Université de Montréal

IMPUTATION EN PRÉSENCE DE DONNÉES CONTENANT DES ZÉROS

par

CHRISTIAN OLIVIER NAMBEU

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)
en STATISTIQUE

décembre 2010

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

IMPUTATION EN PRÉSENCE DE DONNÉES CONTENANT DES ZÉROS

présenté par

CHRISTIAN OLIVIER NAMBEU

a été évalué par un jury composé des personnes suivantes :

Pierre Duchesne

(président-rapporteur)

David Haziza

(directeur de recherche)

Pierre Lafaye de Micheaux

(membre du jury)

Mémoire accepté le:

06 décembre 2010

SOMMAIRE

L'imputation simple est très souvent utilisée dans les enquêtes pour compenser pour la non-réponse partielle. Dans certaines situations, la variable nécessitant l'imputation prend des valeurs nulles un très grand nombre de fois. Ceci est très fréquent dans les enquêtes entreprises qui collectent les variables économiques. Dans ce mémoire, nous étudions les propriétés de deux méthodes d'imputation souvent utilisées en pratique et nous montrons qu'elles produisent des estimateurs imputés biaisés en général. Motivé par un modèle de mélange, nous proposons trois méthodes d'imputation et étudions leurs propriétés en termes de biais. Pour ces méthodes d'imputation, nous considérons un estimateur jackknife de la variance convergent vers la vraie variance, sous l'hypothèse que la fraction de sondage est négligeable. Finalement, nous effectuons une étude par simulation pour étudier la performance des estimateurs ponctuels et de variance en termes de biais et d'erreur quadratique moyenne.

Mots clés : imputation par la régression ; imputation aléatoire ; imputation déterministe ; imputation aléatoire équilibrée ; non-réponse partielle ; jackknife ; estimation de la variance.

SUMMARY

Single imputation is often used in surveys to compensate for item nonresponse. In some cases, the variable requiring imputation contains a large amount of zeroes. This is especially frequent in business surveys that collect economic variables. In this thesis, we study the properties of two imputation procedures frequently used in practice and show that they lead to biased estimators, in general. Motivated by a mixture regression model, we then propose three imputation procedures and study their properties in terms of bias. For the proposed imputation procedures, we consider a jackknife variance estimator that is consistent for the true variance, provided the overall sampling fraction is negligible. Finally, we perform a simulation study to evaluate the performance of point and variance estimators in terms of relative bias and mean square error.

Keywords : Regression imputation ; random imputation ; deterministic imputation ; balanced random imputation ; item nonresponse ; jackknife ; variance estimation.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Tout d'abord, je tiens à remercier mon directeur de recherche M. David Haziza pour m'avoir accompagné tout au long de ma maîtrise et durant la rédaction de ce mémoire.

Je remercie également M. Guillaume Chauvet pour le programme SAS développé pour l'imputation équilibrée aléatoire. Cela m'a permis de gagner un temps fou.

Je dédie ce mémoire à ma mémé chérie à qui je serai éternellement reconnaissant pour son amour inconditionnel et ses constants encouragements à mon égard depuis que j'aie entrepris de poursuivre mes études universitaires aussi loin de la maison.

INTRODUCTION

Les populations contenant beaucoup de zéros sont très fréquentes dans les enquêtes-entreprises. Le traitement de la non-réponse par les méthodes d'imputation classiques peut affecter la qualité des estimateurs ponctuels. Dans ce mémoire, notre but est de proposer une méthode d'imputation capable de produire un estimateur imputé approximativement sans biais et efficace et dont les valeurs imputées sont réalistes pour l'utilisateur des microdonnées. La structure de ce mémoire est la suivante : dans le chapitre 1, nous présentons quelques notions de base en théorie des sondages telles que le plan de sondage, le π-estimateur, l'information auxiliaire, l'estimateur par la régression généralisée et l'échantillonnage à deux phases ; dans le chapitre 2, nous introduisons le concept d'inférence en présence de non-réponse, ensuite nous présentons quelques méthodes d'imputation usuelles couramment utilisées en pratique. Nous présentons aussi la notion de mécanisme, de biais et de variance de non-réponse et nous illustrons enfin ces notions dans le cas de l'imputation par la régression ; au chapitre 3, nous introduisons l'inférence dans le contexte d'une population dont la variable d'intérêt prend la valeur zéro un très grand nombre de fois. Ensuite, nous postulons des hypothèses sur le modèle d'imputation. Nous proposons ensuite plusieurs méthodes d'imputation dont nous étudions les propriétés en termes de biais ; au chapitre 4, nous menons une étude par simulation pour mesurer la performance des estimateurs présentés au chapitre 3 en termes de biais relatif et d'efficacité relative. Finalement, au chapitre 5, nous proposons un estimateur jackknife de la variance convergent vers la vraie variance pour les méthodes d'imputation proposées et nous terminons par étude par simulation pour mesurer la performance de l'estimateur jackknife de la variance proposé.

Chapitre 1

QUELQUES RAPPELS EN THÉORIE DES SONDAGES

Dans ce chapitre, nous présentons les notions élémentaires utilisées en théorie des sondages qui serviront par la suite dans ce mémoire.

1.1. Population et Échantillon

En échantillonnage, nous travaillons avec une population finie U composée de N unités. On écrira

$$U = \{1, ..., i, ...N\}.$$

Nous supposons que toutes les unités de la population sont identifiables d'une manière ou d'une autre sans ambiguité. L'enquête porte souvent sur certaines caractéristiques d'intérêt (ou variables d'intérêt) de la population et on suppose que ces caractéristiques peuvent être mesurées sur chaque unité de la population U. Désignons par y une variable d'intérêt. On cherche à estimer une fonction des paramètres $\theta = f(y_1, ..., y_i, ..., y_N)$. Par exemple, on pourrait s'intéresser à estimer le total $Y = \sum_{i \in U} y_i$ (ou la moyenne $\bar{Y} = Y/N$), le ratio de deux totaux ou encore la variance d'une variable d'intérêt.

Pour estimer θ, nous allons tirer un échantillon s, de taille n, selon un plan de sondage, ce qui nous permettra d'observer $\mathbf{y}_s = (y_1, ..., y_i, ..., y_n)^\top$. L'objectif de la méthode de tirage ne vise pas forcément la sélection d'un échantillon « représentatif » de la population dans le sens où l'échantillon est un modèle réduit de la population car il est parfois nécessaire de tirer des échantillons dans lesquels

certaines unités sont quasi-sélectionnées d'avance (c'est-à-dire ont une très forte probabilité d'être sélectionnée) ; par exemple, voir Tillé (2001). La question qui se pose donc à juste titre est de savoir comment tirer un échantillon.

1.2. Plan de sondage

Le plan de sondage (ou plan d'échantillonnage) est la définition des caractéristiques régissant le mécanisme par lequel nous prélevons aléatoirement un échantillon s dans la population U. On désigne le plan de sondage par $p(.)$ tel que $P(S = s) = p(s)$ est la probabilité de sélection de l'échantillon s. Donc s est une réalisation de la variable aléatoire S. Par souci de simplicité, nous noterons tout simplement $p(s)$ pour désigner le plan de sondage et la probabilité de sélection de l'échantillon s. Autrement dit, le plan de sondage est une loi de probabilité qui associe à chaque échantillon s de l'ensemble de tous les échantillons possibles Ω une probabilité de sélection $p(s)$. En tant que loi de probabilité, le plan de sondage satisfait les conditions suivantes : $\sum_{s\in\Omega} p(s) = 1$ et $p(s) \geq 0$, $\forall s \in \Omega$.
Le plan de sondage joue un rôle crucial en théorie des sondages. L'inférence est menée sous les hypothèses sous-jacentes au plan de sondage car c'est le plan qui modélise l'aléa principal (la sélection des unités dans l'échantillon) du sondage. Généralement, les statisticiens d'enquêtes essaient d'éviter l'usage de modèles aux fins d'inférence ; mais comme nous le verrons au chapitre 2, en présence de non-réponse par exemple, l'utilisation des modèles semble incontournable. Introduisons alors une variable indicatrice pour formaliser ce mécanisme :

$$I_i(s) = \begin{cases} 1, & \text{si l'unité } i \text{ est sélectionnée dans } s, \\ 0, & \text{si l'unité } i \text{ n'est pas sélectionnée dans } s. \end{cases} \tag{1.2.1}$$

En théorie des sondages, l'inférence est menée sous le plan de sondage. Dans ce cas, le vecteur $\mathbf{y} = (y_1, ..., y_i, ..., y_N)^\top$ est traité comme fixe. L'aléa correspond à l'échantillon S. Désignons par π_i la probabilité d'inclusion du premier ordre de l'unité i dans l'échantillon et défini par

$$\pi_i = P(i \in s) = P(I_i(s) = 1) = \sum_{\substack{s\in\Omega \\ s\ni i}} p(s),$$

où $s \ni i$ désigne une réalisation de S contenant l'unité i. De manière similaire, on définit les probabilités d'inclusion du second ordre, π_{ij}, par :

$$\pi_{ij} = P(i \in s, j \in s) = P(I_i(s) = 1, I_j(s) = 1) = \sum_{\substack{s \in \Omega \\ s \ni (i,j)}} p(s).$$

Un exemple de plan de sondage simple est le tirage aléatoire simple sans remise. Le tirage aléatoire simple sans remise est un plan qui assigne à chaque échantillon de taille n, la même probabilité d'être sélectionnée. On a C_n^N échantillons possibles de taille n. On a alors

$$p(s) = \begin{cases} 1/C_n^N, & \text{si } s \text{ est de taille } n, \\ 0, & \text{sinon .} \end{cases} \tag{1.2.2}$$

Pour ce plan, les probabilités d'inclusion du premier et du second ordre sont respectivement données par

$$\pi_i = \frac{C_{n-1}^{N-1}}{C_n^N} = \frac{n}{N}$$

et

$$\pi_{ij} = \frac{C_{n-2}^{N-2}}{C_n^N} = \frac{n(n-1)}{N(N-1)}.$$

Soit θ un paramètre d'intérêt et $\hat{\theta}$ un estimateur de θ. L'espérance de l'estimateur $\hat{\theta}$ sous le plan de sondage est définie selon

$$E_p(\hat{\theta}) = \sum_{s \in \Omega} \hat{\theta}(s) p(s) \tag{1.2.3}$$

et sa variance sous le plan de sondage est donnée par

$$V_p(\hat{\theta}) = \sum_{s \in \Omega} \left(\hat{\theta}(s) - E_p(\hat{\theta}) \right)^2 p(s), \tag{1.2.4}$$

où $\hat{\theta}(s)$ représente la valeur de l'estimateur $\hat{\theta}$ obtenu au moyen de s.

1.3. Le π-estimateur

Proposé par Narain (1951) et Horvitz et Thompson (1952), le π-estimateur (ou estimateur par dilatation) d'un total $Y = \sum_{i \in U} y_i$ est donné par

$$\hat{Y}_\pi \;=\; \sum_{i \in s} w_i y_i, \tag{1.3.1}$$

où $w_i = 1/\pi_i$ désigne le poids de sondage de l'unité i.

Le π-estimateur (1.3.1) est sans biais pour Y sous le plan de sondage si $\pi_i > 0, \forall i \in U$. En effet, notant que $E_p(I_i(s)) = \pi_i$, on a

$$\begin{aligned} E_p(\hat{Y}_\pi) &= E_p\left(\sum_{i\in U} I_i(s) w_i y_i\right) \\ &= \sum_{i\in U} E_p(I_i(s)) w_i y_i \\ &= \sum_{i\in U} y_i. \end{aligned}$$

La variance du π-estimateur est donnée par

$$V_p\left(\hat{Y}_\pi\right) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{1.3.2}$$

Notons que $\pi_{ii} = \pi_i$. Lorsque le plan de sondage est à taille fixe, Yates et Grundy (1953) et Sen (1953) ont montré que la variance (1.3.2) peut également s'écrire comme suit :

$$V_p\left(\hat{Y}_\pi\right) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{1.3.3}$$

La variance de $\hat{Y}_\pi$ donnée par (1.3.2) ou (1.3.3) est inconnue car les valeurs de la variable y ne sont connues que pour l'échantillon s. Il faut donc l'estimer. Un estimateur sans biais de $V_p\left(\hat{Y}_\pi\right)$, noté $v_{HT}(\hat{Y}_\pi)$, est l'estimateur de variance d'Horvitz-Thompson donné par

$$v_{HT}\left(\hat{Y}_\pi\right) = \sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{1.3.4}$$

Si $\pi_{ij} > 0$ pour tout (i, j), alors $E_p\left(v_{HT}\left(\hat{Y}_\pi\right)\right) = V_p\left(\hat{Y}_\pi\right)$.

1.4. Information auxiliaire

L'information auxiliaire est un ensemble de variables permettant d'assister le statisticien d'enquête lors de différentes étapes d'une enquête. On distingue plusieurs types de variables auxiliaires :

– les variables du plan de sondage (en anglais, *design variables*) : elles sont utilisées à l'étape du plan de sondage, et doivent être disponibles pour toutes les unités de la population. Elles peuvent servir par exemple à stratifier la

population ou à construire des plans à probabilités d'inclusion proportionnelles à la taille ;

– les variables de calage : elles sont utilisées à l'étape de l'estimation. Elles doivent être connues pour toutes les unités de l'échantillon seulement mais leur total dans la population doit être connu ;

– les variables utilisées pour le traitement de la non-réponse : elles sont utilisées pour traiter la non-réponse (par repondération ou par imputation). Elles sont minimalement requises pour toutes les unités échantillonnées. Dans ce contexte, les variables auxiliaires serviront à construire les modèles servant à réduire le biais de non-réponse.

1.5. Estimateur par la régression généralisée

L'estimateur par la régression linéaire généralisée du total, noté $\hat{Y}_{GREG}$, est un estimateur utilisant une information auxiliaire à l'étape de l'estimation. On suppose que l'on dispose d'un vecteur $\mathbf{z}$ de q variables auxiliaires, disponible pour toutes les unités dans l'échantillon et que le vecteur des totaux, $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, est connu. L'idée sous-jacente est d'utiliser la relation supposée linéaire entre les variables auxiliaires $\mathbf{z}_i$ et la variable d'intérêt y. On considère ainsi un modèle de la forme

$$\zeta : y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{1.5.1}$$

tel que $E_\zeta(\epsilon_i) = 0$ et

$$Cov_\zeta(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2 c_i, & \text{si} \quad i = j, \\ 0, & \text{si} \quad i \neq j, \end{cases}$$

où $E_\zeta(.)$, $Cov_\zeta(.)$ représentent respectivement l'espérance, la covariance sous le modèle de régression (1.5.1) et c_i un coefficient supposé connu. Le modèle (1.5.1) est fréquemment appelé modèle de superpopulation. Autrement dit, on suppose que la population finie U est une réalisation du modèle de superpopulation (1.5.1). Si l'on observe la variable y pour toutes les unités de la population (cas d'un

recensement), un estimateur de $\boldsymbol{\beta}$, est celui qui minimise

$$\sum_{i\in U}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2/(\sigma^2 c_i), \tag{1.5.2}$$

ce qui mène à l'estimateur des moindres carrés pondérés

$$\begin{aligned}\mathbf{B} &= \left(\sum_{i\in U}\mathbf{x}_i\mathbf{x}_i^\top/c_i\right)^{-1}\left(\sum_{i\in U}\mathbf{x}_i y_i/c_i\right)\\ &= \mathbf{T}^{-1}\mathbf{t},\end{aligned}$$

où $\mathbf{T} = \sum_{i\in U}\mathbf{x}_i\mathbf{x}_i^\top/c_i$ et $\mathbf{t} = \sum_{i\in U}\mathbf{x}_i y_i/c_i$.

Mais le vecteur $\mathbf{B}$ est inconnu puisqu'on ne dispose des valeurs de la variable d'intérêt que pour les unités échantillonnées. On estimera alors $\mathbf{B}$ en estimant chacun des termes séparément, ce qui mène à :

$$\hat{\mathbf{T}}_\pi = \sum_{i\in s} w_i\mathbf{x}_i\mathbf{x}_i^\top/c_i$$

et

$$\hat{\mathbf{t}}_\pi = \sum_{i\in s} w_i\mathbf{x}_i y_i/c_i.$$

Un estimateur de $\mathbf{B}$, noté $\hat{\mathbf{B}}$, est donné par

$$\hat{\mathbf{B}} = \hat{\mathbf{T}}_\pi^{-1}\hat{\mathbf{t}}_\pi.$$

Maintenant, en utilisant la relation (1.5.1), on exprime le total Y comme :

$$\begin{aligned}Y &= \sum_{i\in U}(\mathbf{x}_i^\top\boldsymbol{\beta} + \epsilon_i)\\ &= \sum_{i\in U}\mathbf{x}_i^\top\boldsymbol{\beta} + \sum_{i\in U}\epsilon_i.\end{aligned}$$

L'estimateur GREG de Y est obtenu en estimant les deux termes séparément, et on obtient :

$$\hat{Y}_{GREG} = \sum_{i\in U}\mathbf{x}_i^\top\hat{\mathbf{B}} + \sum_{i\in s} w_i e_i, \tag{1.5.3}$$

où $e_i = y_i - \mathbf{x}_i^\top\hat{\mathbf{B}}$. On retrouve dans la littérature deux autres représentations de l'estimateur GREG :

$$\hat{Y}_{GREG} = \hat{Y}_\pi + \left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)^\top\hat{\mathbf{B}} \tag{1.5.4}$$

et

$$\hat{Y}_{GREG} = \sum_{i \in s} w_i g_i(s) y_i, \tag{1.5.5}$$

où $g_i(s) = 1 + (\mathbf{X} - \hat{\mathbf{X}}_\pi)^\top \hat{\mathbf{T}}_\pi^{-1} \mathbf{x}_i / c_i$.

L'estimateur par la régression généralisée (1.5.3) est asymptotiquement sans biais pour Y sous le plan de sondage. La preuve de ce résultat est obtenue en utilisant un développement de Taylor du premier ordre ; voir, par exemple, Särndal, Swensson et Wretman (1992). On a

$$E_p(\hat{Y}_{GREG}) \approx Y. \tag{1.5.6}$$

La propriété (1.5.6) est satisfaite même dans le cas où le modèle (1.5.1) n'est pas correctement spécifié. Nous ne disposons pas d'une forme simple de la variance de l'estimateur par la régression généralisée. L'utilisation d'un développement de Taylor du premier ordre nous permet d'obtenir une expression de la variance approximative de l'estimateur GREG, notée $AV_p(\hat{Y}_{GREG})$, et donnée par

$$AV_p(\hat{Y}_{GREG}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}, \tag{1.5.7}$$

où $E_i = y_i - \mathbf{x}_i^\top \mathbf{B}$ représente les résidus de la population (en anglais, *census residuals*). Lorsque le modèle (1.5.1) est correctement spécifié, autrement dit lorsque $y_i \approx \mathbf{x}_i^\top \mathbf{B}, \forall i \in U$, la variance de l'estimateur GREG (1.5.7) est petite ; dans le cas contraire, l'estimateur GREG (1.5.3) peut être inefficace. La variance (1.5.7) étant inconnue, on peut l'estimer par

$$v_{HT}(\hat{Y}_{GREG}) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}, \tag{1.5.8}$$

où $e_i = y_i - \mathbf{x}_i^\top \hat{\mathbf{B}}$.

1.6. Échantillonnage à deux phases

Dans certaines situations, on ne dispose pas d'une base de sondage contenant une information auxiliaire utilisable ou de bonne qualité. En l'absence d'information auxiliaire, l'usage de l'échantillonnage à deux phases peut nous permettre

d'en acquérir à moindre coût et permettre ainsi d'obtenir des estimateurs plus précis tels que l'estimateur par la régression généralisée (GREG). L'échantillonnage à deux phases consiste à tirer à la première phase un échantillon s_1 de taille n_1 selon un plan de sondage $p_1(s_1)$. L'information auxiliaire collectée à cette étape sera utilisée au stade de l'estimation à la seconde phase. À la deuxième phase, on tire un échantillon s_2 de taille n_2 selon un plan de sondage $p_2(s_2|s_1)$ de l'échantillon s_1 et on collecte l'information sur la caractéristique d'intérêt y. On désigne par π_{1i} et $\pi_{2i}(s_1)$ respectivement les probabilités d'inclusion du premier ordre de l'unité i à la première et à la deuxième phase. La probabilité π_{1i} est connue pour toute les unités de la population tandis que la probabilité $\pi_{2i}(s_1)$ est connue uniquement pour toutes les unités de l'échantillon s_1.

Dans l'échantillonnage à deux phases, un estimateur sans biais du total est l'estimateur par double dilatation (ou *double expansion*, en anglais) donné par :

$$\hat{Y}_{TP} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\pi_{2i}(s_1)}.$$

Son espérance mathématique est donnée par

$$E_p(\hat{Y}_{TP}) = E_1 E_2(\hat{Y}_{TP}|s_1) = Y,$$

où $E_1(.)$ et $E_2(.|s_1)$ désignent respectivement l'espérance par rapport au plan $p_1(s_1)$ et $p_2(s_2|s_1)$ et sa variance est donnée par

$$\begin{aligned} V_p(\hat{Y}_{TP}) &= V_1 E_2(\hat{Y}_{TP}|s_1) + E_1 V_2(\hat{Y}_{TP}|s_1) \\ &= \sum_{i \in U} \sum_{j \in U} (\pi_{1ij} - \pi_{1i}\pi_{1j}) \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{j1}} \\ &\quad + E_1 \left\{ \sum_{i \in s_1} \sum_{j \in s_1} (\pi_{2ij}(s_1) - \pi_{2i}(s_1)\pi_{2j}(s_1)) \frac{y_i}{\pi_{1i}\pi_{2i}(s_1)} \frac{y_j}{\pi_{1j}\pi_{2j}(s_1)} \right\}, \end{aligned}$$

où π_{1ij} et $\pi_{2ij}(s_1)$ désignent respectivement les probabilités d'inclusion du second ordre à la première et à la deuxième phase et $V_1(.)$ et $V_2(.|s_1)$, désignent respectivement les variances sous les plans $p_1(s_1)$ et $p_2(s_2|s_1)$.

L'échantillonnage à deux phases nous donne une meilleure compréhension de la situation prévalant dans un contexte de non-réponse. En effet, on compare très

fréquemment la situation prévalant en présence de non réponse à celle prévalant dans l'échantillonnage à deux phases. Par analogie, l'échantillon s_1 sélectionné à la première phase d'échantillonnage est vu comme l'échantillon tiré de la population, tandis que le sous-échantillon s_2 prélevé à la deuxième phase, est vu comme l'ensemble de répondants. Dans le chapitre suivant, nous discutons du concept de non-réponse et nous abordons de façon plus générale la question de l'inférence en présence de non-réponse.

