Triple robustesse en présence de données imputées dans les enquêtes

David HAZIZA (*), Valéry DONGMO JIONGO (**), Pierre DUCHESNE (***)

(*)Université de Montréal
(**) Statistique Canada
(***) Université de Montréal

Introduction

En présence de non-réponse, les estimateurs non ajustés pour la non-réponse peuvent être fortement biaisés si les répondants diffèrent des non-répondants au regard des variables étudiées et que les taux de non-réponse sont grands. L'objectif premier des méthodes de traitement est donc de réduire le biais de non-réponse. Dans cet article, nous considérons le problème de la non-réponse partielle, qui est, dans la plupart des cas, traitée par imputation.

En l'absence d'erreurs non dues à l'échantillonnage (par exemple, erreurs de couverture, erreurs de non-réponse, etc), les statisticiens d'enquêtes utilisent des procédures d'estimation qui sont (asymptotiquement) sans biais sous le plan. Autrement dit, la validité de ces estimateurs ne dépend pas de la validité d'un modèle sous-jacent. En présence de non-réponse, l'usage de modèle est incontournable. On distingue deux types de modèles : le modèle de non-réponse qui est un ensemble d'hypothèses à propos du mécanisme (inconnu) de non-réponse et le modèle d'imputation qui est un ensemble d'hypothèse à propos de la distribution de la variable que l'on cherche à imputer.

Dans cet article, nous considérons les procédures d'imputation doublement robustes. Une procédure d'imputation est dite doublement robuste si l'estimateur imputé résultant est asymptotiquement sans biais et consistant lorsque le modèle de non-réponse ou le modèle d'imputation est correctement spécifié. Une procédure doublement robuste conduit donc à un estimateur asymptotiquement sans biais si l'un des deux modèles est correctement spécifié. Dans le contexte des enquêtes, les procédures doublement robustes ont été étudiées, entre autre, par Kott (1994), Haziza et Rao (2006) et Kim et Haziza (2011).

Bien que les procédures d'imputation offrent une certaine protection contre la mauvaise spécification de l'un des deux modèles, les estimateurs imputés sont sensibles à la présence d'unités influentes. Une unité influente fait partie de la population d'intérêt. Ces dernières sont fréquentes lorsque les variables d'intérêt sont fortement asymétriques ; par exemple, les variables de revenu. Les unités influentes ont généralement un impact important sur la volatilité (variance) des estimateurs. Le but sera donc de réduire l'influence des unités qui ont une grande influence, ce qui mènera à des estimateurs biaisés mais stables. Dans cet article, nous proposons des estimateurs robustes à la présence d'unités influentes. Ces derniers ont une forme similaire à l'estimateur robuste proposé par Beaumont, Haziza et

Ruiz-Gazen (2011) dans le cas de données complètes. Les estimateurs proposés dans cet article sont dits triplement robustes car ils présentent la propriété de double robustesse mais ils sont également robustes à la présence d'unités influentes.

1 Notation et cadres de travail

Soit U une population finie de taille N. Nous cherchons à estimer le total dans la population $Y = \sum_{i \in U} y_i$, où y_i désigne la i-ème valeur de la variable d'intérêt y, $i = 1, \cdots, N$. Un échantillon s, de taille n, est sélectionné selon un plan de sondage $p(s)$. Soit $d_i = 1/\pi_i$, le poids de sondage de l'unité i, où π_i désigne sa probabilité d'inclusion d'ordre 1 dans l'échantillon. En l'absence de non-réponse, un estimateur basé sur les données complètes est donné par l'estimateur par dialatation

$$\hat{Y}_\pi = \sum_{i \in U} d_i I_i y_i, \tag{1}$$

où I_i est une variable indicatrice de sélection de l'unité i telle que $I_i = 1$ si $i \in s$ et $I_i = 0$, sinon. L'estimateur (1) est sans biais sous le plan pour Y. Autrement dit, $E_p(\hat{Y}_\pi) = Y$, où $E_p(.)$ désigne l'espérance par rapport au plan de sondage $p(s)$.

