Université de Montréal

# ESTIMATION DE LA VARIANCE DANS LES SONDAGES UTILISANT L'IMPUTATION PAR LE PLUS PROCHE VOISIN

Par

**Nancy Forget**

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

Mai 1999



Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

# ESTIMATION DE LA VARIANCE DANS LES SONDAGES UTILISANT L'IMPUTATION PAR LE PLUS PROCHE VOISIN

présenté par

**Nancy Forget**

A été évalué par un jury composé des personnes suivantes :

Yves Lepage
(président-rapporteur)

Carl Erik Särndal
(directeur de recherche)

Christian Léger
(membre du jury)

Mémoire accepté le:

F.S. 27 Mai 1999

# SOMMAIRE

Dans la majorité des sondages, le statisticien doit faire face au taux élevé de non-réponse que nous observons actuellement dans les firmes de sondage. Sur l'échantillon prélevé à partir d'un plan d'échantillonnage donné, diverses explications et raisons sont à l'origine de ce manque d'information. Une des méthodes à laquelle nous avons recours aujourd'hui est l'imputation. Cette méthode, fréquemment utilisée dans les firmes gouvernementales, a pour but d'insérer des valeurs artificielles aux données manquantes.

Est-ce que les substituts sont assez proches des valeurs manquantes? Voilà la question à se poser. Quelques statisticiens perspicaces et familiers avec ce sujet peuvent, dans plusieurs cas, fournir d'excellentes substitutions. Dans ces cas, l'imputation peut s'avérer meilleure que la technique de repondération des répondants. La forme d'imputation que nous retenons dans ce mémoire est l'imputation par le plus proche voisin. Nous allons tenter d'expliquer l'estimateur de la variance lorsque le but du sondage est d'estimer la moyenne d'une variable sur l'ensemble de la population et que l'imputation par le plus proche voisin a été utilisée pour combler la non-réponse.

Au chapitre 2, nous introduirons l'imputation par la moyenne des répondants. Cette forme élémentaire nous permettra de définir les

notions de base qui nous serviront par la suite à bâtir nos estimateurs de variance pour la méthode d'imputation par le plus proche voisin.

Au chapitre 3, nous verrons que nous ne pouvons ignorer les effets causés par l'imputation. L'accroissement de la variance ainsi qu'un biais non négligeable vont normalement être de la partie. Il est nécessaire d'éliminer ce biais et de fournir clairement aux utilisateurs la composante additionnelle de la variance, réservée uniquement à l'imputation. C'est pourquoi nous avons selon certaines hypothèses, divisé l'estimateur de la variance en trois composantes : l'une due à l'échantillonnage, la seconde, à un terme d'ajustement, qui est la différence entre la variance échantillonnale, dans le cas où il y a 100 % de réponse et la variance échantillonnale dans le cas de la non-réponse, et l'autre, à l'imputation elle-même. Ces trois termes seront traités individuellement.

Au chapitre 4, nous décrirons la population étudiée et les différentes mesures utilisées dans les simulations.

Au chapitre 5 et 6, nous procéderons à des simulations par la méthode de Monte Carlo afin de tester la performance des estimateurs théoriques développés dans ce mémoire. La partie expérimentale se composera de trois estimateurs, soit l'estimateur d'Horvitz-Thompson, l'estimateur par le ratio et celui par la régression. Le plan simple est le plan choisi pour le présent ouvrage.

# REMERCIEMENTS

Dans un premier temps, je tiens à remercier mon directeur de recherche, M. Carl Erik Särndal, pour son encadrement, sa disponibilité, sa patience, sa compréhension, son apport financier, les rapports cordiaux que nous avons entretenus ainsi que pour ses conseils judicieux qui m'ont guidée tout au long de la rédaction de mon mémoire. J'aimerais également remercier Eric Rancourt et Wisner Jocelyn de Statistique Canada pour leurs nombreux conseils concernant l'achèvement de mon mémoire.

En second lieu, j'aimerais remercier mes parents Denise et Jean-Claude pour leur soutien moral et financier. Sans leur appui, mes études universitaires auraient pratiquement été impossibles. De plus, je dis un gros merci à mon grand-père, mes frères Danny et Francis, ma soeur Cathy et mon copain Michel pour leurs encouragements à continuer mes études. Ma grand-mère aurait fort probablement apprécié ce présent ouvrage.

En terminant, je ne peux passer sous silence la contribution de mon employeur, Mme Andrée Demers, et les responsables du laboratoire de statistique, MM. Christian Léger et Miguel Chagnon qui, de leur ouverture d'esprit, m'ont permis d'allier étude et travail afin de compléter ma formation universitaire.

# TABLE DES MATIÈRES

# LISTE DES FIGURES

# LISTE DES TABLEAUX

CHAPITRE 1

# Notions d'échantillonnage

## 1.1. Population et échantillon

Dans notre société, le besoin d'obtenir de l'information statistique semble être insatiable. Des données sont régulièrement amassées pour satisfaire aux demandes d'information sur un ensemble d'éléments communément appelé "population finie". Ce chapitre introduit quelques notions de base sur la théorie de l'échantillonnage qui nous seront utiles tout au long de ce mémoire.

Considérons une population finie $U$ comprenant $N$ éléments étiquetés $k = 1, 2, ..., N$

$$U = \{u_1, u_2, ..., u_N\}$$

Pour simplifier la notation, nous allons représenter le $k^e$ élément par l'étiquette $k$. Nous dénoterons donc la population finie par $U = \{1, ..., k, ..., N\}$. Dans ce mémoire, nous allons considérer la taille de la population $N$ comme étant connue. Cependant, il ne faut pas oublier qu'en pratique, on a souvent recours à des méthodes d'estimation pour évaluer la taille de la population, puisque cette quantité n'est pas toujours disponible.

Nous désignerons par $y$ la variable d'étude et par $x$ la variable auxiliaire. Ainsi, $x_k$ et $y_k$ seront respectivement les valeurs des variables $x$ et $y$ pour la $k^e$ unité de la population $U$. Par exemple, si $U$ est la population des ménages et $y$ la variable qui traduit le revenu, alors $y_k$ est la valeur du revenu du $k^e$ ménage. Nous supposons toujours que les valeurs $y_k$ sur l'ensemble de la population sont inconnues, d'où le nom "variable d'étude". Par opposition, les variables auxiliaires sont faciles à observer, connues sur l'ensemble de la population, et nous supposons qu'elles ont une relation avec la variable d'étude. Dans notre exemple, la variable auxiliaire pourrait être l'âge, le sexe, etc. Fréquemment, elles procurent de meilleurs résultats et sont utilisées pour certains estimateurs, notamment l'estimateur par le ratio, l'estimateur par la régression, etc.

La moyenne et la variance de la variable d'étude d'une population finie (par exemple, la moyenne des revenus des ménages) sont des paramètres importants.

La moyenne de la variable $y$ sur l'ensemble $U$ sera représentée par

$$\bar{y}_U = \frac{1}{N}\sum_{k\in U} y_k$$

Pour alléger la notation, si $A \subseteq U$ est un ensemble d'unités quelconque, nous allons écrire $\sum_{k\in A} y_k$ simplement $\sum_A y_k$. La variance de $y$ sur $U$ sera donnée par

$$S^2_{yU} = \frac{1}{N-1}\sum_U (y_k - \bar{y}_U)^2$$

Toutefois, étant donné que la variable $y$ est inconnue sur l'ensemble de la population, il nous est impossible de calculer les paramètres $\bar{y}_U$

et $S_{yU}$. Nous allons plutôt évaluer ces paramètres en tirant un sous-ensemble de la population $U$, appelé "échantillon" et noté $s$ (où $s \subseteq U$), selon un plan d'échantillonnage. Ainsi, nous observons, lorsque possible, les valeurs prises par la variable d'étude pour les unités de l'échantillon.

## 1.2. Plan d'échantillonnage

Le plan d'échantillonnage est la définition des caractéristiques régissant le mécanisme par lequel nous prélevons aléatoirement un échantillon $s$ dans la population $U$. La probabilité de tirer l'échantillon $s$ est notée $p(s)$. Le plan d'échantillonnage choisi joue un rôle central. Non seulement il détermine les propriétés statistiques essentielles, comme les probabilités d'inclusion de premier et de second ordre, mais il détermine également la distribution des échantillons, la valeur espérée et la variance des quantités aléatoires calculées à partir d'un échantillon.

Imaginons que l'on fixe un plan d'échantillonnage, c'est-à-dire $p(s)$ est fixé. L'inclusion d'un élément $k$ quelconque dans un échantillon est un événement aléatoire, traduit par la variable aléatoire $I_k$ et défini par

$$I_k = \begin{cases} 1 & \text{si } k \in S \\ 0 & \text{sinon} \end{cases} \tag{1.2.1}$$

La probabilité qu'un élément $k$ quelconque soit inclus dans un échantillon est obtenue par

$$\pi_k = P(k \in s) = P(I_k = 1) = \sum_{s \ni k} p(s) \tag{1.2.2}$$

Notons que $\sum_{s \ni k}$ traduit la somme sur les échantillons $s$ dont $k$, une unité fixée, est membre. Nous caractériserons $\{\pi_k : k \in U\}$ *l'ensemble*

*des probabilités d'inclusion d'ordre 1.* Le plan supposé est tel que $\pi_k > 0$, $\forall k \in U$, ainsi, chaque unité a assurément la possibilité d'être sélectionnée dans l'échantillon. Par le même raisonnement, la probabilité que les éléments $k$ et $l$ soient inclus dans l'échantillon est dénotée par $\pi_{kl}$. Alors $\{\pi_{kl} : l, k \in U\}$ constitue *l'ensemble des probabilités d'inclusion d'ordre 2.* Évidemment, nous avons $\pi_{kk} = \pi_k$.

Nous poserons $a_k = 1/\pi_k$ le poids d'échantillonnage de l'unité $k$ et $a_{kl} = 1/\pi_{kl}$ le poids d'échantillonnage conjoint des unités $k$ et $l$.

Un exemple de plan d'échantillonnage est le tirage aléatoire simple sans remise, noté plan SI. Pour ce plan, il faut tout d'abord identifier la taille d'échantillon souhaitée, que l'on note $n$. Par définition, sous le plan SI, les $\binom{N}{n}$ échantillons $s$ de cette taille auront la même probabilité d'être sélectionnés. Nous aurons donc

$$p(s) = \begin{cases} 1 \Big/ \binom{N}{n} & \text{si } s \text{ est de taille } n \\ 0 & \text{sinon} \end{cases}$$

Ceci entraîne que

$$\pi_k = \sum_{s \ni k} p(s) = \binom{N-1}{n-1} \Big/ \binom{N}{n} = \frac{n}{N} \quad \forall k \in U \tag{1.2.3}$$

et

$$\pi_{kl} = \sum_{s \ni k \& l} p(s) = \binom{N-2}{n-2} \Big/ \binom{N}{n} = \frac{n(n-1)}{N(N-1)} \quad \forall k \text{ et } l,\ k \neq l, \in U \tag{1.2.4}$$

Ici, nous nous intéresserons plus particulièrement au plan SI.

## 1.3. L'estimateur d'Horvitz-Thompson (HT)

Sous n'importe quel plan, l'estimateur de base de la théorie de l'échantillonnage est l'estimateur d'Horvitz-Thompson (noté HT),

$$\hat{\bar{y}}_{\mathrm{HT}} = \frac{1}{N}\sum\nolimits_s a_k y_k \tag{1.3.1}$$

où $a_k = \frac{1}{\pi_k}$. Cet estimateur est sans biais pour $\bar{y}_U$ et possède la variance

$$\mathbb{V}(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{1}{N^2}\sum\sum\nolimits_U (a_k a_l / a_{kl} - 1) y_k y_l \tag{1.3.2}$$

où $a_{kl} = \frac{1}{\pi_{kl}}$. Un estimateur sans biais de $\mathbb{V}(\hat{\bar{y}}_{\mathrm{HT}})$ est donné par

$$\widehat{\mathbb{V}}(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{1}{N^2}\sum\sum\nolimits_s (a_k a_l - a_{kl}) y_k y_l \tag{1.3.3}$$

Cet estimateur de variance est sans biais et toujours non négatif si

$$\forall k \neq l, \quad \pi_{kl} > 0 \quad et \quad (\pi_{kl} - \pi_k \pi_l) < 0$$

Pour le plan SI, en utilisant (1.2.3) et (1.2.4), il est facile de vérifier que l'estimateur sans biais de $\bar{y}_U$, la variance ainsi que son estimateur de variance sont donnés respectivement par

$$\hat{\bar{y}}_{\mathrm{HT}} = \frac{1}{N}\sum\nolimits_s a_k y_k = \frac{1}{N}\sum\nolimits_s \frac{N}{n}\, y_k = \bar{y}_s \tag{1.3.4}$$

$$\mathbb{V}(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{1-f}{n} S_{yU}^2 \quad \text{où} \quad S_{yU}^2 = \frac{1}{N-1}\sum\nolimits_U (y_k - \bar{y}_U)^2 \tag{1.3.5}$$

$$\widehat{\mathbb{V}}(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{1-f}{n} S_{ys}^2 \quad \text{où} \quad S_{ys}^2 = \frac{1}{n-1}\sum\nolimits_s (y_k - \bar{y}_s)^2 \tag{1.3.6}$$

## 1.4. L'estimateur par régression généralisé

Dans ce mémoire, nous étudierons l'utilisation de l'information auxiliaire dans un sondage visant à déterminer $\bar{y}_U$. L'information auxiliaire est exprimée à l'aide d'un vecteur auxiliaire noté $\vec{x}$. Sa valeur pour la $k^e$ unité est notée $\vec{x}_k$. Selon la disponibilité des ressources, nous distinguerons deux types de vecteurs auxiliaires : (a) un vecteur auxiliaire avec total connu si $\sum_U \vec{x}_k$ est connu; (b) un vecteur auxiliaire avec valeurs individuellement connues si $\vec{x}_1, \ldots, \vec{x}_N$ sont connues. La raison pour laquelle on veut distinguer ces deux cas est que fréquemment, dans un sondage, l'information $\sum_U \vec{x}_k$ provient d'une source extérieure au sondage, par exemple, un recensement de la population à une date antérieure.

L'information auxiliaire peut être utilisée soit pour la conception du plan d'échantillonnage et pour le tirage de l'échantillon soit pour l'estimation (dans la formule de l'estimateur), soit pour une combinaison des deux.

Un estimateur qui repose sur l'utilisation de l'information auxiliaire, sous la forme d'un total connu d'un vecteur auxiliaire, est l'estimateur par régression généralisé que l'on note GREG (voir Särndal, Swensson et Wretman (1992)). Sa forme générale est donnée par

$$\hat{\bar{y}}_{\text{GREG}} = \frac{1}{N}\left\{(\textstyle\sum_U \vec{x}_k)'\hat{\vec{R}} + \sum_s a_k e_k\right\} \tag{1.4.1}$$

où

$$e_k = y_k - \vec{x}_k'\hat{\vec{R}} \tag{1.4.2}$$

avec

$$\hat{\vec{R}} = \left(\sum_s a_k \vec{x}_k \vec{x}_k'/c_k\right)^{-1} \sum_s a_k \vec{x}_k y_k/c_k \tag{1.4.3}$$

et $\sum_U \vec{x}_k$ est le total connu du vecteur auxiliaire. Le vecteur $\vec{x}_k$ et la valeur $y_k$ sont disponibles pour $k \in s$. Les constantes $c_k$ sont connues $\forall k$ et spécifiées par le statisticien. On peut toujours utiliser $c_k = 1, \forall k$; autrement, les $c_k$ donnent la possibilité de pondérer les observations. Notons qu'une représentation équivalente de (1.4.1) est

$$\hat{\bar{y}}_{\text{GREG}} = \frac{1}{N}\left\{\sum_s a_k y_k + (\sum_U \vec{x}_k - \sum_s a_k \vec{x}_k)' \hat{\vec{R}}\right\} \tag{1.4.4}$$

Cette représentation montre que $\hat{\bar{y}}_{\text{GREG}}$ consiste de l'estimateur HT de $\bar{y}_U$ (voir 1.3.1) plus un terme d'ajustement, corrélé négativement avec l'estimateur HT. Par conséquent, l'estimateur par régression généralisé est davantage plus précis que l'estimateur HT.

On ne dispose pas d'une expression simple pour la variance de (1.4.1). Cependant, la linéarisation par le développement en série de Taylor fournit une excellente approximation de la variance, soit

$$A\mathbb{V}(\hat{\bar{y}}_{\text{GREG}}) = \frac{1}{N^2} \sum\sum_U (a_k a_l/a_{kl} - 1) E_k E_l \tag{1.4.5}$$

où $E_k = y_k - \vec{x}_k' \vec{R}$ avec $\vec{R} = \left(\sum_U \vec{x}_k \vec{x}_k'/c_k\right)^{-1} \sum_U \vec{x}_k y_k/c_k$.

Un estimateur de la variance (voir Särndal, Swensson et Wretman (1992), chap. 6) est donné par

$$\hat{\mathbb{V}}(\hat{\bar{y}}_{\text{GREG}}) = \frac{1}{N^2} \sum\sum_s (a_k a_l - a_{kl})(g_k e_k)(g_l e_l) \tag{1.4.6}$$

où

$$g_k = 1 + (\sum_U \vec{x}_k - \sum_s a_k \vec{x}_k)'(\sum_s a_k \vec{x}_k \vec{x}_k'/c_k)^{-1} \vec{x}_k/c_k \tag{1.4.7}$$

et où $e_k$ est donné par l'équation (1.4.2). à l'aide des poids $g_k$ ci-dessus, nous pouvons obtenir à niveau une représentation équivalente de l'estimateur (1.4.1). Nous avons en effet

$$\hat{\bar{y}}_{\text{GREG}} = \frac{1}{N} \sum_s a_k g_k y_k \tag{1.4.8}$$

Les spécifications suivantes de $\vec{x}_k$ et de $c_k$ seront étudiées en profondeur dans ce mémoire.

**1)** Pour un plan tel que $\sum_s a_k = N$ (par exemple, le plan SI), on a, lorsque $\vec{x}_k = 1$ et $c_k = 1$, $\forall k$, que l'estimateur GREG donné par (1.4.1) devient identique à l'estimateur HT donné par (1.3.4). Il s'ensuit que l'estimateur de la variance de cet estimateur est alors donné par (1.3.6).

**2)** Pour le plan SI, lorsque $\vec{x}_k = x_k$ et $c_k = x_k$, l'estimateur GREG donné par (1.4.1) devient

$$\hat{\bar{y}}_{\text{RA}} = \bar{x}_U \frac{\bar{y}_s}{\bar{x}_s} \tag{1.4.9}$$

On l'appelle "estimateur par le ratio". Si l'on se sert de la forme pondérée donnée par (1.4.8), les poids sont $g_k = \bar{x}_U/\bar{x}_s$, $\forall k \in s$ . La variance approximative (1.4.5) devient

$$\text{A}\mathbb{V}(\hat{\bar{y}}_{\text{RA}}) = \frac{1-f}{n} \frac{1}{N-1} \sum_U (\, y_k - R x_k \,)^2 \tag{1.4.10}$$

où $R = \sum_U y_k / \sum_U x_k$. L'estimateur de la variance (1.4.6) peut s'écrire

$$\hat{\mathbb{V}}(\hat{\bar{y}}_{\text{RA}}) = (\frac{\bar{x}_U}{\bar{x}_s})^2 \frac{1-f}{n} \frac{\sum_s (\, y_k - \hat{R} x_k \,)^2}{n-1} \tag{1.4.11}$$

où $\hat{R} = \sum_s y_k / \sum_s x_k$.

Dans les sections 5.2.2 et 5.4.2, nous continuons à travailler avec le cas $\vec{x}_k = x_k = c_k$.

**3)** Pour le plan SI, lorsque $\vec{x}_k = (1, x_k)'$ et $c_k = 1\ \forall k$, l'estimateur GREG donné par (1.4.1) devient

$$\hat{\bar{y}}_{\text{REG}} = \bar{y}_s + \hat{R}(\bar{x}_U - \bar{x}_s) \tag{1.4.12}$$

où $\hat{R} = \sum_s (x_k - \bar{x}_s)(y_k - \bar{y}_s)/\sum_s (x_k - \bar{x}_s)^2$.

On l'appelle "l'estimateur par la régression simple". Si l'on se sert de la forme pondérée donnée par (1.4.8), les poids sont $g_k = 1 + v_s(x_k - \bar{x}_s)$, où $v_s = n(\bar{x}_U - \bar{x}_s)/\sum_s (x_k - \bar{x}_s)^2, \forall k \in s$. La variance approximative (1.4.5) et l'estimateur de variance (1.4.6) deviennent respectivement

$$A\mathbb{V}(\hat{\bar{y}}_{\text{REG}}) = \frac{1-f}{n} \frac{1}{N-1} \sum_U E_k^2 \tag{1.4.13}$$

où $E_k = y_k - \bar{y}_U - R(x_k - \bar{x}_U)$ avec $R = \frac{\sum_U (x_k - \bar{x}_U)(y_k - \bar{y}_U)}{\sum_U (x_k - \bar{x}_U)^2}$ et

$$\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{REG}}) = \frac{1-f}{n} \frac{1}{n-1} \sum_s [1 + v_s(x_k - \bar{x}_s)]^2 e_k^2 \tag{1.4.14}$$

où $e_k = y_k - \bar{y}_s - \widehat{R}(x_k - \bar{x}_s)$ avec $\widehat{R} = \frac{\sum_s (x_k - \bar{x}_s)(y_k - \bar{y}_s)}{\sum_s (x_k - \bar{x}_s)^2}$.

Dans les sections 5.2.3 et 5.4.3, nous continuons à travailler avec le cas $\vec{x}_k = (1, x_k)'$ et $c_k = 1, \forall k$. Il est important de souligner la nécessité de connaître le total $\sum_U \vec{x}_k = \sum_U (1, x_k)' = (N, \sum_U x_k)'$ pour le calcul de l'estimateur par la régression simple.

## 1.5. Non-réponse et mécanisme de réponse

Étant donné le taux élevé de non-réponse que l'on observe à maintes reprises dans les sondages, l'imputation est une méthode populairement utilisée, surtout dans les enquêtes d'entreprises. Autrefois, la philosophie voulait que les conséquences de l'imputation soient considérées minimes, voire négligeables. Cette affirmation demeure peut-être vraie pour un taux de non-réponse de 1 à 2 pour cent, mais elle s'avère totalement fausse pour un taux de 30 à 40 pour cent et plus que nous observons souvent dans les sondages.

Plusieurs méthodes d'imputation ont été proposées au cours des années dont l'imputation simple et l'imputation multiple. L'imputation simple est, comme son nom l'indique, la substitution d'une seule valeur à chacune des valeurs manquantes. Sur ce, nous ne disposons que d'un seul ensemble de données complétées à partir duquel les estimateurs ponctuels et les estimateurs de la variance doivent être calculés. L'imputation multiple, quant à elle, consiste à imputer plusieurs valeurs, ce qui produit plusieurs ensembles de données complétées. Elle permet l'utilisation des méthodes habituelles d'estimation de variance en les appliquant sur chacun des ensembles de données pour ensuite combiner les résultats. Ceci permet l'obtention d'une estimation de la variance totale, incluant la variance résultant de l'imputation. La variabilité à l'intérieur et entre les ensembles est mesurée à partir de la théorie de l'imputation multiple. Cependant en pratique, l'imputation multiple entraîne des inconvénients, puisque entreposer et maintenir ces ensembles engendrent des coûts. Dans ce mémoire, nous opterons plutôt pour l'imputation simple.

Passons maintenant au mécanisme de réponse. L'échantillon sélectionné $s$, est affecté par la non-réponse, c'est-à-dire que pour certaines des unités choisies, nous ne pouvons observer la valeur de la variable d'étude. L'échantillon $s$ se trouve alors divisé en deux groupes : le groupe des répondants, noté $r$, et celui des non-répondants, dénoté $s - r = o$. La composition de ces groupes est induite par une loi probabiliste appelée "mécanisme de réponse" noté $\mathrm{q}(\bullet|s)$, $\mathrm{q}(r|s)$ étant la probabilité d'obtenir le groupe de répondants $r$, si l'échantillon $s$ a été tiré. Normalement, le mécanisme $\mathrm{q}(r|s)$ est une loi inconnue.

De ce mécanisme découle des probabilités de réponse $\theta_{k|s}$=P(k $\in r|s$ est tiré), $\forall k \in s$. Le mécanisme de réponse sera qualifié de *non confondu* si la probabilité d'obtenir le groupe de répondants $r$, étant donné que l'échantillon $s$ a été tiré, ne dépend pas de la variable d'étude. En d'autres termes, les valeurs prises par la variable d'étude $y$ sont entièrement indépendantes des probabilités de réponse $\theta_{k|s}$. Notons cependant que les probabilités $\theta_{k|s}$, associées à un mécanisme non confondu, pourraient dépendre d'une variable auxiliaire $z$.

Le *mécanisme uniforme* est un exemple simple de mécanisme de réponse non confondu. Ses caractéristiques sont que $\theta_{k|s} = \theta$, $\forall k \in s$ et que les unités répondent indépendamment. Nous obtiendrons alors

$$q(r|s) = \theta^m(1-\theta)^{n-m} \qquad (1.5.1)$$

pour $m = 0, 1, ..., n$, où $m$ désigne la taille de $r$. Nous désignerons par *taux de réponse échantillonnal* la proportion des répondants dans l'échantillon, soit $\frac{m}{n}$.

Nous allons également travailler avec un mécanisme de non-réponse dont les probabilités de réponse sont de la forme

$$\theta_k = \exp^{-c \cdot z_k} \tag{1.5.2}$$

où $c$ est une constante positive et $z_k$ est la valeur pour l'unité $k$ de la variable auxiliaire $z$. Pour ce mécanisme, nous supposons que la probabilité de réponse dépend de $k$, et non de $s$, et que les unités répondent indépendamment les unes des autres. Nous verrons à la fin du chapitre 4 les valeurs attribuées à la constante $c$ pour les différents taux de réponse étudiés dans le présent ouvrage.