Chapitre 2

INFÉRENCE EN PRÉSENCE DE NON-RÉPONSE

2.1. Introduction

Dans la quasi-totalité des enquêtes, on doit faire face à une non-réponse. La non-réponse peut être totale ou partielle. La première est une absence complète d'information sur l'unité échantillonnée et la seconde est une absence d'information sur certaines variables. Si elle n'est pas traitée, la non-réponse peut significativement affecter la qualité des estimateurs en aboutissant à des estimateurs biaisés et peu précis.

Pour traiter la non-réponse, plusieurs approches sont généralement utilisées, parmi lesquelles nous pouvons mentionner l'utilisation des répondants complets, la repondération et l'imputation. L'utilisation des répondants complets consiste à prendre en compte uniquement les répondants pour calculer les estimateurs des quantités d'intérêts. Cependant, cette approche conduit généralement à des estimateurs biaisés, surtout lorsque le taux de réponse est faible. La repondération quant à elle, consiste généralement à ajuster les poids de sondage des répondants pour tenir compte de la suppression des non-répondants. C'est une technique très utilisée en pratique dans les cas de non-réponse totale. Dans ce mémoire, nous mettons l'emphase sur le traitement de la non-réponse partielle par l'imputation. L'imputation est un processus qui produit une valeur de remplacement à une valeur manquante. L'imputation peut être simple ou multiple. L'imputation simple étant la plus utilisée par les statisticiens d'enquête, c'est elle que nous étudierons.

Elle consiste à créer une valeur unique pour remplacer une valeur manquante, ce qui résulte en un fichier complet unique de données. Contrairement à l'imputation simple, l'imputation multiple (Rubin 1987) consiste à créer au moins deux valeurs de remplacement pour la valeur manquante résultant ainsi en au moins deux fichiers de données complets.

Bien qu'elle présente quelques avantages, l'utilisation de l'imputation simple comporte néanmoins plusieurs risques. Parmi ses avantages, on peut citer la création d'un fichier de données complet et l'utilisation des poids de sondage uniques. Quant aux risques, on sait que l'imputation peut aboutir à une inférence invalide si le mécanisme de non-réponse et/ou le modèle d'imputation ne sont pas valides ; elle peut entraîner la distortion des distributions des variables que l'on impute et la distortion de la relation entre les variables. De plus, l'utilisation d'estimateurs standards de la variance peut mener à une sous-estimation importante de la variance de l'estimateur imputé.

L'utilisation de modèles est incontournable dans un contexte de non-réponse ; *a fortiori*, la qualité (biais et variance) des estimateurs obtenus après imputation dépend de la qualité des modèles utilisés. Il s'agit donc de choisir des variables auxiliaires reliées aux variables d'intérêt et/ou à la probabilité de réponse. En présence de non-réponse, une variable z est appelée variable auxiliaire si elle est au moins disponible pour toutes les unités échantillonnées (répondants et non-répondants).

2.2. Terminologie et Définitions.

En présence de non-réponse sur la variable y, l'estimateur (1.3.1) ne peut pas être calculé car toutes les valeurs de la variable y ne sont pas observées. On définit alors l'estimateur imputé suivant :

$$\hat{Y}_I = \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) y_i^*, \tag{2.2.1}$$

où y_i^* est la valeur imputée utilisée pour remplacer la valeur manquante y_i et r_i est la variable indicatrice de réponse égale à 1 si l'unité i a répondu à la variable y et égale à 0 sinon. Pour estimer la moyenne $\bar{Y} = \sum_{i \in U} y_i / N$, on peut simplement

utiliser $\hat{\bar{Y}}_I = \hat{Y}_I/N$ si la taille de la population N est connue ou $\hat{\bar{Y}}_I = \hat{Y}_I/\hat{N}_\pi$, si N est inconnue, où $\hat{N}_\pi = \sum_{i\in s} w_i$ est l'estimateur de la taille de la population N. Désignons par s_r l'ensemble des répondants de taille r et par s_m l'ensemble des non-répondants de taille m. On a $s = s_r \cup s_m$ et $n = r + m$.

On distingue les méthodes d'imputation déterministes des méthodes d'imputation stochastiques. Une méthode d'imputation est dite déterministe losqu'elle produit une valeur fixe étant donné l'échantillon lorsque le processus d'imputation est répété, contrairement aux méthodes dites aléatoires ou stochastiques. Les méthodes d'imputation déterministes ont tendance à distordre la distribution des variables que l'on impute à la difference des méthodes stochastiques qui la préservent. La majorité des méthodes d'imputation peut être représentée par le modèle (Kalton and Kaspryk, 1986) défini ci-dessous :

$$m : y_i = f(\mathbf{z}_i, \boldsymbol{\beta}) + \epsilon_i, \tag{2.2.2}$$

où $E_m(\epsilon_i) = 0$, $E_m(\epsilon_i\epsilon_j) = 0$ pour $i \neq j$, $V_m(\epsilon_i) = \sigma_i^2 = \sigma^2 c_i$, $\mathbf{z}$ est un vecteur de variables auxiliaires disponibles pour toutes les unités échantillonnées et $\boldsymbol{\beta}$ un vecteur de paramètres inconnus et le coefficient c_i une constante supposée connue. Le modèle (2.2.2) est souvent appelé modèle d'imputation.

Dans le cas des méthodes déterministes, la donnée manquante y_i est imputée par $y_i^* = f(\mathbf{z}_i, \hat{\mathbf{B}}_r)$ où $\hat{\mathbf{B}}_r$ est un estimateur de $\boldsymbol{\beta}$ calculé au moyen des unités répondantes.

Les méthodes d'imputation aléatoires peuvent être vues comme des méthodes d'imputation déterministes auxquelles on a ajouté une composante aléatoire e^*. Cette composante aléatoire peut être générée à partir d'une distribution donnée ou bien, comme c'est souvent le cas en pratique, tirée dans l'ensemble des résidus observés correspondant aux unités répondantes. Soit $\tilde{e}_j = e_j - \bar{e}_r$ le résidu standardisé pour le répondant $j \in s_r$, où $e_j = \frac{1}{\hat{\sigma}_j}(y_j - f(\mathbf{z}_j, \hat{\mathbf{B}}_r))$, $\hat{\sigma}_j$ est un estimateur de σ_j et $\bar{e}_r = \sum_{i\in s} w_i r_i e_i / \sum_{i\in s} w_i r_i$. La valeur à imputer est donnée par $y_i^* = f(\mathbf{z}_i, \hat{\mathbf{B}}_r) + \hat{\sigma}_i e_i^*$, où e_i^* est tirée (habituellement avec remise) dans l'ensemble des résidus standardisés.

2.3. Quelques méthodes d'imputation déterministes.

Dans cette section, nous décrivons certaines méthodes d'imputation déterministes fréquemment utilisées en pratique. Il s'agit de l'imputation par la régression linéaire, l'imputation par le ratio, l'imputation par la moyenne, l'imputation par le plus proche voisin (PPV) et l'imputation par la valeur précédente ou historique.

2.3.1. L'imputation par la régression linéaire

Cette méthode d'imputation consiste à remplacer une valeur manquante par la valeur prédite au moyen d'un modèle de régression linéaire. Dans ce cas, $f(\mathbf{z}_i, \boldsymbol{\beta}) = \mathbf{z}_i^\top \boldsymbol{\beta}$ et $c_i = \boldsymbol{\lambda}^\top \mathbf{z}_i$, où $\boldsymbol{\lambda}$ est un vecteur de constantes connu. La valeur imputée est donnée par :

$$y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_r, \qquad i \in s_m, \tag{2.3.1}$$

où $\hat{\mathbf{B}}_r = \left(\sum_{i\in s} w_i r_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)^{-1} \left(\sum_{i\in s} w_i r_i \mathbf{z}_i y_i / c_i\right)$.

2.3.2. L'imputation par le ratio

Cette méthode d'imputation est un cas particulier de l'imputation par la régression. Une seule variable auxiliaire z est utilisée. Dans ce cas, $f(\mathbf{z}_i, \boldsymbol{\beta}) = \beta z_i$ et $c_i = z_i$. Le modèle sous-jacent suppose donc que la variable y est approximativement proportionnelle à la variable z. La valeur imputée est donnée par :

$$y_i^* = \hat{B}_r z_i = \frac{\hat{Y}_r}{\hat{Z}_r} z_i, \qquad i \in s_m, \tag{2.3.2}$$

où $\hat{Y}_r = \sum_{i\in s} w_i r_i y_i$ et $\hat{Z}_r = \sum_{i\in s} w_i r_i z_i$.

2.3.3. L'imputation par la moyenne

L'imputation par la moyenne consiste à remplacer une valeur manquante par la moyenne des répondants. Dans ce cas, $f(\mathbf{z}_i, \boldsymbol{\beta}) = \beta$. C'est un cas particulier de l'imputation par la régression avec $\mathbf{z}_i = 1$ et $c_i = 1$ pour tout i. La valeur imputée est donnée par :

$$y_i^* = \hat{\bar{Y}}_r = \frac{\sum_{i\in s} w_i r_i y_i}{\sum_{i\in s} w_i r_i}, \qquad i \in s_m. \tag{2.3.3}$$

2.3.4. L'imputation par le plus proche voisin

L'imputation PPV consiste à remplacer une valeur manquante par la valeur d'un donneur (choisi parmi les répondants) sur la base d'une fonction de distance notée $dist(.,.)$ et en tenant compte de l'information auxiliaire appropriée. La valeur de remplacement est donnée par :

$$y_i^* = y_j, \qquad i \in s_m, \tag{2.3.4}$$

où $j \in s_r$ et telle que la distance $dist(\mathbf{z}_i, \mathbf{z}_j)$ soit minimale. L'imputation PPV est donc une méthode d'imputation non-paramétrique puisqu'il n'est pas nécessaire de spécifier la forme de $f(\mathbf{z}_i, \boldsymbol{\beta})$, pas plus que la structure de la variance σ_i^2.

2.3.5. L'imputation par la valeur précédente

L'imputation historique consiste à utiliser la valeur observée d'une unité à une occasion précédente pour remplacer la valeur manquante de l'occasion courante. Soit $y_{i,t-1}$ la valeur observée de l'unité i au temps $t-1$. Dans ce cas, on a $f(\mathbf{z}_i, \mathbf{B}) = y_{i,t-1}$. La valeur imputée est donnée par :

$$y_i^* = y_{i,t-1}, \qquad i \in s_m. \tag{2.3.5}$$

2.4. Quelques méthodes aléatoires

Dans cette section, nous décrivons certaines méthodes d'imputation aléatoires couramment utilisées. Il s'agit de l'imputation par la régression aléatoire et l'imputation par *hot-deck* aléatoire.

2.4.1. Imputation par *hot-deck* aléatoire

L'imputation par *hot-deck* aléatoire consiste à tirer au hasard avec remise dans l'ensemble des répondants, s_r, un répondant (donneur) et assigner sa valeur au non-répondant (receveur). La valeur imputée pour remplacer y_i est donnée par

$$y_i^* = y_j, \qquad i \in s_m, \tag{2.4.1}$$

où $j \in s_r$ et tel que $P(y_i^* = y_j) = w_j / \sum_{i \in s_r} w_i$. L'imputation par *hot-deck* aléatoire peut être vue comme l'imputation par la moyenne (2.3.3) à laquelle on a rajouté un résidu aléatoire. Autrement dit, on a $y_i^* = \hat{\bar{Y}}_r + e_i^*$, où $e_i^* = y_i - \hat{\bar{Y}}_r$.

2.4.2. Imputation par la régression aléatoire.

L'imputation par la régression avec résidus est équivalente à l'imputation par la régression (2.3.1) à laquelle on a ajouté un résidu aléatoire tiré (avec remise) dans l'ensemble des résidus standardisés correspondants aux répondants. La valeur imputée est donnée par :

$$y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_r + c_i^{1/2} e_i^*, \tag{2.4.2}$$

où $e_i^* = \tilde{e}_j$, $j \in s_r$ tel que $P(e_i^* = \tilde{e}_j) = w_j / \sum_{i \in s_r} w_i$ et $\tilde{e}_j = c_i^{-1/2}(y_j - \mathbf{z}_j \hat{\mathbf{B}}_r) - \bar{e}_r$, avec $\bar{e}_r = \sum_{i \in s} w_i r_i e_i / \sum_{i \in s} w_i r_i$.

2.5. Mécanisme de non-réponse.

Comme mentionné en section 1.6, la situation prévalant en présence de non-réponse est souvent comparée à celle rencontrée dans l'échantillonnage à deux phases. Dans le contexte de l'échantillonnage à deux phases, le statisticien contrôle le mécanisme (plan de sondage) ayant généré les échantillons de première et de deuxième phase (s_1 et s_2). Autrement dit, il connait les probabilités d'inclusion $\pi_{1i}(s_1)$ et $\pi_{2i}(s_2|s_1)$. Par contre, en présence de non-réponse, le mécanisme aléatoire (appelé mécanisme de non-réponse) ayant généré l'échantillon de répondants s_r est généralement inconnu. Il faudra donc postuler les hypothèses à propos de ce mécanisme. Désignons par $q(s_r|s)$ le mécanisme de non-réponse et par $p_i = P(i \in s_r | s, i \in s)$ la probabilité de réponse de l'unité i à la variable y. On suppose que $p_i > 0, \forall i$ et que $p_{ij} = P(i \in s_r \quad et \quad j \in s_r | s, i \in s, j \in s, i \neq j) = p_i p_j$ c'est-à-dire que les unités répondent indépendamment les unes des autres.

Dans les sous-sections suivantes, nous présentons trois mécanismes de non-réponse souvent rencontrés. Il s'agit des mécanismes de réponse uniforme, ignorable et non-ignorable.

2.5.1. Mécanisme uniforme

La non-réponse est uniforme si la probabilité de réponse p_i est constante pour toutes les unités de la population. Conséquemment, la probabilité de réponse ne dépend ni de la variable d'intérêt ni des variables auxiliaires ($p_i = p, \forall i \in U$). Dans ce cas, on dit que les données sont « Missing Completely At Random »(MCAR).

2.5.2. Mécanisme ignorable et non-ignorable

Un mécanisme de non-réponse ignorable est toujours défini par rapport au modèle d'imputation postulé (Rubin 1976). On dit qu'un mécanisme de non-réponse est ignorable, si après avoir pris en compte l'information auxiliaire appropriée, la probabilité de réponse et l'erreur du modèle d'imputation sont indépendantes. Lorsque le mécanisme de non-réponse est ignorable, on dit que les données sont « Missing At Random ». Dans ce cas, le biais de non-réponse est négligeable si l'on tient compte de l'information auxiliaire appropriée. Par exemple, le mécanisme de non-réponse uniforme est un mécanisme ignorable car la probabilité de réponse est indépendante de l'erreur du modèle d'imputation.

Un mécanisme de non-réponse qui n'est pas ignorable est dit non-ignorable. Par exemple, un mécanisme de non-réponse est non-ignorable si la probabilité de réponse dépend de la variable d'intérêt. Dans ce cas, la probabilité de réponse et le terme d'erreur du modèle d'imputation ne sont pas indépendants. Dans ce cas, les données sont dites « Not Missing At Random »(NMAR). En général, il est impossible de déterminer si l'on se trouve en présence d'un mécanisme de non-réponse ignorable ou non ignorable.

2.6. Cadre de travail pour l'inférence

Deux cadres de travail ont été proposés dans la littérature afin d'étudier les propriétés de l'estimateur imputé (2.2.1). Il s'agit de l'approche basée sur le modèle de non-réponse (NM) étudiée entre autre, par Rao (1990), Rao et Sitter (1995) et Haziza et Rao (2006) et l'approche basée sur le modèle d'imputation (IM) étudiée entre autre, par Särndal (1992) et Deville et Särndal (1994).

Sous l'approche NM, des hypothèses explicites sont postulées à propos du mécanisme de non-réponse. L'inférence est basée sur la distribution conjointe induite par le plan de sondage et le mécanisme de non-réponse. Par exemple, on peut travailler sous l'hypothèse que le mécanisme de non-réponse est uniforme. On parlera de l'approche UNM. Cette approche n'est généralement pas réaliste si elle est considérée globalement sur l'ensemble de la population. En pratique, on suppose très souvent que le mécanisme de non-réponse est uniforme à l'intérieur des classes d'imputation (sous-groupes de la population).

Sous l'approche IM, on suppose que le mécanisme de non-réponse est ignorable et aucune hypothèse supplémentaire n'est faite sur ce mécanisme. Pour mener l'inférence, on postule un modèle d'imputation sur la variable que l'on cherche à imputer. Par exemple, un modèle d'imputation souvent utilisé est l'imputation par la régression linéaire motivé par (2.2.2), avec $f(\mathbf{z}_i, \boldsymbol{\beta}) = \mathbf{z}_i^\top \boldsymbol{\beta}$ et $c_i = \boldsymbol{\lambda}^\top \mathbf{z}_i$. L'inférence est basée sur la distribution conjointe induite par le plan de sondage, le mécanisme de non-réponse et le modèle d'imputation.

2.7. Biais et variance de non-réponse

Avant de définir les concepts de biais et de variance de non-réponse, nous présentons d'abord deux décompositions de l'erreur totale, $\hat{Y}_I - Y$ très souvent utilisées dans le cas des méthodes d'imputation déterministes et des méthodes d'imputation aléatoires.

Dans le cas des méthodes d'imputation déterministes, l'erreur totale $\hat{Y}_I - Y$ peut être décomposée comme suit :

$$\hat{Y}_I - Y = \left(\hat{Y}_\pi - Y\right) + \left(\hat{Y}_I - \hat{Y}_\pi\right), \tag{2.7.1}$$

où $\hat{Y}_\pi$ est le π-estimateur que l'on aurait utilisé en l'absence de non-réponse. On peut remarquer que sous une méthode d'imputation déterministe, l'erreur totale s'exprime donc comme la somme de l'erreur due à l'échantillonnage, $\hat{Y}_\pi - Y$ et de l'erreur due à la non-réponse, $\hat{Y}_I - \hat{Y}_\pi$.

Dans le cas des méthodes d'imputation aléatoires, l'erreur totale $\hat{Y}_I - Y$ est donnée par :

$$\hat{Y}_I - Y = \left(\hat{Y}_\pi - Y\right) + \left(E_I(\hat{Y}_I|s, s_r) - \hat{Y}_\pi\right) + \left(\hat{Y}_I - E_I(\hat{Y}_I|s, s_r)\right), \quad (2.7.2)$$

où $E_I(\hat{Y}_I|s, s_r)$ désigne l'espérance de $\hat{Y}_I$ sous le mécanisme d'imputation dénotée par l'indice I. On peut donc remarquer que sous une méthode d'imputation aléatoire, l'erreur totale s'exprime comme la somme de l'erreur due à l'échantillonnage, $\hat{Y}_\pi - Y$, de l'erreur due à la non-réponse, $E_I(\hat{Y}_I|s, s_r) - \hat{Y}_\pi$ et de l'erreur due à l'imputation, $\hat{Y}_I - E_I(\hat{Y}_I|s, s_r)$.

2.7.1. Biais de non-réponse

Sous l'approche NM, en utilisant la décomposition (2.7.1), le biais de non-réponse dans le cas d'une méthode d'imputation déterministe est donné par :

$$\begin{aligned} Biais\left(\hat{Y}_I\right) &= E_p E_q\left(\hat{Y}_I - Y|s\right) \\ &= E_p E_q\left(\hat{Y}_\pi - Y|s\right) + E_p E_q\left(\hat{Y}_I - \hat{Y}_\pi|s\right) \\ &= E_p B_q(\hat{Y}_I), \end{aligned} \quad (2.7.3)$$

où $B_q(\hat{Y}_I) = E_q(\hat{Y}_I - \hat{Y}_\pi|s)$ est le biais conditionnel de non-réponse et où l'indice q représente l'aléa du mécanisme de non-réponse. La dernière égalité en (2.7.3) découle du fait que

$$E_p E_q\left(\hat{Y}_\pi - Y|s\right) = E_p\left(\hat{Y}_\pi - Y\right) = 0.$$

L'estimateur $\hat{Y}_I$ est dit pq-sans biais (autrement dit, il est sans biais sous la distribution conjointe induite par le plan de sondage et le mécanisme de non-réponse) si $B_q(\hat{Y}_I) = 0$.

Dans le cas d'une méthode d'imputation aléatoire, en utilisant la décomposition

de l'erreur totale donnée en (2.7.2), on obtient le biais de non-réponse suivant :

$$\begin{aligned} Biais\left(\hat{Y}_I\right) &= E\left(\hat{Y}_I - Y\right) \\ &= E_p E_q\left(\hat{Y}_\pi - Y|s\right) + E_p E_q\left(E_I(\hat{Y}_I|s,s_r) - \hat{Y}_\pi|s\right) \\ &\quad + E_p E_q E_I\left(\hat{Y}_I - E_I(\hat{Y}_I|s,s_r)|s,s_r\right) \\ &= E_p E_q E_I\left(\hat{Y}_I - \hat{Y}_\pi|s,s_r\right) \\ &= E_p B_{qI}(\hat{Y}_I), \end{aligned} \tag{2.7.4}$$

où $B_{qI}(\hat{Y}_I) = E_q E_I\left(\hat{Y}_I - \hat{Y}_\pi|s,s_r\right)$ correspond au biais conditionnel de non-réponse, et où l'indice qI désigne conjointement les aléas induits par le mécanisme de non-réponse et par le mécanisme d'imputation. L'estimateur imputé $\hat{Y}_I$ est pqI-sans biais dans le cas d'une méthode d'imputation aléatoire si le biais de non-réponse conditionnel $B_{qI}(\hat{Y}_I) = 0$.