Quand certaines des valeurs de la variable d'intérêt y sont manquantes, un estimateur de Y est l'estimateur imputé :

$$\hat{Y}_I = \sum_{i \in U} d_i r_i I_i y_i + \sum_{i \in U} d_i (1 - r_i) I_i y_i^*, \tag{2}$$

où r_i est une variable indicatrice de réponse de l'unité i telle que $r_i = 1$ si l'unité i a répondu à la variable y et $r_i = 0$, sinon, et y_i^* désigne la valeur imputée pour remplacer la valeur manquante y_i. On note également s_r et s_m les ensembles de répondants et de non-répondants, respectivement.

1.1 Les modèles en présence

Nous présentons maintenant le modèle de non-réponse ainsi que le modèle d'imputation à l'aide desquels nous construirons les valeurs imputées et étudierons les propriétés de l'estimateur (2).

D'une part, nous supposons que les unités répondent indépendamment les unes des autres et que la probabilité de réponse, p_i, de l'unité i peut être modélisée au moyen d'un modèle paramétrique

$$p_i = Prob(r_i = 1) = p(\mathbf{z}_i, \boldsymbol{\alpha}), \tag{3}$$

où $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus et $\mathbf{z}$ est un vecteur de variables auxiliaires disponibles pour toutes les unités échantillonnées (répondants et non-répondants). Le modèle (3) est appelé *modèle de non-réponse*. Un cas particulier de (3) est le modèle logistique

$$p_i = \frac{\exp(\mathbf{z}_i' \boldsymbol{\alpha})}{1 + \exp(\mathbf{z}_i' \boldsymbol{\alpha})}.$$

D'autre part, nous supposons que la variable d'intérêt obéit au modèle suivant :

$$y_i = \mathbf{z}_i' \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, N, \tag{4}$$

où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus et ϵ_i, $i = 1, \ldots, N$, sont des variables aléatoires indépendantes satisfaisant

$$E_m(\epsilon_i) = 0, \quad E_m(\epsilon_i^2) = \sigma^2 c_i, \quad E_m(\epsilon_i \epsilon_j) = 0 \quad i \neq j,$$

et $E_m(.)$ désigne l'espérance par rapport au modèle (4) et $c_i = \gamma(\mathbf{z}_i)$. La fonction $\gamma(.)$ est supposée connue. Le modèle (4) est appelé *modèle d'imputation*.

Dans cet article, nous considérons le cas de l'imputation par la régression déterministe doublement robuste pour laquelle la valeur imputée y_i^* est donnée par

$$y_i^* = \mathbf{z}_i' \hat{\boldsymbol{\beta}}_r \quad i \in s_m, \tag{5}$$

où $\hat{\boldsymbol{\beta}}_r$ est obtenu comme solution de l'équation estimante

$$\sum_{i \in U} d_i r_i I_i \left(\hat{p}_i^{-1} - 1\right) \left(y_i - \mathbf{z}_i' \hat{\boldsymbol{\beta}}_r\right) \mathbf{z}_i = 0, \tag{6}$$

où $\hat{p}_i = p(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$ désigne la probabilité de réponse estimée pour l'unité i et $\hat{\boldsymbol{\alpha}}$ est un estimateur de $\boldsymbol{\alpha}$ (par exemple, l'estimateur du maximum de vraisemblance) de $\boldsymbol{\alpha}$; voir Haziza et Rao (2006) et Kim et Haziza (2011).

1.2 Approches pour l'inférence

Afin d'étudier les propriétés des estimateurs imputés (par exemple, biais et variance), nous considérons deux approches distinctes : l'approche par modèle de non-réponse (NM) et l'approche par modèle d'imputation (IM). Avant de présenter ces deux approches, il convient de décrire les trois sources d'aléa sous-jacentes : (i) le modèle d'imputation qui génère le vecteur $\mathbf{y} = (y_1, ..., y_N)'$; (ii) le plan de sondage qui génère le vecteur $\mathbf{I} = (I_1, ..., I_N)'$ et (iii) le mécanisme de non-réponse qui génère le vecteur $\mathbf{r} = (r_1, ..., r_N)'$.

Dans le cas de l'approche NM, l'inférence est menée par rapport à la distribution conjointe du plan de sondage et du modèle de non-réponse (3). Notons, que le vecteur $\mathbf{y}$ est traité comme fixe. Cette approche a été étudiée, entre autre, par Rao (1996), Shao and Steel (1999), Beaumont (2005), Kim and Park (2006), Haziza and Rao (2006) et Haziza (2009).