## 1.6. Données complétées

L'imputation est la procédure qui consiste à substituer des valeurs artificielles $\hat{y}_k$ aux données manquantes $y_k$. Le fait d'imputer nous donne un jeu de données complétées $\{ y_{\bullet k} : k \in s \}$, où

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \text{ (répondants)} \\ \hat{y}_k & \text{si } k \in o \text{ (non-répondants)} \end{cases}$$

On utilise les données complétées pour le calcul des estimations ponctuelles, plus précisément *l'estimateur HT imputé* et *l'estimateur GREG imputé* qui seront obtenus à partir des formules prévues pour le cas de 100 % de réponse, soit (1.3.1) et (1.4.8) respectivement, et en les calculant sur les données complétées. Nous obtenons ainsi

$$\hat{\bar{y}}_{\bullet \mathrm{HT}} = \frac{1}{N} \sum\nolimits_s a_k y_{\bullet k} = \frac{1}{N} \left( \sum\nolimits_r a_l y_l + \sum\nolimits_o a_k \hat{y}_k \right) \tag{1.6.1}$$

$$\hat{\bar{y}}_{\bullet \mathrm{GREG}} = \frac{1}{N} \sum\nolimits_s a_k g_k y_{\bullet k} = \frac{1}{N} \left( \sum\nolimits_r a_l g_l y_l + \sum\nolimits_o a_k g_k \hat{y}_k \right) \tag{1.6.2}$$

Cependant, comme la valeur imputée ne correspond pas à la valeur réelle $y_k$ (sauf dans des circonstances exceptionnelles), l'imputation a un impact sur les estimations et sur leur précision puisqu'elle a pour effet d'accroître la variance du total estimé. Il est alors important de bien mesurer cet impact pour connaître la qualité réelle des estimations.

Un des problèmes est qu'un estimateur sans biais en l'absence de non-réponse peut être sévèrement biaisé pour différentes formes d'imputation et sous certains mécanismes de réponse.

De plus, si on utilise les méthodes habituelles pour estimer la variance, on ne tient pas compte du fait que des données ont été imputées, c'est- à-dire que l'on traite les données imputées comme des observations réelles. Dans cette situation, la variance sera le plus souvent sous-estimée, car nous ne calculons que la variance résultant de l'échantillonnage. N'oublions pas que la présence de données artificielles perturbe l'estimation de la variance. Pour répondre à ces problèmes, l'estimation de la variance causée par l'imputation se révèle un moyen de premier choix.

## 1.7. Quelques méthodes d'imputation

Il existe plusieurs méthodes d'imputation. Voici une brève description de certaines des méthodes les plus populaires.

Hot Deck : Pour chacune des unités $k \in o$, un donneur $\ell(k)$ est tiré aléatoirement parmi le groupe des répondants $r$, et la valeur de la variable $y$ du donneur est imputée. Nous avons alors $\hat{y}_k = y_{\ell(k)}$. Une variante consiste à tirer le donneur $\ell(k)$ uniquement parmi les

répondants qui précèdent $k$ dans la liste, ceci dans le but d'accélérer le processus d'imputation (voir Provost, Martin (1995)).

Cold Deck : Le donneur est tiré dans une autre source de données, par exemple l'enquête précédente, s'il s'agit d'une enquête répétée périodiquement.

Plus proche voisin : Cette méthode se sert d'une ou plusieurs variables auxiliaires pour l'identification d'un donneur. Dans ce mémoire, nous n'utiliserons qu'une seule variable auxiliaire. Elle est unidimensionnelle et est identifiée par $z$. Une mesure de distance $d(\bullet, \bullet)$ est spécifiée. Ensuite, pour une unité non répondante $k \in o$, le donneur sélectionné est le répondant $\ell$ pour lequel $d(z_k, z_l)$ est minimale parmi tous les $\ell \in r$. La valeur de $y$ pour le donneur est alors imputée, nous avons ainsi $\hat{y}_k = y_{\ell(k)}$. Cette méthode sera décrite à la section 3.2.

Régression : Il faut présupposer l'existence d'un modèle de régression. Les paramètres de ce modèle sont alors calculés à l'aide de l'ensemble des répondants. Ensuite, les données manquantes sont remplacées par leur prédiction selon le modèle de régression. La pertinence de l'imputation dépend directement de la qualité de l'ajustement du modèle.

Dans le cas de l'imputation par régression, on essaie de créer des imputations $\hat{y}_k$ à partir du modèle

$$y_k = \beta' \vec{z}_k + \epsilon_k \tag{1.7.1}$$

où $\vec{z}_k$ est un vecteur de prédicteurs connus pour $k \in s$.

La forme la plus simple d'imputation par régression est celle par la moyenne des répondants. Le modèle présupposé est alors

$$y_k = \mu + \epsilon_k \tag{1.7.2}$$

Ici, l'estimateur naturel de $\mu$ est $\hat{\mu} = \bar{y}_r$, où $\bar{y}_r = \frac{1}{m}\sum_r y_k$, la moyenne des répondants. Alors, pour $k \in o$, la valeur imputée est $\hat{y}_k = \bar{y}_r$. Comme le modèle de régression est très élémentaire, la qualité de cette imputation est souvent remise en question.

## 1.8. Survol de la littérature

Le livre "Model Assisted Survey Sampling" par Särndal, Swensson et Wretman (1992) est un excellent livre pour approfondir les notions mentionnés dans ce chapitre.

De plus, nous nous sommes basés sur plusieurs articles pour examiner les divers aspects de l'imputation par le plus proche voisin et les propriétés des estimateurs qui en découlent. Les articles les plus importants sont Rancourt, Lee et Särndal (1993), Rancourt, Särndal et Lee (1994), Deville et Särndal (1994), Rancourt (1996) et Gagnon, Lee, Provost, Rancourt et Särndal (1997).

CHAPITRE 2

# Imputation par la moyenne des répondants

## 2.1. Introduction

Avant d'étudier l'imputation par le plus proche voisin, nous allons étudier, dans un premier temps, l'imputation par la moyenne des répondants. Il est utile de faire d'abord appel à cette forme d'imputation élémentaire, car cette étude nous introduira les astuces mathématiques pour le calcul d'espérance et de variance. Ce faisant, nous nous en servirons par la suite pour bâtir nos estimateurs de variance selon la méthode d'imputation par le plus proche voisin.

## 2.2. Processus d'imputation

Dans ce chapitre, nous considérons le plan d'échantillonnage SI, pour le tirage de l'échantillon (n unités parmi N), et un mécanisme de réponse uniforme, pour lequel $\theta_{k|s} = \theta$, $\forall k \in s$ et pour tout $s$ et dans lequel les unités répondent de façon indépendante (voir équation 1.5.1).

Pour le plan SI et un taux de réponse de 100 %, l'estimateur habituel de la moyenne de la population est $\bar{y}_s$, la moyenne échantillonnale.

L'espérance et la variance de cet estimateur sont données respectivement par

$$\begin{aligned} \mathbb{E}_{\mathrm{SI}}(\bar{y}_s) &= \bar{y}_U \\ \mathbb{V}_{\mathrm{SI}}(\bar{y}_s) &= \left(\frac{1}{n} - \frac{1}{N}\right) S_{y_U}^2 \end{aligned}$$

L'imputation par la moyenne consiste à affecter la valeur de la moyenne de la variable d'étude sur l'ensemble des répondants, soit $\bar{y}_r$, à chacun des non-répondants. Ainsi, nous obtenons comme données complétées l'ensemble $\{y_{\bullet k} : k \in s\}$ où

$$y_{\bullet k} = \begin{cases} y_k & k \in r \\ \bar{y}_r & k \in o \end{cases} \tag{2.2.1}$$

Notons qu'avec cette méthode d'imputation, la moyenne échantillonnale imputée prend la forme

$$\begin{aligned} \bar{y}_{\bullet s} &= \frac{1}{n}\sum\nolimits_s y_{\bullet k} = \frac{1}{n}\left(\sum\nolimits_r y_k + \sum\nolimits_{s-r} \bar{y}_r\right) \\ &= \frac{1}{n}(m\bar{y}_r + (n-m)\bar{y}_r) = \bar{y}_r \end{aligned}$$

c'est-à-dire, dans ce cas-ci, l'estimateur imputé équivaut à la moyenne des répondants.

## 2.3. L'espérance mathématique de l'estimateur imputé

À la section précédente, nous avons constaté que la moyenne échantillonnale du jeu de données complétées par la méthode de l'imputation par la moyenne des répondants donne simplement $\bar{y}_{\bullet s} = \bar{y}_r$ où $\bar{y}_r$ est

donné par

$$\bar{y}_r = \begin{cases} \frac{1}{m}\sum_r y_k & m \neq 0 \\ C & m = 0 \end{cases}$$

où $C$ est une constante fixée et $m$, le nombre de répondants.

Supposant un plan d'échantillonnage SI, un mécanisme de réponse uniforme et $m \neq 0$, nous allons maintenant démontrer qu'il s'agit d'un estimateur sans biais. À cette fin, il faut calculer l'espérance $\mathbb{E}(\bar{y}_r)$.

Toutefois, vu que l'on utilise un certain plan d'échantillonnage $p$ pour tirer l'échantillon et que nous supposons qu'il existe un certain mécanisme de réponse $q$ régissant la loi de $r$, étant donné $s$, alors l'espérance $\mathbb{E}(\bar{y}_r)$ pourrait, en général, s'écrire

$$\mathbb{E}(\bar{y}_r) = \mathbb{E}_{\mathrm{p}}\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s)$$

Plus particulièrement, dans notre cas, $p$ est le plan SI et $q$ est le mécanisme uniforme donné par (1.5.1). Sous ces conditions, nous allons démontrer à l'aide du lemme suivant que $\mathbb{E}_{\mathrm{p}}\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s) = \bar{y}_U$.

LEMME 2.3.1. *Lorsque l'échantillon $s$ et le nombre de répondants $m \geq 1$ sont fixés, le mécanisme de réponse uniforme donné par* (1.5.1) *se ramène à un tirage SI de $m$ unités parmi les $n$ unités de $s$.*

Démonstration. Il y a $\binom{n}{m}$ sous-ensembles $r$ de $s$ de taille $m$. En utilisant (1.5.1), chacun de ces sous-ensembles a la même probabilité. Donc, pour $m \geq 1$,

$$P(r|s,m) = \frac{1}{\binom{n}{m}}$$

ce qui est conforme à un plan SI pour un tirage de $m$ unités parmi $n$. □

Il s'ensuit du lemme (2.3.1) que

$$\begin{aligned} \mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m,m\geq 1) &= \bar{y}_s \\ \mathbb{V}_{\mathrm{q}}(\bar{y}_r|s,m,m\geq 1) &= (\frac{1}{m}-\frac{1}{n})S_{ys}^2 \end{aligned} \tag{2.3.1}$$

Nous pouvons maintenant facilement démontrer que $\bar{y}_r$ est sans biais pour $\bar{y}_U$ conditionnellement à $m \geq 1$. Indiquons par l'indice $m$ l'espérance par rapport à la loi du nombre de répondants. Par le fait que $\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m)$ ne dépend aucunement de $m$, nous pouvons affirmer que sous le plan SI,

$$\begin{aligned} \mathbb{E}_{\mathrm{p}}\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m\geq 1) &= \mathbb{E}_{\mathrm{p}}\mathbb{E}_{\mathrm{m}}\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m,m\geq 1) = \mathbb{E}_{\mathrm{p}}\mathbb{E}_{\mathrm{m}}(\bar{y}_s|m\geq 1) \\ &= \mathbb{E}_{\mathrm{p}}(\bar{y}_s) = \bar{y}_U \end{aligned}$$

Dans le cas où le nombre de répondants $m$ pourrait être nul, nous aurions sous le mécanisme de réponse (1.5.1)

$$\begin{aligned} \mathbb{E}_{\mathrm{m}}\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m) &= P(m>0)\cdot\bar{y}_s + P(m=0)\cdot C \\ &= \bar{y}_s + P(m=0)\cdot C \\ &= \bar{y}_s + (1-\theta)^n\cdot(C-\bar{y}_s) \end{aligned} \tag{2.3.2}$$

Mais nous savons qu'en pratique, $n$ est presque sans exception assez grand et que $\theta$ n'est pas très près de zéro, ce qui implique que $P(m=0)$ est négligeable. Notons que pour le reste du mémoire, toute espérance par rapport au mécanisme de réponse sera conditionnelle au fait que le nombre de répondants $m$ est plus grand ou égal à 1, mais afin d'alléger la notation cette condition ne sera pas indiquée explicitement.

## 2.4. Estimation de la variance de l'estimateur de moyenne

Sous une approche par modèle de régression, il a été démontré à maintes reprises (voir entre autres Särndal (1992) ou Deville et Särndal (1994)) que la variance de $\bar{y}_r$ pouvait se décomposer comme

$$\mathbb{V}_{\text{TOT}} = \mathbb{V}_{\text{ECH}} + \mathbb{V}_{\text{IMP}}$$

où $\mathbb{V}_{\text{ECH}}$ désigne la part de la variance totale résultant de l'échantillonnage et $\mathbb{V}_{\text{IMP}}$ celle découlant de l'imputation. De plus, $\mathbb{V}_{\text{TOT}}$ peut s'estimer par $\widehat{\mathbb{V}}_{\text{TOT}} = \widehat{\mathbb{V}}_{\text{ECH}} + \widehat{\mathbb{V}}_{\text{IMP}}$, où

$$\widehat{\mathbb{V}}_{\text{ECH}} = \left(\frac{1}{n} - \frac{1}{N}\right) S_{yr}^2 \tag{2.4.1}$$

avec

$$S_{yr}^2 = \frac{1}{m-1} \sum\nolimits_r (y_k - \bar{y}_r)^2 \tag{2.4.2}$$

et

$$\widehat{\mathbb{V}}_{\text{IMP}} = \left(\frac{1}{m} - \frac{1}{n}\right) S_{yr}^2 \tag{2.4.3}$$

d'où

$$\widehat{\mathbb{V}}_{\text{TOT}} = \left(\frac{1}{m} - \frac{1}{N}\right) S_{yr}^2 \tag{2.4.4}$$

Nous allons maintenant démontrer que, sous le plan SI et sous le mécanisme de réponse uniforme, $\widehat{\mathbb{V}}_{\text{TOT}}$ est sans biais pour $\mathbb{V}_{\text{TOT}}$.

Posons $\mathbb{V}_{\text{TOT}} = \mathbb{V}_{\text{pq}}(\bar{y}_r)$, la variance totale de l'estimateur de moyenne, par rapport au plan d'échantillonnage $p$ et au mécanisme de réponse $q$. Cette variance peut se décomposer en deux termes

$$\mathbb{V}_{\text{TOT}} = \mathbb{V}_{\text{p}}[\mathbb{E}_{\text{q}}(\bar{y}_r|s)] + \mathbb{E}_{\text{p}}[\mathbb{V}_{\text{q}}(\bar{y}_r|s)] \tag{2.4.5}$$

où $\mathbb{V}_{\mathrm{p}}[\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s)]$ est la variance résultant de l'échantillonnage, que l'on dénote $\mathbb{V}_{\mathrm{ECH}}$. En supposant un plan SI et un mécanisme uniforme et en utilisant (2.3.1), on a donc

$$\begin{aligned}\mathbb{V}_{\mathrm{ECH}} = \mathbb{V}_{\mathrm{p}}[\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s)] &= \mathbb{V}_{\mathrm{p}}[\mathbb{E}_{\mathrm{m}}\{\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m)\}] \\ &= \mathbb{V}_{\mathrm{p}}[\bar{y}_s] \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2\end{aligned}$$

Le terme $\mathbb{E}_{\mathrm{p}}[\mathbb{V}_{\mathrm{q}}(\bar{y}_r|s)]$ dans (2.4.5) est la variance causée par l'imputation, que l'on dénote $\mathbb{V}_{\mathrm{IMP}}$. Elle se décompose de la manière suivante :

$$\mathbb{E}_{\mathrm{p}}[\mathbb{V}_{\mathrm{q}}(\bar{y}_r|s)] = \mathbb{E}_{\mathrm{p}}[\mathbb{E}_{\mathrm{m}}[\mathbb{V}_{\mathrm{q}}(\bar{y}_r|s,m)] + \mathbb{V}_{\mathrm{m}}\{\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m)\}] \tag{2.4.6}$$

Notons que

$$\mathbb{V}_{\mathrm{m}}\{\mathbb{E}_{\mathrm{q}}(\bar{y}_r|s,m)\} = \mathbb{V}_{\mathrm{m}}(\bar{y}_s) = 0$$

car $\bar{y}_s$ ne dépend pas de $m$. Donc, à chaque fois que nous nous servirons de la représentation de la variance (2.4.6), nous imposerons une valeur nulle à la deuxième composante. Dans le but d'alléger le texte, nous éviterons de mentionner ce terme dans les sections et chapitres suivants. Il s'ensuit que

$$\begin{aligned}\mathbb{V}_{\mathrm{IMP}} &= \mathbb{E}_{\mathrm{p}}\left[\mathbb{E}_{\mathrm{m}}\left(\frac{1}{m} - \frac{1}{n}\right) S_{ys}^2\right] \\ &= \left[\mathbb{E}_{\mathrm{m}}\left(\frac{1}{m}\right) - \frac{1}{n}\right] S_{yU}^2\end{aligned}$$

Le passage à la dernière ligne est devenu possible, puisque le mécanisme de réponse est uniforme et que les probabilités de réponse ne varient pas en fonction de l'échantillon.

Nous allons maintenant démontrer que $\widehat{\mathbb{V}}_{\text{ECH}}$ et $\widehat{\mathbb{V}}_{\text{IMP}}$, tels qu'ils sont donnés par les formules (2.4.1) et (2.4.3), sont des estimateurs sans biais de $\mathbb{V}_{\text{ECH}}$ et $\mathbb{V}_{\text{IMP}}$ respectivement, en supposant toujours le plan SI, le mécanisme uniforme et $m \neq 0$. Pour ce faire, nous allons utiliser le fait suivant :

$$\begin{aligned}
\mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}(S_{yr}^2|s) &= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{m}}\mathbb{E}_{\text{q}}(S_{yr}^2|s,m) \\
&= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{m}}(S_{ys}^2) \\
&= \mathbb{E}_{\text{p}}(S_{ys}^2) = S_{yU}^2
\end{aligned}$$

Nous obtenons finalement

$$\mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}(\widehat{\mathbb{V}}_{\text{ECH}}|s) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2 = \mathbb{V}_{\text{ECH}}$$

et

$$\begin{aligned}
\mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}(\widehat{\mathbb{V}}_{\text{IMP}}|s) &= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{m}}\mathbb{E}_{\text{q}}\left[\left(\frac{1}{m} - \frac{1}{n}\right) S_{yr}^2 \mid s,m\right] \\
&= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{m}}\left[\left(\frac{1}{m} - \frac{1}{n}\right) S_{ys}^2\right] \\
&= \mathbb{V}_{\text{IMP}}
\end{aligned}$$

Nous concluons que $\widehat{\mathbb{V}}_{\text{ECH}}$ et $\widehat{\mathbb{V}}_{\text{IMP}}$ sont bel et bien sans biais et il s'ensuit que $\widehat{\mathbb{V}}_{\text{TOT}} = \widehat{\mathbb{V}}_{\text{ECH}} + \widehat{\mathbb{V}}_{\text{IMP}}$ est sans biais pour $\mathbb{V}_{\text{TOT}}$.

CHAPITRE 3

# Décomposition de la variance

## 3.1. Introduction

La méthode dite *imputation par le plus proche voisin* est utilisée pour combler la non-réponse de plusieurs sondages de Statistique Canada. Par la suite, nous la désignerons par "la méthode PPV". Cette méthode consiste à choisir de manière optimale un donneur parmi les répondants pour chaque unité non-répondante. Alors la donnée imputée sera $\hat{y}_k = y_{\ell(k)}$, où $y_{\ell(k)}$ est la valeur de la variable d'étude pour le donneur $\ell(k)$. De plus, le choix du donneur se fait indépendamment d'un non-répondant à l'autre. Contrairement à l'imputation par la régression, cette valeur existe pour au moins un répondant. L'imputation par la méthode PPV peut ne pas être la substitution parfaite pour les non-répondants, mais son approche est logique.

En assumant une relation linéaire entre le variable d'étude $y$ et les variables concomitantes utilisées pour l'identification du plus proche voisin, cette méthode donne normalement des estimateurs avec un biais léger ou même négligeable.

Nous savons pertinemment qu'une variance calculée à partir des données complétées sous-estime sévèrement la vraie valeur de celle-ci. Fondamentalement, il existe trois approches pour estimer la variance

en présence d'imputation. La plus ancienne, et probablement la mieux connue, est la méthode d'imputation multiple préconisée par Rubin (1977) et Rubin (1987). Il y a eu également l'approche assistée par un modèle (Särndal (1990)) et la technique du jackknife (Rao (1991)). Certains aspects de ces approches ont été testés pour la méthode PPV par différents auteurs, sans toutefois donner un franc succès.

La difficulté rencontrée lors de l'utilisation de la méthode proposée par Rubin était de définir l'imputation multiple appropriée pour la méthode PPV, ce qui s'est traduit par une sous-estimation de la variance (voir Rancourt, Särndal et Lee (1994)). Ces mêmes auteurs ont également utilisé l'approche assistée par un modèle, supposant que la formule pour l'imputation par le ratio pouvait s'appliquer à l'imputation par le plus proche voisin. Cette procédure s'est avérée supérieure à l'approche multiple, sans toutefois enrayer le problème du biais négatif non négligeable (voir Rancourt et al., 1994).

Dans ce mémoire, nous allons développer une technique d'estimation de la variance appropriée pour l'imputation par la méthode PPV. Nous utiliserons l'approche assistée par un modèle, ce qui nous conduira à une estimation valide de la variance lorsque la variable d'étude $y$ et la variable concomitante $z$ sont reliées par une régression linéaire passant par l'origine.

## 3.2. Approche assistée par un modèle

Considérons une valeur imputée par la méthode PPV. En supposant que, pour une unité $k \in o$,

$$\min_{\ell \in r} |z_l - z_k|$$

est atteint pour $\ell = \ell(k)$. Alors la valeur manquante $y_k$ est imputée par la valeur $y_{\ell(k)}$, où l'unité $\ell(k)$ est appelée "le donneur" pour l'unité $k$. Le jeu de données complétées est alors l'ensemble $\{y_{\bullet k} : k \in s\}$ où

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \text{ (répondants)} \\ y_{\ell(k)} & \text{si } k \in o \text{ (non-répondants)} \end{cases}$$

Si nous observons 100 % de réponse, l'estimateur de la moyenne de la population et l'estimateur de la variance correspondant sont donnés respectivement par les formules (1.4.8) et (1.4.6). Comme nous l'avons déjà relevé à la section 1.6, une façon de calculer l'estimateur ponctuel en présence de non-réponse est de se servir de la formule prévue pour 100 % de réponse et de l'appliquer sur le jeu de données complétées. Nous obtenons alors

$$\hat{\bar{y}}_{\bullet \text{GREG}} = \frac{1}{N} \sum\nolimits_s a_k g_k y_{\bullet k}$$

Si l'on suit le même principe pour l'estimateur de la variance correspondant, on aura, en utilisant (1.4.6),

$$\hat{\mathbb{V}}(\hat{\bar{y}}_{\bullet \text{GREG}}) = \frac{1}{N^2} \sum \sum\nolimits_s (a_k a_l - a_{kl})(g_k e_{\bullet k})(g_l e_{\bullet l})$$

Cette façon de procéder est plutôt naïve pour évaluer la variance en présence de non-réponse, car elle ne traduit tout au plus que la variance résultant de l'échantillonnage. Il ne faut pas oublier que l'imputation

ajoute de la variance. Nous présenterons par la suite une approche qui tient compte de la composante de variance causée par l'imputation par la méthode PPV.

## 3.3. Les composantes de la variance

L'erreur totale de $\hat{\bar{y}}_{\bullet\text{GREG}}$ se décompose en erreur d'échantillonnage et en erreur d'imputation,

$$\hat{\bar{y}}_{\bullet\text{GREG}} - \bar{y}_U = (\hat{\bar{y}}_{\text{GREG}} - \bar{y}_U) + (\hat{\bar{y}}_{\bullet\text{GREG}} - \hat{\bar{y}}_{\text{GREG}}) \tag{3.3.1}$$

où $\hat{\bar{y}}_{\text{GREG}}$ et $\hat{\bar{y}}_{\bullet\text{GREG}}$ sont donnés respectivement par les formules (1.4.8) et (1.6.2). La différence $\hat{\bar{y}}_{\text{GREG}} - \bar{y}_U$ est appelée l'erreur due à l'échantillonnage; la différence $\hat{\bar{y}}_{\bullet\text{GREG}} - \hat{\bar{y}}_{\text{GREG}}$ est appelée l'erreur due à l'imputation.

Puisque $\mathbb{E}_{\text{p}}(\hat{\bar{y}}_{\text{GREG}} - \bar{y}_U) \approx 0$ (voir Särndal, Swensson et Wretman (1992), chap. 6), il s'ensuit que le biais de $\hat{\bar{y}}_{\bullet\text{GREG}}$ est donné approximativement par

$$B(\hat{\bar{y}}_{\bullet\text{GREG}}) \approx \mathbb{E}_{\text{p}}(B_{\text{cond}}) \tag{3.3.2}$$

où $B_{\text{cond}} = \mathbb{E}_{\text{q}}\{\hat{\bar{y}}_{\bullet\text{GREG}} - \hat{\bar{y}}_{\text{GREG}}|s\}$.