Sous l'approche IM, en utilisant la décomposition de l'erreur présentée en (2.7.1) et en supposant que le plan de sondage est ignorable (c'est-à-dire que les probabilités d'inclusion dans l'échantillon ne dépendent pas de la variable d'intérêt, une fois qu'on a tenu compte des variables auxiliaires dans l'estimation) ; par exemple, voir Pfeffermann (1993), le biais de non-réponse dans le cas d'une méthode d'imputation détermiste est donné par :

$$\begin{aligned} Biais(\hat{Y}_I) &= E_m E_p E_q\left(\hat{Y}_I - Y|s\right) \\ &= E_m E_p\left(\hat{Y}_\pi - Y\right) + E_m E_p E_q\left(\hat{Y}_I - \hat{Y}_\pi|s,s_r\right) \\ &= E_p E_q E_m\left(\hat{Y}_I - \hat{Y}_\pi|s,s_r\right) \\ &= E_p B_{qm}(\hat{Y}_I), \end{aligned}$$

où $B_{qm}(\hat{Y}_I) = E_q E_m\left(\hat{Y}_I - \hat{Y}_\pi|s,s_r\right)$ est le biais conditionnel de non-réponse sous l'approche IM. L'estimateur $\hat{Y}_I$ est mpq-sans biais (c'est-à-dire qu'il est sans biais sous la distribution conjointe induite par le modèle d'imputation, le plan de sondage et le mécanisme de non-réponse) si $B_{qm}(\hat{Y}_I) = 0$. Notons que l'ordre des espérances a pu être changé parce que le mécanisme de non-réponse et le plan de sondage sont supposés ignorable.

En utilisant la décomposition de l'erreur totale (2.7.2) et en supposant le plan de sondage ignorable, le biais de non-réponse dans le cas d'une méthode d'imputation aléatoire sous l'approche IM est donné par :

$$\begin{aligned} Biais\left(\hat{Y}_I\right) &= E_m E_p E_q E_I\left(\hat{Y}_I - Y|s, s_r\right) \\ &= E_m E_p\left(\hat{Y}_\pi - Y\right) + E_m E_p E_q\left(E_I(\hat{Y}_I|s, s_r) - \hat{Y}_\pi|s, s_r\right) \\ &\quad + E_m E_p E_q E_I\left(\hat{Y}_I - E_I(\hat{Y}_I|s, s_r)|s, s_r\right) \\ &= E_p E_q E_m E_I\left(\hat{Y}_I - \hat{Y}_\pi|s, s_r\right) \\ &= E_p B_{qmI}(\hat{Y}_I), \end{aligned}$$

où $B_{qmI}(\hat{Y}_I) = E_q E_m E_I\left(\hat{Y}_I - \hat{Y}_\pi|s, s_r\right)$. L'estimateur imputé $\hat{Y}_I$ est $mpqI$-sans biais dans le cas d'une méthode d'imputation aléatoire si le biais conditionnel de non-réponse $B_{qmI}(\hat{Y}_I) = 0$. Une fois de plus, notons que l'ordre des espérances a pu être changé parce que le mécanisme de non-réponse et le plan de sondage sont supposés ignorable.

2.7.2. Variance de non-réponse

On commence par exprimer la variance totale de $\hat{Y}_I$ sous l'approche NM, dans le cas d'une méthode d'imputation déterministe. On suppose que $B_q\left(\hat{Y}_I\right) = 0$. La variance totale de $\hat{Y}_I$ est donc donnée par

$$\begin{aligned} V\left(\hat{Y}_I\right) &= E\left(\hat{Y}_I - Y\right)^2 \\ &= E_p E_q\left(\hat{Y}_\pi - Y|s\right)^2 + E_p E_q\left(\hat{Y}_I - \hat{Y}_\pi|s\right)^2 + 2E_p E_q\left((\hat{Y}_I - \hat{Y}_\pi)(\hat{Y}_\pi - Y)|s\right). \end{aligned}$$

Puisque $B_q(\hat{Y}_I) = 0$, on a

$$\begin{aligned} E_p E_q\left((\hat{Y}_I - \hat{Y}_\pi)(\hat{Y}_\pi - Y)|s\right) &= E_p\left((\hat{Y}_\pi - Y)E_q(\hat{Y}_I - \hat{Y}_\pi|s)\right) \\ &= 0. \end{aligned}$$

Il s'ensuit que

$$\begin{aligned} V(\hat{Y}_I) &= V_p(\hat{Y}_\pi) + E_p V_q\left(\hat{Y}_I|s\right) \\ &= V^q_{SAM} + V^q_{NR}, \end{aligned} \tag{2.7.5}$$

où $V^q_{SAM} = V_p(\hat{Y}_\pi)$ désigne la variance due à l'échantillonnage et $V^q_{NR} = E_p V_q\left(\hat{Y}_I|s\right)$ désigne la variance due à la non-réponse.
Dans le cas d'une méthode d'imputation aléatoire, la variance totale comporte une variabilité additionnelle appelée variance due à l'imputation due au mécanisme d'imputation. En utilisant (2.7.2), la variance totale de l'estimateur imputé est donnée par :

$$V(\hat{Y}_I - Y) \;=\; V^q_{SAM} + V^q_{NR} + V^q_I, \tag{2.7.6}$$

où $V^q_{NR} = E_p V_q E_I\left(\hat{Y}_I|s\right)$ est la variance due à la non-réponse et la variance due à l'imputation est $V^q_I = E_p E_q V_I\left(\hat{Y}_I|s\right)$.

Maintenant, nous exprimons la variance totale de $\hat{Y}_I$ sous l'approche IM dans le cas d'une méthode d'imputation déterministe. On suppose que $B_{mq}\left(\hat{Y}_I\right) = 0$. La variance totale de $\hat{Y}_I$ est donc donnée par

$$\begin{aligned} V\left(\hat{Y}_I - Y\right) &= E_m E_p E_q\left(\hat{Y}_I - Y|s\right)^2 \\ &= E_q E_m V_p\left(\hat{Y}_\pi|s, s_r\right) + E_q E_p V_m\left(\hat{Y}_I - \hat{Y}_\pi|s, s_r\right) \\ &\quad + 2E_q E_p Cov_m\left(\hat{Y}_\pi - Y, \hat{Y}_I - \hat{Y}_\pi|s, s_r\right) \\ &= V^m_{SAM} + V^m_{NR} + V^m_{MIX}, \end{aligned} \tag{2.7.7}$$

où $V^m_{SAM} = E_m V_p\left(\hat{Y}_\pi|s, s_r\right)$ est la variance due à l'échantillonnage, $V^m_{NR} = E_q E_p V_m\left(\hat{Y}_I - \hat{Y}_\pi|s, s_r\right)$ est la variance due à la non-réponse et V^m_{MIX} est la covariance entre l'erreur d'échantillonnage et l'erreur d'imputation.
Dans le cas d'une méthode d'imputation aléatoire, en utilisant la décomposition (2.7.2), la variance totale de l'estimateur imputé $\hat{Y}_I$ peut s'exprimer comme suit

$$V\left(\hat{Y}_I - Y\right) \;=\; V^m_{SAM} + V^m_{NR} + V^m_{MIX} + V^m_I, \tag{2.7.8}$$

où $V^m_{NR} = E_p E_q V_m E_I\left(\hat{Y}_I - \hat{Y}_\pi|s, s_r\right)$ désigne la variance due à la non-réponse, $V^m_{MIX} = 2E_q E_p Cov_m E_I\left(\hat{Y}_\pi - Y, \hat{Y}_I - \hat{Y}_\pi|s, s_r\right)$ désigne une composante mixte et $V^m_I = E_p E_q E_m V_I\left(\hat{Y}_I|s, s_r\right)$ désigne la variance due à l'imputation.

2.8. Un cas particulier : L'imputation par la régression

L'utilisation des valeurs imputées (2.3.1) dans (2.2.1) nous mène à l'estimateur imputé suivant :

$$\hat{Y}_I = \hat{Y}_r + (\hat{\mathbf{Z}}_\pi - \hat{\mathbf{Z}}_r)^\top \hat{\mathbf{B}}_r, \tag{2.8.1}$$

où $\hat{Y}_r = \sum_{i\in s} w_i r_i y_i$, $\hat{\mathbf{Z}}_\pi = \sum_{i\in s} w_i \mathbf{z}_i$ et $\hat{\mathbf{Z}}_r = \sum_{i\in s} w_i r_i \mathbf{z}_i$.

En particulier, l'estimateur imputé (2.8.1) dans le cas de l'imputation par le ratio est égal à

$$\hat{Y}_I = \frac{\hat{Y}_r}{\hat{Z}_r}\hat{Z}_\pi,$$

alors qu'il est égal à

$$\hat{Y}_I = \hat{N}_\pi \hat{\bar{Y}}_r,$$

dans le cas de l'imputation par la moyenne.

2.8.1. Biais de l'estimateur du total sous l'imputation par la régression

Sous l'approche UNM, l'estimateur imputé (2.8.1) est asymptotiquement pq-sans biais (Rao, 1990). Il est mpq-sans biais sous l'approche IM (Särndal, 1992). Ces résultats sont également valides pour l'imputation par la moyenne (2.3.3) et l'imputation par le ratio (2.3.2) car ces méthodes sont des cas particuliers de l'imputation par la régression.

Cependant, lorsque le mécanisme de non-réponse n'est pas uniforme, Haziza et Rao (2006) ont montré que l'estimateur imputé (2.8.1) est asymptotiquement pq-biaisé et son biais conditionnel de non-réponse est donné par

$$B_q(\hat{Y}_I) \approx \sum_{i\in s} w_i(1-p_i)(y_i - \mathbf{z}_i^\top \hat{\mathbf{B}}_p), \tag{2.8.2}$$

où $\hat{\mathbf{B}}_p = \left(\sum_{i\in s} w_i p_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)^{-1} \left(\sum_{i\in s} w_i p_i \mathbf{z}_i y_i / c_i\right)$. Une preuve détaillée de ce résultat est donnée en annexe A.

Ces derniers ont proposé une méthode d'imputation par la régression modifiée qui consiste à prendre en compte les probabilités de réponse estimées, $\hat{p}_i$. Par

exemple, les probabilités de réponses estimées $\hat{p}_i$, peuvent être obtenues au moyen d'un modèle logistique. La valeur imputée dans ce cas est donnée par :

$$y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*, \qquad i \in s_m, \tag{2.8.3}$$

où

$$\hat{\mathbf{B}}_r^* = \left(\sum_{i \in s} w_i \frac{(1-\hat{p}_i)}{\hat{p}_i} r_i \mathbf{z}_i \mathbf{z}_i^\top / c_i \right)^{-1} \left(\sum_{i \in s} w_i \frac{(1-\hat{p}_i)}{\hat{p}_i} r_i \mathbf{z}_i y_i / c_i \right).$$

L'estimateur imputé par la régression modifiée est obtenu en remplacant (2.8.3) dans (2.2.1). Cet estimateur est doublement robuste en ce sens qu'il est asymptotiquement sans biais, si l'un des deux modèles (modèle de non-réponse ou modèle d'imputation) est bien spécifié. Autrement dit, les valeurs imputées (2.8.3) garantissent une certaine protection si l'un des deux modèles est mal spécifié.

2.8.2. Variance due à la non-réponse dans le cas de l'imputation par la régression

Sous l'approche NM, la variance due à la non-réponse V_{NR}^q de l'estimateur (2.8.1) est obtenue en utilisant un développement de Taylor du premier ordre, ce qui mène à :

$$V_{NR}^q \approx E_p \left(\sum_{i \in s} w_i^2 p_i (1-p_i) \xi_{ip}^2 \right), \tag{2.8.4}$$

où $\xi_{ip} = \left(1 + c_i^{-1} (\hat{\mathbf{Z}}_\pi - \hat{\mathbf{Z}}_p)^\top \hat{\mathbf{T}}_p^{-1} \mathbf{z}_i\right) E_{ip}$, $E_{ip} = y_i - \mathbf{z}_i^\top \hat{\mathbf{B}}_p$, $\hat{\mathbf{Z}}_p = \sum_{i \in s} w_i p_i \mathbf{z}_i$ et $\hat{\mathbf{T}}_p = \sum_{i \in s} w_i p_i \mathbf{z}_i \mathbf{z}_i^\top / c_i$.

Le terme V_{NR}^q en (2.8.4) est petit lorsque les résidus E_{ip} sont petits. La variance due à l'échantillonnage est égale à la variance du π-estimateur que nous avons déjà calculée et qui est donnée par l'expression (1.3.2). En combinant les deux composantes de la variance (2.7.5), on obtient la variance totale :

$$V(\hat{Y}_I) \approx \sum_{i \in U} \sum_{i \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + E_p \left(\sum_{i \in s} w_i^2 p_i (1-p_i) \xi_i^2 \right). \tag{2.8.5}$$

Sous l'approche IM, pour faciliter le calcul des composantes V_{NR}^m et V_{MIX}^m, on peut réécrire l'estimateur $\hat{Y}_I$ en (2.8.1) comme une fonction linéaire de y_i. Nous

avons alors :

$$\begin{aligned}\hat{Y}_I &= \sum_{i\in s} w_i \left[1 + c_i^{-1}(\hat{\mathbf{Z}}_\pi - \hat{\mathbf{Z}}_r)^\top \hat{\mathbf{T}}_r^{-1}\mathbf{z}_i\right] r_i y_i \\ &= \sum_{i\in s} w_i^* r_i y_i,\end{aligned}$$

où $w_i^* = w_i \left[1 + c_i^{-1}(\hat{\mathbf{Z}}_\pi - \hat{\mathbf{Z}}_r)^\top \hat{\mathbf{T}}_r^{-1}\mathbf{z}_i\right]$, $\hat{\mathbf{Z}}_r = \sum_{i\in s} w_i r_i \mathbf{z}_i$ et $\hat{\mathbf{T}}_r = \sum_{i\in s} w_i r_i \mathbf{z}_i \mathbf{z}_i^\top / c_i$. La composante V_{NR}^m se calcule alors comme suit :

$$\begin{aligned}V_{NR}^m &= E_q E_p V_m \left(\hat{Y}_I - \hat{Y}_\pi | s, s_r\right) \\ &= E_q E_p V_m \left(\sum_{i\in s}(w_i^* r_i - w_i) y_i | s, s_r\right) \\ &= \sigma^2 E_q E_p \left(\sum_{i\in s}(w_i^* r_i - w_i)^2 c_i\right) \qquad (2.8.6)\end{aligned}$$

et la composante V_{MIX}^m comme suit :

$$\begin{aligned}V_{MIX}^m &= 2E_q E_p Cov_m(\hat{Y}_\pi - Y, \hat{Y}_I - \hat{Y}_\pi | s, s_r) \\ &= 2E_q E_p Cov_m \left(\sum_{i\in s}(w_i - 1) y_i, \sum_{i\in s}(w_i^* r_i - w_i) y_i | s, s_r\right) \\ &\quad + 2E_q E_p Cov_m \left(\sum_{i\in U|s} y_i, \sum_{i\in s}(w_i^* r_i - w_i) y_i | s, s_r\right) \\ &= 2\sigma^2 E_q E_p \left(\sum_{i\in s}(w_i - 1)(w_i^* r_i - w_i) c_i\right), \qquad (2.8.7)\end{aligned}$$

où la dernière égalité découle de l'hypothèse d'indépendance du modèle (2.2.2). Puisqu'on sait déjà calculer la variance du π-estimateur, il est facile de déduire l'expression de la variance totale donnée par :

$$\begin{aligned}V(\hat{Y}_I) = E_m &\left(\sum_{i\in U}\sum_{i\in U}(\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right) + \sigma^2 E_q E_p \left(\sum_{i\in s}(w_i^* r_i - w_i)^2 c_i\right) \\ &+ 2\sigma^2 E_q E_p \left(\sum_{i\in s}(w_i - 1)(w_i^* r_i - w_i) c_i\right).\end{aligned}$$

Les expressions (2.8.6) et (2.8.7) montrent que les termes V_{NR}^m et V_{MIX}^m sont petits lorsque le modèle d'imputation est hautement prédictif, car dans ce cas, σ^2 est petit.

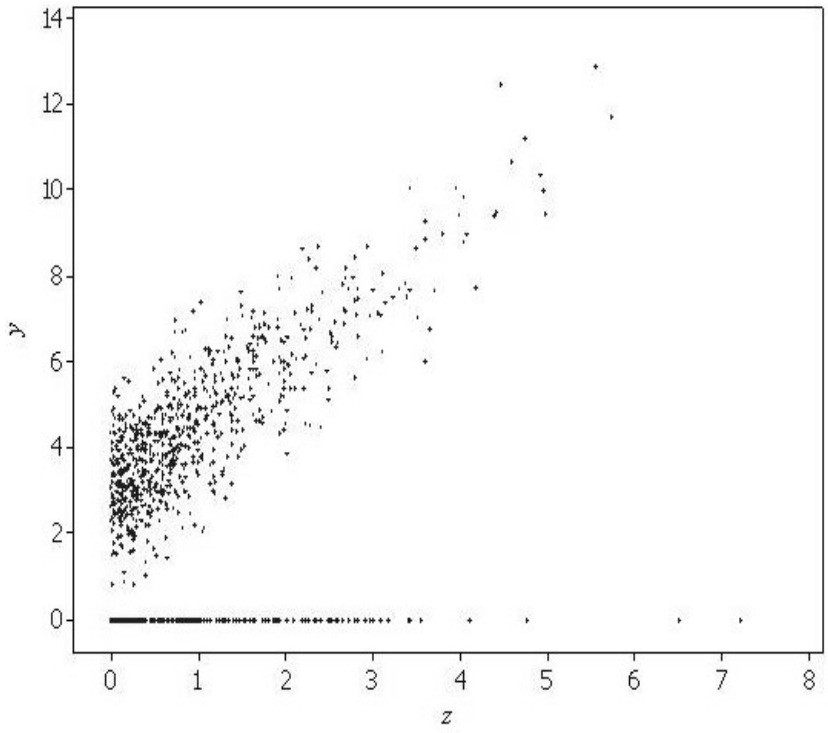
Chapitre 3

ESTIMATION PONCTUELLE POUR DES POPULATIONS CONTENANT DES ZÉROS

3.1. Introduction

Une des manières de minimiser l'impact de la non-réponse sur l'estimation dans les enquêtes est l'utilisation des méthodes d'imputation pour remplacer les valeurs manquantes par des valeurs artificielles obtenues au moyen d'un modèle. On peut choisir de modéliser le mécanisme de non-réponse ou bien de modéliser la variable d'intérêt en se servant d'information auxiliaire de qualité si disponible. Le choix de l'approche utilisée dépend fortement de la nature du problème et de la nature des données collectées. Par exemple, Haziza et Rao (2006) ont décrit le cas de l'enquête sur les dépenses en immobilisation menée à Statistique Canada qui collecte les données sur les investissements qui se font au Canada et dans tous les types d'industries canadiennes. Pour cette enquête, les deux variables principales sont les capitaux immobilisés pour de la nouvelle construction (CC) ainsi que les capitaux immobilisés pour de la nouvelle machinerie et du nouvel équipement (CM). Pour une année donnée, un grand nombre d'entreprises n'a investi aucun montant pour de la nouvelle construction ou du nouvel équipement si bien que le fichier de données avant imputation contient un grand nombre de valeurs égales à zéro. Nous sommes donc en présence de non-réponse et d'un grand nombre de zéros sur les variables d'intérêt. Le graphique 3.1, nous permet de visualiser la situation prévalant dans un tel contexte. La population à l'étude est clairement

FIG. 3.1. Population contenant des zéros



subdivisée en deux sous-populations : la première contient les unités dont les valeurs de la variable d'intérêt sont positives et la seconde contient les unités dont les valeurs de la variable d'intérêt sont égales à zéro. Dans ce mémoire, nous proposons de modéliser la variable d'intérêt en utilisant un modèle de mélange. Nous étudions les propriétés de l'estimateur imputé obtenu dans le cas de l'imputation par la régression.

3.2. MODÈLE D'IMPUTATION

Nous disposons d'une population U finie de taille N composée de deux sous-populations : $U_0 \subset U$, de taille N_0, la population des unités pour lesquelles la valeur de la variable d'intérêt y est égale à zéro, et $U_1 \subset U$, de taille N_1, la population des unités pour lesquelles $y > 0$. On a $U = U_0 \cup U_1$ et $N = N_0 + N_1$. La population U est un mélange de deux populations. Le modèle sous-jacent ayant généré la population U peut donc être décrit comme suit :

$$m : y_i = \delta_i(\mathbf{z}_i^\top \boldsymbol{\beta} + \epsilon_i) + (1 - \delta_i) \times 0, \tag{3.2.1}$$

où $\mathbf{z}_i$ est un vecteur de variables auxiliaires de dimension q et

$$\delta_i = \begin{cases} 1, & \text{si} \quad i \in U_1, \\ 0, & \text{si} \quad i \in U_0. \end{cases}$$

Autrement dit, nous supposons que dans la population U_1, les valeurs de la variable y sont générées à partir d'un modèle de régression linéaire $y_i = \mathbf{z}_i^\top \boldsymbol{\beta} + \epsilon_i$ et que dans la population U_0, les valeurs de y sont toutes égales à zéro. De plus, nous faisons les hypothèses suivantes : $E(\epsilon_i|\delta_i = 1) = 0$, $E(\epsilon_i\epsilon_j|\delta_i = 1, \delta_j = 1, i \neq j) = 0$ et $V(\epsilon_i|\delta_i = 1) = \sigma^2 c_i$. Nous supposons que $c_i = \boldsymbol{\lambda}^\top \mathbf{z}_i$, où $\boldsymbol{\lambda}$ est un vecteur de constantes connu. Désignons par $\phi_i = P(\delta_i = 1)$ la probabilité que l'unité i appartienne à U_1. Il s'ensuit que l'espérance sous le modèle (3.2.1) de la variable d'intérêt y est donné par :

$$\begin{aligned} E_m(y_i) &= E(E(y_i|\delta_i)) \\ &= \phi_i E(\mathbf{z}_i^\top \boldsymbol{\beta} + \epsilon_i|\delta_i = 1) \\ &= \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} \end{aligned} \tag{3.2.2}$$

et sa variance $V_m(y_i)$, est donnée par :

$$\begin{aligned} V_m(y_i) &= E(V(y_i|\delta_i)) + V(E(y_i|\delta_i)) \\ &= \phi_i V(\mathbf{z}_i^\top \boldsymbol{\beta} + \epsilon_i|\delta_i = 1) + \phi_i(1 - \phi_i) E(\mathbf{z}_i^\top \boldsymbol{\beta} + \epsilon_i|\delta_i = 1)^2 \\ &= \sigma^2 \phi_i c_i + \phi_i(1 - \phi_i)(\mathbf{z}_i^\top \boldsymbol{\beta})^2. \end{aligned} \tag{3.2.3}$$

Dans les lemmes 3.2.1 et 3.2.2 présentés ci-dessous, on suppose que la population U de taille N a été générée selon le modèle (3.2.1).