Dans le cas de l'approche IM, l'inférence est menée par rapport à la distribution conjointe du plan de sondage et du modèle d'imputation (4). Notons qu'il n'est pas nécessaire de postuler un modèle de non-réponse explicite (comme, par exemple, le modèle (3)). Cependant, on suppose que le mécanisme de non-réponse est ignorable. Autrement dit, on supposera que

$$E_m\left(y_i \mid \mathbf{z}_i, r_i = 1\right) = E_m\left(y_i \mid \mathbf{z}_i, r_i = 0\right).$$

L'approche IM a été étudiée, entre autre, par Särndal (1992), Shao and Steel (1999), Brick, Kalton and Kim (2004) et Haziza (2009).

1.3 Décomposition de l'erreur totale et biais de non-réponse

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

$$\hat{Y}_I - Y = \left(\hat{Y}_\pi - Y\right) + \left(\hat{Y}_I - \hat{Y}_\pi\right). \tag{7}$$

Le terme $\hat{Y}_\pi - Y$ en (7) désigne l'erreur due à l'échantillonnage alors que le terme $\hat{Y}_I - \hat{Y}_\pi$ désigne l'erreur due à la non-réponse.

Dans le cas de l'approche NM, le biais de l'estimateur imputé $\hat{Y}_I$ est défini selon

$$Biais\left(\hat{Y}_I\right) = E_p E_q\left(\hat{Y}_I - Y|\mathbf{I}\right) = E_p B_q\left(\hat{Y}_I|\mathbf{I}\right),$$

où $B_q\left(\hat{Y}_I|\mathbf{I}\right) = E_q\left(\hat{Y}_I - \hat{Y}_\pi|\mathbf{I}\right)$ désigne le biais conditionnel de non-réponse sous l'approche NM et $E_q(.)$ désigne l'espérance par rapport au modèle de non-réponse (3). Lorsque $E_p B_q\left(\hat{Y}_I|\mathbf{I}\right) = 0$, on dira que l'estimateur imputé $\hat{Y}_I$ est sans biais par rapport à pq.

Dans le cas de l'approche IM, le biais de l'estimateur imputé $\hat{Y}_I$ est défini selon

$$Biais\left(\hat{Y}_I\right) = E_{mpq}\left(\hat{Y}_I - Y\right) = E_{pqm}\left(\hat{Y}_I - Y|\mathbf{I},\mathbf{r}\right) = E_{pq} B_m\left(\hat{Y}_I|\mathbf{I},\mathbf{r}\right),$$

où $B_m\left(\hat{Y}_I|\mathbf{I},\mathbf{r}\right) = E_m\left(\hat{Y}_I - \hat{Y}_\pi|\mathbf{I},\mathbf{r}\right)$ désigne le biais conditionnel de non-réponse sous l'approche IM. Lorsque $E_{pq} B_m\left(\hat{Y}_I|\mathbf{I},\mathbf{r}\right) = 0$, on dira que l'estimateur imputé $\hat{Y}_I$ est sans biais par rapport à mpq.

Sous certaines conditions de régularité, on peut montrer que l'estimateur imputé $\hat{Y}_I$ donné par (2) obtenu au moyen des valeurs imputées (5), est asymptotiquement sans biais et consistant pour Y si le modèle de non-réponse (3) est correctement spécifié et/ou le modèle d'imputation (4) est correctement spécifié ; voir Haziza et Rao (2006) et Kim et Rao (2011). Autrement dit, l'estimateur $\hat{Y}_I$ est doublement robuste. Bien qu'il possède la propriété de double robustesse, ce dernier est sensible à la présence de valeurs influentes. Il s'agira donc de robustifier $\hat{Y}_I$, ce qui nous amène à discuter du concept d'influence d'une unité.

2 Résultats théoriques préliminaires

Avant d'introduire le concept d'influence, nous présentons des résultats théoriques préliminaires qui nous seront utiles dans la suite des choses.

2.1 Approche NM

Le théorème suivant fournit une approximation de l'erreur due à la non-réponse sous l'approche NM.