Il s'ensuit de (3.3.1) que l'erreur quadratique moyenne (EQM) de $\hat{\bar{y}}_{\bullet\text{GREG}}$, dénotée par $\text{EQM}_{pq}(\hat{\bar{y}}_{\bullet\text{GREG}})$, est donnée par

$$\mathrm{EQM}_{\mathrm{pq}}(\hat{\bar{y}}_{\bullet\mathrm{GREG}}) = \mathbb{E}_{\mathrm{p}}\mathbb{E}_{\mathrm{q}}(\hat{\bar{y}}_{\bullet\mathrm{GREG}} - \bar{y}_U)^2 = \mathbb{V}_{\mathrm{ECH}} + \mathbb{V}_{\mathrm{IMP}} + 2\mathbb{V}_{\mathrm{MIX}} \tag{3.3.3}$$

où

$$\mathbb{V}_{\mathrm{ECH}} = \mathbb{E}_{\mathrm{p}}(\,\hat{\bar{y}}_{\mathrm{GREG}} - \bar{y}_U\,)^2 \tag{3.3.4}$$

appelée la variance échantillonnale,

$$\mathbb{V}_{\mathrm{IMP}} = \mathbb{E}_{\mathrm{p}}\mathbb{E}_{\mathrm{q}}(\hat{\bar{y}}_{\bullet\mathrm{GREG}} - \hat{\bar{y}}_{\mathrm{GREG}})^2 \tag{3.3.5}$$

appelée la variance due à l'imputation, et

$$\mathbb{V}_{\mathrm{MIX}} = \mathbb{E}_{\mathrm{p}}\{(\,\hat{\bar{y}}_{\mathrm{GREG}} - \bar{y}_U\,)\,\mathbb{E}_{\mathrm{q}}(\,\hat{\bar{y}}_{\bullet\mathrm{GREG}} - \hat{\bar{y}}_{\mathrm{GREG}}\,)\} \tag{3.3.6}$$

appelée la covariance mixte. Elle mesure la covariance entre l'erreur d'échantillonnage et l'erreur d'imputation.

Il s'ensuit de (3.3.3) que la variance de $\hat{\bar{y}}_{\bullet\mathrm{GREG}}$ est

$$\mathbb{V} = \mathbb{V}_{\mathrm{ECH}} + \mathbb{V}_{\mathrm{IMP}} + 2\mathbb{V}_{\mathrm{MIX}} - \Big(\,\mathbb{E}_{\mathrm{p}}(B_{\mathrm{cond}})\,\Big)^2 \tag{3.3.7}$$

où $B_{\mathrm{cond}}$ est donné par (3.3.2).

Si nous supposons que $B_{\mathrm{cond}} = 0$, alors $\mathbb{V} = \mathrm{EQM}_{\mathrm{pq}}(\hat{\bar{y}}_{\bullet\mathrm{GREG}})$. Sinon, un estimateur de

$$\mathbb{V}_{\mathrm{ECH}} + \mathbb{V}_{\mathrm{IMP}} + 2\mathbb{V}_{\mathrm{MIX}}$$

est un estimateur conservateur de $\mathbb{V}$ (légère surestimation), puisque le terme $(\mathbb{E}_{\mathrm{p}}(B_{\mathrm{cond}}))^2$ est positif. D'après cette justification, nous avons décidé d'enlever ce terme de l'équation (3.3.7) et de nous concentrer sur l'apport des trois autres termes de la variance totale, soit la variance échantillonnale, la variance due à l'imputation et la covariance mixte.

Ces trois composantes sont difficiles à estimer. Pour nous aider dans cette procédure, nous allons utiliser un modèle de régression linéaire entre la variable d'imputation $z$ et la variable d'étude $y$. Pour ce modèle d'imputation, dénoté $\xi$, nous supposons pour $k \in U$, la forme suivante

$$y_k = \beta z_k + \epsilon_k \tag{3.3.8}$$

avec $E_\xi(\epsilon_k) = 0$, $\forall k$ et

$$Cov_\xi(\epsilon_k, \epsilon_l) = \begin{cases} \sigma^2 z_k & \text{si } k = l \\ 0 & \text{si } k \neq l \end{cases}$$

On sait que dans beaucoup d'enquêtes, la structure de variance s'accorde avec cette hypothèse.

L'EQM anticipé tient compte des deux sources de variabilité soit celle associée à la Superpopulation (modèle d'imputation) dont la population finie est un échantillon et celle provenant du plan de sondage. L'EQM anticipé, nous permet de dériver un estimateur consistant selon le plan et bclup (best conditionnal linear unbiaised predictor) sur le modèle. Par EQM anticipé, on sous-entend l'espérance de $\mathrm{EQM}_{\mathrm{pq}}(\hat{\bar{y}}_{\bullet\mathrm{GREG}})$ sous le modèle $\xi$. En utilisant notre hypothèse que le mécanisme de réponse est non confondu, nous trouvons que l'EQM anticipé est donné par

$$\begin{aligned}
\mathbb{E}_\xi \mathrm{EQM}_{\mathrm{pq}}(\hat{\bar{y}}_{\bullet\mathrm{GREG}}) &= \mathbb{E}_\xi(\mathbb{V}_{\mathrm{ECH}}) + \mathbb{E}_\xi \mathbb{E}_{\mathrm{p}} \mathbb{E}_{\mathrm{q}} \Big[ (\hat{\bar{y}}_{\bullet\mathrm{GREG}} - \hat{\bar{y}}_{\mathrm{GREG}})^2 \mid s, r \Big] + \\
&\quad 2\mathbb{E}_\xi \mathbb{E}_{\mathrm{p}} \mathbb{E}_{\mathrm{q}} \Big[ (\hat{\bar{y}}_{\mathrm{GREG}} - \bar{y}_U)(\hat{\bar{y}}_{\bullet\mathrm{GREG}} - \hat{\bar{y}}_{\mathrm{GREG}}) \mid s, r \Big] \\
&= \mathbb{E}_\xi(\mathbb{V}_{\mathrm{ECH}}) + \mathbb{E}_{\mathrm{p}} \mathbb{E}_{\mathrm{q}} \Big[ \mathbb{E}_\xi \{ (\hat{\bar{y}}_{\bullet\mathrm{GREG}} - \hat{\bar{y}}_{\mathrm{GREG}})^2 \mid s, r \} \Big] + \\
&\quad 2\mathbb{E}_{\mathrm{p}} \mathbb{E}_{\mathrm{q}} \Big[ \mathbb{E}_\xi \{ (\hat{\bar{y}}_{\mathrm{GREG}} - \bar{y}_U)(\hat{\bar{y}}_{\bullet\mathrm{GREG}} - \hat{\bar{y}}_{\mathrm{GREG}}) \mid s, r \} \Big]
\end{aligned}$$

Le calcul de l'espérance sous le modèle $\xi$ amène des expressions qui dépendent strictement des valeurs connues $z_k$ et des paramètres inconnus $\beta$ et $\sigma^2$. Pour estimer le terme $\mathbb{V}_{\text{ECH}}$ de l'erreur quadratique moyenne (3.3.3), il faut d'abord définir des estimateurs non biaisés des paramètres $\beta$ et $\sigma^2$ du modèle, basés sur les données pour l'ensemble des répondants $\{\ (y_k, z_k)\ :\ k\ \in\ r\ \}$. Nous suggérons (voir Särndal, Swensson et Wretman (1992), p535)

$$\hat{\sigma}^2 = \frac{1}{1-(\text{CV}_{zr})^2/m}\ \frac{\sum_r (y_k - \hat{\beta} z_k)^2}{\sum_r z_k} \tag{3.3.9}$$

où

$$\text{CV}_{zr} = \sqrt{\frac{\sum_r (z_l - \bar{z}_r)^2}{m-1}}\ \Big/\ \bar{z}_r \tag{3.3.10}$$

et

$$\hat{\beta} = \frac{\sum_r y_k}{\sum_r z_k} \tag{3.3.11}$$

## 3.4. Estimation de la variance totale

Si l'on part de la supposition que $\mathbb{V}_{\text{MIX}}$ dans (3.3.7) est très peu important, un estimateur de la variance totale pourrait être donné par

$$\hat{\mathbb{V}}\ =\ \hat{\mathbb{V}}_{\text{ECH}}\ +\ \hat{\mathbb{V}}_{\text{IMP}}$$

où $\hat{\mathbb{V}}_{\text{ECH}}$ est l'estimateur de la variance due à l'échantillonnage, et $\hat{\mathbb{V}}_{\text{IMP}}$ l'estimateur de la variance due à l'imputation. Dans la section 3.4.1, nous proposons un estimateur intuitif pour $\mathbb{V}_{\text{ECH}}$, noté $\hat{\mathbb{V}}_{\bullet\text{ECH}}$. Cependant, on sait que, pour la méthode PPV, la proposition $\hat{\mathbb{V}}_{\bullet\text{ECH}}$ a tendance à surestimer la composante $\mathbb{V}_{\text{ECH}}$. Pour cette raison, nous allons développer un terme d'ajustement noté $\widehat{diff}$, ayant pour but d'éliminer au moins une partie de cette surestimation. Ce développement est le sujet de la section 3.4.2. La dérivation de $\hat{\mathbb{V}}_{\text{IMP}}$ sera présentée dans la

section 3.4.3. Le raisonnement qui justifie le fait d'omettre le terme $\widehat{\mathbb{V}}_{\text{MIX}}$ de nos calculs sera présenté dans la section 3.4.4.

Suite à ces développements, on aura donc deux estimateurs possibles de la variance totale, soient

$$\widehat{\mathbb{V}}_{\bullet\text{INT}} = \widehat{\mathbb{V}}_{\bullet\text{ECH}} + \widehat{\mathbb{V}}_{\text{IMP}} \tag{3.4.1}$$

et

$$\widehat{\mathbb{V}}_{\bullet} = \widehat{\mathbb{V}}_{\bullet\text{ECH}} - \widehat{diff} + \widehat{\mathbb{V}}_{\text{IMP}} \tag{3.4.2}$$

Dans les simulations du chapitre 5 et 6, ces deux possibilités seront comparées.

### 3.4.1. Dérivation de la variance due à l'échantillonnage

Pour évaluer la variance échantillonnale, nous proposons l'estimateur

$$\widehat{\mathbb{V}}_{\bullet\text{ECH}} = \frac{1-f}{n}\frac{1}{n-1}\sum\nolimits_s g_k^2 e_{\bullet k}^2 \tag{3.4.3}$$

où $e_{\bullet k} = y_{\bullet k} - \vec{x}_k'\vec{\hat{R}}_{\bullet}$ et où $\vec{\hat{R}}_{\bullet}$ est le résultat obtenu lorsque $\vec{\hat{R}}$, donné par (1.4.3), est calculé sur les données complétées, $\{\ y_{\bullet k} : k \in s\ \}$, c'est-à-dire

$$\vec{\hat{R}}_{\bullet} = \left(\sum\nolimits_s a_k \vec{x}_k \vec{x}_k'/c_k\right)^{-1} \sum\nolimits_s a_k \vec{x}_k y_{\bullet k}/c_k \tag{3.4.4}$$

Le raisonnement menant à cette formule est le suivant : nous prenons la formule (1.4.6), laquelle pour le plan SI devient

$$\widehat{\mathbb{V}}_{\text{ECH}} = \frac{1-f}{n}\frac{1}{n-1}\sum\nolimits_s g_k^2 e_k^2 \tag{3.4.5}$$

Nous insérons dans la formule (3.4.5) les données complétées. Le résultat de cette procédure nous donne l'équation (3.4.3).

### 3.4.2. Dérivation du terme d'ajustment

Sous un plan SI, la différence entre la variance échantillonnale calculée à partir des données complétées et la variance échantillonnale dans le cas de 100 % de réponse, dénotée DIFF, s'exprime comme suit:

$$\text{DIFF} = \frac{1-f}{n}\frac{1}{n-1}\left(\sum_s g_k^2 e_{\bullet k}^2 - \sum_s g_k^2 e_k^2\right) \tag{3.4.6}$$

PROPOSITION 3.4.1. *La différence entre la variance échantillonnale calculée à partir des données complétées et la variance échantillonnale dans le cas de 100 % de réponse est donnée par*

$$\begin{aligned}\text{DIFF} \;=\; & \frac{1-f}{n}\frac{1}{n-1}\left(\sum_o g_k^2\delta_k^2 + \sum_s g_k^2\mu_k^2 + 2\sum_o g_k^2\delta_k e_k - \right.\\ & \left. 2\sum_o g_k^2\delta_k\mu_k - 2\sum_s g_k^2 e_k\mu_k\right)\end{aligned} \tag{3.4.7}$$

*où $g_k$ et $e_k$ sont définis respectivement par* (1.4.7) *et* (1.4.2) *et où les autres paramètres se définissent comme suit :*

$$\delta_k = y_{\ell(k)} - y_k \tag{3.4.8}$$

$$\mu_k = \vec{x}_k{}'\vec{\hat{R}}_\bullet - \vec{x}_k{}'\vec{\hat{R}} = \vec{x}_k{}'(\vec{\hat{R}}_\bullet - \vec{\hat{R}})$$

*avec $\vec{\hat{R}}$ et $\vec{\hat{R}}_\bullet$ donnés respectivement par* (1.4.3) *et* (3.4.4).

Démonstration. En nous servant de l'identité suivante,

$$e_{\bullet k} = \begin{cases} e_k - \mu_k & \text{si } k \in r \\ \delta_k + e_k - \mu_k & \text{si } k \in o \end{cases}$$

nous trouvons que

$$
\begin{aligned}
\sum_s g_k^2 e_{\bullet k}^2 - \sum_s g_k^2 e_k^2 &= \sum_r g_k^2 (\, e_k - \mu_k \,)^2 + \sum_o g_k^2 (\delta_k + e_k - \mu_k)^2 - \sum_s g_k^2 e_k^2 \\
&= \sum_r g_k^2 e_k^2 + \sum_r g_k^2 \mu_k^2 - 2 \sum_r g_k^2 e_k \mu_k + \sum_o g_k^2 \delta_k^2 + \\
&\quad \sum_o g_k^2 e_k^2 + \sum_o g_k^2 \mu_k^2 + 2 \sum_o g_k^2 \delta_k e_k - 2 \sum_o g_k^2 \delta_k \mu_k - \\
&\quad 2 \sum_o g_k^2 e_k \mu_k - \sum_r g_k^2 e_k^2 - \sum_o g_k^2 e_k^2 \\
&= \sum_o g_k^2 \delta_k^2 + \sum_s g_k^2 \mu_k^2 + 2 \sum_o g_k^2 \delta_k e_k - 2 \sum_o g_k^2 \delta_k \mu_k - \\
&\quad 2 \sum_s g_k^2 e_k \mu_k
\end{aligned}
$$

L'expression (3.4.7) s'ensuit. □

Notre premier objectif est de développer une procédure d'estimation de variance qui pourra fonctionner bien pour le cas de base, c'est-à-dire le cas du mécanisme de réponse uniforme donné par (1.5.1). Nous appliquerons par la suite la même procédure à un autre mécanisme de réponse, le mécanisme de réponse non uniforme défini par (1.5.2). De ce fait, en utilisant le mécanisme de réponse uniforme, nous allons évaluer les cinq termes de l'équation (3.4.7) par une simulation de Monte Carlo basée sur 1000 répétitions. La population considérée a été la population MU281. Une description détaillée de cette simulation sera donnée dans le chapitre 4 de ce présent mémoire. En calculant l'espérance de Monte Carlo (4.3.1) des 5 termes mentionnés ci-dessus, voici la valeur de l'estimateur Horvitz-Thompson, l'estimateur par le ratio et l'estimateur par la régression de chacun des termes.

TABLEAU 3.1. La valeur des cinq termes de la variable DIFF (Espérance de Monte Carlo, Population MU281, plan SI, estimateur Horvitz-Thompson).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| $\sum_o g_k^2 \delta_k^2$ | 6905,19 | 3088,29 | 1596,82 |
| $\sum_s g_k^2 \mu_k^2$ | 437,28 | 158,59 | 72,73 |
| $2\sum_o g_k^2 \delta_k e_k$ | -5770,77 | -2726,93 | -1421,01 |
| $-2\sum_o g_k^2 \delta_k \mu_k$ | -3,32 | -1,05 | -0,44 |
| $-2\sum_s g_k^2 e_k \mu_k$ | 5,46 | 3,29 | -2,05 |

TABLEAU 3.2. La valeur des cinq termes de la variable DIFF (Espérance de Monte Carlo, Population MU281, plan SI, estimateur par le ratio).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| $\sum_o g_k^2 \delta_k^2$ | 6008,95 | 2856,46 | 1438,25 |
| $\sum_s g_k^2 \mu_k^2$ | 206,94 | 67,94 | 33,94 |
| $2\sum_o g_k^2 \delta_k e_k$ | -4641,50 | -2242,08 | -1094,67 |
| $-2\sum_o g_k^2 \delta_k \mu_k$ | -299,51 | -98,87 | -44,45 |
| $-2\sum_s g_k^2 e_k \mu_k$ | 103,34 | 35,48 | -32,73 |

TABLEAU 3.3. La valeur des cinq termes de la variable DIFF (Espérance de Monte Carlo, Population MU281, plan SI, estimateur par la régression).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| $\sum_o g_k^2 \delta_k^2$ | 6151,75 | 2939,83 | 1460.50 |
| $\sum_s g_k^2 \mu_k^2$ | 332,18 | 120,61 | 48,25 |
| $2\sum_o g_k^2 \delta_k e_k$ | -5216,01 | -2509,74 | -1272,34 |
| $-2\sum_o g_k^2 \delta_k \mu_k$ | -700,44 | -246,46 | -96,89 |
| $-2\sum_s g_k^2 e_k \mu_k$ | -12,62 | -2,47 | -0.89 |

Les chiffres des tableaux 3.1, 3.2 et 3.3 semblent indiquer que les 2 principaux termes de la décomposition sont le premier et le troisième. Analysons maintenant ce qui reste de l'expression (3.4.7) si l'on omet les 3 termes, soit

$$diff = \frac{1-f}{n}\frac{1}{n-1}\left(\mathrm{T}_1 + \mathrm{T}_3\right) \tag{3.4.9}$$

où $\mathrm{T}_1 = \sum_o g_k^2 \delta_k^2$ et $\mathrm{T}_3 = 2\sum_o g_k^2 \delta_k e_k$ .

Remarquons que les paramètres $\delta_k$ et $e_k$, définis par les équations (3.4.8) et (1.4.2), sont inutilisables dans le cas de non-réponse, car ils présupposent que l'on connaisse la variable d'étude $y_k$ pour les non-répondants (ce qui est déraisonnable). Donc, nous ne pouvons pas utiliser directement l'équation (3.4.9). Nous allons chercher à estimer les deux termes $\mathrm{T}_1$ et $\mathrm{T}_3$ du paramètre $diff$. À cette fin, nous évaluerons d'abord les espérances de $\mathrm{T}_1$ et de $\mathrm{T}_3$ sous le modèle

d'imputation $\xi$ donné par (3.3.8). Nous obtenons

$$\begin{aligned}
\mathbb{E}_\xi(\mathrm{T}_1) &= \mathbb{E}_\xi\Big( \sum_o g_k^2 \delta_k^2 \Big) \\
&= \mathbb{E}_\xi\Big( \sum_o g_k^2 (\, \beta d_k \, + \, \epsilon_{\ell(k)} \, - \, \epsilon_k)^2 \Big) \\
&= \beta^2 \sum_o g_k^2 d_k^2 \, + \, \sigma^2 \sum_o g_k^2 (\, z_{\ell(k)} \, + \, z_k \,) \\
&= \beta^2 \sum_o g_k^2 d_k^2 + \sigma^2 \Big( 2 \sum_o g_k^2 z_k \, + \, \sum_o g_k^2 d_k \Big)
\end{aligned}$$

où $d_k \, = \, z_{\ell(k)} \, - \, z_k$ et

$$\begin{aligned}
\mathbb{E}_\xi(\mathrm{T}_3) &= \mathbb{E}_\xi\Big( 2 \sum_o g_k^2 \delta_k e_k \Big) \\
&= \mathbb{E}_\xi\Big( 2 \sum_o g_k^2 (\, \beta d_k \, + \, \epsilon_{\ell(k)} \, - \, \epsilon_k)(\, \beta z_k \, + \, \epsilon_k + h_k \,) \Big)
\end{aligned}$$

où $h_k$ est défini par $-\vec{x}_k{}' T^{-1} \sum_s \frac{a_j(\beta z_j + \epsilon_j)\vec{x}_j}{c_j}$ avec $T = \sum_s \frac{a_j \vec{x}_j \vec{x}_j{}'}{c_j}$. Nous obtenons donc

$$\begin{aligned}
\mathbb{E}_\xi(\mathrm{T}_3) \;=\; & 2\beta^2 \sum_o g_k^2 d_k z_k \, - \, 2\beta^2 \sum_o g_k^2 d_k \Big( \vec{x}_k{}' \, T^{-1} \sum_s \frac{a_j \vec{x}_j z_j}{c_j} \Big) \, + \\
& 2 \, \mathbb{E}_\xi \Big( \sum_o g_k^2 (\epsilon_{\ell(k)} \, - \, \epsilon_k) h_k \Big) \, - \, 2\sigma^2 \sum_o g_k^2 z_k
\end{aligned}$$

En regroupant les termes $\mathrm{T}_1$ et $\mathrm{T}_3$, nous avons

$$\begin{aligned}
\mathbb{E}_\xi(\mathrm{T}_1 \, + \, \mathrm{T}_3) \;=\; & \beta^2 \sum_o g_k^2 d_k^2 \, + \, 2\sigma^2 \sum_o g_k^2 z_k \, + \, \sigma^2 \sum_o g_k^2 d_k \, + \\
& 2\beta^2 \sum_o g_k^2 d_k (\, z_k - \hat{z}_k \,) \, - \, 2\sigma^2 \sum_o g_k^2 z_k \, + \, \mathit{Reste}
\end{aligned}$$

où $\hat{z}_k = \vec{x}_k{}' \, \widehat{\beta}_{xz} = \vec{x}_k{}' \, T^{-1} \, \sum_s \frac{a_j \vec{x}_j z_j}{c_j}$, $\quad \mathit{Reste} = 2\mathbb{E}_\xi(\sum_o g_k^2 (\, \epsilon_{\ell(k)} \, - \, \epsilon_k) h_k)$ et $d_k \, = \, z_{\ell(k)} \, - \, z_k$ .

Nous procédons en supposant que la quantité *Reste* est négligeable. Puisque l'on impute par la méthode PPV, nous savons que les distances $d_k$ sont minimisées et sont soient positives si le donneur est à droite du

receveur ou négatives si le donneur est à gauche du receveur. Donc, nous voyons que la somme pondérée des $d_k$ est aussi négligeable étant donné que l'on somme des termes qui alternent de signe constamment. Puisque $d_k$ est de nature alternante, $d_k^2$ ne l'est plus, d'où la necessité de garder tous les termes comprenant $d_k^2$. Ainsi, l'espérance sous le modèle $\xi$ du terme $diff$, défini par l'équation (3.4.9), se résume par

$$\mathbb{E}_\xi(diff) = \frac{1-f}{n}\frac{1}{n-1}\left(\beta^2 \sum_o g_k^2 d_k^2\right) \tag{3.4.10}$$

Ce résultat nous amène à proposer, pour $diff$, l'estimateur que nous utiliserons par la suite, soit

$$\widehat{diff} = \frac{1-f}{n}\frac{1}{n-1}\left(\widehat{\beta}^2 \sum_o g_k^2 d_k^2\right) \tag{3.4.11}$$

où $\widehat{\beta}$ est donné par (3.3.11).

Si le modèle $\xi$ défini par (3.3.8) est respecté, la variance due à l'échantillonnage de l'estimateur $\widehat{\mathbb{V}}_\bullet$ sera donc estimée par

$$\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}} - \widehat{diff} \tag{3.4.12}$$

où $\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}}$ et $\widehat{diff}$ sont donnés respectivement par (3.4.3) et (3.4.11). Nous n'assurons pas que cet estimateur sera non-négatif.

### 3.4.3. Dérivation de la variance due à l'imputation

La variance due à l'imputation $\mathbb{V}_{\mathrm{IMP}}$ est définie par (3.3.5). Il nous sera utile d'avoir une expression pour son espérance sous le modèle d'imputation $\xi$ donné par (3.3.8).