Lemme 3.2.1. *Sous le modèle (3.2.1), le pseudo estimateur des moindres carrés ordinaires* $\mathbf{B}_1 = \left(\sum_{i \in U_1} \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)^{-1} \left(\sum_{i \in U_1} \mathbf{z}_i y_i / c_i\right)$ *obtenu à partir des observations positives uniquement est sans biais pour* $\boldsymbol{\beta}$.

Démonstration.

$$\begin{aligned} E_m(\mathbf{B}_1) &= E\left(E\left(\mathbf{B}_1|\boldsymbol{\delta}\right)\right) \\ &= E(\mathbf{B}_1) \\ &= E\left(\left(\sum_{i\in U_1}\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)^{-1}\left(\sum_{i\in U_1}\mathbf{z}_i y_i/c_i\right)\right) \\ &= \left(\sum_{i\in U_1}\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)^{-1}\left(\sum_{i\in U_1}\mathbf{z}_i\mathbf{z}_i^\top\boldsymbol{\beta}/c_i\right) \\ &= \boldsymbol{\beta}. \end{aligned}$$

□

Lemme 3.2.2. *Sous le modèle (3.2.1), le pseudo estimateur des moindres carrés ordinaires* $\mathbf{B} = \left(\sum_{i\in U}\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)^{-1}\left(\sum_{i\in U}\mathbf{z}_i y_i/c_i\right)$ *obtenu à partir de toutes les observations est biaisé pour* $\boldsymbol{\beta}$ *et le biais est donné par :*

$$B_m(\mathbf{B}) = -\left(\sum_{i\in U}\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)^{-1}\left(\sum_{i\in U}(1-\phi_i)\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)\boldsymbol{\beta}.$$

Démonstration. En utilisant l'équation (3.2.2), on montre facilement que :

$$\begin{aligned} E_m(\mathbf{B}) &= \left(\sum_{i\in U}\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)^{-1}\left(\sum_{i\in U}\phi_i\mathbf{z}_i\mathbf{z}_i^\top\boldsymbol{\beta}/c_i\right) \\ &= \mathbf{T}^{-1}\mathbf{T}_\phi\boldsymbol{\beta}, \end{aligned}$$

où $\mathbf{T} = \sum_{i\in U}\mathbf{z}_i\mathbf{z}_i^\top/c_i$ et $\mathbf{T}_\phi = \sum_{i\in U}\phi_i\mathbf{z}_i\mathbf{z}_i^\top/c_i$. Le biais de $\mathbf{B}$ est égal à :

$$\begin{aligned} B_m(\mathbf{B}) &= \mathbf{T}^{-1}\mathbf{T}_\phi\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= \left(\mathbf{T}^{-1}\mathbf{T}_\phi - \mathbf{I}_q\right)\boldsymbol{\beta} \\ &= \left(\mathbf{T}^{-1}\mathbf{T}_\phi - \mathbf{T}^{-1}\mathbf{T}\right)\boldsymbol{\beta} \\ &= \mathbf{T}^{-1}\left(\mathbf{T}_\phi - \mathbf{T}\right)\boldsymbol{\beta} \\ &= \mathbf{T}^{-1}\left(\sum_{i\in U}(\phi_i - 1)\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)\boldsymbol{\beta} \\ &= -\mathbf{T}^{-1}\left(\sum_{i\in U}(1-\phi_i)\mathbf{z}_i\mathbf{z}_i^\top/c_i\right)\boldsymbol{\beta}, \end{aligned}$$

où $\mathbf{I}_q$ désigne la matrice identité d'ordre q. □

Sous l'hypothèse que les données proviennent du modèle (3.2.1), nous étudions les propriétes de l'estimateur imputé (2.2.1) en termes de biais dans le cas de cinq méthodes d'imputation. Les deux premières méthodes décrites sont motivées par le modèle de régression traditionnel (2.3.1), tandis que les trois autres, celles que nous proposons, sont motivées par le modèle de mélange (3.2.1). Le but ultime est de proposer une méthode d'imputation qui satisfait simultanément les trois critères suivants :

(a) l'estimateur imputé est asymptotiquement sans biais sous les approches UNM et IM.

(b) les valeurs imputées sont réalistes.

(c) l'estimateur imputé est complètement efficace.

Les méthodes d'imputation satisfaisant le critère (a) sont habituellement appelées des méthodes d'imputation doublement robustes : voir, par exemple, Haziza et Rao (2006) et Kim et Haziza (2010). Le critère (b) suppose que, puisque la population comporte un grand nombre de zéros alors les valeurs imputées devraient refléter cette situation. Les valeurs imputées consisteraient ainsi en un mélange de valeurs nulles et de valeurs positives. Finalement, le critère (c) est satisfait lorsque l'estimateur imputé proposé n'est pas affecté par une variabilité additionnelle produite par la sélection aléatoire des valeurs imputées dans le cas d'une méthode d'imputation aléatoire, voir par exemple, Kim et Fuller (2004). Les méthodes d'imputation déterministes sont donc par défaut efficaces.

3.3. Imputation par la moyenne

Par souci de simplicité, nous étudions d'abord le cas de l'imputation par la moyenne. Nous commençons par décrire deux méthodes d'imputation couramment utilisées en pratique, motivées par le modèle (2.2.2) avec $f(\mathbf{z}_i^\top, \boldsymbol{\beta}) = \beta$: la première consiste à calculer les valeurs imputées uniquement au moyen des unités répondantes dont la variable d'intérêt y est positive et la seconde, à calculer les valeurs imputées au moyen de toutes les unités répondantes. Ensuite,

nous proposons trois méthodes d'imputation motivées par le modèle de mélange (3.2.1).

3.3.1. Imputation par la moyenne déterministe positive

Dans cette section, nous présentons l'expression du biais de l'estimateur imputé sous les approches NM et IM, lorsque les valeurs imputées sont calculées uniquement au moyen des unités répondantes positives. Les valeurs imputées y_i^* sont données par :

$$y_i^* = \hat{\bar{Y}}_{r_1} = \frac{\sum_{i\in s_1} w_i r_i y_i}{\sum_{i\in s_1} w_i r_i}, \qquad i \in s_m, \tag{3.3.1}$$

où $s_1 = s \cap U_1$, désigne le sous-ensemble de l'échantillon s qui contient uniquement les unités échantillonnées dont les valeurs de la variable y sont positives. Dans ce cas, l'estimateur imputé (2.2.1) s'écrit comme suit :

$$\hat{Y}_I = \sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i (1-r_i) \hat{\bar{Y}}_{r_1}. \tag{3.3.2}$$

Résultat 3.3.1. *Le biais conditionnel de non-réponse sous l'approche NM, de l'estimateur imputé (3.3.2), obtenu en utilisant les valeurs imputées (3.3.1) peut être approximé par*

$$B_q(\hat{Y}_I) \approx -\sum_{i\in s} w_i (1-p_i)(y_i - \hat{\bar{Y}}_{p_1}), \tag{3.3.3}$$

où $\hat{\bar{Y}}_{p_1} = \dfrac{\sum_{i\in s_1} w_i p_i y_i}{\sum_{i\in s_1} w_i p_i}$.

Remarque 3.3.1. *En général, le biais de non-réponse (3.3.3) est non-nul même sous le mécanisme de non-réponse UNM. En effet, sous UNM, il se réduit à l'expression suivante :*

$$B_q(\hat{Y}_I) \approx (1-p)(\hat{N} - \hat{N}_1)\hat{\bar{Y}}_1, \tag{3.3.4}$$

où $\hat{\bar{Y}}_1 = \sum_{i\in s_1} w_i y_i / \hat{N}_1$, $\hat{N} = \sum_{i\in s} w_i$ *et* $\hat{N}_1 = \sum_{i\in s_1} w_i$.

Le biais conditionnel (3.3.4) est positif et asymptotiquement nul : (i) en absence de non-réponse, auquel cas, $p_i = 1$ pour tout i ; (ii) si toutes les observations

échantillonnées sont positives, auquel cas $\hat{N}_1 = \hat{N}$.

Résultat 3.3.2. *Le biais conditionnel de non-réponse sous l'approche IM, de l'estimateur imputé (3.3.2), obtenu en utilisant les valeurs imputées (3.3.1), est donné par l'expression suivante :*

$$B_{qm}(\hat{Y}_I) \quad = \quad \beta \sum_{i \in s} w_i (1 - \phi_i)(1 - p_i). \tag{3.3.5}$$

Le biais (3.3.5) est nul si (i) $p_i = 1$ (absence de non-réponse) ou si (ii) $\phi_i = 1$ (toutes les observations sont strictement positives).

L'imputation par la moyenne déterministe positive satisfait au critère (c) naturellement, puisque c'est une méthode d'imputation déterministe. Mais elle ne satisfait pas aux critères (a), car elle mène à un estimateur biaisé, simultanément sous les approches UNM et IM. Et parce que les valeurs imputées (3.3.1) sont toujours positives, elles ne sont pas réalistes ; l'imputation par la moyenne déterministe ne satisfait donc pas au critère (b).

3.3.2. Imputation par la moyenne déterministe

Dans cette section, nous étudions le biais de non-réponse de l'estimateur imputé (2.2.1), sous les approches NM et IM, lorsque les valeurs imputées sont calculées au moyen de tous les répondants. Dans ce cas, les valeurs imputées sont données par (2.3.3).

Résultat 3.3.3. *Sous l'approche NM, le biais conditionnel de non-réponse de l'estimateur imputé (2.2.1), calculé en utilisant les valeurs imputées (2.3.3) est non-nul. En outre, le biais obtenu est un cas particulier du biais (2.8.2), lorsque* $\mathbf{z}_i = 1$ *et* $c_i = 1$ *et il est approximativement égal à :*

$$B_q(\hat{Y}_I) \approx \frac{1}{\hat{\bar{P}}} \sum_{i \in s} w_i (p_i - \hat{\bar{P}})(y_i - \hat{\bar{Y}}), \tag{3.3.6}$$

où $\hat{\bar{P}} = \sum_{i \in s} w_i p_i / \hat{N}$ *et* $\hat{\bar{Y}} = \sum_{i \in s} w_i y_i / \hat{N}$.

L'expression (3.3.6) montre que le biais de non-réponse est approximativement nul lorsque la probabilité p_i n'est pas liée à la variable d'intérêt y, ce qui survient dans au moins deux situations : (i) dans le cas d'un mécanisme de non-réponse uniforme auquel cas on a $p_i = p$, (ii) lorsque $y_i = c$, où c est une constante.

Résultat 3.3.4. *Sous l'approche IM, le biais conditionnel de non-réponse de l'estimateur imputé $\hat{Y}_I$, calculé avec les valeurs imputées (2.3.3), est asymptotiquement égal à :*

$$B_{qm}\left(\hat{Y}_I\right) \approx \frac{\beta}{\hat{\bar{P}}} \sum_{i \in s} w_i (p_i - \hat{\bar{P}})(\phi_i - \hat{\bar{\Phi}}), \tag{3.3.7}$$

où $\hat{\bar{P}} = \sum_{i \in s} w_i p_i / \sum_{i \in s} w_i$ *et* $\hat{\bar{\Phi}} = \sum_{i \in s} w_i \phi_i / \sum_{i \in s} w_i$.

Le biais conditionnel de non-réponse (3.3.7) est asymptotiquement nul lorsque la covariance entre p_i et ϕ_i est nulle, ce qui survient, par exemple lorsque $p_i = p$ (mécanisme de non-réponse uniforme) ou $\phi_i = \phi$ (auquel cas la probabilité pour l'unité i d'avoir une valeur de y positive, est constante). De plus, le biais décroit à mesure que $\hat{\bar{P}}$ croit.

L'imputation par la moyenne déterministe mène à un estimateur biaisé sous l'approche IM ; elle ne satisfait donc pas au critère (a). Elle ne satisfait pas non plus au critère (b), puisque les valeurs imputées produites sont toujours positives. Par contre, elle satisfait au critère (c), puisque c'est une méthode d'imputation déterministe.

3.3.3. Imputation par la moyenne déterministe-ϕ

L'imputation par la moyenne détermimiste-ϕ est une méthode d'imputation motivée par le modèle de mélange (3.2.1), avec $\mathbf{z}_i = 1$ et $c_i = 1$. Lorsque nous supposons que les probabilités ϕ_i sont connues, les valeurs imputées y_i^* sont données par :

$$y_i^* = \phi_i \hat{\bar{Y}}_{r_1}, \qquad i \in s_m. \tag{3.3.8}$$

En utilisant les valeurs imputées (3.3.8), dans l'estimateur imputé (2.2.1) on obtient

$$\hat{Y}_I = \sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i (1-r_i)\phi_i \hat{\hat{Y}}_{r_1}. \tag{3.3.9}$$

Dans la présente section, nous évaluons le biais conditionnel de non-réponse de l'estimateur imputé (2.2.1), sous les approches UNM et IM, lorsque les valeurs imputées (3.3.8) sont utilisées.

Résultat 3.3.5. *Le biais de non-réponse sous l'approche UNM, de l'estimateur imputé (3.3.9), obtenu en utilisant les valeurs imputées (3.3.8) est asymptotiquement nul.*

Démonstration. Par un développement de Taylor du premier ordre, on obtient

$$\begin{aligned}
B_q\left(\hat{Y}_I\right) &= E_q\left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i(1-r_i)\phi_i \hat{\hat{Y}}_{r_1} - \sum_{i\in s} w_i y_i\right) \\
&\approx p\sum_{i\in s} w_i y_i + (1-p)\sum_{i\in s} w_i \phi_i \hat{\hat{Y}}_1 - \sum_{i\in s} w_i y_i \\
&= p\hat{N}_1 \hat{\hat{Y}}_1 + (1-p)\sum_{i\in s} w_i \phi_i \hat{\hat{Y}}_1 - \hat{N}_1 \hat{\hat{Y}}_1 \\
&= \hat{\hat{Y}}_1\left(\hat{N}_1 p + (1-p)\sum_{i\in s} w_i\phi_i - \hat{N}_1\right) \\
&= \hat{\hat{Y}}_1 (1-p)\sum_{i\in s} w_i(\phi_i - \delta_i).
\end{aligned}$$

Sous certaines conditions de régularité, on a $\sum_{i\in s} w_i\delta_i - \sum_{i\in s} w_i\phi_i \xrightarrow{p} 0$.

□

Résultat 3.3.6. *Le biais de non-réponse sous l'approche IM, de l'estimateur imputé (3.3.9), obtenu en utilisant les valeurs imputées (3.3.8) est nul.*

Démonstration. Le biais de l'estimateur imputé (3.3.9) est donné par

$$B_{qm}(\hat{Y}_I) = E_q E_m\left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i(1-r_i)\phi_i \frac{\sum_{i\in s_1} w_i r_i y_i}{\sum_{i\in s_1} w_i r_i} - \sum_{i\in s} w_i y_i\right).$$

En utilisant l'égalité (3.2.2) et le lemme (3.2.1), on obtient

$$\begin{aligned} B_{qm}(\hat{Y}_I) &= E_q\left(\sum_{i\in s} w_i r_i \phi_i \beta + \sum_{i\in s} w_i (1-r_i)\phi_i\beta - \sum_{i\in s} w_i \phi_i \beta | s\right) \\ &= 0. \end{aligned}$$

□

L'imputation par la moyenne déterministe-ϕ satisfait au critère (a), car elle mène à un estimateur imputé asymptotiquement sans biais simultanément sous les approches UNM et IM. Elle satisfait aussi au critère (c), puisque c'est une méthode d'imputation déterministe. Cependant, elle ne satisfait pas au critère (b), car l'ensemble des valeurs imputées qu'elle produit, n'est pas un mélange de valeurs nulles et de valeurs positives.

3.3.4. Imputation par la moyenne aléatoire-ϕ

Dans la présente section, nous présentons une autre méthode d'imputation motivée par le modèle de mélange (3.2.1) avec $\mathbf{z}_i = 1$ et $c_i = 1$. En supposant que les probabilités ϕ_i sont connues, les valeurs imputées, y_i^* sont données par

$$y_i^* = \begin{cases} \hat{\bar{Y}}_{r_1}, & \text{avec probabilité } \phi_i, \\ 0, & \text{avec probabilité } 1-\phi_i\ . \end{cases} \tag{3.3.10}$$

En notant que $E_I(y_i^*) = \phi_i \hat{\bar{Y}}_{r_1}$, on remarque que l'imputation par la moyenne aléatoire-ϕ a comme contrepartie déterministe, l'imputation par la moyenne déterministe-ϕ. Les valeurs manquantes sont imputées par (3.3.10) en générant selon une loi de Bernoulli de paramètre ϕ_i, une variable indicatrice qui vaut 1 lorsqu'on doit imputer par $\hat{\bar{Y}}_{r_1}$ et qui vaut 0, lorsqu'on doit imputer par 0.

Résultat 3.3.7. *Le biais de non-réponse sous l'approche UNM, de l'estimateur imputé (2.2.1), calculé en utilisant les valeurs imputées (3.3.10), est asymptotiquement nul.*

Démonstration. En notant que $E_I(\hat{Y}_I|s, s_r)$ correspond à l'estimateur (3.3.9), on déduit facilement que le biais de non-réponse conditionnel sous l'approche

UNM, de l'estimateur imputé (2.2.1) est asymptotiquement nul, comme dans le cas de l'imputation par la moyenne déterministe-ϕ. □

Résultat 3.3.8. *Le biais de non-réponse conditionnel, sous l'approche IM, de l'estimateur imputé $\hat{Y}_I$, calculé en utilisant les valeurs imputées (3.3.10) est nul.*

Démonstration. En notant une fois de plus que $E_I(\hat{Y}_I|s, s_r)$ correspond à l'estimateur (3.3.9), on déduit par analogie à l'imputation par la moyenne déterministe-ϕ, que l'imputation par la moyenne aléatoire-ϕ mène à un estimateur imputé sans biais sous l'approche IM. □

L'imputation par la moyenne aléatoire-ϕ conduit donc à un estimateur asymptotiquement pqI-sans biais sous l'approche UNM et $mpqI$-sans biais sous l'approche IM, elle satisfait donc au critère (a). De plus, elle satisfait au critère (b) contrairement à sa contrepartie déterministe, car les valeurs imputées qu'elle produit consistent en un mélange de valeurs nulles et de valeurs positives. Cependant, elle ne satisfait pas au critère (c), car elle est sujette à une variabilité additionnelle due à la sélection aléatoire des valeurs imputées. En effet, en notant que

$$V_I(y_i^*) = \phi_i(1-\phi_i)\hat{\bar{Y}}_{r_1}^2,$$

la variance due à l'imputation de $\hat{Y}_I$ est donnée par

$$E_* V_I(\hat{Y}_I|\mathbf{r}) = E_*\left(\sum_{i\in s} w_i^2\phi_i(1-\phi_i)\hat{\bar{Y}}_{r_1}^2|\mathbf{r}\right), \tag{3.3.11}$$

où $E_*(.)$ correspond à $E_pE_q(.)$ sous l'approche NM et à $E_mE_qE_p(.)$ sous l'approche IM. Dans certains cas, la variance (3.3.11) peut être importante et ainsi conduire à un estimateur imputé inefficace. Dans la prochaine section, nous proposons une méthode d'imputation qui consiste à choisir aléatoirement les valeurs imputées de manière à éliminer la variance due à l'imputation (3.3.11).

3.3.5. Imputation par la moyenne équilibrée aléatoire-ϕ

L'imputation par la moyenne équilibrée aléatoire-ϕ a l'avantage de produire un jeu de données plus crédible pour l'utilisateur des microdonnées, dans le sens

qu'elle produit un jeu de données après imputation comportant une certaine proportion de zéros comme dans la population à l'étude. Néanmoins, comme toutes les méthodes d'imputation aléatoire standards, elle conduit à un estimateur imputé dont la variance totale comporte une variabilité additionnelle (variance due à l'imputation) en comparaison à sa contrepartie déterministe.

De la décomposition de l'erreur totale (2.7.2), on peut facilement voir que la variance due à l'imputation dans le cas des méthodes d'imputation aléatoire est nulle si l'erreur due à l'imputation est nulle. Dans le cas de l'estimateur imputé $\hat{Y}_I$, calculé en utilisant les valeurs imputées (3.3.10), ceci revient à poser

$$\begin{aligned} \hat{Y}_I - E_I(\hat{Y}_I|s, s_r) &= \sum_{i\in s} w_i(1-r_i)y_i^* - \sum_{i\in s} w_i(1-r_i)\phi_i\hat{\bar{Y}}_{r_1} \\ &= 0. \end{aligned} \tag{3.3.12}$$

L'imputation par la moyenne équilibrée aléatoire-ϕ consiste donc à tirer les valeurs imputées y_i^* de façon aléatoire comme en (3.3.10), tout en respectant la contrainte d'équilibrage (3.3.12). Une fois que les valeurs imputées sont sélectionnées comme décrit précédemment, l'imputation par la moyenne équilibrée aléatoire-ϕ satisfait simultanément aux critères (a)-(c). En section 3.4.5, nous décrivons un algorithme permettant de sélectionner les valeurs imputées afin d'éliminer la variance due à l'imputation dans le cas de l'imputation par la régression.

3.4. Imputation par la régression linéaire

Nous généralisons dans cette section, les résultats obtenus dans le cas de l'imputation par la moyenne au cas de l'imputation par la régression.

3.4.1. Imputation par la régression déterministe positive

Dans cette section, nous présentons le biais de non-réponse conditionnel sous les approches NM et IM, de l'estimateur imputé (2.2.1), lorsque les valeurs imputées sont calculées uniquement au moyen des répondants positifs. Dans ce cas, les valeurs imputées y_i^* sont données par :

$$y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}, \qquad i \in s_m, \tag{3.4.1}$$

où $\hat{\mathbf{B}}_{r_1} = \left(\sum_{i\in s_1} w_i r_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)^{-1} \left(\sum_{i\in s_1} w_i r_i \mathbf{z}_i y_i / c_i\right)$. En utilisant ces valeurs imputées dans l'estimateur (2.2.1), on obtient

$$\hat{Y}_I = \sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i (1-r_i) \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}. \tag{3.4.2}$$

Résultat 3.4.1. *Le biais conditionnel de non-réponse sous l'approche NM, de l'estimateur imputé (3.4.2), obtenu en utilisant les valeurs imputées (3.4.1) peut être approximé par :*

$$B_q\left(\hat{Y}_I\right) \approx -\sum_{i\in s} w_i (1-p_i)(y_i - \mathbf{z}_i^\top \hat{\mathbf{B}}_{p1}), \tag{3.4.3}$$

où $\hat{\mathbf{B}}_{p_1} = \left(\sum_{i\in s_1} w_i p_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)^{-1} \left(\sum_{i\in s_1} w_i p_i \mathbf{z}_i y_i / c_i\right)$.