Theorème 1. *Sous certaines conditions de régularité, on a :*

$$\hat{Y}_I - \hat{Y}_\pi = \sum_{i\in s} d_i\left(r_i g_i(p) - 1\right)\left\{y_i - \mathbf{z}_i^\top \mathbf{B}_p\right\} + \sum_{i\in s} d_i \nu_i (r_i - p_i) + o_p(Nn^{-1/2}), \tag{8}$$

où l'ordre de grandeur est par rapport à la distribution conjointe du modèle de non-réponse et du plan d'échantillonnage et où

$$g_i(p) = 1 + \left\{N^{-1}\sum_{l\in U}(1-p_l)\mathbf{z}_l^\top\right\}\left\{N^{-1}\sum_{l\in U}(1-p_l)\frac{\mathbf{z}_l\mathbf{z}_l^\top}{c_l}\right\}^{-1}\frac{1-p_i}{p_i}\frac{\mathbf{z}_i}{c_i}$$

et

$$\nu_i = \left\{\sum_{l\in U} p_l \left\{g_l(p)-1\right\}\left(y_l - \mathbf{z}_l^\top \mathbf{B}_p\right)\mathbf{z}_l^\top\right\}\left\{\sum_{l\in U} p_l(1-p_l)\mathbf{z}_l\mathbf{z}_l^\top\right\}^{-1}\mathbf{z}_i$$

avec

$$\mathbf{B}_p = \left\{\sum_{i\in U}(1-p_i)\frac{\mathbf{z}_i\mathbf{z}_i^\top}{c_i}\right\}^{-1}\sum_{i\in U}(1-p_i)\frac{\mathbf{z}_i y_i}{c_i}.$$

2.2 Approche IM

Le théorème suivant fournit une approximation de l'erreur due à la non-réponse sous l'approche IM.

Theorème 2. *Sous certaines conditions de régularité, on a :*

$$\hat{Y}_I - \hat{Y}_\pi = \sum_{i\in s} d_i\left(r_i g_i(r)-1\right)\left\{y_i - \mathbf{z}_i'\boldsymbol{\beta}\right\} + o_p(Nn^{-1/2}), \tag{9}$$

où l'ordre de grandeur est par rapport à la distribution conjointe du modèle d'imputation et du plan d'échantillonnage et où

$$g_i(r) = 1 + \left\{N^{-1}\sum_{l\in U}(1-r_l)\mathbf{z}_l^\top\right\}\left\{N^{-1}\sum_{i\in U}(1-\hat{p}_l)\frac{\mathbf{z}_l\mathbf{z}_l^\top}{c_l}\right\}^{-1}\frac{1-\hat{p}_i}{\hat{p}_i}\mathbf{z}_i.$$

3 Influence d'une unité : utilisation du biais conditionnel

En l'absence de non-réponse, Moreno-Rebollo, Munoz-Reyes and Munoz-Pichardo (1999) ont proposé le biais conditionnel d'une unité comme mesure d'influence ; voir aussi Beaumont, Haziza and Ruiz-Gazen (2011). Le biais conditionnel de l'unité échantillonnée i par rapport à l'estimateur par dilatation $\hat{Y}_\pi$ est défini selon :

$$\begin{aligned} B_i^\pi(I_i=1) &= E_p(\hat{Y}_\pi - Y|I_i=1) \\ &= (d_i-1)y_i + \sum_{\substack{j\in U\\ j\neq i}}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_j \\ &= \sum_{j\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_j. \end{aligned} \tag{10}$$

Le biais conditionnel $B_i^\pi(I_i=1)$ est une mesure de l'influence de l'unité i. Plus une unité a une grande influence, plus elle a un impact important sur la volatilité (variance) d'un estimateur. Notons que $B_i^\pi(I_i=1)=0$ lorsque $\pi_i=1$. Autrement dit, une unité sélectionnée avec probabilité 1 n'a aucune influence sur la volatilité de $\hat{Y}_\pi$. De plus, notons qu'une unité non-échantillonnée peut également avoir une grande influence. Le biais conditionnel de l'unité non-échantillonnée i par rapport à l'estimateur par dilatation $\hat{Y}_\pi$ est défini selon :

$$B_i^\pi(I_i=0) = E_p(\hat{Y}_\pi - Y|I_i=0) = -\frac{1}{d_i-1}B_i^\pi(I_i=1).$$

Cependant, à l'étape de l'estimation, seule l'influence des unités échantillonnées peut être réduite et rien ne peut être fait pour les unités non-échantillonnées.

