathbf{B}}_r = \left[\sum_{i\in s} w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}c_i\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\sum_{i\in s} w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}c_i\mathbf{z}_i\mathbf{y}_i. \tag{16}$$

On a $E_r\left(\tilde{\mathbf{B}}_r|s\right) = \tilde{\mathbf{B}}$ quand $\hat{p}_i \approx p_i$; autrement dit, $\tilde{\mathbf{B}}_r$ est un estimateur conditionnellement approximativement sans biais de $\tilde{\mathbf{B}}$ sous l'approche MNC. Il s'ensuit que l'estimateur imputé $\hat{Y}_I$ en (4) qui utilise les valeurs imputées

$$y_i^* = \mathbf{z}_i'\tilde{\mathbf{B}}_r \tag{17}$$

est approximativement sans biais pour *Y*. Notons que $\tilde{\mathbf{B}}_r$ représente l'estimateur des moindres carrés pondérés de $\boldsymbol{\beta}$ utilisant les poids $\tilde{w}_i = w_i\dfrac{(1-\hat{p}_i)}{\hat{p}_i}$. Notons que la procédure fait en sorte d'ajuster les poids de sondage à la hausse pour les unités qui présentent une faible probabilité de réponse et de les ajuster à la baisse pour les unités qui ont une forte probabilité de réponse. L'estimateur imputé peut être calculé à partir du fichier de données comprenant les poids de sondage w_i et les valeurs de la variable $\tilde{y}_i$. Les variables indicatrices de réponse et les probabilités de réponse estimées $\hat{p}_i$ ne sont pas requises. Le choix des valeurs imputées (17) mène également à un estimateur imputé approximativement sans biais pour Y sous l'approche MI. Par conséquent, ce choix de valeurs imputées mène à un estimateur robuste en ce sens qu'il

est valide sous les deux approches. Deux cas particuliers de (17) sont fréquemment utilisés en pratique: (i) l'imputation par le ratio modifié $(\boldsymbol{\lambda} = 1, \mathbf{z}_i = z_i)$ pour laquelle (17) devient

$$y_i^* = \frac{\sum_{i \in s} \tilde{w}_i a_i y_i}{\sum_{i \in s} \tilde{w}_i a_i z_i} z_i \tag{18}$$

et (ii) l'imputation par la moyenne modifiée $(\boldsymbol{\lambda} = 1, \mathbf{z}_i = 1)$ pour laquelle (17) devient

$$y_i^* = \frac{\sum_{i \in s} \tilde{w}_i a_i y_i}{\sum_{i \in s} \tilde{w}_i a_i}. \tag{19}$$

Sous réponse uniforme avec $\hat{p}_i = \hat{p}$ (approche MU), les valeurs imputées (18) et (19) deviennent $\left(\sum_{i \in s} w_i a_i y_i \Big/ \sum_{i \in s} w_i a_i z_i \right) z_i$ et $\sum_{i \in s} w_i a_i y_i \Big/ \sum_{i \in s} w_i a_i$, respectivement. Ces valeurs imputées sont celles traditionnellement utilisées par les statisticiens d'enquête (voir, Rao et Sitter, 1995).

2.3.2 Choix optimal de $\boldsymbol{\beta}$

Donc cette section, nous étudions le choix optimal de $\boldsymbol{\beta}$ qui est celui qui minimise l'erreur quadratique moyenne de l'estimateur imputé $\hat{Y}_I$ avec les valeurs imputées $y_i^* = \mathbf{z}_i' \boldsymbol{\beta}$. L'erreur quadratique moyenne, conditionnelle de l'estimateur imputé $\hat{Y}_I$ est définie selon

$$\begin{aligned} EQM_r\left(\hat{Y}_I | s\right) &= V_r\left(\hat{Y}_I | s\right) + \left[Biais\left(\hat{Y}_I | s\right)\right]^2 \\ &= \sum_{i \in s} w_i^2 p_i (1 - p_i)(y_i - \mathbf{z}_i' \boldsymbol{\beta})^2 + \left[\sum_{i \in s} w_i (1 - p_i)(y_i - \mathbf{z}_i' \boldsymbol{\beta}) \right]^2, \end{aligned} \tag{20}$$