Démonstration. En appliquant un raisonnement similaire à celui du résultat (2.8.2), on obtient par analogie :

$$\begin{aligned} B_q\left(\hat{Y}_I\right) &= E_q\left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i (1-r_i) \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} - \sum_{i\in s} w_i y_i | s\right) \\ &\approx \sum_{i\in s} \left(w_i p_i y_i + w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{p_1} - w_i p_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{p_1} - w_i y_i\right) \\ &= -\sum_{i\in s} w_i (1-p_i)(y_i - \mathbf{z}_i^\top \hat{\mathbf{B}}_{p_1}). \end{aligned}$$

□

Remarque 3.4.1. *En général, le biais de non-réponse donné par (3.4.3) n'est pas nul. Dans le cas d'un mécanisme de non-réponse uniforme, il se réduit à :*

$$B_q(\hat{Y}_I) = (1-p)(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_1)^\top \hat{\mathbf{B}}_1, \tag{3.4.4}$$

où $\hat{\mathbf{B}}_1 = \left(\sum_{i\in s_1} w_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)^{-1} \left(\sum_{i\in s_1} w_i \mathbf{z}_i y_i / c_i\right)$, $\hat{\mathbf{Z}} = \sum_{i\in s} w_i \mathbf{z}_i$ *et* $\hat{\mathbf{Z}}_1 = \sum_{i\in s_1} w_i \mathbf{z}_i$. *En effet, lorsque* $p_i = p$ *on a :* $\hat{\mathbf{B}}_{p_1} = \hat{\mathbf{B}}_1$. *De plus, en notant que*

$$\begin{aligned} \boldsymbol{\lambda}^\top \left(\sum_{i\in s_1} w_i \mathbf{z}_i \mathbf{z}_i^\top / (\boldsymbol{\lambda}^\top \mathbf{z}_i)\right) \hat{\mathbf{B}}_1 &= \sum_{i\in s_1} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 \\ &= \sum_{i\in s_1} w_i y_i, \end{aligned}$$

on déduit que

$$\begin{aligned} B_q\left(\hat{Y}_I\right) &= -(1-p)\left[\sum_{i\in s} w_i y_i - \sum_{i\in s} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1\right] \\ &= -(1-p)\left[\sum_{i\in s_1} w_i y_i - \sum_{i\in s} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1\right] \\ &= -(1-p)\left[\sum_{i\in s_1} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 - \sum_{i\in s} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1\right] \\ &= -(1-p)\left(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}\right)^\top \hat{\mathbf{B}}_1. \end{aligned}$$

Lorsque $\mathbf{z}_i = 1$ *et* $c_i = 1$, *l'expression (3.4.4) se ramène au cas de la moyenne dont l'expression est donnée en (3.3.4).*

Résultat 3.4.2. *Le biais conditionnel de non-réponse sous l'approche IM, de l'estimateur imputé (3.4.2), obtenu en utilisant les valeurs imputées (3.4.1) est donné par l'expression suivante :*

$$B_{qm}(\hat{Y}_I) = \sum_{i\in s} w_i (1-p_i)(1-\phi_i) \mathbf{z}_i^\top \boldsymbol{\beta}. \tag{3.4.5}$$

Démonstration.

$$B_{qm}\left(\hat{Y}_I\right) = E_q E_m \left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i (1-r_i) \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} - \sum_{i\in s} w_i y_i | s, s_r\right).$$

Du lemme (3.2.1), on a que $E_m(\hat{\mathbf{B}}_{r_1}|s) = \boldsymbol{\beta}$ et on déduit finalement le biais,

$$\begin{aligned} B_{qm}\left(\hat{Y}_I\right) &= E_q\left(\sum_{i\in s} w_i r_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} + \sum_{i\in s} w_i(1-r_i)\mathbf{z}_i^\top \boldsymbol{\beta} - \sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta}|s\right) \\ &= E_q\left(\sum_{i\in s}(w_i r_i \phi_i + w_i - w_i r_i - w_i \phi_i)\mathbf{z}_i^\top \boldsymbol{\beta}|s\right) \\ &= E_q\left(\sum_{i\in s} w_i(1-r_i)(1-\phi_i)\mathbf{z}_i^\top \boldsymbol{\beta}|s\right) \\ &= \sum_{i\in s} w_i(1-p_i)(1-\phi_i)\mathbf{z}_i^\top \boldsymbol{\beta}. \end{aligned}$$

□

Le biais conditionnel (3.4.5) est non-nul en général. Il est nul dans les situations extrèmes où tous les p_i sont égaux à l'unité et/ou tous les ϕ_i sont égaux à l'unité. Dans le cas où $\mathbf{z}_i = 1$ et $c_i = 1$, l'expression (3.4.5) se réduit au cas de la moyenne dont l'expression est donnée en (3.3.5).

L'imputation par la régression déterministe positive ne satisfait pas au critère (a), car elle mène à un estimateur biaisé sous les approches UNM et IM ; ni au critère (b), car les valeurs imputées produites sont toujours positives. Néanmoins, elle satisfait au critère (c), car c'est une méthode d'imputation déterministe.

3.4.2. Imputation par la régression déterministe

Dans la présente section, nous présentons le biais de l'estimateur imputé (2.2.1), lorsque les valeurs imputées sont calculées au moyen de toutes les unités répondantes sous approches NM et IM. Dans ce cas, les valeurs imputées sont données par l'équation (2.3.1).

Résultat 3.4.3. *Sous l'approche NM, le biais de non-réponse conditionnel de l'estimateur imputé (2.2.1) calculé avec les valeurs imputées obtenues en tenant compte de toutes les unités répondantes, est identique à celui donné en (2.8.2).*

Résultat 3.4.4. *En ignorant les termes d'ordre supérieur dans le développement de Taylor du premier ordre, le biais conditionnel de non-réponse sous l'approche IM, de l'estimateur imputé obtenu en utilisant toutes les unités répondantes dans le calcul des valeurs imputées est donné par*

$$B_{qm}(\hat{Y}_I) \approx \boldsymbol{\lambda}^\top \hat{\mathbf{T}}_p^{-1} \hat{\mathbf{T}}_\pi \left(\hat{\mathbf{T}}_{p\phi} - \hat{\mathbf{T}}_\pi^{-1} \hat{\mathbf{T}}_p \hat{\mathbf{T}}_\phi \right) \boldsymbol{\beta}, \tag{3.4.6}$$

où $\hat{\mathbf{T}}_p = \left(\sum_{i\in s} w_i p_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)$, $\hat{\mathbf{T}}_\phi = \left(\sum_{i\in s} w_i \phi_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)$ *et*
$\hat{\mathbf{T}}_{p\phi} = \left(\sum_{i\in s} w_i p_i \phi_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)$.

Démonstration.

$$\begin{aligned}
B_{qm}\left(\hat{Y}_I\right) &= E_q E_m \left(\hat{Y}_I - \hat{Y} | s, s_r\right) \\
&= E_q E_m \left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i (1-r_i) \mathbf{z}_i^\top \hat{\mathbf{B}}_r - \sum_{i\in s} w_i y_i | s, s_r \right) \\
&= E_q \left(\sum_{i\in s} w_i r_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} + \sum_{i\in s} w_i (1-r_i) \mathbf{z}_i^\top E_m \left(\hat{\mathbf{B}}_r | s, s_r\right) - \sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | s \right).
\end{aligned}$$

En effet, du lemme 2 on a

$$\begin{aligned}
E_m \left(\hat{\mathbf{B}}_r | s, s_r\right) &= E_m \left(\hat{\mathbf{T}}_r^{-1} \hat{\mathbf{t}}_r | s, s_r\right) \\
&= \hat{\mathbf{T}}_r^{-1} E_m \left(\hat{\mathbf{t}}_r | s, s_r\right) \\
&= \hat{\mathbf{T}}_r^{-1} \hat{\mathbf{T}}_{r\phi} \boldsymbol{\beta},
\end{aligned}$$

où $\hat{\mathbf{T}}_{r\phi} = \sum_{i\in s} w_i r_i \phi_i \mathbf{z}_i^\top \mathbf{z}_i / c_i$.

Finalement, en utilisant un développement de Taylor du premier ordre, on a

$$\begin{aligned}
B_{qm}(\hat{Y}_I) &= E_q \left(\sum_{i\in s} w_i r_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} + \sum_{i\in s} w_i (1-r_i) \mathbf{z}_i^\top \hat{\mathbf{T}}_r^{-1} \hat{\mathbf{T}}_{r\phi} \boldsymbol{\beta} - \sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | s \right) \\
&= E_q \left(\boldsymbol{\lambda}^\top \hat{\mathbf{T}}_{r\phi} \boldsymbol{\beta} + \boldsymbol{\lambda}^\top \hat{\mathbf{T}}_\pi \hat{\mathbf{T}}_r^{-1} \hat{\mathbf{T}}_{r\phi} \boldsymbol{\beta} - \boldsymbol{\lambda}^\top \hat{\mathbf{T}}_r \hat{\mathbf{T}}_r^{-1} \hat{\mathbf{T}}_{r\phi} \boldsymbol{\beta} - \boldsymbol{\lambda}^\top \hat{\mathbf{T}}_\phi \boldsymbol{\beta} | s \right) \\
&= E_q \left(\boldsymbol{\lambda}^\top \left[\hat{\mathbf{T}}_\pi \hat{\mathbf{T}}_r^{-1} \hat{\mathbf{T}}_{r\phi} - \hat{\mathbf{T}}_\phi \right] \boldsymbol{\beta} | s \right) \\
&\approx \boldsymbol{\lambda}^\top \left[\hat{\mathbf{T}}_\pi \hat{\mathbf{T}}_p^{-1} \hat{\mathbf{T}}_{p\phi} - \hat{\mathbf{T}}_\phi \right] \boldsymbol{\beta} \\
&= \boldsymbol{\lambda}^\top \hat{\mathbf{T}}_\pi \hat{\mathbf{T}}_p^{-1} \left[\hat{\mathbf{T}}_{p\phi} - \hat{\mathbf{T}}_p \hat{\mathbf{T}}_\pi^{-1} \hat{\mathbf{T}}_\phi \right] \boldsymbol{\beta}.
\end{aligned}$$

□

Remarque 3.4.2. *Si* $\mathbf{z}_i = 1$ *et* $c_i = 1$*, le biais conditionnel (3.4.6) se réduit au biais de non-réponse (3.3.7) de l'imputation par la moyenne. Comme dans le cas de la moyenne, lorsque les* ϕ_i *sont constants* ($\phi_i = \phi$) *et/ou* p_i *sont constants, le biais conditionnel de non-réponse (3.4.6) est nul.*

L'imputation par la régression déterministe ne satisfait pas au critère (a) car elle mène à un estimateur biaisé sous l'approche IM. Les valeurs imputées produites par l'imputation par la régression déterministe ne sont pas réalistes dans le sens qu'elles ne sont jamais nulles, donc le critère (b) n'est pas satisfait. Étant donné que nous sommes en présence d'une méthode d'imputation déterministe, le critère (c) est donc satisfait.

3.4.3. Imputation par la régression déterministe-ϕ

L'imputation par la régression déterministe-ϕ est une méthode d'imputation motivée par le modèle de mélange (3.2.1). Lorsque les probabilités ϕ_i sont supposées connues, les valeurs imputées y_i^* sont données par

$$y_i^* = \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}, \qquad i \in s_m. \tag{3.4.7}$$

En remplaçant (3.4.7) dans l'estimateur imputé (2.2.1), on obtient

$$\hat{Y}_I = \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}. \tag{3.4.8}$$

Dans la présente section, nous évaluons le biais conditionnel de non-réponse de l'estimateur imputé (2.2.1), sous les approches UNM et IM, lorsque les valeurs imputées (3.4.7) sont utilisées.

Résultat 3.4.5. *Le biais conditionnel de non-réponse sous l'approche UNM, de l'estimateur imputé* $\hat{Y}_I$ *(3.4.8) obtenu en utilisant les valeurs imputées (3.4.7) est asymptotiquement nul.*

Démonstration. Le biais conditionnel de non-réponse de l'estimateur imputé $\hat{Y}_I$ est donné par

$$B_q\left(\hat{Y}_I\right) = E_q\left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i(1-r_i)\phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} - \sum_{i\in s} w_i y_i | s\right).$$

Par un développement de Taylor du premier ordre, on obtient

$$B_q\left(\hat{Y}_I\right) \approx p\sum_{i\in s} w_i y_i + (1-p)\sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 - \sum_{i\in s} w_i y_i.$$

En considérant les égalités suivantes

$$\boldsymbol{\lambda}^\top\left(\sum_{i\in s} w_i \delta_i \mathbf{z}_i \mathbf{z}_i^\top/(\boldsymbol{\lambda}^\top \mathbf{z}_i)\right)\hat{\mathbf{B}}_1 = \sum_{i\in s} w_i \delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 = \sum_{i\in s} w_i y_i,$$

on a

$$\begin{aligned} B_q\left(\hat{Y}_I\right) &\approx p\sum_{i\in s} w_i \delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 + (1-p)\sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 - \sum_{i\in s} w_i \delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 \\ &= \left((1-p)\sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top - (1-p)\sum_{i\in s} w_i \delta_i \mathbf{z}_i^\top\right)\hat{\mathbf{B}}_1 \\ &= (1-p)\sum_{i\in s} w_i(\phi_i - \delta_i)\mathbf{z}_i^\top \hat{\mathbf{B}}_1. \end{aligned}$$

Sous certaines conditions de régularité, nous avons $\sum_{i\in s} w_i \delta_i \mathbf{z}_i^\top - \sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \xrightarrow{p} \mathbf{0}$.

□

Résultat 3.4.6. *Le biais conditionnel de non-réponse sous l'approche IM, de l'estimateur imputé $\hat{Y}_I$ (3.4.8) obtenu en utilisant les valeurs imputées (3.4.7) est nul.*

Démonstration.

$$\begin{aligned} B_{qm}\left(\hat{Y}_I\right) &= E_q E_m \left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i(1-r_i)y_i^* - \sum_{i\in s} w_i y_i | s, s_r\right) \\ &= E_q E_m \left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i(1-r_i)\phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} - \sum_{i\in s} w_i y_i | s, s_r\right) \\ &= E_q \left(\sum_{i\in s} w_i r_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | s\right) + E_q E_m \left(\sum_{i\in s} w_i(1-r_i)\delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} | s, s_r\right) \\ &\quad - E_q \left(\sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | s\right). \end{aligned}$$

En utilisant le résultat du lemme (3.2.1), on obtient facilement que

$$E_m \left(\sum_{i\in s} w_i(1-r_i)\phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} | s, s_r\right) = \sum_{i\in s} w_i(1-r_i)\phi_i \mathbf{z}_i^\top \boldsymbol{\beta}.$$

Finalement, le biais est donné par

$$\begin{aligned} B_{qm}\left(\hat{Y}_I\right) &= E_q \left(\sum_{i\in s} w_i r_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} + \sum_{i\in s} w_i(1-r_i)\phi_i \mathbf{z}_i^\top \boldsymbol{\beta} - \sum_{i\in s} w_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | s\right) \\ &= 0. \end{aligned}$$

□

Comme dans le cas de l'imputation par la moyenne déterministe-ϕ, l'imputation par la régression déterministe-ϕ mène à un estimateur approximativement sans biais sous les approches UNM et IM. Elle satisfait donc au critère (a). Elle satisfait aussi au critère (c) étant donné que c'est une méthode d'imputation déterministe. Par contre, elle ne satisfait pas au critère (b), car l'ensemble des valeurs imputées produites n'est pas un mélange de valeurs nulles et de valeurs positives.

3.4.4. Imputation par la régression aléatoire-ϕ

Nous présentons dans la présente section, une méthode d'imputation aléatoire par la régression motivée par le modèle (3.2.1). Nous supposons que les valeurs de ϕ_i sont connues et nous imputons les valeurs manquantes par

$$y_i^* = \begin{cases} \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}, & \text{avec probabilité } \phi_i, \\ 0, & \text{avec probabilité } 1-\phi_i, \end{cases} \tag{3.4.9}$$

pour $i \in s_m$ et où $\hat{\mathbf{B}}_{r_1} = \left(\sum_{i\in s_1} w_i r_i \mathbf{z}_i \mathbf{z}_i^\top / c_i\right)^{-1} \left(\sum_{i\in s_1} w_i \mathbf{z}_i^\top y_i / c_i\right)$.
En notant que $E_I(y_i^*) = \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}$, on peut visualiser l'imputation par la régression déterministe-ϕ comme la contrepartie déterministe de l'imputation par la régression aléatoire-ϕ. Les valeurs manquantes sont imputées par (3.4.9), en générant selon une loi de Bernoulli de paramètre ϕ_i une variable qui vaut 1 lorsqu'on doit imputer par $\mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}$ et qui vaut 0, lorsqu'on doit imputer par 0.

Résultat 3.4.7. *Le biais conditionnel de non-réponse sous l'approche UNM, de l'estimateur imputé $\hat{Y}_I$ calculé en utilisant les valeurs imputées (3.4.9) est asymptotiquement nul.*

Démonstration. En notant que $E_I(\hat{Y}_I|s, s_r)$ correspond à l'estimateur déterministe (3.4.8), on conclu par analogie que le biais de non-réponse sous l'imputation par la régression aléatoire-ϕ est approximativement nul sous l'approche UNM. □

Résultat 3.4.8. *Sous l'approche IM, l'estimateur imputé $\hat{Y}_I$ obtenu par la méthode d'imputation par la régression aléatoire, est sans biais lorsque les données proviennent du modèle de mélange (3.2.1).*

Démonstration. En notant une fois de plus que $E_I(\hat{Y}_I|s, s_r)$ correspond à l'estimateur déterministe (3.4.8), on conclut que le biais de non-réponse sous l'imputation par régression aléatoire-ϕ est nul sous l'approche IM. □

L'imputation par la régression aléatoire-ϕ mène à un estimateur asymptotiquement *pqI*- sans biais sous l'approche UNM et *mpqI*-sans biais sous l'approche IM. Donc, elle satisfait au critère (a). De plus, elle satisfait aussi au critère (b), car les valeurs imputées consiste en un mélange de valeurs nulles et de valeurs positives. Par contre, elle ne satisfait pas au critère (c), car elle est sujette à une variabilité additionnelle due à la sélection aléatoire des valeurs imputées. En effet, en notant que

$$V_I(y_i^*) = \phi_i(1-\phi)\mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1},$$

la variance due à l'imputation de $\hat{Y}_I$ est donnée par

$$E_* V_I(\hat{Y}_I) = E_* \left(\sum_{i \in s} w_i^2 \phi_i (1 - \phi_i) (\mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1})^2 \right), \tag{3.4.10}$$

où $E_*(.)$ correspond à $E_p E_q(.)$ sous l'approche NM et à $E_m E_q E_p(.)$ sous l'approche IM. Dans certains cas, la variance (3.4.10) peut être importante et ainsi conduire à un estimateur imputé inefficace. Dans la prochaine section, nous proposons une méthode d'imputation qui consiste à sélectionner les valeurs imputées (3.4.9) de manière à éliminer la variance due à l'imputation (3.4.10).

3.4.5. Imputation par la régression équilibrée aléatoire-ϕ

La méthode d'imputation aléatoire décrite en section 3.4.4, a l'avantage de produire un jeu de données plus crédible pour l'utilisateur des microdonnées, dans le sens qu'elle produit un jeu de données après imputation comportant une certaine proportion de zéros comme dans la population à l'étude. Néanmoins, comme toutes les méthodes d'imputation aléatoire, elle produit un estimateur dont la variance totale comporte une variabilité additionnelle (variance due à l'imputation) en comparaison à sa contrepartie déterministe.

De la décomposition de l'erreur totale donnée en (2.7.2), on peut facilement voir que la variance due à l'imputation dans le cas des méthodes d'imputation aléatoire est nulle si l'erreur due à l'imputation est nulle. Dans le cas de l'estimateur imputé, obtenu par la méthode d'imputation par la régression aléatoire-ϕ, dont les valeurs imputées sont données par (3.4.9), ceci est équivalent à poser

$$\begin{aligned} \hat{Y}_I - E_I(\hat{Y}_I | s, s_r) &= \sum_{i \in s} w_i (1 - r_i) y_i^* - \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} \\ &= 0. \end{aligned} \tag{3.4.11}$$

La méthode d'imputation par la régression équilibrée aléatoire-ϕ, consiste essentiellement à sélectionner les valeurs imputées comme en (3.4.9), tout en respectant la contrainte (3.4.11). Pour ce faire, nous utilisons l'algorithme présenté dans Chauvet, Deville et Haziza (2010). On sélectionne dans s_m un sous-échantillon

s_*, avec des probabilités d'inclusion ϕ_i, en équilibrant sur la variable

$$x_i = w_i \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}.$$

Pour un individu $i \in s_m$, on prend alors $y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}$ si $i \in s_*$, et $y_i^* = 0$ sinon. Les probabilités individuelles de tirage pour l'imputation sont bien respectées, et l'équation d'équilibrage s'écrit :

$$\begin{array}{rlcl} & \sum_{i \in s_*} \frac{x_i}{\phi_i} & = & \sum_{i \in s_m} x_i \\ \Leftrightarrow & \sum_{i \in s_*} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} & = & \sum_{i \in s_m} w_i \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} \\ \Leftrightarrow & \sum_{i \in s_*} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} & = & \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}. \end{array} \tag{3.4.12}$$

Notons maintenant que

$$\begin{aligned} \sum_{i \in s} w_i (1 - r_i) y_i^* &= \sum_{i \in s_m} w_i y_i^* \\ &= \sum_{i \in s_*} w_i y_i^* + \sum_{i \in s_m \setminus s_*} w_i y_i^* \\ &= \sum_{i \in s_*} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}, \end{aligned} \tag{3.4.13}$$

où la dernière égalité découle du fait que $y_i^* = 0$, $\forall i \in s_m \setminus s_*$. En utilisant les équations (3.4.12) et (3.4.13), on obtient donc

$$\sum_{i \in s} w_i (1 - r_i) y_i^* = \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1},$$

et en injectant cette dernière expression dans (2.2.1), on a

$$\begin{array}{rcl} \hat{Y}_I & = & \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} \\ & = & E_I \left(\hat{Y}_I | s, s_r \right). \end{array}$$

Cette méthode permet donc bien d'annuler la variance due à l'imputation.