Dans le cas du plan de Poisson, du plan stratifié aléatoire simple sans remise et des plans à grande entropie, Beaumont, Haziza and Ruiz-Gazen (2011) ont montré que l'erreur due à l'échantillonnage peut s'écrire (exactement ou approximativement) comme :

$$\hat{Y}_\pi - Y = \sum_{i \in s} B_i^\pi \left(I_i = 1\right) + \sum_{i \in U-s} B_i^\pi \left(I_i = 0\right).$$

Autrement dit, le biais conditionnel d'une unité peut s'interpréter comme la contribution d'une unité à l'erreur due à l'échantillonnage.

Dans les prochaines sous-sections, nous définissons deux mesures d'influence par rapport à l'estimateur imputé $\hat{Y}_I$: l'une sous l'approche NM et l'autre sous l'approche IM.

3.1 Biais conditionnel d'une unité sous l'approche NM

Dans cette section, nous définissons l'influence d'une unité répondante sous l'approche NM par rapport à l'estimateur imputé $\hat{Y}_I$ obtenu au moyen des valeurs imputées (5). Le biais conditionnel de l'unité répondante i est défini selon

$$B_{qi}^I \left(I_i = 1, r_i = 1\right) = E_{pq}\left(\hat{Y}_I - Y \middle| I_i = 1, r_i = 1\right). \tag{11}$$

Il découle de la décomposition (7) que le biais conditionnel de l'unité répondante i peut s'écrire comme

$$B_{qi}^I \left(I_i = 1, r_i = 1\right) = E_p\left(\hat{Y}_\pi - Y \middle| I_i = 1\right) + E_{pq}\left(\hat{Y}_I - \hat{Y}_\pi \middle| I_i = 1, r_i = 1\right). \tag{12}$$

Le premier terme à droite de l'égalité (12) représente la contribution (ou l'impact) de l'unité i à l'erreur due à l'échantillonnage, $\hat{Y}_\pi - Y$, alors que le deuxième terme représente la contribution de l'unité répondante i à l'erreur due à la nonréponse, $\hat{Y}_I - \hat{Y}_\pi$, sous l'approche NM. Le biais conditionnel (13) peut donc s'interpréter comme la contribution de l'unité répondante i à l'erreur totale, $\hat{Y}_I - Y$, sous l'approche NM. En ignorant les termes d'ordre inférieur, il découle de (8) que le biais conditionnel en (11) peut être approximé par

$$B_{qi}^I \left(I_i = 1, r_i = 1\right) \approx \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) y_j + d_i(1 - p_i)\left\{\nu_i + g_i(p)\left(y_i - \mathbf{z}_i' \mathbf{B}_p\right)\right\}. \tag{13}$$

L'unité i a une grande contribution à l'erreur due à la non-réponse lorsque (i) son poids de sondage est grand et/ou (ii) sa probabilité de réponse est petite et/ou (iii) le facteur $g_i(p)$ est grand et/ou (iv) son résidu $y_i - \mathbf{z}_i' \mathbf{B}_p$ est grand. Lorsque $p_i = 1$, notons que le deuxième terme à droite de l'égalité en (13) est égal à 0. Dans ce cas, l'unité i n'a aucune influence sur l'erreur due à la non-réponse.

3.2 Biais conditionnel d'une unité sous l'approche IM

Dans cette section, nous définissons l'influence d'une unité répondante sous l'approche IM par rapport à l'estimateur imputé $\hat{Y}_I$ obtenu au moyen des valeurs imputées (5). Le biais conditionnel de l'unité répondante i est défini selon

$$B_{mi}^I \left(y_i, I_i = 1, r_i = 1\right) = E_{mpq}\left(\hat{Y}_I - Y \middle| y_i, I_i = 1, r_i = 1\right). \tag{14}$$

Il découle de la décomposition (7) que le biais conditionnel de l'unité répondante i peut s'écrire comme

$$B_{mi}^{I}\left(y_i, I_i=1, r_i=1\right)=E_m E_p\left(\hat{Y}_\pi-Y\middle|y_i, I_i=1\right)+E_q E_p E_m\left(\hat{Y}_I-\hat{Y}_\pi\middle|y_i, I_i=1, r_i=1\right). \quad (15)$$