où $V_r(.|s)$ désigne la variance conditionnelle par rapport au mécanisme de non-réponse. Nous cherchons donc $\boldsymbol{\beta}$ pour lequel $EQM_r\left(\hat{Y}_I | s\right)$ est minimum. Ce choix optimal, $\tilde{\mathbf{B}}_{opt}$, est généralement relativement complexe mais dans le cas particulier de l'imputation par le ratio, $\tilde{\mathbf{B}}_{opt}$ est égal à

$$\tilde{\mathbf{B}}_{opt} = \frac{\sum_{i \in s} w_i (1 - p_i) y_i \sum_{i \in s} w_i (1 - p_i) z_i + \sum_{i \in s} w_i^2 p_i (1 - p_i) y_i z_i}{\left[\sum_{i \in s} w_i (1 - p_i) z_i \right]^2 + \sum_{i \in s} w_i^2 p_i (1 - p_i) z_i^2}. \tag{21}$$

Supposons que les poids de sondages w_i sont tels que $\max_i \left(\frac{n}{N} w_i \right) = O(1)$ et qu'il existe une constante C telle que $C < p_i$. Alors,

$$\tilde{\mathbf{B}}_{opt} = \frac{\sum_{i\in s} w_i(1-p_i)y_i}{\sum_{i\in s} w_i(1-p_i)z_i} + O\left(\frac{1}{n}\right)$$
$$= \tilde{\mathbf{B}}_{opt} + 0\left(\frac{1}{n}\right),$$

où $\tilde{\mathbf{B}}$ est donné en (15). Autrement dit, le choix $\tilde{\mathbf{B}}$ est presque optimal pour de grandes tailles d'échantillon.

2.4 Imputation par la régression aléatoire

L'imputation aléatoire peut être vue comme une imputation déterministe à laquelle on ajoute un résidu aléatoire. Soit s_r et s_m les ensembles des répondants et des non-répondants, respectivement. Soit $e_j = c_j^{1/2}\left(y_j - \mathbf{z}_j'\hat{\mathbf{B}}_r\right)$ les résidus standardisés correspondant au répondant $j \in s_r$. Posons $e_i^* = e_j$ tel que $P\left(e_i^* = e_j\right) = \frac{w_i}{\sum_{l\in s} w_l a_l}$. L'imputation par la régression aléatoire utilise les valeurs imputées

$$y_i^* = \mathbf{z}_j'\hat{\mathbf{B}}_r + \varepsilon_i^*,$$

où $\varepsilon_i^* = c_i^{1/2}\left(e_i^* - \bar{e}_r\right)$ et $\bar{e}_r = \sum_{i\in s} w_j a_j e_j \Big/ \sum_{i\in s} w_j a_j$. Désignons par $E_*(.)$ l'espérance par rapport au mécanisme d'imputation aléatoire. On a $E_*\left(\varepsilon_i^*\right) = 0$ et $E_*\left(\hat{Y}_I\right)$ est égal à (8). L'estimateur imputé $\hat{Y}_I$ est donc approximativement sans biais pour Y sous les approches MU et MI. Cependant, l'estimateur $\hat{Y}_I$ est biaisé sous l'approche MNC. Nous proposons donc une version aléatoire de l'imputation par la régression modifiée qui mènera à un estimateur approximativement sans biais sous l'approche MNC. Soit $\tilde{e}_j = c_j^{1/2}\left(y_j - \mathbf{z}_j'\tilde{\mathbf{B}}_r\right)$ et $\tilde{e}_i^* = \tilde{e}_j$ tel que $P\left(\tilde{e}_i^* = \tilde{e}_j\right) = \tilde{w}_j \Big/ \sum_{l\in s} \tilde{w}_l a_l$ où $\tilde{\mathbf{B}}_r$ est donné en (16) et $\tilde{w}_i = w_i\left(1-\hat{p}_i\right)/\hat{p}_i$. L'imputation par la régression aléatoire modifiée utilise les valeurs imputées