PROPOSITION 3.4.2. *Sous le modèle d'imputation $\xi$ et sous le mécanisme de réponse non confondu, l'espérance de $\mathbb{V}_{\text{IMP}}$ est donnée par*

$$\mathbb{E}_\xi \mathbb{V}_{\text{IMP}} = \mathbb{E}_p \mathbb{E}_q \left[ \sigma^2 \sum\nolimits_r S_l^2 z_l + \sigma^2 \sum\nolimits_o W_k^2 z_k + \beta^2 ( \sum\nolimits_o W_k d_k )^2 \right] \tag{3.4.13}$$

*où $W_k = a_k g_k$ et $S_l = \sum_{o_l} W_k$ avec $o_l = \{k : k \in o$ et $k$ utilise $\ell$ comme donneur$\}$.*

Démonstration. En utilisant l'hypothèse que le mécanisme de réponse est non confondu, la dérivation de l'équation (3.4.13) passe par les étapes suivantes :

$$\begin{aligned} \mathbb{E}_\xi \mathbb{V}_{\text{IMP}} &= \mathbb{E}_\xi \mathbb{E}_p \mathbb{E}_q ( \hat{\bar{y}}_{\bullet \text{GREG}} - \hat{\bar{y}}_{\text{GREG}} )^2 \\ &= \mathbb{E}_\text{p} \mathbb{E}_\text{q} \mathbb{E}_\xi ( \sum\nolimits_s a_k g_k y_{\bullet k} - \sum\nolimits_s a_k g_k y_k )^2 \\ &= \mathbb{E}_\text{p} \mathbb{E}_\text{q} \mathbb{E}_\xi ( \sum\nolimits_s W_k y_{\bullet k} - \sum\nolimits_s W_k y_k )^2 \end{aligned}$$

où $W_k = a_k g_k$. Il s'ensuit que

$$\mathbb{E}_\xi \mathbb{V}_{\text{IMP}} = \mathbb{E}_\text{p} \mathbb{E}_\text{q} \mathbb{V}_\xi ( \sum\nolimits_s W_k y_{\bullet k} - \sum\nolimits_s W_k y_k ) + \mathbb{E}_\text{p} \mathbb{E}_\text{q} \mathbb{E}_\xi{}^2 ( \sum\nolimits_s W_k y_{\bullet k} - \sum\nolimits_s W_k y_k )$$

Remarquons que

$$\begin{aligned} \sum\nolimits_s W_k y_{\bullet k} - \sum\nolimits_s W_k y_k &= \sum\nolimits_r W_\ell y_\ell + \sum\nolimits_o W_k y_{\ell(k)} - \sum\nolimits_s W_k y_k \\ &= \sum\nolimits_o W_k y_{\ell(k)} - \sum\nolimits_o W_k y_k \\ &= \sum\nolimits_o W_k ( y_{\ell(k)} - y_k ) \end{aligned}$$

Calculons maintenant les deux termes de $\mathbb{E}_\xi \mathbb{V}_{\text{IMP}}$. L'évaluation de la variance selon le modèle d'imputation $\xi$ nous donne

$$\begin{aligned}
\mathbb{V}_\xi\Big(\sum_s W_k y_{\bullet k} - \sum_s W_k y_k\Big) &= \mathbb{V}_\xi\Big(\sum_o W_k(y_{\ell(k)} - y_k)\Big) \\
&= \mathbb{V}_\xi\Big(\sum_o W_k(\beta z_{\ell(k)} + \epsilon_{\ell(k)} - \beta z_k - \epsilon_k)\Big) \\
&= \mathbb{V}_\xi\Big(\sum_o W_k(\beta z_{\ell(k)} - \beta z_k) + \sum_o W_k(\epsilon_{\ell(k)} - \epsilon_k)\Big) \\
&= \mathbb{V}_\xi\Big(\sum_o W_k(\epsilon_{\ell(k)} - \epsilon_k)\Big) \\
&= \mathbb{V}_\xi\Big(\sum_r S_l \epsilon_l - \sum_o W_k \epsilon_k\Big)
\end{aligned}$$

où $S_l = \sum_{o_l} W_k$. Nous obtenons finalement pour ce terme

$$\begin{aligned}
\mathbb{V}_\xi\Big(\sum_s W_k y_{\bullet k} - \sum_s W_k y_k\Big) &= \mathbb{V}_\xi\Big(\sum_r S_l \epsilon_l - \sum_o W_k \epsilon_k\Big) \\
&= \sigma^2\,(\textstyle\sum_r S_l^2 z_l + \sum_o W_k^2 z_k)
\end{aligned}$$

L'évaluation du carré de la moyenne selon le modèle d'imputation $\xi$ nous donne

$$\begin{aligned}
{\mathbb{E}_\xi}^2\Big(\sum_s W_k y_{\bullet k} - \sum_s W_k y_k\Big) &= {\mathbb{E}_\xi}^2\Big(\sum_o W_k(y_{\ell(k)} - y_k)\Big) \\
&= {\mathbb{E}_\xi}^2\Big(\sum_o W_k(\beta z_{\ell(k)} - \beta z_k) + \sum_o W_k(\epsilon_{\ell(k)} - \epsilon_k)\Big) \\
&= \beta^2(\textstyle\sum_o W_k d_k)^2
\end{aligned}$$

où $\mathbb{E}_\xi(\epsilon_k) = 0, \forall k$ . En agglomérant ces deux termes, voici le produit final :

$$\mathbb{E}_\xi \mathbb{V}_{\text{IMP}} = \mathbb{E}_{\text{p}} \mathbb{E}_{\text{q}} \Big[\sigma^2 \textstyle\sum_r S_l^2 z_l + \sigma^2 \sum_o W_k^2 z_k + \beta^2 (\sum_o W_k d_k)^2\Big]$$

□

Or, la formule (3.4.13) fait appel à deux paramètres inconnus, soit la variance $\sigma^2$ et le paramètre $\beta$. Le terme en $\beta^2$ dépend des différences $d_k = z_{\ell(k)} - z_k, \forall k \in o$ . Nous proposons de remplacer le $\sigma^2$ inconnu

par son estimateur sans biais sous le modèle d'imputation. Pour ce qui est du terme $\beta^2(\sum_o W_k d_k)^2$, nous proposons de l'estimer par zéro. Encore une fois, puisque nous imputons sous la méthode PPV, la somme $\sum_o W_k d_k$ équivaut à une somme pondérée des distances $d_k$ qui sont soient positif ou négatif. La somme $\sum_o W_k d_k$ doit donc être négligeable ou petite. Nous proposons donc, pour nos simulations de Monte Carlo, la formule suivante pour l'estimation de la variance due à l'imputation

$$\widehat{\mathbb{V}}_{\text{IMP}} = \hat{\sigma}^2 \sum_r S_l^2 z_l + \hat{\sigma}^2 \sum_o W_k^2 z_k \tag{3.4.14}$$

où $\hat{\sigma}^2$ est estimée par la formule (3.3.9).

### 3.4.4. Dérivation de la covariance mixte

À ce point, nous avons défini trois termes majeurs pour estimer la variance de l'estimateur de la moyenne de la population, soit l'estimateur de la variance due à l'échantillonnage, l'estimateur de la variance due à l'imputation et l'estimateur DIFF, qui est la différence entre la variance échantillonnale dans le cas 100 % de réponse et la variance échantillonnale dans le cas de non-réponse. Parlons maintenant du terme mixte de l'équation (3.3.3).

PROPOSITION 3.4.3. *Sous n'importe quel plan d'échantillonnage $p$ et sous n'importe quel mécanisme de réponse non confondu $q$, l'espérance sous le modèle $\xi$ de la covariance mixte est donnée par*

$$\begin{aligned}\mathbb{E}_\xi(\mathbb{V}_{\text{MIX}}) \approx \mathbb{E}_p\mathbb{E}_q \Big\{ & \sigma^2 \sum_r W_l S_l z_l - \sigma^2 \sum_o W_k^2 z_k - \sigma^2 \sum_r S_l z_l + \\ & \sigma^2 \sum_o W_k z_k \Big\}\end{aligned} \tag{3.4.15}$$

La démonstration est présentée à l'appendice A.

En nous basant sur (3.4.15), nous proposons d'estimer $\mathbb{V}_{\text{MIX}}$ par

$$\widehat{\mathbb{V}}_{\text{MIX}} = \left\{ \sum_r W_l S_l z_l - \sum_o W_k^2 z_k - \sum_r S_l z_l + \sum_o W_k z_k \right\} \hat{\sigma}^2 \tag{3.4.16}$$

Nous remarquons souvent dans la littérature que ce terme est négligeable (voir Rancourt, Särndal et Lee (1994) et Gagnon, Lee, Provost, Rancourt et Särndal (1997)). D'après les résultats empiriques calculés à partir des simulations de Monte Carlo (technique expliquée au chapitre 4), nous remarquons effectivement dans les chapitres 5 et 6, que cette quantité est négligeable.

À la suite du raisonnement des sections 3.4.1 à 3.4.4, nous proposons d'utiliser les estimateurs de la variance totale donnés par (3.4.1) et (3.4.2). Ces variances seront estimées sous un plan SI et à l'aide de trois estimateurs : l'estimateur HT, l'estimateur par le ratio et l'estimateur par la régression.

Finalement, afin de savoir si nous avons réussi à bien estimer chacune des composantes $\widehat{\mathbb{V}}_{\bullet\text{ECH}}$, $\widehat{\mathbb{V}}_{\text{IMP}}$ et $\widehat{\mathbb{V}}_{\text{MIX}}$, nous allons évaluer par la méthode de Monte Carlo, les approximations des composantes $\mathbb{V}_{\text{ECH}}$, $\mathbb{V}_{\text{IMP}}$ et $\mathbb{V}_{\text{MIX}}$ qui sont définies respectivement par (3.3.4), (3.3.5) et (3.3.6). Nous dénotons ces approximations par $\mathbb{V}^*_{\text{ECH}}$, $\mathbb{V}^*_{\text{IMP}}$ et $\mathbb{V}^*_{\text{MIX}}$. Ces dernières sont obtenues en calculant respectivement leur moyenne sur les 10 000 répétitions. Ces quantités peuvent être calculées puisque la population utilisée pour nos simulations est complètement connue. Les comparaisons seront faites dans les simulations du chapitre 5 et 6.

CHAPITRE 4

# Description des conditions de simulation

## 4.1. Présentation de la population MU281

Le jeu de données utilisé pour les simulations provient du livre de Särndal, Swensson et Wretman (1992), "Model Assisted Survey Sampling", Appendice B. Il s'agit d'un jeu de données, dénoté MU284, contenant, pour 284 municipalités de Suède, différentes variables socio-économiques. Concrètement, par municipalité, nous entendons une ville et sa région avoisinante. Ces municipalités varient largement en taille de population. Ici, nous allons considérer ce même jeu de données, excluant les trois plus grosses municipalités, selon la variable P75 qui définit la population de l'année 1975. Les municipalités exclues sont Stockholm, Göteborg et Malmö. La population obtenue est appelée population MU281.

Les variables que nous utiliserons pour l'identification du plus proche voisin, pour la variable auxiliaire et pour la variable d'étude, sont dans l'ordre :

- z=P75, représente la population d'une municipalité en 1975;
- x=RMT85, représente les revenus des taxes municipales pour une municipalité en 1985;
- y=REV84, représente les valeurs immobilières pour une municipalité en 1984.

Si nous regardons le graphique de la population des municipalités en 1975 versus les revenus des taxes municipales pour chacune des municipalités en 1985, nous nous apercevons qu'il y a trois municipalités divergentes.

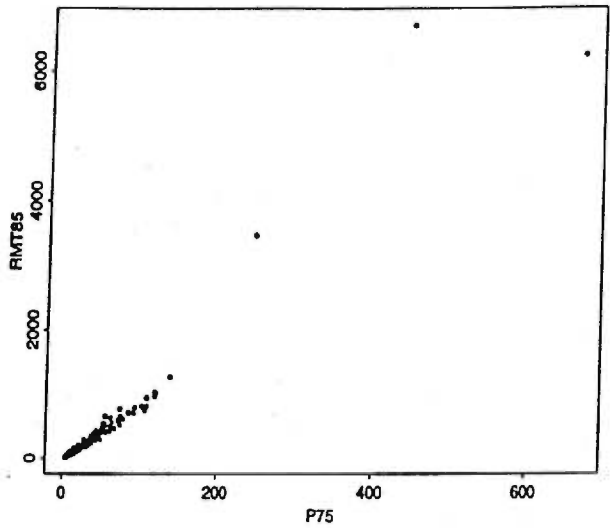


FIGURE 4.1. Nuage de points de la population MU284 avec $z$=P75 et $x$=RMT85.

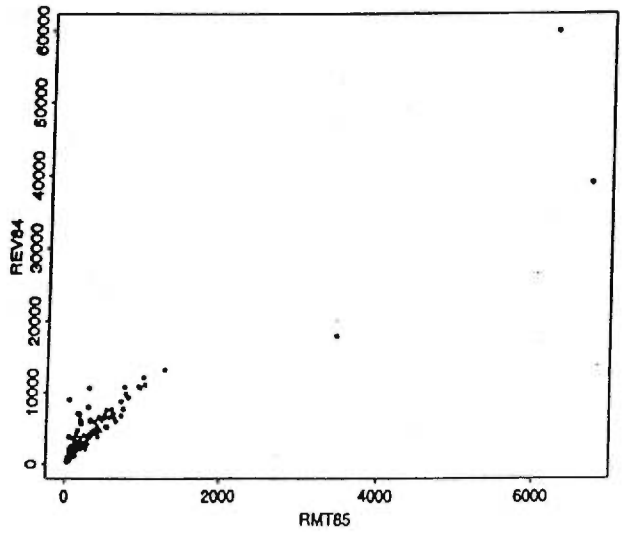


FIGURE 4.2. Nuage de points de la population MU284 avec $x$=RMT85 et $y$=REV84.

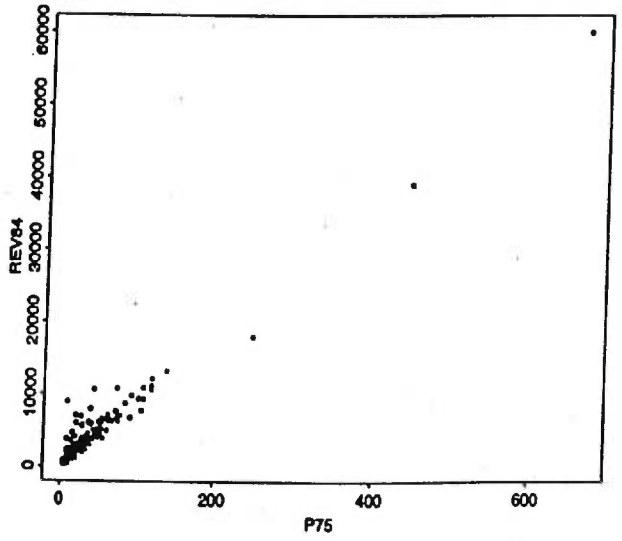


**FIGURE 4.3. Nuage de points de la population MU284 avec $z$=P75 et $y$=REV84.**

Dans ces figures (les nuages de points de la population MU284), les trois grosses unités sont identifiées par les trois plus grandes municipalités de Suède. Nous avons décidé de les rayer de notre étude afin de faciliter l'interprétation de nos résultats de simulation et d'éviter des biais dans les résultats. À l'aide du jeu de données MU281 et des variables décrites antérieurement, nous étudions pour les simulations de Monte Carlo, le plan d'échantillonnage simple avec deux estimateurs distincts :

- Estimateur par le ratio : MU281 avec $y$=REV84, $x$=RMT85 et $z$=P75;
- Estimateur par la régression : MU281 avec $y$=REV84, $x$=RMT85 et $z$=P75.

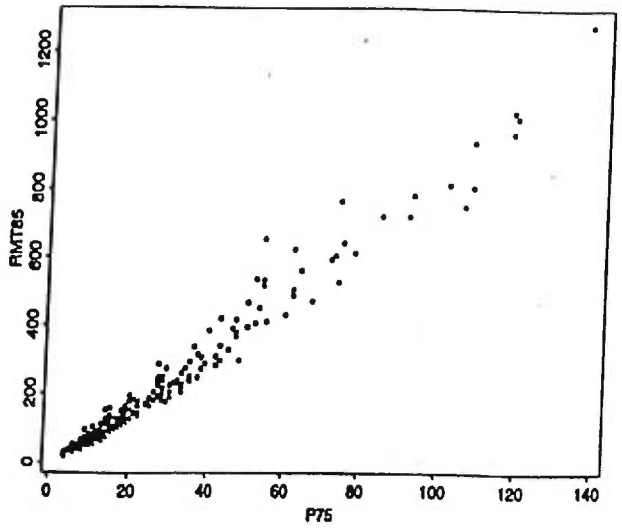


FIGURE 4.4. Nuage de points de la population MU281 avec $z$=P75 et $x$=RMT85.

Lorsque nous observons le nuage de points ci-dessus de la population MU281, on constate qu'il s'agit d'une version agrandie d'une partie de la figure 4.1, car les trois grosses unités ne sont pas incluses dans cette population.

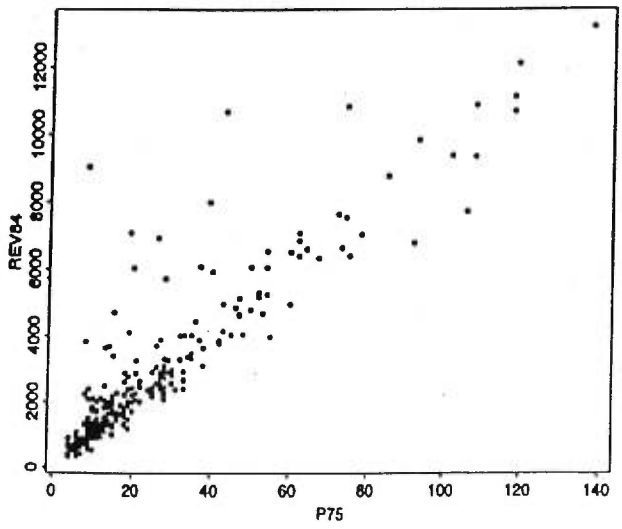


FIGURE 4.5. Nuage de points de la population MU281 avec $z$=P75 et $y$=REV84.

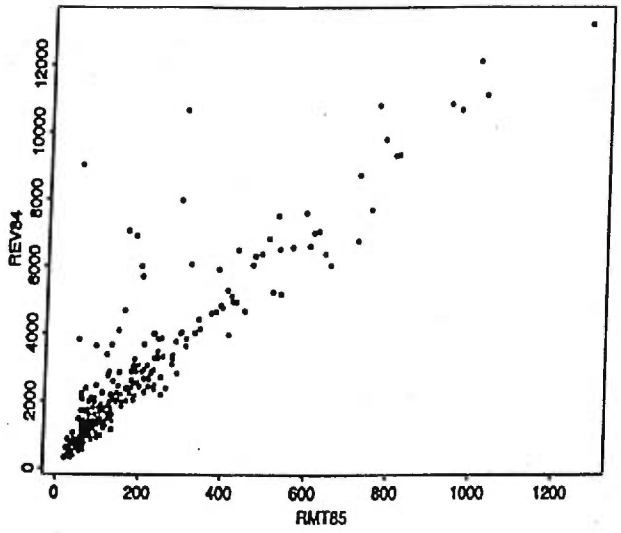


FIGURE 4.6. Nuage de points de la population MU281 avec $x$=RMT85 et $y$=REV84.

Le tableau 4.1 nous présente une analyse descriptive des variables $x$, $y$ et $z$ utilisées pour la population MU281. On y présente, pour chacune des variables, la moyenne, la variance et le coefficient de variation, noté CV et défini comme l'écart-type divisé par la moyenne. Le tableau 4.2 présente les coefficients de corrélation entre les variables.

TABLEAU 4.1. Population MU281. Analyse descriptive des variables utilisées.

| Variable | Moyenne | Variance | CV |
|---|---|---|---|
| $x$=RMT85 | 189,05 | 40 008,40 | 1,058 |
| $y$=REV84 | 2694,83 | 5 707 842,50 | 0,887 |
| $z$=P75 | 24,26 | 540,29 | 0,958 |

TABLEAU 4.2. Population MU281. Coefficients de corrélation entre les variables utilisées.

| Variables | $\rho$ |
|---|---|
| $x$=RMT85, $y$=REV84 | 0,9074 |
| $x$=RMT85, $z$=P75 | 0,9870 |
| $y$=REV84, $z$=P75 | 0,9021 |

De plus, à l'aide du logiciel Splus, nous avons vérifié les hypothèses du modèle de régression de $y$ sur $z$ donné par (3.3.8) et celles du modèle de régression $y$ sur $x$ pour les estimateurs. L'analyse conclut que pour les deux modèles, l'hypothèse d'une ordonnée à l'origine n'est pas satisfaite pour $\alpha = 1\%$. Par contre, la pente s'est avérée significative pour les deux modèles.

Les résultats de simulation qui se rattachent à cette population seront présentés au chapitre 6.

## 4.2. Présentation de la population MU281-yratioz

La population MU281-yratioz est obtenue de la population MU281 en soustrayant de la variable d'intérêt $y$, l'ordonnée à l'origine du modèle de régression $y$ sur $z$. Nous avons par conséquent, une population qui satisfait le modèle d'imputation $\xi$ donné par (3.3.8). Contrairement à la population MU281, l'hypothèse d'une ordonnée à l'origine du modèle de régression de $y$ sur $x$ est satisfaite pour $\alpha = 1\%$. Par contre, il est toujours vrai d'affirmer que la pente est significative pour les deux modèles.

Les résultats de simulation qui se rattachent à cette population seront présentés au chapitre 5.

## 4.3. Description des différentes mesures utilisées

Pour la simulation de Monte Carlo, nous avons réalisé 10 000 ensembles de répondants $r$. Un ensemble $r$ est réalisé en 2 étapes. D'abord, nous tirons un échantillon $s$ de taille $n = 100$ par le plan SI défini à la section 1.2. Pour cet échantillon, nous réalisons un ensemble de répondants $r$ en utilisant un mécanisme de réponse spécifié. Dans la section 5.2, nous utilisons le mécanisme uniforme défini par (1.5.1) et dans la section 5.4, nous utilisons trois différents mécanismes de type (1.5.2) qui seront expliqués à la section 4.3.

Pour chaque $r$, nous calculons la valeur prise par chacun des estimateurs qui nous intéressent, soit l'estimateur HT ($\hat{\bar{y}}_{\text{HT}}$ et $\hat{\bar{y}}_{\bullet\text{HT}}$), l'estimateur par le ratio ($\hat{\bar{y}}_{\text{RA}}$ et $\hat{\bar{y}}_{\bullet\text{RA}}$) et l'estimateur par la régression ($\hat{\bar{y}}_{\text{REG}}$ et $\hat{\bar{y}}_{\bullet\text{REG}}$). Si $\hat{\bar{y}}_U$ note un de ces six estimateurs, nous calculons

- l'espérance de Monte Carlo de $\hat{\bar{y}}_U$;
- la variance de Monte Carlo de $\hat{\bar{y}}_U$;
- l'espérance de Monte Carlo de l'estimateur de la variance de $\hat{\bar{y}}_U$;
- le taux de recouvrement de Monte Carlo de $\hat{\bar{y}}_U$.

Regardons plus en détail les différentes mesures calculées.

L'espérance de Monte Carlo de $\hat{\bar{y}}_U$ est définie par la moyenne des 10 000 estimateurs ponctuels calculés et est donnée par

$$\text{EMC}(\hat{\bar{y}}_U) = \frac{1}{10\,000} \sum_{j=1}^{10\,000} \hat{\bar{y}}_{U_j} \tag{4.3.1}$$

où $\hat{\bar{y}}_{U_j}$ est la valeur de l'estimateur ponctuel $\hat{\bar{y}}_U$ pour le j-ème échantillon; j=1,...,10 000.

La variance de Monte Carlo de $\hat{\bar{y}}_U$ est définie par la variance des 10 000 estimateurs ponctuels calculés et est donnée par

$$\text{VMC}(\hat{\bar{y}}_U) = \frac{1}{10\,000 - 1} \sum_{j=1}^{10\,000} (\hat{\bar{y}}_{U_j} - \text{EMC}(\hat{\bar{y}}_U))^2$$

L'espérance de Monte Carlo de l'estimateur de la variance correspond à la moyenne des 10 000 estimateurs de variance et est donnée par

$$\text{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_U)) = \frac{1}{10\,000} \sum_{j=1}^{10\,000} \widehat{\mathbb{V}}(\hat{\bar{y}}_{U_j})$$

où $\widehat{\mathbb{V}}(\hat{\bar{y}}_{U_j})$ est la valeur de l'estimateur de variance $\widehat{\mathbb{V}}(\hat{\bar{y}}_U)$ pour le j-ème échantillon; j=1,...,10 000. En théorie, $\text{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_U))$ devrait être près de $\text{VMC}(\hat{\bar{y}}_U)$ si l'estimateur de variance $\widehat{\mathbb{V}}(\hat{\bar{y}}_U)$ est sans biais.

De plus, nous calculerons d'autres espérances de Monte Carlo. Ces calculs sont toujours effectués de la même manière, c'est-à-dire si $A$ est la variable aléatoire étudiée et $A_j$ sa valeur pour le j-ième échantillon, alors l'espérance de Monte Carlo sera définie par

$$\text{EMC}(A) = \frac{1}{10\,000} \sum_{j=1}^{10\,000} A_j$$

Le taux de recouvrement de Monte Carlo de $\hat{\bar{y}}_U$ est le nombre de fois (en pourcentage) que $\bar{y}_U$ se retrouve dans l'intervalle de confiance de l'estimateur, et ce, pour les 10 000 échantillons tirés, c'est-à-dire

$$\text{TRMC}(\hat{\bar{y}}_U) = 100 \cdot \frac{1}{10\,000} \sum_{j=1}^{10\,000} I_j$$

où

$$I_j = \begin{cases} 1 & \text{si } \bar{y}_U \in \left( \hat{\bar{y}}_{U_j} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathbb{V}}(\hat{\bar{y}}_{U_j})} \right) \\ 0 & \text{sinon} \end{cases}$$

et où la constante $z_{1-\alpha/2}$ est telle que $\int_{z_{1-\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \alpha/2$. Il est à souligner que, pour la simulation, nous avons choisi de travailler avec un niveau de confiance théorique de 95 %, donc $z_{1-\alpha/2}$ sera égal à 1,96.

## 4.4. Mécanisme de réponse non uniforme

Pour les simulations, nous avons utilisé deux différents mécanismes de réponse. Le premier est le mécanisme uniforme (1.5.1) décrit au chapitre 1 et l'autre est un mécanisme non uniforme, décrit dans le même chapitre, équation (1.5.2).

Pour ces deux mécanismes de réponse, nous avons effectué des simulations avec quatre taux de réponse, soit 60 %, 80 %, 90 % et 100 %. Tirer un échantillon de répondants parmi l'échantillon $s$ ne nécessite aucun calcul préalable lorsque l'on utilise un mécanisme de réponse uniforme. Cependant, pour ce qui est du mécanisme non uniforme, il a fallu programmer une fonction qui nous permettait de calculer la constante c dans (1.5.2) telle que le taux de réponse pour la population,

$$\frac{1}{N} \sum\nolimits_U \theta_k \tag{4.4.1}$$

soit égale à une valeur fixée.