Le premier terme à droite de l'égalité (15) représente la contribution (ou l'impact) de l'unité i à l'erreur due à l'échantillonnage, $\hat{Y}_\pi - Y$, alors que le deuxième terme représente la contribution de l'unité répondante i à l'erreur due à la nonréponse, $\hat{Y}_I - \hat{Y}_\pi$ sous l'approche IM. Encore une fois, le biais conditionnel peut donc s'interpréter comme la contribution de l'unité répondante i à l'erreur totale, $\hat{Y}_I - Y$, sous l'approche IM. En ignorant les termes d'ordre inférieur, il découle de (9) que le biais conditionnel en (15) peut être approximé par

$$\begin{aligned} B_{mi}^{I}\left(y_i, I_i=1, r_i=1\right) &\approx E_m\left\{\sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right) y_j\middle|y_i\right\} \\ &+ E_q\left[d_i\left(g_i(r)-1\right)\left\{y_i-z_i'\boldsymbol{\beta}\right)\right\}\middle|r_i=1\right]. \end{aligned} \quad (16)$$

L'unité i a une grande contribution à l'erreur due à la non-réponse lorsque (i) son poids de sondage est grand et/ou (ii) le facteur $g_i(r)$ est grand et/ou (iii) son résidu $y_i - \mathbf{z}_i'\boldsymbol{\beta}$ est grand.

4 Estimateur imputé robuste à la présence de valeurs influentes

En l'absence de non-réponse, Beaumont, Haziza et Ruiz-Gazen (2011) ont proposé une version robuste de l'estimateur par dilatation :

$$\hat{Y}_\pi^R = \hat{Y}_\pi - \sum_{i\in s}\hat{B}_i^\pi(I_i=1)+\sum_{i\in s}\psi_c\left(\hat{B}_i^\pi(I_i=1)\right), \quad (17)$$

où $\hat{B}_i^\pi(I_i=1)$ est un estimateur de $B_i^\pi(I_i=1)$ obtenu en remplaçant les paramètres inconnus par des estimateurs robustes et $\psi_c(.)$ est une fonction dont le rôle est de réduire l'influence des unités qui ont une grande influence. Une fonction $\psi_c(.)$ populaire est la fonction de Huber donnée par

$$\psi_c(t)=\begin{cases} c & \text{if } t>c \\ t & \text{if } |t|\le c \\ -c & \text{if } t<-c \end{cases}$$

où c est une constante à déterminer. Nous faisons les remarques suivantes : (i) l'estimateur $\hat{Y}_\pi^R$ est consistant au sens de Cochran. Autrement dit, lorsque $s=U$, on a $\hat{Y}_\pi^R = Y$. (ii) Lorsque $c\to\infty$, l'estimateur robuste $\hat{Y}_\pi^R$ tend vers l'estimateur non-robuste $\hat{Y}_\pi$. (iii) Dans le cas stratifié aléatoire simple sans remise, $\hat{Y}_\pi^R$ coïncide (à un facteur près) avec l'estimateur de Kokic et Bell (1994).

4.1 Estimateur triplement robuste sous l'approche NM

Suivant Beaumont, Haziza et Ruiz-Gazen (2011), une version robuste de $\hat{Y}_I$ sous l'approche NM est donnée par

$$\hat{Y}_I^R = \hat{Y}_I - \sum_{i\in s_r}\hat{B}_{qi}^I(I_i=1,r_i=1)+\sum_{i\in s_r}\psi_c\left(\hat{B}_{qi}^I(I_i=1,r_i=1)\right), \quad (18)$$

où $\hat{B}^I_{qi}(I_i = 1, r_i = 1)$ est un estimateur de $B^I_{qi}(I_i = 1, r_i = 1)$ donné par (13) obtenu en remplaçant les paramètres inconnus par des estimateurs robustes. Nous faisons les remarques suivantes : (i) En l'absence de non-réponse, $s_r = s$, l'estimateur (18) coïncide avec l'estimateur robuste (17). (ii) Lorsque $c \to \infty$, l'estimateur imputé robuste, $\hat{Y}^R_I$, tend vers l'estimateur imputé doublement robuste $\hat{Y}_I$.

4.2 Estimateur triplement robuste sous l'approche IM

De manière similaire, une version robuste de $\hat{Y}_I$ sous l'approche IM est donnée par

$$\hat{Y}^R_I = \hat{Y}_I - \sum_{i \in s_r} \hat{B}^I_{mi}(y_i, I_i = 1, r_i = 1) + \sum_{i \in s_r} \psi_c \left(y_i, \hat{B}^I_{mi}(I_i = 1, r_i = 1) \right), \quad (19)$$

où $\hat{B}^I_{mi}(y_i, I_i = 1, r_i = 1)$ est un estimateur de $B^I_{mi}(y_i, I_i = 1, r_i = 1)$ donné par (16) obtenu en remplaçant les paramètres inconnus par des estimateurs robustes. Les remarques faites à propos de l'estimateur robuste (18) s'appliquent également à l'estimateur robuste (19).