L'imputation par la régression équilibrée aléatoire-ϕ satisfait simultanément aux critères (a)-(c). Elle produit un estimateur imputé approximativement sans biais sous les approches UNM et IM. Les valeurs imputées produites sont réalistes car elles consistent en un mélange de valeurs nulles et de valeurs positives. Finalement, la variance due à l'imputation est nulle.

3.5. Estimation des ϕ_i

En pratique, les probabilités ϕ_i sont inconnues et doivent donc être estimées. Nous supposons que

$$\phi_i = f(\boldsymbol{\mu}_i; \boldsymbol{\alpha}),$$

pour une certaine fonction $f(.)$ et $\boldsymbol{\mu}_i$ est un vecteur de variables auxiliaires de l'unité i, disponible pour toutes les unités échantillonnées. Les méthodes paramétriques telle que la régression logistique requièrent que $f(.)$ soit spécifiée. Une alternative aux méthodes paramétriques, est l'utilisation des méthodes non paramétriques qui ne nécessitent pas une spécification de $f(.)$, par exemple la méthode par noyau et la méthode par régression par polynômes locaux (voir, par exemple, Wand et Jones, 1995), et qui de ce fait, sont plus robustes à une mauvaise spécification du modèle. Un choix approprié de variables auxiliaires à inclure dans le modèle est primordial afin de prédire correctement la variable δ. Si un modèle paramétrique est utilisé, un estimateur de ϕ_i est donné par $\hat{\phi}_i = f(\boldsymbol{\mu}_i; \hat{\boldsymbol{\alpha}})$, où $\hat{\boldsymbol{\alpha}}$ est un estimateur convergent de $\boldsymbol{\alpha}$.

Chapitre 4

ÉTUDE PAR SIMULATIONS

4.1. Introduction

Dans le présent chapitre, nous effectuons une étude par simulations afin de comparer la performance des méthodes d'imputation décrites au chapitre 3 en termes de biais et d'erreur quadratique moyenne. Toutes les simulations présentées dans ce chapitre ont été effectuées à l'aide de la version 9.1 du logiciel SAS, sous le système d'exploitation Windows XP. En section 4.2, nous décrivons la façon dont nous avons généré les populations ainsi que les mécanismes utilisés pour générer les zéros et la non-réponse à la variable d'intérêt. Finalement, nous présentons les résultats de l'étude par simulation en section 4.3.

4.2. Description de l'étude par simulations et des populations

Nous avons généré trois populations finies de taille $N = 1000$, chacune constituée de deux variables : une variable d'intérêt y et une variable auxiliaire z. Cette dernière a été générée selon une loi gamma de paramètre de position $\alpha = 4$ et de paramètre d'échelle $\beta = 25$. Les valeurs de y ont été générées à partir du modèle

$$y = 2z + \epsilon, \tag{4.2.1}$$

où les erreurs ϵ_i ont été générées à partir d'une loi normale de moyenne 0 et de variance σ^2 choisie de manière à obtenir un coefficient de détermination R^2 entre y et z d'environ 0.36 pour la population 1, d'environ 0.5 pour la population 2 et

d'environ 0.7 pour la population 3. Dans chacune de ces trois populations, nous avons généré les zéros à la variable d'intérêt y selon trois mécanismes aléatoires distincts menant à des proportions de zéros d'environ 25% et 50%. Les trois mécanismes sont décrits ci-dessous :

(1) mécanisme-ϕ 1 : les zéros sont répartis uniformément dans toute la population. Par exemple, dans le cas où l'on a une proportion de zéros égale à 25%, on pose $\phi_i = 0.75$ pour l'unité i.

(2) mécanisme-ϕ 2 : la probabilité ϕ_i pour l'unité i est calculée à partir d'un modèle logistique

$$log\left(\frac{\phi_i}{1-\phi_i}\right) = \lambda_o + \lambda_1 z_i \tag{4.2.2}$$

où les paramètres λ_0 et λ_1 sont choisis, de sorte, qu'en moyenne la probabilité des ϕ_i soit de 0.5 et de 0.75.

(3) mécanisme-ϕ 3 : la probabilité ϕ_i est à nouveau calculée à partir d'un modèle logistique similaire à celui du mécanisme-ϕ 2, à la différence que les probabilités ϕ_i dépendent cette fois-ci de la variable y au lieu de z.

Une fois la probabilité ϕ_i obtenue, nous avons généré une variable δ_i selon une loi de Bernoulli de paramètre ϕ_i. Lorsque $\delta_i = 0$, nous avons posé $y_i = 0$ et lorsque $\delta_i = 1$, nous avons conservé la valeur de y_i générée à partir du modèle (4.2.1). A ce stade, nous disposons de 18 populations générées selon le modèle 3.2.1 avec $\mathbf{z}_i = z_i$ et $c_i = z_i$.

Dans chaque population, nous avons tiré $R = 10000$ échantillons de taille $n = 200$ selon un plan d'échantillonnage aléatoire simple sans remise. La non-réponse à la variable y a été ensuite générée dans tous les échantillons selon les trois mécanismes aléatoires décrits ci-dessous :

(1) mécanisme-p 1 : la probabilité de réponse p_i est constante et égale à 0.7 pour toutes les unités de la population ;

(2) mécanisme-p 2 : la probabilité de réponse p_i de l'unité i est calculée à partir du modèle logistique $log\left(\frac{p_i}{1-p_i}\right) = \lambda_0 + \lambda_1 z_i$ en choisissant les valeurs des coefficients λ_0 et λ_1 de sorte que le taux de non-réponse moyen soit de 0.7 ;

(3) mécanisme-p 3 : Ce mécanisme est similaire à celui du mécanisme-p 2 à la différence que les probabilités de réponse p_i, dépendent cette fois-ci de la variable y au lieu de la variable z.

Une fois la probabilité p_i obtenue, nous avons généré la variable indicatrice de réponse r_i selon une loi de Bernoulli de paramètre p_i. Lorsque $r_i = 0$, nous avons généré une valeur manquante à la variable y pour l'unité i. Nous étions intéressé à estimer la moyenne de la population $\bar{Y} = \sum_{i \in U} y_i/N$, dans le cas des cinq méthodes d'imputation par le ratio suivantes :

(a) l'imputation par le ratio déterministe positif (RDP) dont les valeurs valeurs imputées sont données par $y_i^* = z_i \dfrac{\sum_{i \in s_1} w_i r_i y_i}{\sum_{i \in s_1} w_i r_i z_i}$;

(b) l'imputation par le ratio déterministe (RD) dont les valeurs imputées sont données par $y_i^* = z_i \dfrac{\sum_{i \in s} w_i r_i y_i}{\sum_{i \in s} w_i r_i z_i}$;

(c) l'imputation par le ratio déterministe-ϕ (RD-ϕ) dont les valeurs imputées sont données par $y_i^* = \hat{\phi}_i z_i \dfrac{\sum_{i \in s_1} w_i r_i y_i}{\sum_{i \in s_1} w_i r_i z_i}$;

(d) l'imputation par le ratio aléatoire-ϕ (RAL-ϕ) dont les valeurs imputées sont données par

$$y_i^* = \begin{cases} z_i \dfrac{\sum_{i \in s_1} w_i r_i y_i}{\sum_{i \in s_1} w_i r_i z_i}, & \text{avec probabilité } \hat{\phi}_i, \\ 0, & \text{avec probabilité } 1 - \hat{\phi}_i; \end{cases}$$

(e) l'imputation par le ratio équilibrée aléatoire-ϕ (REA-ϕ), dont les valeurs imputées sont choisies aléatoirement comme sous RAL-ϕ, en plus de satisfaire à l'équation d'équilibrage donnée en (3.4.11).

Les probabilités ϕ_i sous RAL-ϕ, RD-ϕ et REA-ϕ ont été estimées à l'aide d'un modèle logistique de la forme suivante :

$$\phi_i = exp(\boldsymbol{\mu}_i^\top \boldsymbol{\alpha})/(1 + exp(\boldsymbol{\mu}_i^\top \boldsymbol{\alpha})), \tag{4.2.3}$$

où $\boldsymbol{\mu}_i$ désigne un vecteur de variables auxiliaires et $\boldsymbol{\alpha}$ un vecteur de paramètres. Nous avons utilisé le modèle (4.2.3) pour estimer ϕ_i avec les paramètres $\boldsymbol{\mu}_i = 1$ et

$\boldsymbol{\alpha} = \alpha_0$ pour le mécanisme-ϕ 1, $\boldsymbol{\mu}_i = (1, z_i)^\top$ et $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$ pour le mécanisme-ϕ 2 et pour le mécanisme-ϕ 3.

Pour chacune des méthodes d'imputation, nous avons calculé l'estimateur imputé de la moyenne de la manière suivante :

$$\hat{\bar{Y}}_I = \frac{1}{n}\left(\sum_{i\in s} r_i y_i + \sum_{i\in s}(1-r_i)y_i^*\right). \tag{4.2.4}$$

Afin de mesurer le biais de $\hat{\bar{Y}}_I$, nous avons utilisée le biais relatif Monte Carlo (en %) donné par

$$RB_{MC}(\hat{\bar{Y}}_I) = \frac{1}{R}\sum_{r}^{R}\frac{(\hat{\bar{Y}}_{I(r)} - \bar{Y})}{\bar{Y}} * 100\% \tag{4.2.5}$$

et pour mesurer la variabilité de $\hat{\bar{Y}}_I$, nous avons utilisé l'erreur quadratique moyenne Monte Carlo donnée par

$$EQM_{MC}(\hat{\bar{Y}}_I) = \frac{1}{R}\sum_{r}^{R}(\hat{\bar{Y}}_{I(r)} - \bar{Y})^2, \tag{4.2.6}$$

où $\hat{\bar{Y}}_{I(r)}$ est l'estimateur imputé de $\bar{Y}$ calculé dans le r-ième échantillon.

Dans la première partie des simulations, nous avons évalué la performance des méthodes d'imputation RDP, RD et RAL-ϕ en termes de biais relatif et d'efficacité relative définie comme

$$RE = \frac{EQM_{MC}\left(\hat{\bar{Y}}_I^{(.)}\right)}{EQM_{MC}\left(\hat{\bar{Y}}_I^{(RD)}\right)}, \tag{4.2.7}$$

où $\hat{\bar{Y}}_I^{(RD)}$ désigne l'estimateur imputé sous RD et (.) désigne soit RDP ou RAL-ϕ.

Dans la seconde partie, nous avons évalué la performance des méthodes d'imputation proposées RD-ϕ, RAL-ϕ et REA-ϕ en termes de biais relatif et d'efficacité relative également, mais cette fois, nous nous sommes restreint à un sous-ensemble des scénarios décrits précédemment. Nous avons considéré les populations comportant une proportion de zéros de 50% avec un coefficient de détermination entre y et z de 0.5 et pour lesquelles les mécanisme-ϕ 1 et mécanisme-ϕ 2 ont été utilisés. Désignons par $\hat{\bar{Y}}_I^{(REA-\phi)}$ et $\hat{\bar{Y}}_I^{(RD-\phi)}$ les estimateurs obtenus sous

les méthodes d'imputation REA-ϕ et RD-ϕ respectivement. Nous avons comparé l'efficacité relative de ces estimateurs par rapport à $\hat{\bar{Y}}_I^{(RAL-\phi)}$ choisi comme référence et en utilisant RE donné par (4.2.7).

4.3. Résultats des simulations

Les résultats de l'étude par simulations reliés aux méthodes d'imputation RDP, RD et RAL-ϕ sont présentés dans les tableaux 4.1-4.3 et ceux des méthodes d'imputation RD-ϕ, RAL-ϕ et REA-ϕ sont présentés dans le tableau 4.4.

Sous l'imputation RDP, nous avons remarqué que l'estimateur imputé est lourdement biaisé. Nous avons aussi remarqué que le biais relatif a augmenté à mesure que la proportion de zéros augmentait dans la population. Par exemple, du Tableau 4.3, nous avons observé que le biais relatif est presque toujours supérieur à 5% et pouvait aller au delà de 20%. Du même tableau, nous avons observé que le biais relatif (RE) est passé de 9.76% (3.02) à une proportion de zéros de 25% pour atteindre 27.05% (8.69) à une proportion de zéros 50% dans la population.

Sous l'imputation RD, nous avons observé que l'estimateur imputé est approximativement sans biais lorsque les probabilités ϕ_i et p_i ne sont pas reliées : ce qui survient lorsqu'au moins l'un des mécanismes aléatoires (mécanisme-p ou mécanisme-ϕ) est uniforme. C'est le cas sous le mécanisme-p 1 et sous le mécanisme-ϕ 1. Il est apparu des tableaux 4.1-4.3 que le biais relatif dans cette situation était inférieur à 1% dans la quasi totalité des scénarios. *A contrario*, lorsque les probabilités ϕ_i et p_i dépendaient simultanément de z, donc étaient reliées (mécanime-ϕ 2 et mécanisme-p 2) alors l'estimateur imputé sous l'imputation RD était biaisé. Nous avions aussi observé, que cet estimateur est biaisé lorsque le mécanisme de non-réponse était non-ignorable (mécanisme-p 3), ce qui aurait été le cas même si la population ne contenait pas de zéros. Néanmoins, sous le mécanisme-p 3, le biais relatif a eu tendance à diminuer à mesure que le coefficient de détermination augmentait. Ce qui n'est pas surprenant, car le modèle d'imputation s'enrichissait à mesure que R^2 augmentait. Par ailleurs, nous

avons observé que le biais a augmenté à mesure que la proportion de zéros augmentait dans la population : par exemple du tableau 4.2 sous le mécanisme-p 2 et sous mécanisme-ϕ 2, que le biais relatif est passé de 3.76% à 7.53% lorsque la proportion de zéros est passée de 25% à 50% respectivement.

Sous l'imputation RAL-ϕ, nous avons observé que l'estimateur imputé est approximativement sans biais sous le mécanisme-ϕ 1 (à l'exception du mécanisme-ϕ 3, comme on pouvait s'y attendre). Sous le mécanisme-ϕ 2, l'estimateur imputé a montré un biais négligeable sous les mécanisme-p 1 et mécanisme-p 2 simultanément, contrairement à l'imputation RD qui a mené à un estimateur biaisé sous le mécanisme-p 2. Par exemple, du tableau 4.3, nous avons observé un biais relatif de 0.80% comparativement à 7.70% pour l'imputation RD lorsque la proportion de zéros dans la population était de 50%. Sous le mécanisme-p 3, l'estimateur imputé a présenté un biais simultanément sous l'imputation RD et sous l'imputation RAL-ϕ. Néanmoins, le biais sous RAL-ϕ était significativement plus petit que le biais obtenu sous RD. Par exemple, pour une proportion de zéros de 50%, nous avons observé du tableau 4.3 que RD affichait un biais relatif de 10.28% tandis que celui présenté par RAL-ϕ était de 2.93%. Donc, le modèle d'imputation (3.2.1) motivant RAL-ϕ semble mieux décrire le vrai modèle de la population que le modèle de régression traditionnel, motivant l'imputation RD. Sous le mécanisme-ϕ 3, nous obtenions des résultats similaires à ceux obtenus sous le mécanisme-ϕ 2, à l'exception du cas où $R^2 = 0.36$, l'estimateur imputé présentait un léger biais.

Nous passons maintenant à la comparaison des RE. Sous l'imputation RDP, les valeurs de RE étaient toutes supérieures à l'unité ; ce qui peut être expliqué par la portion importante que prend le biais dans le calcul des EQM. Nous avons observé en outre que lorsque les estimateurs imputés obtenus sous RD et sous RAL-ϕ sont approximativement sans biais, les valeurs de RE prises par RAL-ϕ, sont toutes supérieures à l'unité, montrant ainsi que l'imputation RD est plus efficace que RAL en termes de EQM. Ce qui peut être en partie expliqué par le fait qu'à la différence de l'imputation RD, l'imputation RAL-ϕ comporte une

composante de variance additionnelle : la variance due à l'imputation causée par la sélection aléatoire des valeurs imputées.

Finalement, nous discutons des résultats présentés dans le Tableau 4.4. Les estimateurs obtenus sous les méthodes d'imputation RAL-ϕ, RD-ϕ et REA-ϕ ont montré un biais approximativement nul dans tous les scénarios considérés à l'exception du cas où le mécanisme de non-réponse était non-ignorable (mécanisme-p 3) comme on s'y attendait. En termes d'efficacité relative (RE), les estimateurs imputés sous RD-ϕ et sous REA-ϕ ont des RE toujours inférieures à un. Donc les méthodes d'imputation REA-ϕ et RD-ϕ sont plus efficaces que RAL-ϕ ; ce qui n'est pas surprenant car l' estimateur imputé obtenu sous RD-ϕ et sous RAL-ϕ n'a pas de composante de variance due à l'imputation. Nous avions aussi remarqué que même lorsque les estimateurs étaient biaisés (mécanisme-p 3), on observait un gain en termes de précision sous l'imputation REA-ϕ. Par exemple, sous l'imputation REA-ϕ, RE était inférieure à l'unité même lorsque le mécanisme de non-réponse était non-ignorable (mécanisme-p 3) et son biais relatif était presque égal à celui de l'imputation RAL-ϕ.

TAB. 4.1. Biais relatif Monte Carlo et efficacité relative pour $R^2 = 0.36$

			RDP		RD		RAL-ϕ	
mécanisme-ϕ	proportion de 0	mécanisme-p_i	RB (en %)	RE	RB (en %)	RE	RB (en %)	RE
1	25%	1	9.99	2.56	-0.05	1	-0.02	1.11
		2	6.20	1.62	-0.93	1	-1.19	1.04
		3	24.02	2.23	14.96	1	14.88	1
	50%	1	27.23	7.18	0.03	1	0.12	1.14
		2	19.15	4.32	-0.06	1	-0.74	1.03
		3	39.86	5.04	14.79	1	14.12	0.97
2	25%	1	5.58	1.55	0.10	1	0.11	1.04
		2	5.73	1.48	2.45	1	0.30	0.91
		3	20.37	1.51	16.01	1	13.69	0.78
	50%	1	18.99	4.19	-0.08	1	0.00	1.01
		2	19.86	2.95	7.87	1	0.85	0.58
		3	33.14	2.88	17.40	1	10.20	0.52
3	25%	1	7.30	2.18	-0.11	1	-0.09	1.09
		2	3.89	1.33	-0.22	1	-2.89	1.29
		3	21.33	1.22	19.11	1	18.71	0.97
	50%	1	20.34	6.41	0.04	1	0.11	1.13
		2	15.39	3.50	3.93	1	-1.97	0.87
		3	32.85	2.01	22.14	1	18.12	0.73

TAB. 4.2. Biais relatif Monte Carlo et efficacité relative pour $R^2 = 0.5$

mécanisme-ϕ	proportion de 0	mécanisme-p	RDP RB (en %)	RDP RE	RD RB (en %)	RD RE	RAL-ϕ RB (en %)	RAL-ϕ RE
1	25%	1	9.9	2.69	-0.01	1	0.00	1.11
		2	6.58	1.79	-0.38	1	-0.65	1.02
		3	19.08	2.37	11.01	1	10.88	0.99
	50%	1	31.34	8.83	0.04	1	0.05	1.16
		2	21.54	5.2	-0.47	1	0.35	1.03
		3	37.52	6.54	10.5	1	11.15	1.09
2	25%	1	6.13	1.73	-0.02	1	0.01	1.03
		2	7.23	1.61	3.76	1	0.79	0.79
		3	17.36	1.58	13.07	1	10.04	0.69
	50%	1	16.57	3.98	-0.23	1	-0.17	1.02
		2	17.8	2.8	7.53	1	0.89	0.58
		3	26.86	2.62	14.7	1	7.91	0.49
3	25%	1	6.87	2.06	-0.06	1	-0.05	1.08
		2	4.89	1.54	1.19	1	-1.72	1.10
		3	18.12	1.28	15.74	1	14.80	0.91
	50%	1	19.99	6.12	-0.09	1	-0.01	1.12
		2	16.50	3.34	5.37	1	-0.63	0.72
		3	30.11	2.13	19.50	1	14.81	0.66

TAB. 4.3. Biais relatif Monte Carlo et efficacité relative pour $R^2 = 0.7$

mécanisme-ϕ	proportion de 0	mécanisme-p	RDP RB (en %)	RDP RE	RD RB (en %)	RD RE	RAL-ϕ RB (en %)	RAL-ϕ RE
1	25%	1	9.76	3.02	-0.03	1	-0.09	1.13
		2	7.34	2.22	0.43	1	0.06	1.03
		3	12.35	2.81	5.01	1	4.60	0.96
	50%	1	27.05	8.69	-0.11	1	-0.09	1.18
		2	19.82	5.30	0.56	1	-0.23	1.01
		3	25.67	6.41	4.98	1	4.16	0.97
2	25%	1	4.80	1.63	-0.03	1	-0.01	1.05
		2	6.44	1.57	3.76	1	1.16	0.75
		3	10.26	1.59	7.28	1	4.32	0.61
	50%	1	18.15	4.98	-0.11	1	-0.06	1.04
		2	18.98	3.17	7.70	1	0.80	0.54
		3	22.36	3.11	10.28	1	2.93	0.46
3	25%	1	5.81	1.92	0.07	1	0.08	1.06
		2	6.14	1.61	3.35	1	-0.40	0.79
		3	12.30	1.37	10.06	1	7.84	0.71
	50%	1	16.94	5.18	-0.05	1	-0.03	1.06
		2	16.99	2.77	7.93	1	-0.14	0.48
		3	23.14	2.17	14.41	1	7.15	0.43

TAB. 4.4. Biais relatif et efficacité relative des méthodes d'imputation proposées

mécanisme ϕ	proportion 0	mécanisme p	RAL-ϕ RB(en %)	RAL-ϕ RE	REA-ϕ RB(en %)	REA-ϕ RE	RD-ϕ RB(en %)	RD-ϕ RE
1	50%	1	0,05	1	0,08	0,87	0,08	0,87
		2	0,35	1	0,28	0,91	0,29	0,91
		3	11,15	1	11,16	0,95	11,18	0,95
2	50%	1	-0,17	1	-0,15	0,95	-0,15	0,94
		2	0,89	1	0,89	0,96	0,89	0.95
		3	7,91	1	7,89	0,97	7,88	0,96

Chapitre 5

ESTIMATION DE LA VARIANCE

5.1. Introduction

Dans ce chapitre, nous considérons le problème de l'estimation de la variance en présence de données imputées. Les méthodes de rééchantillonnage, comme par exemple, le *jackknife* trouvent leur justification dans le cadre de travail renversé comme nous le verrons à la section 5.4. La structure de ce chapitre est la suivante : en section 5.2, nous présentons quelques points justifiants pourquoi il est important d'estimer la variance. En section 5.3, nous présentons le cadre de travail renversé habituellement utilisé pour l'estimation de la variance. En section 5.4, nous présentons la technique du *jackknife*, en section 5.5, nous expliquons comment estimer la variance dans le cas de l'estimateur par le ratio en utilisant la technique du *jackknife* et en section 5.6, nous menons une étude par simulation pour évaluer la performance de l'estimateur de *jackknife* proposé.