Pour un taux de 60 %, 80 % et 90 % de réponse, nous avons obtenu respectivement

$$\theta_k = \exp^{-0,02618419 \cdot z_k} \tag{4.4.2}$$

$$\theta_k = \exp^{-0,01016476 \cdot z_k} \tag{4.4.3}$$

$$\theta_k = \exp^{-0,00455522 \cdot z_k} \tag{4.4.4}$$

Nous avons strictement utilisé le logiciel Splus pour programmer les simulations de ce mémoire (voir annexes B et C pour la présentation des programmes).

CHAPITRE 5

# Résultats des simulations pour le cas de base

## 5.1. Résultats sur la population

Pour le cas de base, nous utiliserons une population qui satisfait le modèle d'imputation $\xi$ donné par la (3.3.8) et le mécanisme de réponse uniforme pour sélectionner nos répondants. Contrairement à la population MU281 qui ne satisfait pas le modèle d'imputation, nous allons utiliser pour toutes les simulations rapportées dans ce chapitre, la population MU281-yratioz. Avant d'entamer la section des résultats de simulation, nous présentons, dans le tableau 5.1, les résultats d'une analyse descriptive de la moyenne de la population MU281-yratioz et de la variance de l'estimateur HT, du ratio et de la régression sous un plan SI, avec n=100. En utilisant les formules (1.3.5), (1.4.10) et (1.4.13) correspondant aux formules de variance de la population, nous avons obtenu

TABLEAU 5.1. Population MU281-yratioz, plan SI, n=100. Variance sur la population de l'estimateur d'Horvitz-Thompson, l'estimateur par le ratio et l'estimateur par la régression.

| | |
|---|---|
| $\bar{y}_U$ | 2249,74 |
| $\mathbb{V}(\hat{\bar{y}}_{\mathrm{HT}})$ | 36 765,82 |
| $A\mathbb{V}(\hat{\bar{y}}_{\mathrm{RA}})$ | 6781,67 |
| $A\mathbb{V}(\hat{\bar{y}}_{\mathrm{REG}})$ | 6491,36 |

Remarquons que la variance de $\hat{\bar{y}}_{\mathrm{RA}}$ et la variance de $\hat{\bar{y}}_{\mathrm{REG}}$ sont nettement inférieures à la variance de $\hat{\bar{y}}_{\mathrm{HT}}$. Ceci s'explique par la forte relation entre $x$ et $y$ rendue évidente par la figure 4.6. Par contre, pour ce qui est de la variance de $\hat{\bar{y}}_{\mathrm{RA}}$ et de la variance de $\hat{\bar{y}}_{\mathrm{REG}}$, elles sont pratiquement semblables. Nous pouvons donc s'attendre à ce que les résultats soient meilleurs pour l'estimateur par le ratio et l'estimateur par la régression que pour l'estimateur HT.

## 5.2. Résultats pour 100 % de réponse dans l'échantillon

Peu importe l'estimateur choisi, le biais relatif (tel qu'il est mesuré par la simulation de Monte Carlo) dans le cas 100 % de réponse est défini par

$$\mathrm{Birel}(\widehat{\mathbb{V}}) = 100 \cdot \frac{\mathrm{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_U)) \; - \; \mathrm{VMC}(\hat{\bar{y}}_U)}{\mathrm{VMC}(\hat{\bar{y}}_U)} \tag{5.2.1}$$

De plus, la quantité $\mathbb{V}^*_{\mathrm{ECH}}$ que nous retrouvons dans les tableaux 5.2, 5.3 et 5.4 est définie comme la moyenne de Monte Carlo de (3.3.4). Présentons maintenant les résultats obtenus pour l'estimateur HT, l'estimateur par le ratio et l'estimateur par la régression par la méthode de Monte Carlo dans le cas où il n'y a pas de non-réponse dans l'échantillon.

Pour ce qui est de l'estimateur HT, il possède les propriétés mentionnées à la section 1.3. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.3.4) et (1.3.6). Les résultats pour le cas 100 % de réponse sont présentés dans le tableau suivant :

TABLEAU 5.2. Estimateur Horvitz-Thompson pour 100 % de réponse (Population MU281-yratioz, plan SI, n=100, mécanisme uniforme).

| | |
|---|---|
| $\text{EMC}(\hat{\bar{y}}_{\text{HT}})$ | 2247,56 |
| $\text{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{HT}}))$ | 36 671,48 |
| $\mathbb{V}^*_{\text{ECH}}$ | 36 315,91 |
| $\text{VMC}(\hat{\bar{y}}_{\text{HT}})$ | 36 284,72 |
| $\text{TRMC}(\hat{\bar{y}}_{\text{HT}})$ | 94,0 % |
| $\text{Birel}(\widehat{\mathbb{V}})$ | 1,07 % |

Nous obtenons un biais de simulation (2247,56 - 2249,74) / 2249,74= -0,10 % pour l'estimateur ponctuel et de (36 284,72 - 36 765,82)/36 765,82 = -1,31 % pour la variance Monte Carlo $\text{VMC}(\hat{\bar{y}}_{\text{HT}})$. La différence entre $\text{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{HT}}))$ et $\mathbb{V}^*_{\text{ECH}}$ est très minime (0,98 %). De plus, le taux de recouvrement se trouve très près du taux théorique de 95 %.

Pour ce qui est de l'estimateur par le ratio, il possède les propriétés mentionnées à la section 1.4. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.9) et (1.4.11). Les résultats dans le cas 100 % de réponse sont présentés dans le tableau 5.3.

TABLEAU 5.3. Estimateur par le ratio pour 100 % de réponse (Population MU281-yratioz, plan SI, n=100, mécanisme uniforme).

| | |
|---|---|
| $\text{EMC}(\hat{\bar{y}}_{\text{RA}})$ | 2252,52 |
| $\text{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{RA}}))$ | 6999,12 |
| $\mathbb{V}^*_{\text{ECH}}$ | 7069,69 |
| $\text{VMC}(\hat{\bar{y}}_{\text{RA}})$ | 6988,47 |
| $\text{TRMC}(\hat{\bar{y}}_{\text{RA}})$ | 91,7 % |
| $\text{Birel}(\widehat{\mathbb{V}})$ | 0,15 % |

Encore une fois, les résultats obtenus sont excellents. Le biais relatif de l'estimateur ponctuel est de $(2252,52-2249,74)/2249,74 = 0{,}12$ %; celui de l'estimateur de la variance est de $(6988,47-6781,67)/6781,67 = 3{,}05$ %. De plus, la différence entre $\mathrm{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\mathrm{RA}}))$ et $\mathbb{V}^*_{\mathrm{ECH}}$ est encore très minime (-1,00 %).

Pour ce qui est de l'estimateur par la régression, il possède les propriétés mentionnées à la section 1.4. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.12) et (1.4.14). Les résultats dans le cas 100 % de réponse sont présentés dans le tableau 5.4.

TABLEAU 5.4. Estimateur par la régression pour 100 % de réponse (Population MU281-yratioz, plan SI, n=100, mécanisme uniforme).

| | |
|---|---|
| $\mathrm{EMC}(\hat{\bar{y}}_{\mathrm{REG}})$ | 2251,24 |
| $\mathrm{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\mathrm{REG}}))$ | 6557,33 |
| $\mathbb{V}^*_{\mathrm{ECH}}$ | 6697,38 |
| $\mathrm{VMC}(\hat{\bar{y}}_{\mathrm{REG}})$ | 6691,82 |
| $\mathrm{TRMC}(\hat{\bar{y}}_{\mathrm{REG}})$ | 91,5 % |
| $\mathrm{Birel}(\widehat{\mathbb{V}})$ | -1,83 % |

Les résultats donnés par la méthode de Monte Carlo sont aussi bons que ceux pour l'estimateur HT et pour l'estimateur par le ratio. Le biais relatif de l'estimateur ponctuel est de $(2251,24-2249,74)/2249,74 = 0{,}07\%$; celui de l'estimateur de la variance est de $(6691,82-6491,36)/6491,36 = 3{,}09$ %, ce qui est bon. Encore une fois, nous pouvons constater que la différence entre $\mathrm{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\mathrm{REG}}))$ et $\mathbb{V}^*_{\mathrm{ECH}}$ est très minime (différence de -2,09 %).

## 5.3. Résultats pour la non-réponse avec mécanisme de réponse uniforme

### 5.3.1. Formules utilisées dans les simulations

Peu importe l'estimateur choisi, le biais relatif (tel qu'il est mesuré par la simulation de Monte Carlo) ainsi que la variance totale sont définis respectivement par

$$\text{Birel}(\widehat{\mathbb{V}}_{\bullet}) = 100 \cdot \frac{\text{EMC}(\widehat{\mathbb{V}}_{\bullet}) - \text{VMC}(\hat{\bar{y}}_U)}{\text{VMC}(\hat{\bar{y}}_U)}$$

$$\text{Birel}(\widehat{\mathbb{V}}_{\bullet\text{INT}}) = 100 \cdot \frac{\text{EMC}(\widehat{\mathbb{V}}_{\bullet\text{INT}}) - \text{VMC}(\hat{\bar{y}}_U)}{\text{VMC}(\hat{\bar{y}}_U)} \tag{5.3.1}$$

et

$$\text{EMC}(\widehat{\mathbb{V}}_{\bullet}) = \text{EMC}(\widehat{\mathbb{V}}_{\bullet\text{ECH}}) + \text{EMC}(\widehat{\mathbb{V}}_{\text{IMP}}) - \text{EMC}(\widehat{diff})$$

$$\text{EMC}(\widehat{\mathbb{V}}_{\bullet\text{INT}}) = \text{EMC}(\widehat{\mathbb{V}}_{\bullet\text{ECH}}) + \text{EMC}(\widehat{\mathbb{V}}_{\text{IMP}}) \tag{5.3.2}$$

avec $\widehat{\mathbb{V}}_{\bullet\text{ECH}}$, l'estimateur de variance due à l'échantillonnage donné par (3.4.3), $\widehat{\mathbb{V}}_{\text{IMP}}$, l'estimateur de variance due à l'imputation donnée par (3.4.14) et $\widehat{diff}$, le terme d'ajustement donné par (3.4.11). En plus d'évaluer ces termes pour chacune des simulations du chapitre 5 et 6, nous allons évaluer les composantes $\mathbb{V}^*_{\text{ECH}}$, $\mathbb{V}^*_{\text{IMP}}$ et $\mathbb{V}^*_{\text{MIX}}$ qui sont définies respectivement comme l'espérance de Monte Carlo de (3.3.4), (3.3.5) et (3.3.6).

### 5.3.2. Résultats de simulation utilisant l'estimateur d'Horvitz-Thompson

Les formules utilisées dans les simulations pour cet estimateur sont données par (1.3.4) et (1.3.6). Le biais relatif ainsi que la variance totale, dans le cas 60 %, 80 % et 90 % de réponse, sont définis respectivement par (5.3.1) et (5.3.2). Nous présentons dans le tableau 5.5 les résultats de simulation avec non-réponse selon ce mécanisme de réponse.

TABLEAU 5.5. Estimateur Horvitz-Thompson pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281-yratioz, plan SI, n=100, mécanisme uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 2237,02 | 2243,67 | 2246,47 |
| VMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 49 610,75 | 41 406,12 | 38 767,19 |
| EMC($\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}}$) | 36 675,76 | 36 488,94 | 36 514,10 |
| $\mathbb{V}^*_{\mathrm{IMP}}$ | 13 435,60 | 5268,72 | 2431,06 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$) | 13 172,31 | 5075,81 | 2325,21 |
| $\mathbb{V}^*_{\mathrm{MIX}}$ | -68,32 | -93,21 | 4,08 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{MIX}}$) | -44,06 | -14,04 | -5,67 |
| EMC($\widehat{diff}$) | 327,92 | 100,46 | 40,33 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 49 520,15 | 41 464,29 | 38 798,99 |
| TRMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 92,7 % | 93,5 % | 93,9 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | -0,18 % | 0,14 % | 0.08 % |
| Birel($\widehat{\mathbb{V}}_{\bullet \mathrm{INT}}$) | 0,48 % | 0,38 % | 0,19 % |

En regardant les résultats de Birel($\widehat{\mathbb{V}}_{\bullet}$) pour l'estimateur HT, nous observons que notre estimateur proposé, $\widehat{\mathbb{V}}_{\bullet}$, possède un biais égal à -0,18 %, 0,14 % et 0.08 % pour les taux de réponse de 60 %, 80 % et

90 % respectivement. De plus, le taux de recouvrement est excellent. Nous remarquons qu'en valeur absolue, Birel($\widehat{\mathbb{V}}_{\bullet}$) est inférieur à Birel($\widehat{\mathbb{V}}_{\bullet \text{INT}}$) pour tous les taux de réponse. Nous remarquons également que $\mathbb{V}^*_{\text{MIX}}$ et EMC($\widehat{\mathbb{V}}_{\text{MIX}}$) sont négligeables pour les trois taux de réponse. Finalement, si nous comparons notre estimateur EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) avec $\mathbb{V}^*_{\text{IMP}}$, nous remarquons que notre estimateur traduit bien la variance due à l'imputation (moins de 4,3 % de différence). À la lumière de ces résultats, l'approche suggérée semble bien fonctionner.

### 5.3.3. Résultats de simulation utilisant l'estimateur par le ratio

Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.9) et (1.4.11). Le biais relatif ainsi que la variance totale, dans le cas 60 %, 80 % et 90 % de réponse, sont définis respectivement par (5.3.1) et (5.3.2). Toujours en utilisant le plan SI et ce mécanisme de réponse, nous avons obtenu pour l'estimateur par le ratio les résultats de simulation présentés dans le tableau 5.6 de la page suivante.

Le tableau 5.6 montre que le biais relatif de notre estimateur de variance $\widehat{\mathbb{V}}_{\bullet}$, donné par (5.3.2), est de 0,27 %, 1,74 % et 0,31 % pour un taux de réponse respectivement de 60 %, 80 % et 90 %. Nous remarquons que Birel($\widehat{\mathbb{V}}_{\bullet}$) est inférieur à Birel($\widehat{\mathbb{V}}_{\bullet \text{INT}}$) pour tous les taux de réponse. Nous remarquons également que $\mathbb{V}^*_{\text{MIX}}$ et EMC($\widehat{\mathbb{V}}_{\text{MIX}}$) sont négligeables. Finalement, si nous comparons notre estimateur EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) avec $\mathbb{V}^*_{\text{IMP}}$, nous remarquons que notre estimateur traduit bien la variance due à l'imputation (moins de 5,0 % de différence). Nous considérons que ces résultats sont excellents.

TABLEAU 5.6. Estimateur par le ratio pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281-yratioz, plan SI, n=100, mécanisme uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet RA}$) | 2242,12 | 2248,55 | 2252,10 |
| VMC($\hat{\bar{y}}_{\bullet RA}$) | 20 761,87 | 12 281,05 | 9550,64 |
| EMC($\widehat{\mathbb{V}}_{\bullet ECH}$) | 8036,93 | 7419,13 | 7242,04 |
| $\mathbb{V}^*_{IMP}$ | 13 827,99 | 5389,13 | 2492,57 |
| EMC($\widehat{\mathbb{V}}_{IMP}$) | 13 124,01 | 5182,80 | 2380,33 |
| $\mathbb{V}^*_{MIX}$ | 149,04 | 13,35 | 32,10 |
| EMC($\widehat{\mathbb{V}}_{MIX}$) | -3,94 | -1,29 | -0,44 |
| EMC($\widehat{diff}$) | 343,91 | 106,81 | 42,50 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 20 817,04 | 12 495,12 | 9579,87 |
| TRMC($\hat{\bar{y}}_{\bullet RA}$) | 91,1 % | 92,1 % | 92,7 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | 0,27 % | 1,74 % | 0,31 % |
| Birel($\widehat{\mathbb{V}}_{\bullet INT}$) | 1,92 % | 2,61 % | 0,75 % |

### 5.3.4. Résultats de simulation utilisant l'estimateur par la régression

L'estimateur par la régression possède les propriétés mentionnées à la page 10. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.12) et (1.4.14). Toujours en utilisant le plan SI et le mécanisme de réponse uniforme, nous avons obtenu pour l'estimateur par le régression les résultats suivants :

TABLEAU 5.7. Estimateur par la régression pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281-yratioz, plan SI, n=100, mécanisme uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet \text{REG}}$) | 2239,97 | 2247,05 | 2249,94 |
| VMC($\hat{\bar{y}}_{\bullet \text{REG}}$) | 20 321,85 | 12 057,70 | 9160,55 |
| EMC($\widehat{\mathbb{V}}_{\bullet \text{ECH}}$) | 7307,68 | 6862,46 | 6799,37 |
| $\mathbb{V}^*_{\text{IMP}}$ | 13 849,64 | 5380,44 | 2495,75 |
| EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) | 13 306,44 | 5269,93 | 2396,20 |
| $\mathbb{V}^*_{\text{MIX}}$ | 99,44 | -19,71 | 16,47 |
| EMC($\widehat{\mathbb{V}}_{\text{MIX}}$) | -16,46 | -5,37 | -2,16 |
| EMC($\widehat{diff}$) | 458,04 | 133,16 | 53,95 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 20 156,08 | 11 999,23 | 9141,63 |
| TRMC($\hat{\bar{y}}_{\bullet \text{GREG}}$) | 90,3 % | 91,4 % | 92,6 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | -0,82 % | -0,48 % | -0,21 % |
| Birel($\widehat{\mathbb{V}}_{\bullet \text{INT}}$) | 1,44 % | 0,62 % | 0,38 % |

En regardant les résultats pour l'estimateur $\hat{\bar{y}}_{\bullet \text{REG}}$, nous relevons un biais relatif de notre estimateur de variance $\widehat{\mathbb{V}}_{\bullet}$ égal à -0,82 %, -0,48 % et -0,21 % pour un taux de réponse respectivement de 60 %, 80 % et 90 %. Nous remarquons qu'en valeur absolue, Birel($\widehat{\mathbb{V}}_{\bullet}$) est inférieur à Birel($\widehat{\mathbb{V}}_{\bullet \text{INT}}$) pour tous les taux de réponse. Nous remarquons également que $\mathbb{V}^*_{\text{MIX}}$ et EMC($\widehat{\mathbb{V}}_{\text{MIX}}$) sont négligeables. Finalement, si nous comparons notre estimateur EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) avec $\mathbb{V}^*_{\text{IMP}}$, nous remarquons que notre estimateur traduit bien la variance due à l'imputation (moins de 4,0 % de différence). Nous considérons que ces résultats sont excellents.

## 5.4. Résumé des résultats

Nous remarquons dans les tableaux 5.2, 5.3, 5.4, 5.5, 5.6 et 5.7 que EMC($\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{GREG}})$) se situe très près de EMC($\widehat{\mathbb{V}}_{\bullet\text{ECH}}$) pour les trois estimateurs. Le résultat similaire n'est pas surprenant puisque ces deux estimateurs de Monte Carlo traduisent la variance échantillonnale de l'expérience.

De plus, en comparant VMC($\hat{\bar{y}}_{\text{GREG}}$) et VMC($\hat{\bar{y}}_{\bullet\text{GREG}}$) pour chacun de ces trois estimateurs, nous voyons clairement qu'il aurait été faux de nous baser seulement sur la variance échantillonnale pour obtenir l'évaluation de la variance totale, comme nous procèdons dans le cas de 100 % de réponse. L'apport de la variance due à l'imputation est importante et ne peut être proscrit des calculs. Nous observons également que le terme d'ajustement $diff$ améliore les résultats puisqu'il approche les deux variances $\widehat{\mathbb{V}}_{\bullet}$ et VMC($\hat{\bar{y}}_{\bullet\text{GREG}}$) le plus près possible et nous remarquons que Birel($\widehat{\mathbb{V}}_{\bullet}$) est inférieur à Birel($\widehat{\mathbb{V}}_{\bullet\text{INT}}$) pour les trois taux de réponse et les trois estimateurs. De plus, nous avons observé que l'espérance de Monte Carlo de la variance mixte et celle de l'estimateur de la variance mixte sont toujours négligeables peu importe l'estimateur choisi et le taux de réponse.

Les taux de recouvrement de $\bar{y}_U$ oscillent entre 90,3 % et 94,0 % lorsque l'on observe les tableaux de la population MU281-yratioz. D'autre part, pour l'ensemble des résultats obtenus, EMC($\hat{\bar{y}}_U$) se situe très près de $\bar{y}_U$. Le biais relatif de chacun des estimateurs et pour chacun des taux de réponse se situe en deça de 1 %.

Finalement, comme prévu, lorsque l'on utilise le mécanisme de réponse uniforme, l'estimateur par le ratio et l'estimateur par la régression traduisent le plus fidèlement la population de Suède MU281-yratioz décrite au chapitre 4. Ils favorisent de meilleurs résultats par rapport à l'estimateur HT puisqu'il fait appel à une variable auxiliaire.

CHAPITRE 6

# Résultats des simulations pour les cas déviant du cas de base

## 6.1. Résultats sur la population MU281

Pour tester la robustesse de nos estimateurs développés au chapitre 3, nous allons considérer des cas déviant du cas de base présenté au chapitre 5. Dans ce présent chapitre, nous utiliserons une population qui ne satisfait pas le modèle d'imputation $\xi$ donné par l'équation (3.3.8) et nous utiliserons un mécanisme de réponse uniforme et un mécanisme de réponse non uniforme pour sélectionner les répondants. Dans toutes les simulations rapportées dans ce chapitre, la population MU281 est utilisée. Avant d'entamer la section des résultats de simulation, nous présentons, dans le tableau 6.1, les résultats d'une analyse descriptive de la moyenne de la population MU281 et de la variance de l'estimateur HT, du ratio et de la régression sous un plan SI, avec n=100. En utilisant les formules (1.3.5), (1.4.10) et (1.4.13) correspondant aux formules de variance sur la population, nous avons obtenu

TABLEAU 6.1. Population MU281, plan SI, n=100. Variance sur la population de l'estimateur d'Horvitz-Thompson, l'estimateur par le ratio et l'estimateur par la régression.

| | |
|---|---|
| $\bar{y}_U$ | 2694,83 |
| $\mathbb{V}(\hat{\bar{y}}_{\mathrm{HT}})$ | 36 765,82 |
| $A\mathbb{V}(\hat{\bar{y}}_{\mathrm{RA}})$ | 9497,89 |
| $A\mathbb{V}(\hat{\bar{y}}_{\mathrm{REG}})$ | 6491,36 |

Remarquons que la variance de $\hat{\bar{y}}_{\mathrm{RA}}$ est nettement inférieure à la variance de $\hat{\bar{y}}_{\mathrm{HT}}$. Ceci s'explique par la forte relation entre $x$ et $y$ rendue évidente par la figure 4.6. De plus, nous avons vu au chapitre 4 que l'analyse de régression ne rejettait pas l'hypothèse d'une ordonnée à l'origine pour $\alpha = 1$ %. En regardant le tableau précédent, nous voyons bien que l'existence d'une ordonnée à l'origine est plausible puisque la variance de $\hat{\bar{y}}_{\mathrm{REG}}$ réalise une amélioration impressionnante sur la variance de $\hat{\bar{y}}_{\mathrm{RA}}$. Nous pouvons donc s'attendre à ce que les résultats soient meilleurs pour l'estimateur par la régression que pour l'estimateur HT et l'estimateur par le ratio.

## 6.2. Résultats pour 100 % de réponse dans l'échantillon

Peu importe l'estimateur choisi, le biais relatif (tel qu'il est mesuré par la simulation de Monte Carlo) dans le cas 100 % de réponse est défini par (5.2.1). De plus, la quantité $\mathbb{V}^*_{\mathrm{ECH}}$ que nous retrouvons dans les tableaux 6.2, 6.3 et 6.4 est définie comme la moyenne de Monte Carlo de (3.3.4). Présentons maintenant les résultats obtenus pour l'estimateur HT, l'estimateur par le ratio et l'estimateur par la régression par la méthode de Monte Carlo dans le cas où il n'y a pas de non-réponse dans l'échantillon.

Pour ce qui est de l'estimateur HT, il possède les propriétés mentionnées à la section 1.3. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.3.4) et (1.3.6). Les résultats pour le cas 100 % de réponse sont présentés dans le tableau suivant :

TABLEAU 6.2. Estimateur d'Horvitz-Thompson pour 100 % de réponse (Population MU281, plan SI, n=100, mécanisme uniforme).

| | |
|---|---|
| EMC($\hat{\bar{y}}_{\text{HT}}$) | 2693,76 |
| EMC($\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{HT}})$) | 36 655,37 |
| $\mathbb{V}^*_{\text{ECH}}$ | 36 103,95 |
| VMC($\hat{\bar{y}}_{\text{HT}}$) | 36 116,93 |
| TRMC($\hat{\bar{y}}_{\text{HT}}$) | 93,9 % |
| Birel($\widehat{\mathbb{V}}$) | 1,49 % |

Nous obtenons un biais de simulation (2693,76 - 2694,83) / 2694,83= -0,04 % pour l'estimateur ponctuel et de (36 116,93 - 36 765,82)/36 765,82 = -1,76 % pour l'estimateur de variance. De plus, le taux de recouvrement est excellent, le taux observé de 93,9 % se trouvant très près du taux théorique de 95 %. Finalement, nous observons que EMC($\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{HT}})$) est très près de $\mathbb{V}^*_{\text{ECH}}$ (différence de -1,50 %).

Pour ce qui est de l'estimateur par le ratio, il possède les propriétés mentionnées à la section 1.4. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.9) et (1.4.11). Les résultats dans le cas 100 % de réponse sont présentés dans le tableau 6.3.

TABLEAU 6.3. Estimateur par le ratio pour 100 % de réponse (Population MU281, plan SI, n=100, mécanisme uniforme).

| | |
|---|---|
| EMC($\hat{\bar{y}}_{\text{RA}}$) | 2700,19 |
| EMC($\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{RA}})$) | 9670,58 |
| $\mathbb{V}^*_{\text{ECH}}$ | 9700,34 |
| VMC($\hat{\bar{y}}_{\text{RA}}$) | 9934,20 |
| TRMC($\hat{\bar{y}}_{\text{RA}}$) | 92,6 % |
| Birel($\widehat{\mathbb{V}}$) | 0,65 % |

Encore une fois, les résultats obtenus sont bons. Le biais relatif de l'estimateur ponctuel est de $(2700,19 - 2694,83) / 2694,83 = 0{,}20$ %; celui de l'estimateur de la variance est de $(9934,20 - 9497,89)/9497,89 = 4{,}59$ %. Finalement, nous observons que EMC($\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{RA}})$) est très près de $\mathbb{V}^*_{\text{ECH}}$ (différence de -0,31 %).