5 Étude par simulation

Dans cette section, nous présentons les résultats d'une étude par simulation dont le but était de comparer les estimateurs imputés robustes et non robustes à la présence de valeurs influentes, en termes de biais et d'éfficacité.

Nous avons généré une population de taille $N = 10000$ comprenant deux variables : une variable d'intérêt y et une variable auxilaire z. D'abord, la variable z a été générée d'une loi normale de moyenne 10 et de variance 25. Étant donné z, la variable y a été générée selon le modèle de mélange

$$y_i = (1 - A_i) y_{0i} + A_i y_{1i}, \quad i = 1, \ldots, N, \quad (20)$$

où A_i est une variable dichotomique telle que $A_i = 1$ avec probabilité $\lambda \in (0, 1)$ et $A_i = 0$ avec probabilité $1 - \lambda$ et

$$\begin{aligned} y_{0i} &= 3 + z_i + \epsilon_{0i}, \quad i = 1, \ldots, N, \\ y_{1i} &= 5 z_i + \epsilon_{1i}, \quad i = 1, \ldots, N. \end{aligned}$$

Les erreurs ϵ_{0i} et ϵ_{1i} ont été générées d'une loi normale de moyenne 0 et de variance 1. Nous avons utilisé $\lambda = 0.05$.

De la population, nous avons selectionné $T = 1000$ échantillons selon un plan aléatoire imple sans remise de taille $n = 100$. Dans chaque échantillon tiré, nous avons assigné une probabilité de réponse, p_i, à l'unité i selon :

$$p_i = \frac{\exp^{(2.5 - 0.2 z_i)}}{1 + \exp^{(2.5 - 0.2 z_i)}}, \quad i = 1, \ldots, N. \quad (21)$$

Les paramètres en (21) ont été choisis de manière à obtenir un taux de réponse global approximativement égal à 65%. Finalement, une variable indicatrice r_i pour l'unité i a été générée aléatoirement d'une distribution de Bernoulli de paramètre p_i, $i = 1, \cdots, n$.

Afin de contruire les valeurs imputées, nous avons considéré 3 scénarios :

(1) Scenario 1 : le modèle de non-réponse et le modèle d'imputation sont bien spécifiés. On a d'abord obtenu $\hat{p}_i = p(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$ avec $\mathbf{z}_i = (1, z_i)'$. Les valeurs manquantes à la variable y ont été ensuite imputées selon (5) avec $\mathbf{z}_i = (1, z_i)'$ et $c_i = 1$.

(2) Scenario 2 : le modèle de non-réponse est mal spécifié alors que le modèle d'imputation est bien spécifié. On a d'abord obtenu $\hat{p}_i = p(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$ avec $\mathbf{z}_i = 1$. Les valeurs manquantes à la variable y ont été ensuite imputées selon (5) avec $\mathbf{z}_i = (1, z_i)'$ et $c_i = 1$.

(3) Scenario 3 : le modèle de non-réponse est bien spécifié alors que le modèle d'imputation est mal spécifié. On a obtenu $\hat{p}_i = p(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$ avec $\mathbf{z}_i = (1, z_i)'$. Les valeurs manquantes à la variable y ont été ensuite imputées selon (5) avec $\mathbf{z}_i = 1$ et $c_i = 1$.

Pour chaque scénario, nous avons calculé 2 estimateurs : (i) l'estimateur $\hat{Y}_I$ (non-robuste à la présence de valeurs influentes) donné par (2) avec les valeurs imputées (5). (ii) L'estimateur robuste $\hat{Y}_I^R$ donné par (19), où la constante c est celle qui minimise son erreur quadratique moyenne monte carlo. Pour l'estimateur $\hat{Y}_I^R$, nous avons estimé le biais conditionnel (16) par

$$\begin{aligned}\hat{B}_{mi}^I(y_i, I_i = 1, r_i = 1) &= \frac{N}{N-1}\left(\frac{N}{n} - 1\right)\left(y_i - mediane\left\{\tilde{y}_i, i \in s_r\right\}\right) \\ &+ \frac{N}{n}\left(\hat{g}_i(r) - 1\right)\left(y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\beta}}_r^{rob}\right),\end{aligned}$$