5.2. Estimation de la variance

Lorsque l'estimateur est complexe ou qu'il est construit à partir d'un plan de sondage complexe, obtenir une forme explicite de la variance peut s'avérer être une tâche fastidueuse. Que nous disposions d'une forme explicite pour la variance ou pas, elle est généralement impossible à évaluer directement et il faudra donc l'estimer. Comme mentionné dans Gagnon, Lee, Provost, Rancourt et Särndal (1997) l'estimation de la variance permet

(1) de fournir une mesure de qualité des estimateurs ;

(2) d'aider à tirer des conclusions correctes ;

(3) aux agences statistiques d'informer les utilisateurs de la qualité des données.

5.3. Cadre de travail renversé

Le cadre renversé a été proposé par Fay (1991) et une méthode d'estimation de la variance sous ce cadre de travail a été développée par Shao et Steel (1999). Sous le cadre de travail renversé, l'ordre du processus d'échantillonnage et du mécanisme de non-réponse est inversé. En effet, le cadre de travail renversé consiste à supposer que la population U est d'abord divisée aléatoirement à l'aide du mécanisme de non-réponse ; ce qui mène à une population de répondants U_r et une population de non-répondants U_m. Ensuite, un échantillon s est tiré de la population ainsi divisée selon un plan de sondage $p(s)$, ce qui mène à l'ensemble des répondants s_r et celui des non-répondants s_m. Le cadre de travail renversé facilite généralement le problème de l'estimation de la variance mais requiert une hypothèse supplémentaire : le fait de répondre ou non est indépendant de l'échantillon tiré.

5.3.1. L'approche NM

Dans la présente section, nous présentons une méthode d'estimation de variance proposée par Shao et Steel (1999) sous le cadre de travail renversé dans le cas d'une méthode d'imputation déterministe et d'une imputation aléatoire sous l'approche NM. Désignons par $\mathbf{r} = (r_1, ..., r_i, ..., r_N)^\top$ le vecteur des variables indicatrices de réponse.

Sous l'approche NM, la variance totale dans le cas d'une méthode d'imputation déterministe est donnée par :

$$\begin{aligned} V(\hat{Y}_I) &= E_q V_p(\hat{Y}_I|\mathbf{r}) + V_q E_p(\hat{Y}_I|\mathbf{r}) \\ &= V_1^q + V_2^q, \end{aligned} \tag{5.3.1}$$

où $V_1^q = E_q V_p(\hat{Y}_I|\mathbf{r})$ et $V_2^q = V_q E_p(\hat{Y}_I|\mathbf{r})$.

Afin d'estimer $V(\hat{Y}_I)$ en (5.3.1), il suffit d'estimer séparément les termes V_1^q et V_2^q. Pour estimer V_1^q, il suffit de déterminer un estimateur asymptotiquement p-sans biais de $V_p(\hat{Y}_I|\mathbf{r})$ que nous noterons par v_1. Autrement dit, on aura : $E_p(v_1|\mathbf{r}) \approx V_p(\hat{Y}_I|\mathbf{r})$. Par conséquent, l'estimateur v_1 sera asymptotiquement pq-sans biais pour V_1^q, c'est à dire que $E_q E_p(v_1|\mathbf{r}) \approx V_1^q$.
Notons que l'estimateur v_1 peut être obtenu au moyen de méthodes d'estimation classiques de la variance puisqu'il s'agit d'estimer ici la variance due à l'échantillonnage conditionnellement au vecteur $\mathbf{r}$. Les logiciels standards d'estimation de la variance utilisés en absence de non-réponse peuvent donc servir à estimer V_1^q. Pour estimer V_2^q, il suffit de déterminer un estimateur asymptotiquement pq-sans biais de $V_q E_p(\hat{Y}_I|\mathbf{r})$, que nous dénoterons par v_2^q. Pour calculer v_2^q, il nous faudra estimer les probabilités de réponse du modèle de réponse postulé.
Sous l'hypothèse que la fraction de sondage est négligeable, le terme V_2^q est négligeable devant le terme V_1^q. En effet, sous certaines conditions de régularité, il est possible de montrer que V_1^q est $O(\frac{N^2}{n})$ et que V_2^q est $O(N)$. La contribution de V_2^q à la variance totale, $\frac{V_2^q}{V_1^q+V_2^q}$ est donc $O(n/N)$, qui est négligeable si la fraction de sondage n/N est négligeable. Autrement dit, si n/N est négligeable, un estimateur de la variance totale donnée par (5.3.1) est donnée par $v_R = v_1$. (Shao et Steel, 1999). Nous notons que cet estimateur est asymptotiquement sans biais pour la variance totale et que sa validité ne dépend pas de celle du mécanisme de non-réponse postulé.

Dans le cas d'une méthode d'imputation stochastique, la variance totale de l'estimateur imputé $\hat{Y}_I$ s'écrit comme suit :

$$\begin{aligned} V(\hat{Y}_I) &= E_q V_p E_I(\hat{Y}_I|\mathbf{r}) + E_q E_p V_I(\hat{Y}_I|\mathbf{r}) + V_q E_p E_I(\hat{Y}_I|\mathbf{r}) \qquad (5.3.2) \\ &= \tilde{V}_1^q + \tilde{V}_I^q + \tilde{V}_2^q, \end{aligned}$$

où $\tilde{V}_1^q = E_q V_p E_I(\hat{Y}_I|\mathbf{r})$, $\tilde{V}_I^q = E_q E_p V_I(\hat{Y}_I|\mathbf{r})$ et $\tilde{V}_2^q = V_q E_p E_I(\hat{Y}_I|\mathbf{r})$. Un estimateur de la variance totale peut est obtenu en estimant chacun des termes en (5.3.2). Pour estimer $\tilde{V}_1^q$, nous pouvons tout simplement estimer $V_p E_I(\hat{Y}_I|\mathbf{r})$ par un estimateur asymptotiquement pI-sans biais, que nous dénotons par $\tilde{v}_1$. Pour l'estimation de $\tilde{V}_I^q$, nous n'avons besoin que d'estimer $V_I(\hat{Y}_I|\mathbf{r})$ par un estimateur

approximativement I-sans biais, que nous dénotons par v_I. Finalement, pour estimer $\tilde{V}_2^q$, nous devons trouver un estimateur asymptotiquement pqI-sans biais de $V_q E_p E_I(\hat{Y}_I|\mathbf{r})$, que nous dénotons par $\tilde{v}_2^q$. En outre, pour obtenir $\tilde{v}_2^q$, nous avons besoin d'estimer les probabilités de réponse. Par ailleurs, sous certaines conditions de régularité, le terme $\tilde{V}_1^q$ est $O(\frac{N^2}{n})$, le terme $\tilde{V}_I^q$ est $O(\frac{N^2}{n})$ et le terme $\tilde{V}_2^q$ est $O(N)$. Le terme $\tilde{V}_2^q$ est négligeable par rapport à $\tilde{V}_1^q + \tilde{V}_I^q$, si la fraction de sondage n/N est négligeable (Shao et Steel, 1999). Un estimateur approximativement sans biais de la variance totale, sous l'hypothèse que la fraction de sondage est négligeable est donné par : $\tilde{v}_1 + v_I$.

5.3.2. L'approche IM

Dans la présente section, nous présentons une méthode d'estimation de la variance développée par Shao et Steel, (1999) sous l'approche IM, dans le cas d'une méthode d'imputation déterministe et dans le cas d'une méthode d'imputation aléatoire.

Sous l'approche IM, la variance totale de l'estimateur imputé (2.2.1) dans le cas d'une méthode d'imputation déterministe est donnée par :

$$\begin{aligned} V(\hat{Y}_I - Y) &= E_m E_q V_p(\hat{Y}_I|\mathbf{r}) + E_q V_m E_p(\hat{Y}_I - Y|\mathbf{r}) \\ &+ V_q E_m E_p(\hat{Y}_I - Y|\mathbf{r}) \qquad (5.3.3) \\ &= V_1^m + V_2^m + V_3^m, \end{aligned}$$

où $V_1^m = E_m E_q V_p(\hat{Y}_I|\mathbf{r})$, $V_2^m = E_q V_m E_p(\hat{Y}_I - Y|\mathbf{r})$ et $V_3^m = V_q E_m E_p(\hat{Y}_I - Y|\mathbf{r})$. Sous l'hypothèse que $\hat{Y}_I$ est asymptotiquement mpq-sans biais, nous avons $E_m E_p(\hat{Y}_I - Y|\mathbf{r}) \approx 0$. Nous omettrons donc le troisième terme à droite de l'égalité (5.3.3) dans l'estimation de la variance totale. Pour estimer V_1^m, nous n'avons besoin que d'estimer $V_p(\hat{Y}_I|\mathbf{r})$. L'estimation de V_1^m est donc similaire à celle de V_1^q. Pour estimer V_2^m, il suffit d'estimer $V_m E_p(\hat{Y}_I - Y|\mathbf{r})$ par un estimateur approximativement mp-sans biais que nous dénotons par v_2^m. L'estimation de V_2^m nécessite l'estimation des paramètres inconnus du modèle d'imputation. Shao et Steel (1999) ont montré que sous certaines conditions de régularité, le terme V_1^m est $O(N^2/n)$ et le terme V_2^m est $O(N)$ et donc, si la fraction de sondage n/N est

négligeable, la contribution de V_2^m à la variance totale est aussi négligeable. De ce fait, l'estimateur de la variance totale dans le cas d'une méthode d'imputation déterministe sous l'approche IM est identique à celui obtenu sous l'approche NM.

Dans le cas d'une méthode d'imputation aléatoire, la variance totale de l'estimateur imputé $\hat{Y}_I$ est donnée par :

$$\begin{aligned} V(\hat{Y}_I - Y) &= E_m E_q V_p E_I(\hat{Y}_I|\mathbf{r}) + E_m E_q E_p V_I(\hat{Y}_I|\mathbf{r}) \\ &+ E_q V_m E_p E_I(\hat{Y}_I - Y|\mathbf{r}) + V_q E_m E_p E_I(\hat{Y}_I - Y|\mathbf{r}) \qquad (5.3.4) \\ &= \tilde{V}_1^m + \tilde{V}_I^m + \tilde{V}_2^m + \tilde{V}_3^m, \end{aligned}$$

où $\tilde{V}_1^m = E_m E_q V_p E_I(\hat{Y}_I|\mathbf{r})$, $\tilde{V}_I^m = E_m E_q E_p V_I(\hat{Y}_I|\mathbf{r})$, $\tilde{V}_2^m = E_q V_m E_p E_I(\hat{Y}_I - Y|\mathbf{r})$ et $\tilde{V}_3^m = V_q E_m E_p E_I(\hat{Y}_I - Y|\mathbf{r})$.
Sous l'hypothèse que $\hat{Y}_I$ est asymptotiquement $mpqI$-sans biais, nous omettrons $\tilde{V}_3^m$ dans le calcul de la variance totale (5.3.4). Pour estimer $\tilde{V}_1^m$, il suffit d'estimer la variance conditionnelle $V_p E_I(\hat{Y}_I|\mathbf{r})$ par un estimateur asymptotiquement pI-sans biais. Pour estimer $\tilde{V}_I^m$, il suffit d'estimer $V_I(\hat{Y}_I|\mathbf{r})$ par un estimateur approximativement I-sans biais. L'estimation de $\tilde{V}_1^m$ et de $\tilde{V}_I^m$ sous l'approche IM est donc identique à celle présentée sous l'approche NM. Pour estimer $\tilde{V}_2^m$, il faudra estimer $V_m E_p E_I(\hat{Y}_I - Y|\mathbf{r})$ par un estimateur approximativement mpI-sans biais que nous noterons $\tilde{v}_2^m$. Notons finalement que pour estimer $\tilde{V}_2^m$, nous aurions besoin d'estimer certains paramètres inconnus du modèle d'imputation postulé. Si la fraction de sondage est négligeable, la contribution de $\tilde{V}_2^m$ à la variance totale est négligeable sous certaines conditions de régularité (Shao et Steel, 1999). L'estimateur de la variance totale dans ce cas est donc similaire à celui proposé sous l'approche NM dans le cas d'une méthode d'imputation aléatoire.

5.4. Le *jackknife*

Il existe de nombreuses méthodes de rééchantillonnage. Dans ce mémoire, nous présentons une technique fréquemment utilisée dans les enquêtes : le *jackknife*. Les méthodes de rééchantillonnage sont populaires en pratique car contrairement aux méthodes de linéarisation, elles ne requièrent pas les probabilités d'inclusion du second ordre, π_{ij}, qui peuvent être difficiles à obtenir pour certains plans de

sondage et elles ne requièrent pas un développement particulier pour chaque paramètre d'intérêt.

Le *jackknife* proposé par Quenouille (1949) est une méthode de rééchantillonnage servant généralement à l'estimation de la variance. Dans un premier temps, nous décrivons la technique en l'absence de non-réponse. On procèdera selon les grandes étapes suivantes :

(1) Enlever l'unité j de l'échantillon.

(2) Ajuster les poids de sondage w_i pour obtenir les poids ajustés $w_{i(j)}$, où $w_{i(j)}$ est donné par

$$w_{i(j)} = \begin{cases} \frac{n}{n-1} w_i, & \text{si } i \neq j, \\ 0, & \text{sinon.} \end{cases}$$

(3) Calculer l'estimateur $\hat{Y}_{\pi(j)}$ de la même manière que $\hat{Y}_\pi$ mais avec les poids ajustés $w_i(j)$. On définit alors $\hat{Y}_{\pi(j)} = \sum_{i \in s} w_{i(j)} y_i$.

(4) Replacer l'unité enlevée à l'étape (1).

(5) Répéter les étapes (1)-(4) jusqu'à ce que toutes les unités aient été enlevées chacune à leur tour.

L'estimateur *jackknife* de la variance de $\hat{Y}_\pi$ est donnée par

$$v_J(\hat{Y}_\pi) = \frac{n-1}{n} \sum_{j \in s} \left(\hat{Y}_{\pi(j)} - \hat{Y}_\pi \right)^2. \tag{5.4.1}$$

Dans le cas de l'échantillonnage aléatoire simple sans remise, on a $\hat{Y}_\pi = N\hat{\bar{Y}}$, où $\hat{\bar{Y}} = \sum_{i \in s} y_i/n$. Dans ce cas, l'estimateur (5.4.1) devient

$$v_J(\hat{Y}_\pi) = N^2 \frac{s_y^2}{n}, \tag{5.4.2}$$

où $s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{\bar{Y}})^2$. L'estimateur (5.4.2) est l'estimateur classique de la variance dans le cas de l'échantillonnage aléatoire simple avec remise.

En présence de non-réponse sur la variable y, l'estimateur *jackknife* (5.4.1) obtenu en remplaçant $\hat{Y}_\pi$ par l'estimateur imputé (2.2.1) résulte généralement en une sous-estimation de la variance totale, particulièrement lorsque le taux de

réponse est faible. Pour contrer ce problème, Rao et Shao (1992) ont proposé un estimateur *jackknife* ajusté qui tient compte de la non-réponse. La différence avec le *jackknife* usuel est que lorsqu'une unité répondante $j \in s_r$ est enlevée, les valeurs imputées y_i^* sont ajustées pour tenir compte de l'impact de l'élimination de j sur l'ensemble des valeurs imputées. Mais lorsqu'une unité non-répondante, $j \in s_m$, est enlevée, les valeurs imputées sont laissées telles quelles. Le *jackknife* de Rao et Shao peut être motivé par l'idée que l'ajustement crée une variabilité additionnelle dans l'échantillon *jackknife* de sorte que la part de la variance ignorée par le *jackknife* usuel est capturée (Shao et Chen, 2001).

Haziza et Picard (2009) ont montré que dans le cas d'une méthode d'imputation déterministe, l'estimateur *jackknife* de Rao-Shao est en fait un estimateur de $V_p(\hat{Y}_I|\mathbf{r})$. L'estimateur *jackknife* est donc un estimateur asymptotiquement sans biais de V_1 dans le cas d'un plan de sondage avec remise, où V_1 est une notation générique que l'on utilise lorsqu'une propriété s'applique à V_1^q et à V_1^m. L'estimateur *jackknife* de Rao et Shao peut donc être obtenu en appliquant une méthode standard de *jackknife* (c'est-à-dire une méthode *jackknife* utilisée en l'absence de non-réponse) puisqu'il s'agit d'estimer $V_p(\hat{Y}_I|\mathbf{r})$. Il n'est donc pas nécessaire d'ajuster les valeurs imputées. En outre, le *jackknife* proposé par Rao et Shao, estime convenablement la variance totale de l'estimateur imputé (2.2.1) lorsque la fraction de sondage est négligeable.

5.5. Application du *jackknife* à l'imputation par le ratio.

Dans la présente section, nous proposons d'utiliser la technique du *jackknife* afin d'estimer la variance de l'estimateur imputé de la moyenne $\bar{Y} = \sum_{i \in U} y_i/N$. Par souci de simplicité, on se restreint au cas de l'imputation par le ratio. La généralisation à l'imputation par la régression linéaire est relativement aisée.

Nous considérons deux méthodes d'imputation distinctes : l'imputation par le ratio aléatoire-ϕ pour laquelle les valeurs imputées sont données en (3.4.9) avec

$\mathbf{z}_i = z_i$ et $c_i = z_i$ et l'imputation par le ratio équilibrée aléatoire-ϕ, dont les valeurs imputées sont sélectionnées comme en (3.4.9) avec $\mathbf{z}_i = z_i$ et $c_i = z_i$, tout en respectant l'équation d'équilibrage (3.4.11). Nous désignons par $\hat{\bar{Y}}_I^{(RAL-\phi)}$ et $\hat{\bar{Y}}_I^{(REA-\phi)}$ les estimateurs de la moyenne obtenus sous chacune de ces méthodes d'imputation, respectivement. Nous supposons aussi dans la présente section, que la fraction de sondage n/N est négligeable, si bien que l'on peut ignorer le terme $\tilde{V}_2^q$ (sous l'approche NM) ou le terme $\tilde{V}_2^m$ (sous l'approche IM) dans le calcul de la variance des estimateurs $\hat{\bar{Y}}_I^{(RAL-\phi)}$ et $\hat{\bar{Y}}_I^{(REA-\phi)}$.

Il suffit donc d'estimer tout simplement

$$V(\hat{\bar{Y}}_I^{(RAL-\phi)}) = E_q V_p E_I \left(\hat{\bar{Y}}_I^{(RAL-\phi)}|\mathbf{r}\right) + E_q E_p V_I \left(\hat{\bar{Y}}_I^{(RAL-\phi)}|\mathbf{r}\right) \tag{5.5.1}$$

et

$$V(\hat{\bar{Y}}_I^{(REA-\phi)}) = E_q V_p E_I \left(\hat{\bar{Y}}_I^{(REA-\phi)}|\mathbf{r}\right). \tag{5.5.2}$$

La variance (5.5.2) n'a pas de composante due à l'imputation comme la variance (5.5.1), car les valeurs imputées sont choisies de manière à respecter la contrainte (3.4.11) afin d'éliminer la variance due à l'imputation.

Pour estimer $E_q E_p V_I \left(\hat{\bar{Y}}_I^{RAL-\phi}|\mathbf{r}\right)$ il suffit d'estimer $V_I \left(\hat{\bar{Y}}_I^{RAL}|\mathbf{r}\right)$ par :

$$v_I(\hat{\bar{Y}}_I^{(RAL-\phi)}) = \frac{1}{(\sum_{i\in s} w_i)^2}\left(\sum_{i\in s} w_i^2(1-r_i)\hat{\phi}_i(1-\hat{\phi}_i)z_i^2\left(\frac{\sum_{i\in s} w_i r_i \delta_i y_i}{\sum_{i\in s} w_i r_i \delta_i z_i}\right)^2\right),$$

puisque $V_I(y_i^*) = \hat{\phi}_i(1-\hat{\phi}_i)(z_i\hat{B}_{r_1})^2$.

La méthode d'estimation de $E_q V_p E_I \left(\hat{\bar{Y}}_I^{RAL-\phi}|\mathbf{r}\right)$ et $E_q V_p E_I \left(\hat{\bar{Y}}_I^{REA-\phi}|\mathbf{r}\right)$ est similaire. Nous utilisons la technique de *jackknife* présentée dans Haziza et Picard (2009). Par exemple, dans le cas de l'estimateur $\hat{\bar{Y}}_I^{RAL-\phi}$, il suffit d'estimer $V_p E_I \left(\hat{\bar{Y}}_I^{RAL-\phi}\right)$. Pour ce faire, il suffit d'abord de créer une nouvelle colonne dans le fichier de données correspondant aux valeurs imputées $E_I(y_i^*) = \hat{\phi}_i z_i \hat{B}_{r_1}$. Dans

ce cas, on a

$$\begin{aligned} E_I(\hat{\tilde{Y}}_I^{(RAL-\phi)}) &= \frac{1}{\sum_{i\in s} w_i}\left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i(1-r_i)\hat{\phi}_i z_i\left(\frac{\sum_{i\in s} w_i r_i \delta_i y_i}{\sum_{i\in s} w_i r_i \delta_i z_i}\right)\right) \\ &\equiv \tilde{\tilde{Y}}_I^{(RAL-\phi)}. \end{aligned}$$

Définissons $\tilde{\tilde{Y}}_{I(j)}^{(RAL-\phi)}$ par

$$\tilde{\tilde{Y}}_{I(j)}^{(RAL-\phi)} = \frac{1}{\sum_{i\in s} w_{i(j)}}\left(\sum_{i\in s} w_{i(j)} r_i y_i + \sum_{i\in s} w_{i(j)}(1-r_i)\hat{\phi}_{i(j)} z_i\left(\frac{\sum_{i\in s} w_{i(j)} r_i \delta_i y_i}{\sum_{i\in s} w_{i(j)} r_i \delta_i z_i}\right)\right),$$

où $\hat{\phi}_{i(j)}$ est l'estimateur de ϕ_i lorsque la j^{ieme} unité a été enlevée.
L'estimateur *jackknife* de $V_p E_I\left(\hat{\tilde{Y}}_I^{(RAL-\phi)}|\mathbf{r}\right)$ est donné par

$$v_J(\tilde{\tilde{Y}}_I^{(RAL-\phi)}) = \frac{n-1}{n}\sum_{i\in s}\left(\tilde{\tilde{Y}}_{I(j)}^{(RAL-\phi)} - \tilde{\tilde{Y}}_I^{(RAL-\phi)}\right)^2. \tag{5.5.3}$$

Par un raisonnement similaire, on peut déduire un estimateur *jackknife* $v_J(\tilde{\tilde{Y}}_I^{REA-\phi})$ de $V_p E_I\left(\hat{\tilde{Y}}_I^{REA-\phi}\right)$. Un estimateur *jackknife* de la variance totale de $\hat{\tilde{Y}}_I^{RAL-\phi}$ est donc donné par

$$\tilde{v}_t(\hat{\tilde{Y}}_I^{(RAL-\phi)}) = v_J(\tilde{\tilde{Y}}_I^{(RAL-\phi)}) + v_I(\bar{Y}_I^{(RAL-\phi)}) \tag{5.5.4}$$

et celui de la variance totale de $\hat{\tilde{Y}}_I^{(REA-\phi)}$ est donné par

$$\tilde{v}_t(\hat{\tilde{Y}}_I^{(REA-\phi)}) = v_J(\tilde{\tilde{Y}}_I^{(REA-\phi)}). \tag{5.5.5}$$

5.6. Étude par simulations

La présente étude par simulations a pour objectif d'évaluer la qualité des estimateurs de variance discutée à la section (5.5) en termes de biais et d'erreur quadratique moyenne. Nous avons généré une population de taille $N = 5000$ constituée de deux variables : une variable d'intérêt y et une variable auxiliaire z. Nous avons généré z selon une loi gamma de paramètre de position $\alpha = 4$ et de paramètre d'échelle $\beta = 25$. Les valeurs de y ont été obtenues à partir du modèle $y = 2z + \epsilon$, où les erreurs ϵ_i ont été générées d'une loi normale de moyenne 0 et de variance σ^2 choisie afin d'obtenir un coefficient de détermination $R^2 = 0.5$ entre y et z.