$$y_i^* = \mathbf{z}_i\tilde{\mathbf{B}}_r + \tilde{\varepsilon}_i^*,$$

où $\tilde{\varepsilon}_i^* = c_i^{-1/2}\left(\tilde{e}_i^* - \tilde{e}_r\right)$ et $\tilde{e}_r = \sum_{j\in s} \tilde{w}_j a_j \tilde{e}_j \Big/ \sum_{j\in s} \tilde{w}_j a_j$. On a $E_*\left(\tilde{\varepsilon}_i^*\right) = 0$ et $E_*\left(\hat{Y}_I\right)$ coïncide avec l'estimateur imputé $\hat{Y}_I$ dans le cas de l'imputation par la régression déterministe modifiée. Par conséquent, l'estimateur $\hat{Y}_I$ est approximativement sans biais sous les approches MNC et MI.

3. Estimation de la variance

Dans cette section, nous développons un estimateur de variance pour l'estimateur imputé $\hat{Y}_I$ sous l'approche renversée de Fay (1991). La variance totale de $\hat{Y}_I$ est donnée par

$$V\left(\hat{Y}_I\right) = E_r V_p\left(\hat{Y}_I \middle| \mathbf{a}\right) + V_r E_p\left(\hat{Y}_I \middle| \mathbf{a}\right), \tag{22}$$

où $\hat{Y}_I$ est donné en (4) avec $y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r$ et $\mathbf{a} = \left(a_1, ..., a_N\right)'$ désigne le vecteur des variables indicatrices de réponse (voir Shao et Steel, 1999). Un estimateur de la variance totale $V\left(\hat{Y}_I\right)$ en (22) est donnée par $v_t = v_1 + v_2$, où v_1 est un estimateur de $V_p\left(\hat{Y}_I \middle| \mathbf{a}\right)$ et v_2 est un estimateur de $V_r\left[E_p\left(\hat{Y}_I \middle| \mathbf{a}\right)\right]$. L'estimateur v_1 ne dépend pas de la validité du mécanisme de non-résponse ou de celle du modèle d'imputation; il est donc valide sous les approches MNC et MI.

Dans le cas de l'imputation par la régression aléatoire modifiée, $y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r + \tilde{\varepsilon}_i^*$, la variance totale de l'estimateur imputé $\hat{Y}_I$ est donnée par

$$V\left(\hat{Y}_I\right) = E_r V_p E_*\left(\hat{Y}_I \middle| \mathbf{a}\right) + E_r E_p V_*\left(\hat{Y}_I \middle| \mathbf{a}\right) + V_r E_p E_*\left(\hat{Y}_I \middle| \mathbf{a}\right), \tag{23}$$

où $V_*(.)$ désigne la variance par rapport au mécanisme d'imputation aléatoire. Puisque $E_*\left(\hat{Y}_I \middle| \mathbf{a}\right)$ coïncide avec l'estimateur imputé obtenu dans le cas de l'imputation par la régression déterministe modifiée, la composante $E_r V_p E_*\left(\hat{Y}_I \middle| \mathbf{a}\right)$ en (23) est estimée par v_1 obtenu dans le cas de l'imputation déterministe. De manière similaire, la composante $V_r E_p E_*\left(\hat{Y}_I \middle| \mathbf{a}\right)$ est estimée par v_2 obtenu dans le cas de l'imputation déterministe. La contribution additionnelle à la variance due au mécanisme d'imputation aléatoire est donnée par $E_r E_p V_*\left(\hat{Y}_I \middle| \mathbf{a}\right)$ que l'on estimera par $v_* = V_*\left(\hat{Y}_I \middle| \mathbf{a}\right)$. La variance totale $V\left(\hat{Y}_I\right)$ en (23) est donc estimée par $v_t = v_1 + v_* + v_2$.