où

$$\hat{g}_i(r) = 1 + \left\{\frac{N}{n}\sum_{l\in s}(1 - r_l)\mathbf{z}_l^\top\right\}\left\{\frac{N}{n}\sum_{l\in s} r_l \frac{1-\hat{p}_l}{\hat{p}_l}\mathbf{z}_l\mathbf{z}_l^\top\right\}^{-1}\frac{1-\hat{p}_i}{\hat{p}_i}\mathbf{z}_i$$

et l'estimateur $\hat{\boldsymbol{\beta}}_r^{rob}$ est solution de

$$N^{-1}\sum_{i\in s} r_i d_i\left(\hat{p}_i^{-1} - 1\right)\psi_d\left(\frac{y_i - \mathbf{z}_i^\top\boldsymbol{\beta}}{\sigma c_i^{1/2}}\right)\frac{\mathbf{z}_i}{\sigma c_i^{1/2}} = 0.$$

La constante d a été fixée à 1.345.

Nous avons calculé deux mesures monte carlo : (i) le biais relatif monte carlo (en %) donné par

$$BR(\hat{Y}) = T^{-1}\sum_{t=1}^{T}\frac{\left(\hat{Y}_t - Y\right)}{Y} \times 100,$$

où $\hat{Y}$ est une notation générique pouvant désigner soit $\hat{Y}_I$, soit $\hat{Y}_I^R$ et $\hat{Y}_t$ désigne l'estimateur $\hat{Y}$ pour l'échantillon t, $t = 1, \cdots, T$; (ii) l'efficacité relative monte carlo (en %), définie par

$$ER(\hat{Y}_I^R) = \frac{MSE(\hat{Y}_I^R)}{MSE(\hat{Y}_I)} \times 100,$$

où

$$MSE(\hat{Y}) = \sum_{t=1}^{T}\left(\hat{Y}_t - Y\right)^2.$$

Les résultats son présentés au tableau 1. Il est clair que l'estimateur $\hat{Y}_I$ est doublement robuste puisque son biais est négligeable pour les 3 scénarios. L'estimateur $\hat{Y}_I^R$, quant à lui, exhibe une efficacité relative d'environ $66\% - 68\%$, ce qui illustre sa résistance à la présence d'unités influentes.

TABLE 1 – Biais relatifs Monte Carlo (%) et efficacité relative Monte Carlo (%) des estimateurs avec $n = 100$

Estimateurs	BR (%)	ER (%)
Scénario 1		
$\hat{Y}_I$	-0.01	100.0
$\hat{Y}_I^R$	-3.2	68.2
Scénario 2		
$\hat{Y}_I$	-0.02	100
$\hat{Y}_I^R$	-3.2	66.2
Scénario 3		
$\hat{Y}_I$	-0.02	100
$\hat{Y}_I^R$	-3.2	66.0

REFERENCES

Beaumont, J.-F. (2005). Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach. *Journal of the Royal Statistical Society B*, **67**, 445–458.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2011). A Unified Approach to Robust Estimation in Finite Population Sampling. *En révision.*

Brick, J. M., Kalton, G. and Kim, J. K. (2004). Variance Estimation with Hot-Deck Imputation Using a Model. *Survey Methodology*, **30**, 57–66.

Haziza, D. (2009). Imputation and Inference in the presence of missing data. In C. R. Rao and D. Pfefferman (Editors), *Handbook of Statistics, Sample Surveys, Design Methods and Applications*, Vol. 29A, pp. 215–246.

Haziza, D. et Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, **32**, 53–64.

Kim, J.K. and Haziza, D. (2010). Doubly robust inference with missing data. *Submitted for publication.*

Kim, J.K. and Park, H. (2006). Imputation using response probability. *The Canadian Journal of Statistics*, **34**, 171–182.

Kokic, P.N. and Bell, P.A. (1994). Optimal winzorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, **10**, 419–435.

Kott, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of American Statistical Association*, **89**, 693–696.

Moreno-Rebollo, J. L., Munoz-Reyes, A. et Munoz-Pichardo, J. (1999). Influence diagnostic in survey sampling : conditional bias. *Biometrika*, **86**, 923–928.

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

Särndal, C. E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241–252.

Shao, J. et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, **94**, 254–65.