Pour ce qui est de l'estimateur par la régression, il possède les propriétés mentionnées à la section 1.4. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.12) et (1.4.14). Les résultats dans le cas 100 % de réponse sont présentés dans le tableau 6.4.

TABLEAU 6.4. Estimateur par la régression pour 100 % de réponse (Population MU281, plan SI, n=100, mécanisme uniforme).

| | |
|---|---|
| EMC($\hat{\bar{y}}_{\text{REG}}$) | 2694,95 |
| EMC($\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{REG}})$) | 6501,79 |
| $\mathbb{V}^*_{\text{ECH}}$ | 6452,41 |
| VMC($\hat{\bar{y}}_{\text{REG}}$) | 6555,06 |
| TRMC($\hat{\bar{y}}_{\text{REG}}$) | 91,1 % |
| Birel($\widehat{\mathbb{V}}$) | -2,97 % |

Les résultats donnés par la méthode de Monte Carlo sont aussi bons que ceux de l'estimateur HT et de l'estimateur par le ratio. Le biais relatif de l'estimateur ponctuel est de $(2694,95-2694,83)/2694,83$ =0,01%; celui de l'estimateur de la variance est de $(6555,06-6491,36)/6491,36 =$ 0,98 %, ce qui est excellent. Finalement, nous observons que $\mathrm{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\mathrm{REG}}))$ est très près de $\mathbb{V}^*_{\mathrm{ECH}}$ (différence de -0,77 %).

## 6.3. Résultats pour la non-réponse avec mécanisme de réponse uniforme

Comme nous l'avons mentionné au chapitre 5, peu importe l'estimateur choisi, le biais relatif tel qu'il est mesuré par la simulation de Monte Carlo ainsi que la variance totale sont définis respectivement par (5.3.1) et (5.3.2). Nous allons également évaluer $\widehat{\mathbb{V}}_{\bullet\mathrm{ECH}}$, l'estimateur de variance due à l'échantillonnage donné par (3.4.3), $\widehat{\mathbb{V}}_{\mathrm{IMP}}$, l'estimateur de variance due à l'imputation donnée par (3.4.14), $\widehat{diff}$, le terme d'ajustement donné par (3.4.11) et nous allons également évaluer les composantes $\mathbb{V}^*_{\mathrm{ECH}}$, $\mathbb{V}^*_{\mathrm{IMP}}$ et $\mathbb{V}^*_{\mathrm{MIX}}$ qui sont définies respectivement comme l'espérance de Monte Carlo de (3.3.4), (3.3.5) et (3.3.6).

### 6.3.1. Résultats de simulation utilisant l'estimateur d'Horvitz-Thompson

Les formules utilisées dans les simulations pour cet estimateur sont données par (1.3.4) et (1.3.6). Nous présentons dans le tableau 6.5 les résultats de cette simulation selon ce mécanisme de réponse.

TABLEAU 6.5. Estimateur d'Horvitz-Thompson pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281, plan SI, n=100, mécanisme uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 2683,82 | 2689,45 | 2692,59 |
| VMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 49 542,21 | 41 151,63 | 38 601,18 |
| EMC($\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}}$) | 35 678,75 | 36 273,72 | 36 490,63 |
| $\mathbb{V}^*_{\mathrm{IMP}}$ | 13 695,10 | 5614,18 | 2605,30 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$) | 14 341,00 | 5877,30 | 2710,72 |
| $\mathbb{V}^*_{\mathrm{MIX}}$ | -28,91 | -71,33 | 40,14 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{MIX}}$) | -47,67 | -15,69 | -6,35 |
| EMC($\widehat{diff}$) | 473,05 | 144,04 | 57,84 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 49 946,70 | 42 006,98 | 39 038,09 |
| TRMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 93,0 % | 93,5 % | 93,9 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | 0,82 % | 2,08 % | 1,13 % |
| Birel($\widehat{\mathbb{V}}_{\bullet \mathrm{INT}}$) | 1,77 % | 2,43 % | 1,28 % |

En regardant les résultats, nous observons que Birel($\widehat{\mathbb{V}}_{\bullet}$) et Birel($\widehat{\mathbb{V}}_{\bullet \mathrm{INT}}$) sont très peu élevés. Nous voyons également qu'en ajoutant EMC($\widehat{diff}$), les résultats sont améliorés. De plus, nous remarquons que notre estimateur EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$) est près de $\mathbb{V}^*_{\mathrm{IMP}}$ pour les trois taux de réponse (moins de 4,7 % de différence). Finalement, $\mathbb{V}^*_{\mathrm{MIX}}$ et EMC($\widehat{\mathbb{V}}_{\mathrm{MIX}}$) sont négligeables et le taux de recouvrement est excellent. À la lumière de ces résultats, l'approche suggérée semble bien fonctionner.

### 6.3.2. Résultats de simulation utilisant l'estimateur par le ratio

Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.9) et (1.4.11). Toujours en utilisant le plan SI et ce mécanisme de réponse, nous avons obtenu pour l'estimateur par le ratio les résultats de simulation présentés dans le tableau 6.6.

TABLEAU 6.6. Estimateur par le ratio pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281, plan SI, n=100, mécanisme uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet \mathrm{RA}}$) | 2687,46 | 2694,90 | 2698,38 |
| VMC($\hat{\bar{y}}_{\bullet \mathrm{RA}}$) | 23 980,56 | 15 349,19 | 12 488,07 |
| EMC($\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}}$) | 11 031.73 | 10 225,10 | 9969,63 |
| $\mathbb{V}^*_{\mathrm{IMP}}$ | 14 527,99 | 5789,13 | 2652,57 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$) | 15 082,30 | 6007,63 | 2765,36 |
| $\mathbb{V}^*_{\mathrm{MIX}}$ | 107,45 | -4,93 | 24,83 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{MIX}}$) | -2,15 | -0,86 | -0,23 |
| EMC($\widehat{diff}$) | 500,99 | 151,83 | 60,25 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 25 613,04 | 16 080,90 | 12 674,74 |
| TRMC($\hat{\bar{y}}_{\bullet \mathrm{RA}}$) | 92,7 % | 93,3 % | 93,2 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | 6,80 % | 4,77 % | 1,49 % |
| Birel($\widehat{\mathbb{V}}_{\bullet \mathrm{INT}}$) | 8,90 % | 5,76 % | 1,98 % |

Le tableau 6.6 montre que le biais relatif de notre estimateur de variance $\widehat{\mathbb{V}}_{\bullet}$, est de 6,80 %, 4,77 %, 1,49 % pour un taux de réponse respectivement de 60 %, 80 % et 90 %. Même si ces pourcentages ne sont pas très près de 0, nous considérons que ces résultats sont

encourageants, compte tenu du fait que notre procédure d'estimation de variance est issue d'un raisonnement assez complexe ayant eu recours à plusieurs approximations. De plus, l'estimateur par le ratio ne reflète pas la population utilisée puisque nous savons que l'hypothèse d'une ordonnée à l'origine s'est révélée non satisfaite pour $\alpha = 1$ % (voir chapitre 4). Nous remarquons également que l'ajout de EMC($\widehat{diff}$) dans le calcul de la variance améliore les résultats puisque Birel($\widehat{\mathbb{V}}_{\bullet}$) est inférieur à Birel($\widehat{\mathbb{V}}_{\bullet \text{INT}}$). Les quantités $\mathbb{V}^*_{\text{MIX}}$ et EMC($\widehat{\mathbb{V}}_{\text{MIX}}$) sont encore négligeables peu importe le taux de réponse. Finalement, si nous comparons notre estimateur EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) avec $\mathbb{V}^*_{\text{IMP}}$, nous remarquons que notre estimateur traduit bien la variance due à l'imputation (moins de 4,3 % de différence).

### 6.3.3. Résultats de simulation utilisant l'estimateur par la régression

L'estimateur par la régression possède les propriétés mentionnées à la page 10. Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.12) et (1.4.14). Toujours en utilisant le plan SI et le mécanisme de réponse uniforme, nous avons obtenu pour l'estimateur par le régression les résultats de simulation présentés dans le tableau 6.7 de la page suivante.

TABLEAU 6.7. Estimateur par la régression pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281, plan SI, n=100, mécanisme uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| $\text{EMC}(\hat{\bar{y}}_{\bullet\text{REG}})$ | 2683,72 | 2691,85 | 2694,86 |
| $\text{VMC}(\hat{\bar{y}}_{\bullet\text{REG}})$ | 20 843,02 | 11 939,86 | 9146,58 |
| $\text{EMC}(\widehat{\mathbb{V}}_{\bullet\text{ECH}})$ | 7233,70 | 6833,43 | 6695,14 |
| $\mathbb{V}^*_{\text{IMP}}$ | 14 220,50 | 5580,65 | 2593,18 |
| $\text{EMC}(\widehat{\mathbb{V}}_{\text{IMP}})$ | 14 740,75 | 5753,10 | 2624,06 |
| $\mathbb{V}^*_{\text{MIX}}$ | 113,24 | -30,15 | 20,03 |
| $\text{EMC}(\widehat{\mathbb{V}}_{\text{MIX}})$ | -17,42 | -6,02 | -2,22 |
| $\text{EMC}(\widehat{diff})$ | 655,46 | 204,55 | 79,38 |
| $\text{EMC}(\widehat{\mathbb{V}}_{\bullet})$ | 21 318,99 | 12 381,98 | 9239,82 |
| $\text{TRMC}(\hat{\bar{y}}_{\bullet\text{GREG}})$ | 91,0 % | 92,1 % | 91,7 % |
| $\text{Birel}(\widehat{\mathbb{V}}_{\bullet})$ | 2,28 % | 3,70 % | 1,02 % |
| $\text{Birel}(\widehat{\mathbb{V}}_{\bullet\text{INT}})$ | 5,43 % | 5,42 % | 1,89 % |

En regardant les résultats pour l'estimateur $\hat{\bar{y}}_{\bullet\text{GREG}}$, nous relevons un biais relatif de notre estimateur de variance $\widehat{\mathbb{V}}_{\bullet}$ égal à 2,28 %, 3,70 % et 1,02 % pour un taux de réponse respectivement de 60 %, 80 % et 90 %. Nous voyons bien que le biais relatif a diminué si l'on compare ces résultats avec l'estimateur par le ratio. Nous remarquons encore que $\text{Birel}(\widehat{\mathbb{V}}_{\bullet})$ est inférieur à $\text{Birel}(\widehat{\mathbb{V}}_{\bullet\text{INT}})$ peu importe le taux de réponse. De plus, il est encore vrai d'affirmer que $\mathbb{V}^*_{\text{MIX}}$ et $\text{EMC}(\widehat{\mathbb{V}}_{\text{MIX}})$ sont négligeables. Finalement, si nous comparons notre estimateur $\text{EMC}(\widehat{\mathbb{V}}_{\text{IMP}})$ avec $\mathbb{V}^*_{\text{IMP}}$, nous remarquons que notre estimateur traduit bien la variance due à l'imputation (moins de 3,7 % de différence). Nous considérons ces résultats excellents.

## 6.4. Résumé des résultats pour un mécanisme de réponse uniforme

Nous remarquons dans les tableaux 6.2 à 6.7 que $\text{EMC}(\widehat{\mathbb{V}}(\hat{\bar{y}}_{\text{GREG}}))$ se situe encore très près de $\text{EMC}(\widehat{\mathbb{V}}_{\bullet\text{ECH}})$ pour les trois estimateurs.

De plus, en comparant $\text{VMC}(\hat{\bar{y}}_{\text{GREG}})$ et $\text{VMC}(\hat{\bar{y}}_{\bullet\text{GREG}})$ pour chacun de ces trois estimateurs, nous voyons bien que l'apport de la variance due à l'imputation est importante. Nous observons également que le terme d'ajustement $diff$ améliore les résultats puisqu'il approche les deux variances $\widehat{\mathbb{V}}_{\bullet}$ et $\text{VMC}(\hat{\bar{y}}_{\bullet\text{GREG}})$ le plus près possible et nous remarquons que $\text{Birel}(\widehat{\mathbb{V}}_{\bullet})$ est inférieur à $\text{Birel}(\widehat{\mathbb{V}}_{\bullet\text{INT}})$ pour les trois taux de réponse et les trois estimateurs. Les taux de recouvrement de $\bar{y}_U$ oscillent entre 91,0 % et 93,9 % lorsque l'on observe les tableaux de la population MU281. D'autre part, pour l'ensemble des résultats obtenus, $\text{EMC}(\hat{\bar{y}}_U)$ se situe très près de $\bar{y}_U$. Le biais relatif de chacun des estimateurs et pour chacun des trois taux de réponse se situe en deça de 1 %.

Finalement, lorsque l'on utilise le mécanisme de réponse uniforme, l'estimateur par la régression est celui traduisant le plus fidèlement la population de Suède MU281 décrite au chapitre 4. Il favorise de meilleurs résultats par rapport à l'estimateur HT et l'estimateur par le ratio, puisqu'il fait appel à une variable auxiliaire et présuppose une ordonnée à l'origine. Par contre, en terme de biais relatif, les estimateurs de variances proposés au chapitre 3 réagissent bien dans les simulations de Monte Carlo pour l'estimateur Horvitz-Thompson et l'estimateur par la régression.

## 6.5. Résultats pour la non-réponse avec mécanisme de réponse non uniforme

### 6.5.1. Résultats de simulation utilisant l'estimateur d'Horvitz-Thompson

En utilisant à la fois le plan SI et les mécanismes de réponse non uniformes présentés à la section 4.3, nous avons obtenu pour l'estimateur HT les résultats présentés au tableau 6.8.

TABLEAU 6.8. Estimateur d'Horvitz-Thompson pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281, plan SI, n=100, mécanisme non uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 2534,44 | 2665,50 | 2686,75 |
| VMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 73 341,92 | 48 355,01 | 41 586,00 |
| EMC($\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}}$) | 26 287,76 | 33 877,23 | 35 856,16 |
| $\mathbb{V}^*_{\mathrm{IMP}}$ | 61 610,19 | 12 570,23 | 3872,26 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$) | 50 007,41 | 15 029,33 | 5956,91 |
| $\mathbb{V}^*_{\mathrm{MIX}}$ | -1328,86 | -239,11 | 133,12 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{MIX}}$) | -781,40 | -172,15 | -47,93 |
| EMC($\widehat{diff}$) | 8946,65 | 1700,10 | 449,98 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 67 348,52 | 47 206,46 | 41 363,09 |
| TRMC($\hat{\bar{y}}_{\bullet \mathrm{HT}}$) | 83,0 % | 91,5 % | 92,6 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | -8,17 % | -2,38 % | -0,54 % |
| Birel($\widehat{\mathbb{V}}_{\bullet \mathrm{INT}}$) | 4,63 % | 6,70 % | 1,02 % |

En examinant attentivement le tableau 6.8, nous constatons qu'au moment où l'on doit recourir à l'imputation pour un taux élevé de non-réponse, ce mécanisme de réponse donne de mauvais résultats en terme de biais relatif par rapport au mécanisme 1.5.2. Contrairement au mécanisme de réponse uniforme, nous remarquons une différence beaucoup plus marquée entre notre estimateur $\mathrm{EMC}(\widehat{\mathbb{V}}_{\mathrm{IMP}})$ et $\mathbb{V}^*_{\mathrm{IMP}}$ pour le mécanisme de réponse non uniforme. Notre estimateur semble traduire moins bien la variance due à l'imputation.

D'un autre côté, $\mathbb{V}^*_{\mathrm{MIX}}$ et $\mathrm{EMC}(\widehat{\mathbb{V}}_{\mathrm{MIX}})$ sont encore négligeables par rapport à $\mathrm{EMC}(\widehat{\mathbb{V}}_{\bullet})$. Nous remarquons également que $\mathrm{Birel}(\widehat{\mathbb{V}}_{\bullet}) < \mathrm{Birel}(\widehat{\mathbb{V}}_{\bullet\mathrm{INT}})$ pour les taux de réponse de 80 % et 90 %. À 60 % de réponse, nous observons que c'est le contraire qui se produit. Cet effet est expliqué par le fait que notre estimateur $\mathrm{EMC}(\widehat{\mathbb{V}}_{\mathrm{IMP}})$ est sous-estimé d'environ 20,5 % pour ce taux de réponse.

### 6.5.2. Résultats de simulation utilisant l'estimateur par le ratio

Les formules utilisées dans les simulations pour cet estimateur sont données par (1.4.9) et (1.4.11). Le biais relatif ainsi que la variance totale, dans le cas 60 %, 80 % et 90 % de réponse, sont définis respectivement par (5.3.1) et (5.3.2). Toujours en utilisant le plan SI et ce mécanisme de réponse, nous avons obtenu pour l'estimateur par le ratio les résultats de simulation présentés dans le tableau 6.9.

TABLEAU 6.9. Estimateur par le ratio pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281, plan SI, n=100, mécanisme non uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet \mathrm{RA}}$) | 2542,06 | 2660,35 | 2688,07 |
| VMC($\hat{\bar{y}}_{\bullet \mathrm{RA}}$) | 53 368,38 | 19 775,25 | 13 283,04 |
| EMC($\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}}$) | 16 358,65 | 12 128,67 | 10 695,20 |
| $\mathbb{V}^*_{\mathrm{IMP}}$ | 62 916,33 | 11 373,67 | 3881,65 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$) | 49 838,44 | 15 131,04 | 6036,90 |
| $\mathbb{V}^*_{\mathrm{MIX}}$ | 897,25 | -77,11 | -146,82 |
| EMC($\widehat{\mathbb{V}}_{\mathrm{MIX}}$) | 2,14 | -3,45 | -1,75 |
| EMC($\widehat{diff}$) | 8552,37 | 1715,99 | 456,35 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 57 644,73 | 25 543,71 | 16 275,76 |
| TRMC($\hat{\bar{y}}_{\bullet \mathrm{RA}}$) | 87,0 % | 94,3 % | 94,3 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | 8,01 % | 29,17 % | 22,53 % |
| Birel($\widehat{\mathbb{V}}_{\bullet \mathrm{INT}}$) | 24,04 % | 37,85 % | 25,97 % |

Le tableau 6.9 montre que le biais de $\widehat{\mathbb{V}}_{\bullet}$ est de 8,01 %, 29,17 % et 22,53 % pour les taux de réponse de 60 % à 90 %. Contrairement au mécanisme de réponse uniforme, nous remarquons une différence beaucoup plus prononcée entre notre estimateur EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$) et $\mathbb{V}^*_{\mathrm{IMP}}$ pour le mécanisme de réponse non uniforme. Le biais élevé que nous observons aux taux de réponse de 80 % et 90 % est principalement dû à une surestimation de notre estimateur EMC($\widehat{\mathbb{V}}_{\mathrm{IMP}}$), ce qui, par conséquent, surestime notre estimateur de variance totale EMC($\widehat{\mathbb{V}}_{\bullet}$). Encore une fois, pour ce mécanisme de réponse non uniforme, notre estimateur développé semble traduire moins bien la variance due à l'imputation.

Finalement, nous observons que Birel($\widehat{\mathbb{V}}_{\bullet}$) est inférieur à Birel($\widehat{\mathbb{V}}_{\bullet \text{INT}}$) pour tous les taux de réponse. Encore une fois, la covariance mixte s'est avérée négligeable.

### 6.5.3. Résultats de simulation utilisant l'estimateur par la régression

Analysons maintenant les résultats obtenus pour l'estimateur par la régression sous un mécanisme de réponse non uniforme pour les taux de réponse de 60 %, 80 % et 90 %.

TABLEAU 6.10. Estimateur par la régression pour les taux de réponse de 60 %, 80 % et 90 % (Population MU281, plan SI, n=100, mécanisme non uniforme).

| | Taux de réponse | | |
|---|---|---|---|
| | 60 % | 80 % | 90 % |
| EMC($\hat{\bar{y}}_{\bullet \text{REG}}$) | 2462,73 | 2582,56 | 2611,07 |
| VMC($\hat{\bar{y}}_{\bullet \text{REG}}$) | 48 044,15 | 17 149,18 | 11 227,31 |
| EMC($\widehat{\mathbb{V}}_{\bullet \text{ECH}}$) | 8648,27 | 7685,83 | 7685,83 |
| $\mathbb{V}^*_{\text{IMP}}$ | 60 158,40 | 11 288,19 | 4205,94 |
| EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) | 50 265,96 | 15 136,07 | 6052,74 |
| $\mathbb{V}^*_{\text{MIX}}$ | 300,33 | 63,04 | 40,09 |
| EMC($\widehat{\mathbb{V}}_{\text{MIX}}$) | -91,56 | -41,34 | -17,51 |
| EMC($\widehat{diff}$) | 8689,58 | 1968,48 | 567,54 |
| EMC($\widehat{\mathbb{V}}_{\bullet}$) | 50 224,65 | 20 853,42 | 13 171,13 |
| TRMC($\hat{\bar{y}}_{\bullet \text{GREG}}$) | 83,2 % | 92,2 % | 93,3 % |
| Birel($\widehat{\mathbb{V}}_{\bullet}$) | -4,54 % | 21,60 % | 17,31 % |
| Birel($\widehat{\mathbb{V}}_{\bullet \text{INT}}$) | 22,63 % | 33,08 % | 22,37 % |

À la section précédente, nous avons remarqué que l'estimateur par la régression était celui qui se comportait le mieux selon l'approche proposée pour définir la variance totale (5.3.2). Le tableau 6.10 montre que le biais de $\widehat{\mathbb{V}}_\bullet$ est de -4,54 %, 21,60 % et 17,31 % pour un taux de réponse respectivement de 60 %, 80 % et 90 %. Si l'on compare les résultats entre l'estimateur par le ratio et celui par la régression sous ce même mécanisme, nous retrouvons que les rapports ont diminué. Le biais relatif élevé aux taux de 80 % et 90 % de réponse est principalement dû au fait que nous avons trop sur-estimé $\mathrm{EMC}(\widehat{\mathbb{V}}_{\mathrm{IMP}})$. En fait, entre notre estimateur et $\mathbb{V}^*_{\mathrm{IMP}}$, nous trouvons une différence de 15 136,07 - 11 288,19 = 3847,88 pour le taux 80 % de réponse et une différence de $6052,74 - 4205,94 = 1846,80$ pour le taux 90 % de réponse. Nous remarquons également qu'en valeur absolue, $\mathrm{Birel}(\widehat{\mathbb{V}}_\bullet)$ est inférieur à $\mathrm{Birel}(\widehat{\mathbb{V}}_{\bullet\mathrm{INT}})$ pour tous les taux de réponse. Encore une fois, $\mathbb{V}^*_{\mathrm{MIX}}$ et $\mathrm{EMC}(\widehat{\mathbb{V}}_{\mathrm{MIX}})$ se sont avérés négligeables.

## 6.6. Résumé des résultats pour un mécanisme de réponse non uniforme

Lorsque nous parcourons les tableaux 6.8, 6.9 et 6.10, nous nous apercevons que les résultats pour un mécanisme de réponse non uniforme se sont avérés davantage inférieurs par rapport aux résultats de simulation antérieurs utilisant le mécanisme uniforme.

Nous remarquons que $\mathrm{VMC}(\hat{\bar{y}}_{\mathrm{GREG}})$ est nettement éloigné par rapport à $\mathrm{EMC}(\widehat{\mathbb{V}}_{\bullet\mathrm{ECH}})$ - $\mathrm{EMC}(\widehat{diff})$, plus particulièrement pour le cas 60 % de réponse de l'estimateur HT. Nous constatons une différence de 26 287,76 - 36 116,93 = -9829,17 (soit -27,21 %) entre les estimateurs qui

définissent la variance échantillonnale de l'expérience, alors que sous le mécanisme de réponse uniforme, nous n'avons trouvé qu'une différence minime de 36 775,76 - 36 116,93 = 658,83 pour un pourcentage de -1,82%.

La même chose se produit au sujet de la variance due à l'imputation. Elle prend énormément d'amplitude pour le taux 60 % de réponse. En comparaison avec le mécanisme uniforme, EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) passe de 14 341,00 à 50 007,41 pour l'estimateur HT, de 15 082,30 à 49 838,44 pour l'estimateur par le ratio et de 14 740,75 à 50 265,96 pour l'estimateur par la régression. Nous avons également remarqué que sous ce mécanisme de réponse, notre estimateur développé EMC($\widehat{\mathbb{V}}_{\text{IMP}}$) traduit moins bien la variance due à l'imputation.

De plus, nous ne sommes pas surpris de constater que EMC($\widehat{diff}$) fonctionne moins bien dans le mécanisme non uniforme puisque nous avons pris la décision d'éliminer trois des cinq termes de (3.4.7) en se basant uniquement sur le cas de base, soit le mécanisme de réponse uniforme (voir les tableaux 3.1, 3.2 et 3.3). Cependant, il est encore vrai d'affirmer que l'estimateur de variance ajustée $\widehat{\mathbb{V}}_{\bullet\text{ECH}} - \widehat{diff} + \widehat{\mathbb{V}}_{\text{IMP}}$ fonctionne mieux en terme de biais relatif que l'estimateur intuitif $\widehat{\mathbb{V}}_{\bullet\text{ECH}} + \widehat{\mathbb{V}}_{\text{IMP}}$.

Les différences sur le plan de la variance totale entre ces deux mécanismes sont la conséquence de la variance due à l'imputation. Malgré que le terme d'ajustement $diff$ soit plus élevé sous un mécanisme non uniforme, la différence entre la variance de Monte Carlo des données complétées VMC($\hat{\bar{y}}_{\text{HT}\bullet}$) et la définition de la variance totale donnée par la formule (5.3.2) reste quand même énorme.