3.1 Cas des probabilités de réponse connues

Dans cette section, nous supposons que les probabilités de réponse p_i sont connues. Dans un premier temps, nous considérons le cas de l'imputation par la régression déterministe modifiée en section 3.1.1. Le cas de l'imputation par la régression aléatoire modifiée est étudié en section 3.1.2.

3.1.1 Imputation par la régression déterministe modifiée

Dans le cas de l'imputation par la régression déterministe modifiée, l'estimateur imputé $\hat{Y}_I$ peut s'écrire comme

$$\hat{Y}_{Ip} = \sum_{i \in s} w_i a_i y_i + \left(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r\right)' \tilde{\mathbf{B}}_{rp}, \tag{24}$$

où $$\tilde{\mathbf{B}}_{rp} = \left[\sum_{i\in s} w_i a_i \frac{(1-p_i)}{p_i} c_i \mathbf{z}_i \mathbf{z}_i'\right]^{-1} \sum_{i\in s} w_i a_i \frac{(1-p_i)}{p_i} c_i \mathbf{z}_i y_i. \quad (25)$$

Un développement en série de Taylor permet d'obtenir

$$\hat{Y}_{Ip} - Y \approx \sum_{i\in s} w_i \tilde{\xi}_{ip}, \quad (26)$$

où

$$\tilde{\xi}_{ip} = a_i y_i + (1-a_i)\mathbf{z}_i'\tilde{\mathbf{B}}_{rp} + \left(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r\right)'\tilde{\mathbf{T}}_p^{-1} a_i \frac{(1-p_i)}{p_i} c_i \mathbf{z}_i \left(y_i - \mathbf{z}_i'\tilde{\mathbf{B}}_{rp}\right)$$

et $\tilde{\mathbf{T}}_p = \sum_{i\in s} w_i a_i \frac{(1-p_i)}{p_i} c_i \mathbf{z}_i \mathbf{z}_i'$. Désignons l'estimateur de la variance de $\hat{Y}$ en (3) par $v(y)$. Il s'ensuit que la composante v_1 de la variance est donnée par

$$v_1 = v\left(\tilde{\xi}_p\right), \quad (27)$$

obtenue en remplaçant y_i par $\tilde{\zeta}_{ip}$ dans l'expression de $v(y)$. Afin d'obtenir la deuxième composante v_2, notons d'abord que

$$E_p\left(\hat{Y}_{Ip}|\mathbf{a}\right) \approx \sum_{i\in U} a_i y_i + \sum_{i\in U}(1-a_i)\tilde{\mathbf{B}}_{Np},$$

où $\tilde{\mathbf{B}}_{Np} = \left[\sum_{i\in U} a_i \frac{(1-p_i)}{p_i} c_i \mathbf{z}_i \mathbf{z}_i'\right]^{-1} \sum_{i\in U} a_i \frac{(1-p_i)}{p_i} c_i \mathbf{z}_i y_i$. Un développement par série de Taylor permet d'obtenir

$$V_r E_p\left(\hat{Y}_{Ip}|\mathbf{a}\right) \approx \sum_{i\in U} p_i(1-p_i)\delta_i^2, \quad (28)$$

où

$$\delta_i = 1 + \frac{(1-p_i)}{p_i} c_i \left(\mathbf{Z} - \mathbf{Z}_r\right)' \mathbf{T}_p^{-1} \mathbf{z}_i \left(y_i - \mathbf{z}_i'\tilde{\mathbf{B}}_{Np}\right), \quad (29)$$

$\mathbf{Z} = \sum_{i\in U} \mathbf{z}_i, \mathbf{Z}_r = \sum_{i\in U} a_i \mathbf{z}_i$ et $\mathbf{T}_p = \sum_{i\in U} a_i \frac{(1-p_i)}{p_i} c_i \mathbf{z}_i \mathbf{z}_i'$. La composante v_2 est obtenue en estimant les quantités inconnues en (28), ce qui mène à

$$v_2 = \sum_{i\in s} w_i a_i (1-p_i)\tilde{\delta}_i^2,$$

où

$$\tilde{\delta}_i = 1 + \frac{(1-p_i)}{p_i} c_i \left(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r\right)' \tilde{\mathbf{T}}_p^{-1} \mathbf{z}_i \left(y_i - \mathbf{z}_i'\tilde{\mathbf{B}}_{rp}\right).$$

Un estimateur, v_t, de la variance totale est finalement obtenu en additionnant (27) et (29).