Nous avons considéré deux scénarios : le premier a consisté à étudier les estimateurs de variance proposés lorsque le mécanisme aléatoire ayant généré les zéros sur la variable d'intérêt y était uniforme ($\phi_i = \phi$ est constant, $\forall i \in U$) et le second nous a permis d'étudier les estimateurs de variance lorsque la probabilité ϕ_i dépendait de la variable auxiliaire z_i.

Pour générer les zéros à la variable y dans le cas d'un mécanisme uniforme, nous avons généré une variable indicatrice δ_i selon une loi de Bernoulli de paramètre ϕ_i égale à 0.9 et 0.75. Dans ce cas, la proportion de zéros générée était de 10% et de 25% respectivement. Lorsque les ϕ_i n'étaient pas uniformes, nous avions obtenu les probabilités ϕ_i selon un modèle logistique :

$$log\left(\frac{\phi_i}{1-\phi_i}\right) = \lambda_0 + \lambda_1 z_i,$$

où les valeurs de λ_0 et λ_1 ont été choisies de sorte que la proportion de zéros soit d'environ 10% et de 25%. Ensuite la variable δ_i qui indique qu'une unité i a une valeur de y positive a été générée selon une loi de Bernoulli de paramètre ϕ_i. Finalement, dans chacune des populations, nous avons affecté la valeur zéro à la variable y lorsque la variable δ_i indiquait zéro et nous l'avons laissé inchangée dans le cas contraire.

A partir de chacune des populations créées, nous avons tiré $R = 10000$ échantillons de taille 150 selon le plan d'échantillonnage aléatoire simple sans remise. La fraction de sondage est donc égale à 0.03 pour l'ensemble des échantillons et peut être considérée comme négligeable. La non-réponse sur la variable d'intérêt y a été générée selon le mécanisme uniforme de sorte que le taux de non-réponse moyen p soit de 70% dans chaque échantillon. La variable r_i qui indique si une unité i a répondu ou pas a été générée selon une loi de Bernoulli de paramètre p.

Les estimateurs de variance proposés sont donnés par (5.5.5) et (5.5.4). Pour mesurer la précision de ces estimateurs, nous avons calculé le biais relatif et l'erreur quadratique moyenne Monte Carlo.

Définissons l'espérance Monte Carlo d'un estimateur $\hat{\theta}$ par :

$$E_{MC}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{(r)},$$

où $\hat{\theta}^{(r)}$ dénote l'estimateur $\hat{\theta}$ du r^{ieme} échantillon.
Le biais relatif Monte Carlo de $\tilde{v}_t(\hat{\bar{Y}}_I^{RAL-\phi})$ est donné par :

$$RB_{MC}(\tilde{v}_t(\hat{\bar{Y}}_I^{RAL-\phi})) = 100 \times \frac{E_{MC}\left(\tilde{v}_t(\hat{\bar{Y}}_I^{RAL-\phi}) - V_{MC}(\hat{\bar{Y}}_I^{RAL-\phi})\right)}{V_{MC}(\hat{\bar{Y}}_I^{RAL-\phi})} \quad (5.6.1)$$

et l'erreur quadratique moyenne Monte Carlo par :

$$EQM_{MC}(\tilde{v}_t(\hat{\bar{Y}}_I^{RAL-\phi})) = E_{MC}\left(\tilde{v}_t(\hat{\bar{Y}}_I^{RAL-\phi}) - V_{MC}(\hat{\bar{Y}}_I^{RAL-\phi})\right)^2, \quad (5.6.2)$$

où $V_{MC}(\hat{\bar{Y}}_I^{RAL-\phi}) = E_{MC}\left(\hat{\bar{Y}}_I^{RAL-\phi} - E_{MC}(\hat{\bar{Y}}_I^{RAL-\phi})\right)^2$. Le calcul du biais relatif et de l'erreur quadratique moyenne Monte Carlo de $\tilde{v}_t(\hat{\bar{Y}}_I^{REA-\phi})$ est effectué de façon similaire, en remplaçant $\hat{\bar{Y}}_I^{RAL-\phi}$ dans les équations (5.6.1) et (5.6.2) par $\hat{\bar{Y}}_I^{REA-\phi}$.

Le biais relatif et l'erreur quadratique moyenne Monte Carlo des estimateurs de la variance donnés en (5.5.4) et (5.5.5) sont présentés dans le tableau 5.1. Ce tableau indique que les estimateurs *jackknife* de la variance ont bien performé dans tous les scénarios avec un biais relatif absolu de 6%. En termes de EQM, il apparait que l'estimateur *jackknife* de la variance sous l'imputation REA-ϕ a été plus stable que celui de l'imputation RAL-ϕ. Ceci est dû au terme additionnel de la variance sous l'imputation RAL-ϕ dû à la sélection aléatoire des valeurs imputées.

TAB. 5.1. Monte Carlo RB (en %) et EQM (en parenthèses) de l'estimateur *jackknife* de la variance

		RB (EQM)	
mécanisme-ϕ	proportion de 0	RAL-ϕ	REA-ϕ
1	10%	1.53 (871.87)	0.85 (779.71)
	25%	-2.02 (1042.74)	-3.16 (877.85)
2	10%	1.69 (846.26)	1.67 (804.02)
	25%	5.23 (1004.93)	5.27 (943.52)

CONCLUSION

Le but de ce mémoire était de proposer une méthode d'imputation permettant de traiter le problème de non-réponse à une variable d'intérêt qui prend des valeurs nulles un très grand nombre de fois. Nous avons traité le cas de l'imputation par la régression. Nous recherchions une méthode d'imputation satisfaisant les trois critères suivants :

(a) l'estimateur imputé est asymptotiquement sans biais sous les approches UNM et IM.

(b) les valeurs imputées sont réalistes.

(c) l'estimateur imputé est complètement efficace.

Nous avons commencé par étudier les propriétés de deux méthodes d'imputation par la régression : l'imputation par la régression déterministe positive et l'imputation par la régression déterministe très couramment utilisée en pratique. Nous montrons au chapitre trois, que ces deux méthodes d'imputation ne satisfont en général que le critère (c), mais pas les critères (a) et (b). Les résultats de l'étude par simulations du chapitre 4, confirment ces résultats. Nous avons ensuite proposé trois méthodes d'imputation, motivées par un modèle de mélange par la régression : la première, l'imputation par la régression déterministe-ϕ satisfait aux critères (a) et (c) mais pas au critère (b) ; la seconde, l'imputation par la régression aléatoire-ϕ, quant à elle, satisfait aux critères (a) et (b) uniquement ; finalement, nous avons proposé l'imputation équilibrée aléatoire-ϕ par la régression. Cette méthode est particulièrement attractive dans le sens où elle satisfait simultanément les trois critères (a)-(c). Nous avons également proposé un estimateur *jackknife* de la variance qui ne requiert pas un ajustement des valeurs

imputées comme dans la technique de *jackknife* proposée par Rao et Shao. L'estimateur *jackknife* de la variance proposée est asympotiquement sans biais et convergent vers la vraie variance de l'estimateur imputé sous les approches NM ou IM, en notant que la fraction de sondage est négligeable.

Notons que l'imputation par la régression aléatoire-ϕ que nous avons proposée, ne préserve pas la distribution de la variable imputée des unités positives, car sur le sous-ensemble des unités positives, nous appliquons une méthode d'imputation déterministe standard. Une alternative aurait été l'utilisation d'une méthode d'imputation par la régression avec résidus aléatoires, dont les valeurs imputées sont données par (2.4.2). A la suite de cela, nous aurions pu proposer une méthode d'imputation par la régression équilibrée satisfaisant à deux contraintes d'équilibrage : la première aurait porté sur les résidus aléatoires correspondants aux répondants positifs ; et la seconde, aurait porté sur la sélection aléatoire des valeurs imputées, comme dans le cas de l'imputation équilibrée proposée.

BIBLIOGRAPHIE

Chauvet, G., Deville, J.C. et Haziza, D. (2010), On random balanced imputation in surveys. *Soumis pour publication.*

Chen, H.L., Rao, J.N.K. et Sitter, R.R. (2000), Efficient random imputation for missing survey data in complex survey. *Statistica Sinica*, 10, pp. 1153-1169.

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

Deville, J.C. (2006), Random imputation using balanced sampling. *Presentation to the JointStatistical Meeting of the American Statistical Association, Seattle, USA.*

Deville, J.C. et Särndal, C.E. (1994), Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 23, pp. 33-40.

Deville, J.C. et Tillé, Y. (2004), Efficient balanced sampling : the cube method. *Biometrika*, 91, pp. 893-912.

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

Fuller, W.A. et Kim, J.K. (2005), Hot-deck imputation for the response model survey methodology. *Survey Methodology*, 31, pp. 139-149.

Haziza, D. et Picard, F. (2009), On doubly robust point and variance estimation in the presence of imputed data. *Soumis pour publication.*

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

Isaki, C.T. et Fuller, W.A. (1982), Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, pp. 89-96.

Kalton, G. et Kasprzyk, D. (1986), The treatment of survey missing data. *Survey Methodology*, 12, pp. 1-16.

Kalton, G. et Kish, L. (1981), Two efficient random imputation procedures. *Proceedings of the Survey Research Methods*, American Statistical Association, pp. 146-151.

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

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

Kim, J.K. et Haziza, D. (2010), Doubly robust inference with missing data in survey sampling. *Rapport technique.*

Narain, R.D. (1951), On Sampling without replacement with varying probabilities, *Journal of the Indian Society of Agricultural Statistics,* 4, pp. 169-75.

Pfefferman, D. (1993), The role of sampling weights when modeling survey data, *International Statistical Review,* 61, pp. 317-337.

Quenouille, M.H. (1949), Problems in plane sampling, *Annals of Mathematical Statistics,* 20, pp. 355-375.

Rancourt, E., Gagnon, F., Lee, H., Provost, M., et Särndal, C.E. (1997), Estimation of variance in presence of imputation, *Proceedings of the Statistics Canada Symposium 1997, New Directions in Surveys and Censuses,Statistics Canada,* pp. 273-277.

Rao, J.N.K. (1990), Variance estimation under imputation for missing data. *Rapport technique*, Statistics Canada, Ottawa.

Rao, J.N.K. et Shao, J. (1992), On variance estimation under imputation for missing data. *Biometrika*, 79, pp. 811-822.

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

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

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

Särndal, C.E. (1992), Methods for estimating the precision of surveys estimates when imputation has been used. *Survey Methodology*, 18, pp. 241-252.

Särndal, C.E., Swensson, B. et Wretman, J. (1992), *Model Assisted Survey Sampling*, New York : Springer-Verlag.

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

Sen, A.R. (1953), On the estimate of the variance in sampling with varying probabilities, *Journal of the Indian Society of Agricultural Statistics*, 5, pp. 119-27.

Tillé, Y. (2001), *Théorie des Sondages*, Paris : Dunod.

Wand, M.P. et Jones, M.C. (1995), *Kernel Smoothing*, London : Chapman and Hall.

Yates, F. et Grundy, P.M. (1953), Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society B* 15, pp. 253-61.

Annexe A

PREUVES DETAILLÉES DE QUELQUES RÉSULTATS

A.1. Biais sous NM toutes les observations.

Sous l'approche NM, le biais (2.8.2) de l'estimateur imputé (2.8.1) est

$$B_q(\hat{Y}_I) \approx \sum_{i\in s} w_i(1-p_i)(y_i - \mathbf{z}_i^\top \hat{\mathbf{B}}_p),$$

où $\hat{\mathbf{B}}_p = \left(\sum_{i\in s} w_i p_i \mathbf{z}_i \mathbf{z}_i^\top/(\boldsymbol{\lambda}^\top \mathbf{z}_i)\right)^{-1} \left(\sum_{i\in s} w_i p_i \mathbf{z}_i y_i/(\boldsymbol{\lambda}^\top \mathbf{z}_i)\right)$.

Démonstration.

$$\begin{aligned} B_q(\hat{Y}_I) &= E_q(\hat{Y}_I - \hat{Y}|s) \\ &= E_q\left(\sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i(1-r_i)\mathbf{z}_i^\top \hat{\mathbf{B}}_r - \sum_{i\in s} w_i y_i|s\right) \\ &= E_q\left(\hat{Y}_r - (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)^\top \hat{\mathbf{B}}_r - \hat{Y}|s\right), \end{aligned}$$

où $\hat{Y}_r = \sum_{i\in s} w_i r_i y_i$, $\hat{\mathbf{Z}} = \sum_{i\in s} w_i \mathbf{z}_i$ et $\hat{\mathbf{Z}}_r = \sum_{i\in s} r_i w_i \mathbf{z}_i$.

Par un développement en série de Taylor de premier ordre, on a :

$$\begin{aligned} \hat{\mathbf{B}}_r - \hat{\mathbf{B}}_p &= \hat{\mathbf{T}}_r^{-1}\hat{\mathbf{t}}_r - \hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p \\ &\approx \sum_{i\in s} r_i \xi_i - \sum_{i\in s} p_i \xi_i, \end{aligned}$$

où

$$\begin{aligned}\xi_i &= \frac{\partial \hat{\mathbf{B}}_r}{\partial r_i}\Big|_{\mathbf{r=p}} \\ &= \left(\frac{\partial \hat{\mathbf{T}}_r^{-1}}{\partial r_i}\right)\hat{\mathbf{t}}_r\Big|_{\mathbf{r=p}} + \hat{\mathbf{T}}_r^{-1}\left(\frac{\partial \hat{\mathbf{t}}_r}{\partial r_i}\right)\Big|_{\mathbf{r=p}} \\ &= -\hat{\mathbf{T}}_p^{-1}\frac{w_i\mathbf{z}_i\mathbf{z}_i^\top}{\boldsymbol{\lambda}^\top\mathbf{z}_i}\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p + \hat{\mathbf{T}}_p^{-1}\frac{w_i z_i y_i}{\boldsymbol{\lambda}^\top\mathbf{z}_i}. \qquad \text{(A.1.1)}\end{aligned}$$

On note que :

$$\begin{aligned}\sum_{i\in s} p_i\xi_i &= -\hat{\mathbf{T}}_p^{-1}\sum_{i\in s}\frac{w_i p_i\mathbf{z}_i\mathbf{z}_i^\top}{\boldsymbol{\lambda}^\top\mathbf{z}_i}\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p + \hat{\mathbf{T}}_p^{-1}\sum_{i\in s}\frac{w_i p_i z_i y_i}{\boldsymbol{\lambda}^\top\mathbf{z}_i} \\ &= -\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{T}}_p\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p + \hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p \\ &= -\hat{\mathbf{T}}_p^{-1}\mathbf{I}_q\hat{\mathbf{t}}_p + \hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p \\ &= 0,\end{aligned}$$

où $\mathbf{I}_q$ est la matrice identité de rang q.

Nous avons alors,

$$\begin{aligned}\hat{\mathbf{B}}_r - \hat{\mathbf{B}}_p &\approx -\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{T}}_r\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p + \hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_r \\ &= -\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{T}}_r\hat{\mathbf{B}}_p + \hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_r. \qquad \text{(A.1.2)}\end{aligned}$$

Ce qui implique que,

$$\begin{aligned}E_q\left(\hat{\mathbf{B}}_r - \hat{\mathbf{B}}_p|s\right) &\approx -\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{T}}_p\hat{\mathbf{B}}_p + \hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p \\ &= -\hat{\mathbf{B}}_p + \hat{\mathbf{B}}_p,\end{aligned}$$

où la dernière égalité découle du fait que :

$$E_q(\hat{\mathbf{T}}_r|s) = \hat{\mathbf{T}}_p$$

et

$$E_q(\hat{\mathbf{t}}_r|s) = \hat{\mathbf{t}}_p.$$

On déduit alors facilement que :

$$E_q\left(\hat{\mathbf{B}}_r|s\right) \approx \hat{\mathbf{B}}_p. \qquad \text{(A.1.3)}$$

En appliquant une fois de plus une linéarisation de Taylor du premier ordre, on a :

$$\hat{\mathbf{Z}}_r^\top \hat{\mathbf{B}}_r - \hat{\mathbf{Z}}_p^\top \hat{\mathbf{B}}_p \approx \sum_{i\in s} r_i\xi_i - \sum_{i\in s} p_i\xi_i,$$

où

$$\begin{aligned}
\xi_i &= \frac{\partial \hat{\mathbf{Z}}_r^\top \hat{\mathbf{B}}_r}{\partial r_i}\Big|_{\mathbf{r=p}} \\
&= \left(\frac{\partial \hat{\mathbf{Z}}_r^\top}{\partial r_i}\right)\hat{\mathbf{B}}_r\big|_{\mathbf{r=p}} + \hat{\mathbf{Z}}_r^\top\left(\frac{\partial \hat{\mathbf{B}}_r}{\partial r_i}\right)\Big|_{\mathbf{r=p}} \quad \text{(par la dérivation en chaîne)} \\
&= w_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p + \hat{\mathbf{Z}}_p^\top\left(-\hat{\mathbf{T}}_p^{-1}\frac{w_i\mathbf{z}_i\mathbf{z}_i^\top}{\boldsymbol{\lambda}^\top\mathbf{z}_i}\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}}_p + \hat{\mathbf{T}}_p^{-1}\frac{w_i z_i y_i}{\boldsymbol{\lambda}^\top\mathbf{z}_i}\right) \quad \text{(par (A.1.1))}.
\end{aligned}$$

On trouve facilement que :

$$\sum_{i\in s} p_i\xi_i = \sum_{i\in s} p_i w_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p.$$

Il s'ensuit du résultat précédent et de (A.1.2) que :

$$\hat{\mathbf{Z}}_r^\top \hat{\mathbf{B}}_r - \hat{\mathbf{Z}}_p^\top \hat{\mathbf{B}}_p \approx \sum_{i\in s} r_i w_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p + \hat{\mathbf{Z}}_p^\top\left(-\hat{\mathbf{T}}_p^{-1}\hat{\mathbf{T}}_r\hat{\mathbf{B}}_p + \hat{\mathbf{T}}_p^{-1}\hat{\mathbf{t}_r}\right) - \sum_{i\in s} w_i p_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p.$$

De ce qui précède, on déduit que :

$$\begin{aligned}
E_q\left(\hat{\mathbf{Z}}_r^\top \hat{\mathbf{B}}_r|s\right) &\approx \hat{\mathbf{Z}}_p^\top\hat{\mathbf{B}}_p + \sum_{i\in s} p_i w_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p + 0 - \sum_{i\in s} w_i p_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p \qquad \text{(par (A.1.3))} \\
&= \hat{\mathbf{Z}}_p^\top\hat{\mathbf{B}}_p.
\end{aligned}$$

Finalement,

$$\begin{aligned}
B_q\left(\hat{Y}_I\right) &\approx \hat{Y}_p + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_p)^\top\hat{\mathbf{B}}_p - \hat{Y} \\
&= \sum_{i\in s}(w_i p_i y_i + w_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p - w_i p_i\mathbf{z}_i^\top\hat{\mathbf{B}}_p - w_i y_i) \\
&= -\sum_{i\in s} w_i(1-p_i)(y_i - \mathbf{z}_i^\top\hat{\mathbf{B}}_p),
\end{aligned}$$

□

où $\hat{Y}_p = \sum_{i\in s} w_i p_i y_i$.

Remarque A.1.1. *Lorsque le mécanisme de non-réponse est uniforme ($p_i = p$) le biais en (2.8.2) est approximativement nul. L'estimateur imputé est dit être asymptotiquement sans biais.*

Pour s'en convaincre, il suffit de considérer les deux points suivants :

(1) $\hat{\mathbf{B}}_p = \hat{\mathbf{B}} = \left(\sum_{i\in s} w_i \mathbf{z}_i \mathbf{z}_i^\top / (\boldsymbol{\lambda}^\top \mathbf{z}_i)\right)^{-1} \left(\sum_{i\in s} w_i \mathbf{z}_i y_i / (\boldsymbol{\lambda}^\top \mathbf{z}_i)\right)$

(2) $\boldsymbol{\lambda}^\top \left(\sum_{i\in s} w_i \mathbf{z}_i \mathbf{z}_i^\top / (\boldsymbol{\lambda}^\top \mathbf{z}_i)\right) \hat{\mathbf{B}} = \sum_{i\in s} w_i \mathbf{z}_i^\top \hat{\mathbf{B}} = \sum_{i\in s} w_i y_i$

Et on conclura :

$$\sum_{i\in s} w_i (y_i - \mathbf{z}_i^\top \hat{\mathbf{B}}) \quad = \quad 0.$$