Cependant, le taux de recouvrement de chacun des estimateurs testés sous le taux de réponse de 60 % est excessivement bas. Ils sont dans l'ordre de 83,0 % pour l'estimateur HT, 87,0 % pour l'estimateur par le ratio et 83,2 % pour l'estimateur par la régression, alors qu'ils devraient globalement se situer aux alentours de 95 %, comme dans le cas du mécanisme de réponse uniforme.

D'autre part, pour l'ensemble des résultats obtenus, $\text{EMC}(\hat{\bar{y}}_{\bullet\text{GREG}})$ se situe assez près de $\bar{y}_U$. L'erreur relative de chacun des estimateurs et pour chacun des taux de réponse se situe en deça de 6,5 %. Finalement, $\mathbb{V}^*_{\text{MIX}}$ et $\text{EMC}(\widehat{\mathbb{V}}_{\text{MIX}})$ se sont avérés négligeables par rapport à $\text{EMC}(\widehat{\mathbb{V}}_{\bullet})$ pour les trois estimateurs et les trois taux de réponse.

## 6.7. Conclusion

Dans le chapitre 3, nous avons développé sous certaines hypothèses et sous un mécanisme de réponse non confondu des estimateurs de variance. Ces estimateurs ne sont pas sans biais à cause de la complexité du problème. Il nous a été nécessaire d'utiliser des approximations. Malgré cela, l'étude expérimentale nous a confirmé que ces estimateurs fonctionnent assez bien, surtout pour un mécanisme uniforme. Ceci est un résultat important. Par contre, pour le mécanisme non uniforme (1.5.2), nous avons remarqué que notre estimateur développé $\text{EMC}(\widehat{\mathbb{V}}_{\text{IMP}})$ traduit moins bien la variance due à l'imputation. Finalement, quant au choix du meilleur estimateur pour ce mécanisme de réponse, il y a deux possibilités : si nous préconisons le biais relatif minimal, c'est l'estimateur Horvitz-Thompson qui serait le meilleur choix. Par contre, si nous prônons la variance minimale, c'est vers l'estimateur par la régression qu'il faudrait se tourner.

Pour récapituler, voici l'estimateur de la variance totale obtenu sous l'hypothèse d'un mécanisme de réponse non confondu et du modèle d'imputation (3.3.8)

$$\widehat{\mathbb{V}}_{\bullet} = \widehat{\mathbb{V}}_{\bullet \mathrm{ECH}} + \widehat{\mathbb{V}}_{\mathrm{IMP}} - \widehat{diff}$$

où

$$\widehat{\mathbb{V}}_{\bullet \mathrm{ECH}} = \frac{1-f}{n}\frac{1}{n-1}\sum\nolimits_s g_k^2 e_{\bullet k}^2$$

$$\widehat{\mathbb{V}}_{\mathrm{IMP}} = \hat{\sigma}^2 \sum\nolimits_r S_l^2 z_l + \hat{\sigma}^2 \sum\nolimits_o W_k^2 z_k$$

$$\widehat{diff} = \frac{1-f}{n}\frac{1}{n-1}\left(\hat{\beta}^2 \sum\nolimits_o g_k^2 d_k^2\right)$$

où $\hat{\beta}$, $\hat{\sigma}^2$, $g_k$ et $e_{\bullet k}$ sont définis respectivement par (3.3.11), (3.3.9), (1.4.7) et (3.4.3). Notons également que $W_k = a_k g_k$ avec $a_k = \frac{1}{\pi_k}$, $d_k = z_{\ell(k)} - z_k$ et $S_l = \sum_{o_l} W_k$ avec $o_l = \{k : k \in o \text{ et } k \text{ utilise } \ell \text{ comme donneur}\}$.

APPENDICE A

# La dérivation de l'expression pour la covariance mixte

Nous avons, sous un mécanisme de réponse non confondu,

$$
\begin{aligned}
\mathbb{E}_\xi(\mathbb{V}_{\text{MIX}}) &= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}\Big\{\mathbb{E}_\xi\Big[\,(\,\hat{\bar{y}}_{\text{GREG}} - \bar{y}_U\,)\,(\,\hat{\bar{y}}_{\bullet\text{GREG}} - \hat{\bar{y}}_{\text{GREG}}\,)\,\Big]\Big\} \\
&= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}\Big\{\mathbb{E}_\xi\Big[\Big(\textstyle\sum_s W_k y_k - \sum_U y_k\Big)\Big(\sum_s W_k y_{\bullet k} - \sum_s W_k y_k\Big)\Big]\Big\} \\
&= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}\Big\{\mathbb{E}_\xi\Big[\Big(\textstyle\sum_s W_k y_k - \sum_U y_k\Big)\Big(\sum_o W_k\,(\,y_{\ell(k)} - y_k\,)\Big)\Big]\Big\} \\
&= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}\Big\{\mathbb{E}_\xi\Big[\Big(\textstyle\sum_s W_k(\beta z_k + \epsilon_k) - \sum_U (\beta z_k + \epsilon_k)\Big) \\
&\qquad \Big(\textstyle\sum_o W_k(\beta z_{\ell(k)} - \beta z_k + \epsilon_{\ell(k)} - \epsilon_k)\Big)\Big]\Big\} \\
&= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}\Big\{\mathbb{E}_\xi\Big[\Big(\textstyle\sum_s W_k(\beta z_k + \epsilon_k) - \sum_U (\beta z_k + \epsilon_k)\Big)\Big(\sum_o W_k(\beta z_{\ell(k)} - \beta z_k)\Big) + \\
&\qquad \mathbb{E}_\xi\Big[\Big(\textstyle\sum_s W_k(\beta z_k + \epsilon_k) - \sum_U (\beta z_k + \epsilon_k)\Big)\Big(\sum_o W_k(\,\epsilon_{\ell(k)} - \epsilon_k\,)\Big)\Big]\Big\} \\
&= \mathbb{E}_{\text{p}}\mathbb{E}_{\text{q}}\Big\{(\,\beta\textstyle\sum_s W_k z_k\,)\,(\,\beta\sum_o W_k d_k\,) - (\,\beta\sum_U z_k\,)\,(\,\beta\sum_o W_k d_k\,) + \\
&\qquad \mathbb{E}_\xi\Big[\,(\,\textstyle\sum_s W_k\epsilon_k\,)\,(\,\sum_o W_k\epsilon_{\ell(k)}\,) - (\,\sum_s W_k\epsilon_k\,)\sum_o W_k\epsilon_k\,) - \\
&\qquad (\,\textstyle\sum_U \epsilon_k\,)\,(\,\sum_o W_k\epsilon_{\ell(k)}\,) + (\,\sum_U \epsilon_k\,)(\,\sum_o W_k\epsilon_k\,)\Big]\Big\}
\end{aligned}
$$

Comme nous l'avons souligné à la fin de la section 3.5, nous pouvons donner l'approximation de l'expression $\sum_o W_k d_k$ par 0. Donc, on obtient

$$\mathbb{E}_\xi\ (\mathbb{V}_{\text{MIX}}) \approx \mathbb{E}_\text{p}\mathbb{E}_\text{q}\Big\{\mathbb{E}_\xi\Big[\,(\textstyle\sum_s W_k\epsilon_k)\,(\sum_o W_k\epsilon_{\ell(k)}) - (\sum_s W_k\epsilon_k)\,(\sum_o W_k\epsilon_k) - (\sum_U \epsilon_k)\,(\sum_o W_k\epsilon_{\ell(k)}) + (\sum_U \epsilon_k)(\sum_o W_k\epsilon_k)\,\Big]\Big\}$$

En posant $S_l = \sum_{o_l} W_k$ avec $o_l = \{k : k \in o$et $k$ utilise $\ell$ comme donneur$\}$, nous obtenons finalement

$$\begin{aligned}\mathbb{E}_\xi\ (\mathbb{V}_{\text{MIX}}) &\approx \mathbb{E}_\text{p}\mathbb{E}_\text{q}\Big\{\mathbb{E}_\xi\Big[\,(\textstyle\sum_r W_l\epsilon_l + \sum_o W_k\epsilon_k\,)\,(\,\sum_r S_l\epsilon_l\,) - (\sum_r W_l\epsilon_l + \\ &\qquad \textstyle\sum_o W_k\epsilon_k)\,(\sum_o W_k\epsilon_k) - (\sum_r \epsilon_l + \sum_{U-r}\epsilon_k)\,(\sum_r S_l\epsilon_l) + \\ &\qquad (\textstyle\sum_o \epsilon_k + \sum_{U-o}\epsilon_k)\,(\sum_o W_k\epsilon_k)\,\Big]\Big\} \\ &= \mathbb{E}_\text{p}\mathbb{E}_\text{q}\Big\{\sigma^2\textstyle\sum_r W_l S_l z_l - \sigma^2\sum_o W_k^2 z_k - \sigma^2\sum_r S_l z_l + \sigma^2\sum_o W_k z_k\Big\}\end{aligned} \tag{A.0.1}$$

APPENDICE B

# Simulations de Monte Carlo pour l'estimateur par le ratio

## B.1. Mécanisme de réponse uniforme

Voici le programme Splus permettant de faire une simulation de Monte Carlo. Nous considérons le plan d'échantillonnage SI et l'estimateur par le ratio. Dans cet exemple, le taux de réponse est de 60 %.

```
options(object.size = 50920016)
nb.echantillon ⇐ 100
nb.monte.carlo ⇐ 1000
N ⇐ 281
alpha ⇐ 0.05
moyenne.pop ⇐ mean(mu284.dat.mod[ ,"Rev84"])
```

**Initialisations**

```
y.var ⇐ NULL
y.mean ⇐ NULL
TRMC.100.reponse ⇐ 0

y.variance.complete60 ⇐ NULL
y.variance.complete80 ⇐ NULL
y.variance.complete90 ⇐ NULL
```

vmix60 $\Leftarrow$ NULL

vmix80 $\Leftarrow$ NULL

vmix90 $\Leftarrow$ NULL

estvmix60 $\Leftarrow$ NULL

estvmix80 $\Leftarrow$ NULL

estvmix90 $\Leftarrow$ NULL

estvimp60 $\Leftarrow$ NULL

estvimp80 $\Leftarrow$ NULL

estvimp90 $\Leftarrow$ NULL

estvech $\Leftarrow$ NULL

y.mean.complete60 $\Leftarrow$ NULL

y.mean.complete80 $\Leftarrow$ NULL

y.mean.complete90 $\Leftarrow$ NULL

variance.impute.complete60 $\Leftarrow$ NULL

variance.impute.complete80 $\Leftarrow$ NULL

variance.impute.complete90 $\Leftarrow$ NULL

TRMC.complete60 $\Leftarrow$ 0

TRMC.complete80 $\Leftarrow$ 0

TRMC.complete90 $\Leftarrow$ 0

prob.inclusion60 $\Leftarrow$ 0.60

prob.inclusion80 $\Leftarrow$ 0.80

prob.inclusion90 $\Leftarrow$ 0.90

f ⇐ nb.echantillon/N

diffnew60 ⇐ NULL

diffnew80 ⇐ NULL

diffnew90 ⇐ NULL

**For(i = 1:nb.monte.carlo,**

**{**

repeat

{

**s** ⇐ mu284.dat.mod[(sample(1:281, nb.echantillon, replace = F)), c("Rmt85", "Rev84", "P75")]

**s** ⇐ cbind(s,1:length(dimnames(s)[[1]]))

**dimnames(s)** ⇐ list(dimnames(s)[[1]], c("Rmt85", "Rev84", "P75", "Numero"))

**s.y** ⇐ s[ ,"Rev84"]

**s.x** ⇐ s[ ,"Rmt85"]

**s.z** ⇐ s[ ,"P75"]

estimateur par le ratio pour 100 % réponse

**R.hat** ⇐ mean(s.y) / mean(s.x)

**y.mean[i]** ⇐ mean(mu284.dat.mod[,"Rmt85"]) * R.hat

**y.var[i]** ⇐ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[,"Rmt85"])$^2$ / mean(s.x)$^2$ * 1/(nb.echantillon - 1) * sum((s.y - R.hat * s.x)$^2$)

if (moyenne.pop ≥ y.mean[i] - qnorm(1-alpha/2) * sqrt(y.var[i]) & moyenne.pop ≤ y.mean[i] + qnorm(1-alpha/2) * sqrt(y.var[i]))

{**TRMC.100.reponse** $\Leftarrow$ TRMC.100.reponse + 1/nb.monte.carlo}

estimateur par le ratio pour la non-réponse

**unif** $\Leftarrow$ runif(dimnames(s)[[1]], 0, 1)
**r60** $\Leftarrow$ s[unif $\leq$ prob.inclusion60, 1:4]
**r80** $\Leftarrow$ s[unif $\leq$ prob.inclusion80, 1:4]
**r90** $\Leftarrow$ s[unif $\leq$ prob.inclusion90, 1:4]

**long.r60** $\Leftarrow$ length(dimnames(r60)[[1]])
**long.r80** $\Leftarrow$ length(dimnames(r80)[[1]])
**long.r90** $\Leftarrow$ length(dimnames(r90)[[1]])

**r60** $\Leftarrow$ cbind(r60,1:long.r60)
**r80** $\Leftarrow$ cbind(r80,1:long.r80)
**r90** $\Leftarrow$ cbind(r90,1:long.r90)

**dimnames(r60)** $\Leftarrow$ list(dimnames(r60)[[1]], c("Rmt85", "Rev84", "P75", "Numero.de.s", "NO.de.r60"))
**dimnames(r80)** $\Leftarrow$ list(dimnames(r80)[[1]], c("Rmt85", "Rev84", "P75", "Numero.de.s", "NO.de.r80"))
**dimnames(r90)** $\Leftarrow$ list(dimnames(r90)[[1]], c("Rmt85", "Rev84", "P75", "Numero.de.s", "NO.de.r90"))

**r60.x** $\Leftarrow$ r60[, "Rmt85"]
**r80.x** $\Leftarrow$ r80[, "Rmt85"]
**r90.x** $\Leftarrow$ r90[, "Rmt85"]
**r60.y** $\Leftarrow$ r60[, "Rev84"]
**r80.y** $\Leftarrow$ r80[, "Rev84"]

**r90.y** $\Leftarrow$ r90[, "Rev84"]

**r60.z** $\Leftarrow$ r60[, "P75"]

**r80.z** $\Leftarrow$ r80[, "P75"]

**r90.z** $\Leftarrow$ r90[, "P75"]

**numero60** $\Leftarrow$ s[dimnames(r60)[[1]] ,4]

**numero80** $\Leftarrow$ s[dimnames(r80)[[1]] ,4]

**numero90** $\Leftarrow$ s[dimnames(r90)[[1]] ,4]

**o60** $\Leftarrow$ s[-numero60,1:4]

**o80** $\Leftarrow$ s[-numero80,1:4]

**o90** $\Leftarrow$ s[-numero90,1:4]

**o60.y** $\Leftarrow$ o60[,"Rev84"]

**o80.y** $\Leftarrow$ o80[,"Rev84"]

**o90.y** $\Leftarrow$ o90[,"Rev84"]

**o60.z** $\Leftarrow$ o60[,"P75"]

**o80.z** $\Leftarrow$ o80[,"P75"]

**o90.z** $\Leftarrow$ o90[,"P75"]

**o60.x** $\Leftarrow$ o60[,"Rmt85"]

**o80.x** $\Leftarrow$ o80[,"Rmt85"]

**o90.x** $\Leftarrow$ o90[,"Rmt85"]

if (length(dimnames(o60)[[1]]) $>$ 1) break

if (length(dimnames(o80)[[1]]) $>$ 1) break

if (length(dimnames(o90)[[1]]) $>$ 1) break

}

**long.o60** ⇐ length(dimnames(o60)[[1]])
**long.o80** ⇐ length(dimnames(o80)[[1]])
**long.o90** ⇐ length(dimnames(o90)[[1]])

**r60.matrix** ⇐ matrix( r60[, "P75"], byrow=T, ncol=long.r60, nrow=long.o60)
**r80.matrix** ⇐ matrix( r80[, "P75"], byrow=T, ncol=long.r80, nrow=long.o80)
**r90.matrix** ⇐ matrix( r90[, "P75"], byrow=T, ncol=long.r90, nrow=long.o90)

**o60.matrix** ⇐ matrix( o60[, "P75"], byrow=F, ncol=long.r60, nrow=long.o60)
**o80.matrix** ⇐ matrix( o80[, "P75"], byrow=F, ncol=long.r80, nrow=long.o80)
**o90.matrix** ⇐ matrix( o90[, "P75"], byrow=F, ncol=long.r90, nrow=long.o90)

**distance.minimiser60** ⇐ abs ( o60.matrix - r60.matrix)
**distance.minimiser80** ⇐ abs ( o80.matrix - r80.matrix)
**distance.minimiser90** ⇐ abs ( o90.matrix - r90.matrix)

**dimnames(distance.minimiser60)** ⇐ list( paste("o60",1:long.o60),
paste("r60", 1:long.r60))
**dimnames(distance.minimiser80)** ⇐ list( paste("o80",1:long.o80),
paste("r80", 1:long.r80))
**dimnames(distance.minimiser90)** ⇐ list( paste("o90",1:long.o90),
paste("r90", 1:long.r90))
**distance.minimiser60** ⇐ rbind(distance.minimiser60, 1:long.r60)
**distance.minimiser80** ⇐ rbind(distance.minimiser80, 1:long.r80)
**distance.minimiser90** ⇐ rbind(distance.minimiser90, 1:long.r90)

**minimum60** ⇐ t(apply(distance.minimiser60, 1, min))
**minimum80** ⇐ t(apply(distance.minimiser80, 1, min))
**minimum90** ⇐ t(apply(distance.minimiser90, 1, min))

**l.k60** ⇐ NULL

**l.k80** ⇐ NULL

**l.k90** ⇐ NULL

**z.l60** ⇐ NULL

**z.l80** ⇐ NULL

**z.l90** ⇐ NULL

**o.impute60** ⇐ o60

**o.impute80** ⇐ o80

**o.impute90** ⇐ o90

**Debut de la boucle FOR**

date()

**for(j in 1:long.o60)**

{ **tempo60** ⇐ r60 [distance.minimiser60[long.o60 + 1, distance.minimiser60[j,]==
minimum60[j]] , c("Rev84", "NO.de.r60", "P75")]

if( length( tempo60[, "Rev84"] )> 1 ) **tempo260** ⇐ tempo60[sample(1:dim(tempo60)[1],
size=1), ] } else {**tempo260** ⇐ tempo60 }

**o.impute60[j,2]** ⇐ tempo260[,"Rev84"]

**l.k60[j]** ⇐ tempo60[as.numeric(dimnames(tempo60)[[1]]) ==
as.numeric(dimnames(tempo260)[[1]]), "NO.de.r60"]

**z.l60[j]** ⇐ tempo60[as.numeric(dimnames(tempo60)[[1]]) ==
as.numeric(dimnames(tempo260)[[1]]), "P75"]

}

**for(j in 1:long.o80)**

{ **tempo80** ⇐ r80 [distance.minimiser80[long.o80 + 1, distance.minimiser80[j,]==

```
minimum80[j]] , c("Rev84", "NO.de.r80", "P75")]
if( length( tempo80[, "Rev84"] ) > 1 ) tempo280 <= tempo80[sample(1:dim(tempo80)[1],
size=1), ] } else {tempo280 <= tempo80 }
o.impute80[j,2] <= tempo280[,"Rev84"]
l.k80[j] <= tempo80[as.numeric(dimnames(tempo80)[[1]]) ==
as.numeric(dimnames(tempo280)[[1]]), "NO.de.r80"]
z.l80[j] <= tempo80[as.numeric(dimnames(tempo80)[[1]]) ==
as.numeric(dimnames(tempo280)[[1]]), "P75"] }

for(j in 1:long.o90)
{ tempo90 <= r90 [distance.minimiser90[long.o90 + 1, distance.minimiser90[j,]==
minimum90[j]] , c("Rev84", "NO.de.r90", "P75")]
if( length( tempo90[, "Rev84"] ) > 1 ) tempo290 <= tempo90[sample(1:dim(tempo90)[1],
size=1), ]  else {tempo290 <= tempo90
o.impute90[j,2] <= tempo290[,"Rev84"]
l.k90[j] <= tempo90[as.numeric(dimnames(tempo90)[[1]]) ==
as.numeric(dimnames(tempo290)[[1]]), "NO.de.r90"]
z.l90[j] <= tempo90[as.numeric(dimnames(tempo90)[[1]]) ==
as.numeric(dimnames(tempo290)[[1]]), "P75"]

o.impute60 <= cbind(o.impute60,l.k60)
o.impute80 <= cbind(o.impute80,l.k80)
o.impute90 <= cbind(o.impute90,l.k90)

names(o.impute60) <= c("Rmt85", "Rev84", "P75", "No.de.o60","l.k60")
names(o.impute80) <= c("Rmt85", "Rev84", "P75", "No.de.o80","l.k80")
names(o.impute90) <= c("Rmt85", "Rev84", "P75", "No.de.o90","l.k90")

l.k.sort60 <= sort(l.k60)
```

**l.k.sort80** $\Leftarrow$ sort(l.k80)
**l.k.sort90** $\Leftarrow$ sort(l.k90)

**names(l.k.sort60)** $\Leftarrow$ c(as.character(l.k.sort60))
**names(l.k.sort80)** $\Leftarrow$ c(as.character(l.k.sort80))
**names(l.k.sort90)** $\Leftarrow$ c(as.character(l.k.sort90))

**l.k.unique60** $\Leftarrow$ unique(l.k60)
**l.k.unique80** $\Leftarrow$ unique(l.k80)
**l.k.unique90** $\Leftarrow$ unique(l.k90)

**l.k.unique.sort60** $\Leftarrow$ sort(l.k.unique60)
**l.k.unique.sort80** $\Leftarrow$ sort(l.k.unique80)
**l.k.unique.sort90** $\Leftarrow$ sort(l.k.unique90)

**names(l.k.unique.sort60)** $\Leftarrow$ c(as.character(l.k.unique.sort60))
**names(l.k.unique.sort80)** $\Leftarrow$ c(as.character(l.k.unique.sort80))
**names(l.k.unique.sort90)** $\Leftarrow$ c(as.character(l.k.unique.sort90))

**non.donneur60** $\Leftarrow$ r60[ -l.k.unique60, "NO.de.r60"]
**non.donneur80** $\Leftarrow$ r80[ -l.k.unique80, "NO.de.r80"]
**non.donneur90** $\Leftarrow$ r90[ -l.k.unique90, "NO.de.r90"]

**names(non.donneur60)** $\Leftarrow$ c(as.character(non.donneur60))
**names(non.donneur80)** $\Leftarrow$ c(as.character(non.donneur80))
**names(non.donneur90)** $\Leftarrow$ c(as.character(non.donneur90))

**ordre60** $\Leftarrow$ sort( c(l.k.sort60, non.donneur60))
**ordre80** $\Leftarrow$ sort( c(l.k.sort80, non.donneur80))

**ordre90** $\Leftarrow$ sort( c(l.k.sort90, non.donneur90))

**F.l60** $\Leftarrow$ summary(as.factor(ordre60))
**F.l80** $\Leftarrow$ summary(as.factor(ordre80))
**F.l90** $\Leftarrow$ summary(as.factor(ordre90))

**F.l60** [ is.na(match(names(F.l60),names(l.k.unique.sort60)))] $\Leftarrow$ 0
**F.l80** [ is.na(match(names(F.l80),names(l.k.unique.sort80)))] $\Leftarrow$ 0
**F.l90** [ is.na(match(names(F.l90),names(l.k.unique.sort90)))] $\Leftarrow$ 0

**o.x.impute60** $\Leftarrow$ o.impute60[, "Rmt85"]
**o.x.impute80** $\Leftarrow$ o.impute80[, "Rmt85"]
**o.x.impute90** $\Leftarrow$ o.impute90[, "Rmt85"]

**o.y.impute60** $\Leftarrow$ o.impute60[, "Rev84"]
**o.y.impute80** $\Leftarrow$ o.impute80[, "Rev84"]
**o.y.impute90** $\Leftarrow$ o.impute90[, "Rev84"]

**o.z.impute60** $\Leftarrow$ o.impute60[, "P75"]
**o.z.impute80** $\Leftarrow$ o.impute80[, "P75"]
**o.z.impute90** $\Leftarrow$ o.impute90[, "P75"]

**s.impute60** $\Leftarrow$ rbind( r60[,-c(4:5)],o.impute60[,-c(4:5)])
**s.impute80** $\Leftarrow$ rbind( r80[,-c(4:5)],o.impute80[,-c(4:5)])
**s.impute90** $\Leftarrow$ rbind( r90[,-c(4:5)],o.impute90[,-c(4:5)])

**s.impute60** $\Leftarrow$ s.impute60[dimnames(s)[[1]],]
**s.impute80** $\Leftarrow$ s.impute80[dimnames(s)[[1]],]
**s.impute90** $\Leftarrow$ s.impute90[dimnames(s)[[1]],]

**s.y.impute60** $\Leftarrow$ s.impute60[, "Rev84"]
**s.y.impute80** $\Leftarrow$ s.impute80[, "Rev84"]
**s.y.impute90** $\Leftarrow$ s.impute90[, "Rev84"]

**s.x.impute60** $\Leftarrow$ s.impute60[, "Rmt85"]
**s.x.impute80** $\Leftarrow$ s.impute80[, "Rmt85"]
**s.x.impute90** $\Leftarrow$ s.impute90[, "Rmt85"]

**s.z.impute60** $\Leftarrow$ s.impute60[, "P75"]
**s.z.impute80** $\Leftarrow$ s.impute80[, "P75"]
**s.z.impute90** $\Leftarrow$ s.impute90[, "P75"]