3.1.2 Imputation par la régression aléatoire modifiée

D'abord, notons que

$$V_*\left(y_i^*\right) = c_j \sum_{i\in s} w_i a_i \frac{\left(1-p_i\right)}{p_i}\left(\tilde{e}_i - \tilde{e}_r\right)^2 \Big/ \sum_{i\in s} w_i a_i \frac{\left(1-p_i\right)}{p_i} \equiv \tilde{s}_e^2$$

et $Cov_*\left(y_i^*, y_j^*\right) = 0, \text{si } i \neq j.$ La composante v_* et donc donnée par

$$v_* = \sum_{i\in s} w_i^2 \left(1-a_i\right) V_*\left(y_i^*\right) = \sum_{i\in s} w_i^2 \left(1-a_i\right)\tilde{s}_e^2. \qquad (30)$$

Un estimateur, v_t, de la variance totale est finalement obtenue en additionnant (27), (29) et (30).

3.2 Cas des probabilités de réponse inconnues

Dans cette section, nous utilisons la méthode de Binder (1983) pour développer la composante v_1 lorsque les probabilités de réponse sont estimées. Supposons que p_i satisfait le modèle de régression logistique en (2). Posons $f\left(\mathbf{u}_i'\boldsymbol{\eta}\right) = e^{\mathbf{u}_i'\boldsymbol{\eta}}$. Les probabilités de réponse estimées sont données par

$$\hat{p}_i = \frac{f\left(\mathbf{u}_i'\hat{\boldsymbol{\eta}}\right)}{1+f\left(\mathbf{u}_i'\hat{\boldsymbol{\eta}}\right)},$$

où $\hat{\boldsymbol{\eta}}$ est un estimateur convergent de $\boldsymbol{\eta}$. Soit $\boldsymbol{\theta} = \left(\boldsymbol{\eta}_N', \mathbf{B}_N', Y\right)'$, où $\boldsymbol{\eta}_N$ et $\mathbf{B}_N$ sont les paramètres de population finie correspondant à $\boldsymbol{\eta}$ et $\boldsymbol{\beta}$, respectivement. Un estimateur de $\boldsymbol{\theta}$, donné par $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\eta}}_N', \tilde{\mathbf{B}}_r', \hat{Y}_I\right)'$, peut être exprimé comme solution du système d'équations estimantes suivant :

$$\hat{\mathbf{S}}\left(\boldsymbol{\theta}\right) = \mathbf{0},$$

où $\hat{\mathbf{S}}\left(\boldsymbol{\theta}\right) = \left(\hat{\mathbf{S}}_1\left(\boldsymbol{\theta}\right), \hat{\mathbf{S}}_2\left(\boldsymbol{\theta}\right), \hat{S}_3\left(\boldsymbol{\theta}\right)\right)'$ avec

$$\hat{\mathbf{S}}_1\left(\boldsymbol{\theta}\right) = \sum_{i\in s} w_i \mathbf{u}_i \left(a_i - f\left(\mathbf{u}_i'\boldsymbol{\eta}_N\right)\right) = \mathbf{0},$$

$$\hat{\mathbf{S}}_2\left(\boldsymbol{\theta}\right) = \sum_{i\in s} w_i a_i \mathbf{z}_i \frac{\left(1-f\left(\boldsymbol{\mu}_i'\boldsymbol{\eta}\right)\right)}{f\left(\boldsymbol{\mu}_i'\boldsymbol{\eta}\right)} c_i \left(y_i - \mathbf{z}_i'\mathbf{B}_N\right) = \mathbf{0}$$

et

$$\hat{S}_3\left(\boldsymbol{\theta}\right) = Y - \sum_{i\in s} w_i \mathbf{z}_i'\mathbf{B}_N - \sum_{i\in s} w_i a_i \left(y_i - \mathbf{z}_i'\mathbf{B}_N\right) = \mathbf{0}.$$

Soit $\hat{\mathbf{J}}(\boldsymbol{\theta}) = \partial \hat{\mathbf{S}}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ la matrice carrée des dérivées partielles de dimension $(k+l+1)$. On a

$$\mathbf{V}(\hat{\boldsymbol{\theta}}|\mathbf{a}) = \left[\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})\right]\boldsymbol{\Sigma}(\boldsymbol{\theta})\left[\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})\right]',$$

où $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ désigne la matrice carrée de variance-covariance de dimension $(k+l+1)$. Lorsque $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ est remplacée par un estimateur convergent $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, , on obtient un estimateur $\mathbf{V}(\hat{\boldsymbol{\theta}}|\mathbf{a})$ donné par

$$\mathbf{v}(\hat{\boldsymbol{\theta}}) = \left[\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})\right]\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})\left[\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})\right]'.$$

Rappelons que notre but est d'obtenir la composante v_1. Soit $\hat{\mathbf{b}}$ la dernière ligne de $\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})$ évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Alors

$$\mathbf{v}_1 = \hat{\mathbf{b}}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})\hat{\mathbf{b}}'.$$

Pour obtenir le terme v_2, supposons que les poids de sondage w_i satisfont $\max_i\left(\frac{n}{N}w_i\right) = O(1)$ et qu'il existe une constante positive C telle que $C < p_i$. De plus, supposons que $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = O_p\left(n^{-1/2}\right)$. Par développement en série de Taylor, on a

$$\hat{Y}_I = \hat{Y}_{Ip} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\sum_{i\in U} p_i^{-1}\left(y_i - \tilde{\mathbf{B}}_a\right)\frac{\partial f(\mathbf{u}_i'\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + O_p\left(Nn^{-1}\right),$$

où

$$\tilde{\mathbf{B}}_a = \left[\sum_{i\in U}(1-a_i)c_i\mathbf{z}_i\mathbf{z}_i\right]^{-1}\sum_{i\in U}(1-a_i)c_i\mathbf{z}_i y_i.$$

Si $\frac{\partial f(\mathbf{u}_i'\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}$ est borné uniformément, on a

$$E_p\left(\hat{Y}_I\right) = E_p\left(\hat{Y}_{Ip}\right) + O_p\left(Nn^{-1/2}\right).$$

La composante $V_r E_p\left(\hat{Y}_{Ip}|\mathbf{a}\right)$ est approximativement donnée par (28) et son estimateur v_2 est alors donné par (29).

Bibliographie

[1] Binder, D. A., "On the variances of asymptotically normal estimators from complex surveys", *International Statistical Review*, vol 51, pp 279-292, 1983.

[2] Deville, J. C., Särndal, C. E., "Variance estimation for the regression imputed Horvitz-Thompson estimator", *Journal of Official Statistics*, vol 10, pp 381-394, 1994.

[3] Fay, R. E., "A design-based perspective on missing data variance", *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, pp 420-440, 1991.

[4] Rao, J. N. K.,"On variance estimation with imputed survey data", *Journal of the American Statistical Association*, vol 91, pp 499-506, 1996.

[5] Rao, J. N. K.,"Variance estimation under imputation for missing data", *Technical report, Statistics Canada, Ottawa,* 1990.

[6] Rao, J. N. K., Sitter, R. R.,"Variance estimation under two-phse sampling with application to imputation for missing data", *Biometrika*, vol 82, pp 453-460, 1995.

[7] Särndal, C. E., "Methods for estimating the precision of survey estimates when imputation has been used", *Survey Methodology*, vol 18, pp 241-252, 1992.

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