**R.hat.impute60** $\Leftarrow$ ( sum(r60.y) + sum(o.y.impute60) ) / sum(s.x.impute60)
**R.hat.impute80** $\Leftarrow$ ( sum(r80.y) + sum(o.y.impute80) ) / sum(s.x.impute80)
**R.hat.impute90** $\Leftarrow$ ( sum(r90.y) + sum(o.y.impute90) ) / sum(s.x.impute90)

**y.mean.complete60[i]** $\Leftarrow$ mean(mu284.dat.mod[,"Rmt85"]) * R.hat.impute60
**y.mean.complete80[i]** $\Leftarrow$ mean(mu284.dat.mod[,"Rmt85"]) * R.hat.impute80
**y.mean.complete90[i]** $\Leftarrow$ mean(mu284.dat.mod[,"Rmt85"]) * R.hat.impute90

**estvech[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop)$^2$

**estvimp60[i]** $\Leftarrow$ (y.mean.complete60[i]-y.mean[i])$^2$
**estvimp80[i]** $\Leftarrow$ (y.mean.complete80[i]-y.mean[i])$^2$
**estvimp90[i]** $\Leftarrow$ (y.mean.complete90[i]-y.mean[i])$^2$

**estvmix60[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete60[i]-y.mean[i])
**estvmix80[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete80[i]-y.mean[i])

**estvmix90[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete90[i]-y.mean[i])

**y.variance.complete60[i]** $\Leftarrow$ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[, "Rmt85"])$^2$ / mean(s.x.impute60)$^2$ * 1/(nb.echantillon - 1) * sum( (s.y.impute60 - R.hat.impute60 * s.x.impute60)$^2$ )
**y.variance.complete80[i]** $\Leftarrow$ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[, "Rmt85"])$^2$ / mean(s.x.impute80)$^2$ * 1/(nb.echantillon - 1) * sum( (s.y.impute80 - R.hat.impute80 * s.x.impute80)$^2$ )
**y.variance.complete90[i]** $\Leftarrow$ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[ , "Rmt85"])$^2$ / mean(s.x.impute90)$^2$ * 1/(nb.echantillon - 1) * sum( (s.y.impute90 - R.hat.impute90 * s.x.impute90)$^2$ )

**beta60** $\Leftarrow$ sum (r60.y) / sum (r60.z)
**beta80** $\Leftarrow$ sum (r80.y) / sum (r80.z)
**beta90** $\Leftarrow$ sum (r90.y) / sum (r90.z)

**CV60** $\Leftarrow$ sqrt ( sum( (r60.z - mean(r60.z))$^2$) / (long.r60 - 1) ) / mean(r60.z)
**CV80** $\Leftarrow$ sqrt ( sum( (r80.z - mean(r80.z))$^2$) / (long.r80 - 1) ) / mean(r80.z)
**CV90** $\Leftarrow$ sqrt ( sum( (r90.z - mean(r90.z))$^2$) / (long.r90 - 1) ) / mean(r90.z)

**hatsigma60** $\Leftarrow$ (1/( 1-CV60$^2$/(long.r60))) * (sum ((r60.y - beta60 * r60.z)$^2$)/ sum(r60.z) )
**hatsigma80** $\Leftarrow$ (1/( 1-CV80$^2$/(long.r80))) * (sum((r80.y - beta80 * r80.z)$^2$)/ sum(r80.z) )
**hatsigma90** $\Leftarrow$ (1/( 1-CV90$^2$/(long.r90))) * (sum((r90.y - beta90 * r90.z)$^2$)/ sum(r90.z) )

**vmix60[i]** $\Leftarrow$ 2*hatsigma60 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute60))$^2$ * sum((r60[o.impute60[,"l.k60"],"P75"] - o.impute60[,"P75"])) - 2*hatsigma60 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute60) ) * sum((r60[o.impute60[,"l.k60"],"P75"] - o.impute60[,"P75"]))

**vmix80[i]** $\Leftarrow$ 2*hatsigma80 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute80))$^2$ * sum((r80[o.impute80[,"l.k80"],"P75"] - o.impute80[,"P75"])) - 2*hatsigma80 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute80) ) * sum((r80[o.impute80[,"l.k80"],"P75"] - o.impute80[,"P75"]))

**vmix90[i]** $\Leftarrow$ 2*hatsigma90 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute90))$^2$ * sum((r90[o.impute90[,"l.k90"],"P75"] - o.impute90[,"P75"])) - 2*hatsigma90 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute90) ) * sum((r90[o.impute90[,"l.k90"],"P75"] - o.impute90[,"P75"]))

**variance.impute.complete60[i]** $\Leftarrow$ hatsigma60 * (1 /nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute60))$^2$ * ( 2 * sum(o.z.impute60) + sum( F.l60 * (F.l60 - 1) * r60.z ) )

**variance.impute.complete80[i]** $\Leftarrow$ hatsigma80 * (1 /nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute80))$^2$ * ( 2 * sum(o.z.impute80) + sum( F.l80 * (F.l80- 1) * r80.z ) )

**variance.impute.complete90[i]** $\Leftarrow$ hatsigma90 * (1 /nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute90))$^2$ * ( 2 * sum(o.z.impute90) + sum( F.l90 * (F.l90 - 1) * r90.z ) )

if (moyenne.pop $\geq$ y.mean.complete60[i] - qnorm(1-alpha/2)*sqrt(y.variance.complete60[i] + variance.impute.complete60[i] ) & moyenne.pop $\leq$ y.mean.complete60[i]

+ qnorm(1-alpha/2)*sqrt(y.variance.complete60[i] + variance.impute.complete60[i]))
{**TRMC.complete60** $\Leftarrow$ TRMC.complete60 + 1/nb.monte.carlo}
if (moyenne.pop $\geq$ y.mean.complete80[i] - qnorm(1-alpha/2)*sqrt(y.variance.complete80[i]
+ variance.impute.complete80[i] ) & moyenne.pop $\leq$ y.mean.complete80[i]
+ qnorm(1-alpha/2)*sqrt(y.variance.complete80[i] + variance.impute.complete80[i]))
**TRMC.complete80** $\Leftarrow$ TRMC.complete80 + 1/nb.monte.carlo
if (moyenne.pop $\geq$ y.mean.complete90[i] - qnorm(1-alpha/2)*sqrt(y.variance.complete90[i]
+ variance.impute.complete90[i] ) & moyenne.pop $\leq$ y.mean.complete90[i]
+ qnorm(1-alpha/2)*sqrt(y.variance.complete90[i] + variance.impute.complete90[i]))
**TRMC.complete90** $\Leftarrow$ TRMC.complete90 + 1/nb.monte.carlo

**diffnew60[i]** $\Leftarrow$ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * ( mean(mu284.dat.mod[,
"Rmt85"]) / mean(s.impute60[, "Rmt85"]))$^2$ * (beta60$^2$ *
sum((r60[o.impute60[,"l.k60"],"P75"] - o.impute60[,"P75"])$^2$))
**diffnew80[i]** $\Leftarrow$ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * ( mean(mu284.dat.mod[,
"Rmt85"]) / mean(s.impute80[, "Rmt85"]))$^2$ * (beta80$^2$ *
sum((r80[o.impute80[,"l.k80"],"P75"] - o.impute80[,"P75"])$^2$))
**diffnew90[i]** $\Leftarrow$ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * ( mean(mu284.dat.mod[,
"Rmt85"]) / mean(s.impute90[, "Rmt85"]))$^2$ * (beta90$^2$ *
sum((r90[o.impute90[,"l.k90"],"P75"] - o.impute90[,"P75"])$^2$))

**if (i%%100 = 0) {cat(i, ") ", date(), "\n")**
**}, grain.size = 15 }**
date()
**EMC.vmix60** $\Leftarrow$ mean(vmix60)
**EMC.vmix80** $\Leftarrow$ mean(vmix80)
**EMC.vmix90** $\Leftarrow$ mean(vmix90)
**EMC.estvmix60** $\Leftarrow$ mean(estvmix60)
**EMC.estvmix80** $\Leftarrow$ mean(estvmix80)

**EMC.estvmix90** $\Leftarrow$ mean(estvmix90)

**EMC.moyenne** $\Leftarrow$ mean(y.mean)
**VMC.moyenne.100.reponse** $\Leftarrow$ var(y.mean)
**EMCV.variance.echantillonnale.100.reponse** $\Leftarrow$ mean(y.var)
**EMC.Vech** $\Leftarrow$ mean(estvech)

**EMC.moyenne.impute60** $\Leftarrow$ mean(y.mean.complete60)
**EMC.moyenne.impute80** $\Leftarrow$ mean(y.mean.complete80)
**EMC.moyenne.impute90** $\Leftarrow$ mean(y.mean.complete90)

**VMC.moyenne.complete60** $\Leftarrow$ var(y.mean.complete60)
**VMC.moyenne.complete80** $\Leftarrow$ var(y.mean.complete80)
**VMC.moyenne.complete90** $\Leftarrow$ var(y.mean.complete90)

**EMCV.variance.echantillonnale.complete60** $\Leftarrow$ mean(y.variance.complete60)
**EMCV.variance.echantillonnale.complete80** $\Leftarrow$ mean(y.variance.complete80)
**EMCV.variance.echantillonnale.complete90** $\Leftarrow$ mean(y.variance.complete90)

**EMCV.variance.impute.complete60** $\Leftarrow$ mean(variance.impute.complete60)
**EMCV.variance.impute.complete80** $\Leftarrow$ mean(variance.impute.complete80)
**EMCV.variance.impute.complete90** $\Leftarrow$ mean(variance.impute.complete90)

**EMC.estvimp60** $\Leftarrow$ mean(estvimp60)
**EMC.estvimp80** $\Leftarrow$ mean(estvimp80)
**EMC.estvimp90** $\Leftarrow$ mean(estvimp90)

**EMCV.diffnew60** $\Leftarrow$ mean(diffnew60)
**EMCV.diffnew80** $\Leftarrow$ mean(diffnew80)

**EMCV.diffnew90** $\Leftarrow$ mean(diffnew90)

**v.chapeau.completenew60** $\Leftarrow$ EMCV.variance.impute.complete60 + EMCV.variance.echantillonnale.complete60 - EMCV.diffnew60
**v.chapeau.completenew80** $\Leftarrow$ EMCV.variance.impute.complete80 + EMCV.variance.echantillonnale.complete80 - EMCV.diffnew80
**v.chapeau.completenew90** $\Leftarrow$ EMCV.variance.impute.complete90 + EMCV.variance.echantillonnale.complete90 - EMCV.diffnew90

**differencenew60** $\Leftarrow$ v.chapeau.completenew60 - VMC.moyenne.complete60
**differencenew80** $\Leftarrow$ v.chapeau.completenew80 - VMC.moyenne.complete80
**differencenew90** $\Leftarrow$ v.chapeau.completenew90 - VMC.moyenne.complete90

**Birel60** $\Leftarrow$ ( differencenew60 / VMC.moyenne.complete60 )
**Birel80** $\Leftarrow$ ( differencenew80 / VMC.moyenne.complete80 )
**Birel90** $\Leftarrow$ ( differencenew90 / VMC.moyenne.complete90 )

**R.hat.u** $\Leftarrow$ mean( mu284.dat.mod[, "Rev84"] ) / mean( mu284.dat.mod[, "Rmt85"] )
**estimateur.variance.population** $\Leftarrow$ (1/nb.echantillon - 1/N) * 1 / (N-1) * sum( (mu284.dat.mod[, "Rev84"] - R.hat.u * mu284.dat.mod[, "Rmt85"])$^2$ )

**Faire afficher les résultats**
moyenne.pop
EMC.moyenne
EMC.moyenne.impute60
EMC.moyenne.impute80
EMC.moyenne.impute90

TRMC.100.reponse
TRMC.complete60
TRMC.complete80
TRMC.complete90

estimateur.variance.population
VMC.moyenne.100.reponse
EMCV.variance.echantillonnale.100.reponse
EMC.Vech

estimateur.variance.population
VMC.moyenne.complete60
VMC.moyenne.complete80
VMC.moyenne.complete90
EMCV.variance.echantillonnale.complete60
EMCV.variance.echantillonnale.complete80
EMCV.variance.echantillonnale.complete90

EMCV.variance.impute.complete60
EMCV.variance.impute.complete80
EMCV.variance.impute.complete90

EMC.estvimp60
EMC.estvimp80
EMC.estvimp90

EMC.vmix60
EMC.vmix80

EMC.vmix90

EMC.estvmix60
EMC.estvmix80
EMC.estvmix90

EMCV.diffnew60
EMCV.diffnew80
EMCV.diffnew90

v.chapeau.completenew60
v.chapeau.completenew80
v.chapeau.completenew90

differencenew60
differencenew80
differencenew90

## B.2. Mécanisme de réponse non uniforme

Voici le programme Splus qui nous a permis de faire une simulation de Monte Carlo. Nous considérons toujours le plan d'échantillonnage SI et l'estimateur par le ratio. Dans cet exemple, le taux de réponse est de 60 %.

```
options(object.size = 50920016)
nb.echantillon ⇐ 100
nb.monte.carlo ⇐ 1000
N ⇐ 281
alpha ⇐ 0.05
moyenne.pop ⇐ mean(mu284.dat.mod[ ,"Rev84"])
```

Initialisations

```
y.var ⇐ NULL
y.mean ⇐ NULL
TRMC.100.reponse ⇐ 0

y.variance.complete60 ⇐ NULL
y.variance.complete80 ⇐ NULL
y.variance.complete90 ⇐ NULL

y.mean.complete60 ⇐ NULL
y.mean.complete80 ⇐ NULL
y.mean.complete90 ⇐ NULL

variance.impute.complete60 ⇐ NULL
variance.impute.complete80 ⇐ NULL
```

variance.impute.complete90 ⇐ NULL

TRMC.complete60 ⇐ 0
TRMC.complete80 ⇐ 0
TRMC.complete90 ⇐ 0

vmix60 ⇐ NULL
vmix80 ⇐ NULL
vmix90 ⇐ NULL
estvmix60 ⇐ NULL
estvmix80 ⇐ NULL
estvmix90 ⇐ NULL

estvimp60 ⇐ NULL
estvimp80 ⇐ NULL
estvimp90 ⇐ NULL

estvech ⇐ NULL

prob.inclusion60 ⇐ 0.60
prob.inclusion80 ⇐ 0.80
prob.inclusion90 ⇐ 0.90

f ⇐ nb.echantillon/N
diffnew60 ⇐ NULL
diffnew80 ⇐ NULL
diffnew90 ⇐ NULL

**For(i = 1:nb.monte.carlo,**

{

repeat

{

**s** $\Leftarrow$ mu284.dat.mod[(sample(1:281, nb.echantillon, replace = F)), c("Rmt85", "Rev84", "P75")]

**s** $\Leftarrow$ cbind(s,1:length(dimnames(s)[[1]]))

**dimnames(s)** $\Leftarrow$ list(dimnames(s)[[1]], c("Rmt85", "Rev84", "P75", "Numero"))

**s.y** $\Leftarrow$ s[ ,"Rev84"]

**s.x** $\Leftarrow$ s[ ,"Rmt85"]

**s.z** $\Leftarrow$ s[ ,"P75"]

estimateur par le ratio pour 100 % réponse

**R.hat** $\Leftarrow$ mean(s.y) / mean(s.x)

**y.mean[i]** $\Leftarrow$ mean(mu284.dat.mod[,"Rmt85"]) * R.hat

**y.var[i]** $\Leftarrow$ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[,"Rmt85"])$^2$ / mean(s.x)$^2$ * 1/(nb.echantillon - 1) * sum((s.y - R.hat * s.x)$^2$)

if (moyenne.pop $\geq$ y.mean[i] - qnorm(1-alpha/2) * sqrt(y.var[i]) &
moyenne.pop $\leq$ y.mean[i] + qnorm(1-alpha/2) * sqrt(y.var[i]))

{**TRMC.100.reponse** $\Leftarrow$ TRMC.100.reponse + 1/nb.monte.carlo }

estimateur par le ratio pour la non-réponse

**theta60** $\Leftarrow$ exp ( -0.02618419 * s.z)

**theta80** $\Leftarrow$ exp ( -0.01016476 * s.z)

**theta90** $\Leftarrow$ exp ( -0.004555223 * s.z)

**unif** $\Leftarrow$ runif(dimnames(s)[[1]], 0, 1)

**r60** $\Leftarrow$ s[unif $\leq$ theta60, 1:4]

**r80** $\Leftarrow$ s[unif $\leq$ theta80, 1:4]

**r90** $\Leftarrow$ s[unif $\leq$ theta90, 1:4]

**long.r60** $\Leftarrow$ length(dimnames(r60)[[1]])

**long.r80** $\Leftarrow$ length(dimnames(r80)[[1]])

**long.r90** $\Leftarrow$ length(dimnames(r90)[[1]])

**r60** $\Leftarrow$ cbind(r60,1:long.r60)

**r80** $\Leftarrow$ cbind(r80,1:long.r80)

**r90** $\Leftarrow$ cbind(r90,1:long.r90)

**dimnames(r60)** $\Leftarrow$ list(dimnames(r60)[[1]], c("Rmt85", "Rev84", "P75", "Numero.de.s", "NO.de.r60"))

**dimnames(r80)** $\Leftarrow$ list(dimnames(r80)[[1]], c("Rmt85", "Rev84", "P75", "Numero.de.s", "NO.de.r80"))

**dimnames(r90)** $\Leftarrow$ list(dimnames(r90)[[1]], c("Rmt85", "Rev84", "P75", "Numero.de.s", "NO.de.r90"))

**r60.x** $\Leftarrow$ r60[, "Rmt85"]

**r80.x** $\Leftarrow$ r80[, "Rmt85"]

**r90.x** $\Leftarrow$ r90[, "Rmt85"]

**r60.y** $\Leftarrow$ r60[, "Rev84"]

**r80.y** $\Leftarrow$ r80[, "Rev84"]

**r90.y** $\Leftarrow$ r90[, "Rev84"]

**r60.z** $\Leftarrow$ r60[, "P75"]

**r80.z** $\Leftarrow$ r80[, "P75"]

**r90.z** $\Leftarrow$ r90[, "P75"]

```
numero60 <= s[dimnames(r60)[[1]] ,4]
numero80 <= s[dimnames(r80)[[1]] ,4]
numero90 <= s[dimnames(r90)[[1]] ,4]

o60 <= s[-numero60,1:4]
o80 <= s[-numero80,1:4]
o90 <= s[-numero90,1:4]

o60.y <= o60[,"Rev84"]
o80.y <= o80[,"Rev84"]
o90.y <= o90[,"Rev84"]

o60.z <= o60[,"P75"]
o80.z <= o80[,"P75"]
o90.z <= o90[,"P75"]

o60.x <= o60[,"Rmt85"]
o80.x <= o80[,"Rmt85"]
o90.x <= o90[,"Rmt85"]

if (length(dimnames(o60)[[1]]) > 1) break
if (length(dimnames(o80)[[1]]) > 1) break
if (length(dimnames(o90)[[1]]) > 1) break
}
long.o60 <= length(dimnames(o60)[[1]])
long.o80 <= length(dimnames(o80)[[1]])
long.o90 <= length(dimnames(o90)[[1]])

r60.matrix <= matrix( r60[, "P75"], byrow=T, ncol=long.r60, nrow=long.o60)
```

**r80.matrix** ⇐ matrix( r80[, "P75"], byrow=T, ncol=long.r80, nrow=long.o80)
**r90.matrix** ⇐ matrix( r90[, "P75"], byrow=T, ncol=long.r90, nrow=long.o90)

**o60.matrix** ⇐ matrix( o60[, "P75"], byrow=F, ncol=long.r60, nrow=long.o60)
**o80.matrix** ⇐ matrix( o80[, "P75"], byrow=F, ncol=long.r80, nrow=long.o80)
**o90.matrix** ⇐ matrix( o90[, "P75"], byrow=F, ncol=long.r90, nrow=long.o90)

**distance.minimiser60** ⇐ abs ( o60.matrix - r60.matrix)
**distance.minimiser80** ⇐ abs ( o80.matrix - r80.matrix)
**distance.minimiser90** ⇐ abs ( o90.matrix - r90.matrix)

**dimnames(distance.minimiser60)** ⇐ list( paste("o60",1:long.o60),
paste("r60", 1:long.r60))
**dimnames(distance.minimiser80)** ⇐ list( paste("o80",1:long.o80),
paste("r80", 1:long.r80))
**dimnames(distance.minimiser90)** ⇐ list( paste("o90",1:long.o90),
paste("r90", 1:long.r90))
**distance.minimiser60** ⇐ rbind(distance.minimiser60, 1:long.r60)
**distance.minimiser80** ⇐ rbind(distance.minimiser80, 1:long.r80)
**distance.minimiser90** ⇐ rbind(distance.minimiser90, 1:long.r90)

**minimum60** ⇐ t(apply(distance.minimiser60, 1, min))
**minimum80** ⇐ t(apply(distance.minimiser80, 1, min))
**minimum90** ⇐ t(apply(distance.minimiser90, 1, min))

**l.k60** ⇐ NULL
**l.k80** ⇐ NULL
**l.k90** ⇐ NULL

**z.l60** ⇐ NULL

**z.l80** ⇐ NULL

**z.l90** ⇐ NULL

**o.impute60** ⇐ o60

**o.impute80** ⇐ o80

**o.impute90** ⇐ o90

<u>Debut de la boucle FOR</u>

date()

**for(j in 1:long.o60)**

{ **tempo60** ⇐ r60 [distance.minimiser60[long.o60 + 1, distance.minimiser60[j,]== minimum60[j]] , c("Rev84", "NO.de.r60", "P75")]

if( length( tempo60[, "Rev84"] )> 1 ) **tempo260** ⇐ tempo60[sample(1:dim(tempo60)[1], size=1), ] } else {**tempo260** ⇐ tempo60 }

**o.impute60[j,2]** ⇐ tempo260[,"Rev84"]

**l.k60[j]** ⇐ tempo60[as.numeric(dimnames(tempo60)[[1]]) == as.numeric(dimnames(tempo260)[[1]]), "NO.de.r60"]

**z.l60[j]** ⇐ tempo60[as.numeric(dimnames(tempo60)[[1]]) == as.numeric(dimnames(tempo260)[[1]]), "P75"]

}

**for(j in 1:long.o80)**

{ **tempo80** ⇐ r80 [distance.minimiser80[long.o80 + 1, distance.minimiser80[j,]== minimum80[j]] , c("Rev84", "NO.de.r80", "P75")]

if( length( tempo80[, "Rev84"] ) > 1 ) **tempo280** ⇐ tempo80[sample(1:dim(tempo80)[1], size=1), ] } else {**tempo280** ⇐ tempo80 }

**o.impute80[j,2]** ⇐ tempo280[,"Rev84"]

**l.k80[j]** ⇐ tempo80[as.numeric(dimnames(tempo80)[[1]]) ==

```
as.numeric(dimnames(tempo280)[[1]]), "NO.de.r80"]
z.l80[j] <= tempo80[as.numeric(dimnames(tempo80)[[1]]) ==
as.numeric(dimnames(tempo280)[[1]]), "P75"] }

for(j in 1:long.o90)
{ tempo90 <= r90 [distance.minimiser90[long.o90 + 1, distance.minimiser90[j,]==
minimum90[j]] , c("Rev84", "NO.de.r90", "P75")]
if( length( tempo90[, "Rev84"] ) > 1 ) tempo290 <= tempo90[sample(1:dim(tempo90)[1],
size=1), ]  else {tempo290 <= tempo90
o.impute90[j,2] <= tempo290[,"Rev84"]
l.k90[j] <= tempo90[as.numeric(dimnames(tempo90)[[1]]) ==
as.numeric(dimnames(tempo290)[[1]]), "NO.de.r90"]
z.l90[j] <= tempo90[as.numeric(dimnames(tempo90)[[1]]) ==
as.numeric(dimnames(tempo290)[[1]]), "P75"]

o.impute60 <= cbind(o.impute60,l.k60)
o.impute80 <= cbind(o.impute80,l.k80)
o.impute90 <= cbind(o.impute90,l.k90)

names(o.impute60) <= c("Rmt85", "Rev84", "P75", "No.de.o60","l.k60")
names(o.impute80) <= c("Rmt85", "Rev84", "P75", "No.de.o80","l.k80")
names(o.impute90) <= c("Rmt85", "Rev84", "P75", "No.de.o90","l.k90")

l.k.sort60 <= sort(l.k60)
l.k.sort80 <= sort(l.k80)
l.k.sort90 <= sort(l.k90)

names(l.k.sort60) <= c(as.character(l.k.sort60))
names(l.k.sort80) <= c(as.character(l.k.sort80))
```

**names(l.k.sort90)** $\Leftarrow$ c(as.character(l.k.sort90))

**l.k.unique60** $\Leftarrow$ unique(l.k60)
**l.k.unique80** $\Leftarrow$ unique(l.k80)
**l.k.unique90** $\Leftarrow$ unique(l.k90)

**l.k.unique.sort60** $\Leftarrow$ sort(l.k.unique60)
**l.k.unique.sort80** $\Leftarrow$ sort(l.k.unique80)
**l.k.unique.sort90** $\Leftarrow$ sort(l.k.unique90)

**names(l.k.unique.sort60)** $\Leftarrow$ c(as.character(l.k.unique.sort60))
**names(l.k.unique.sort80)** $\Leftarrow$ c(as.character(l.k.unique.sort80))
**names(l.k.unique.sort90)** $\Leftarrow$ c(as.character(l.k.unique.sort90))

**non.donneur60** $\Leftarrow$ r60[ -l.k.unique60, "NO.de.r60"]
**non.donneur80** $\Leftarrow$ r80[ -l.k.unique80, "NO.de.r80"]
**non.donneur90** $\Leftarrow$ r90[ -l.k.unique90, "NO.de.r90"]

**names(non.donneur60)** $\Leftarrow$ c(as.character(non.donneur60))
**names(non.donneur80)** $\Leftarrow$ c(as.character(non.donneur80))
**names(non.donneur90)** $\Leftarrow$ c(as.character(non.donneur90))

**ordre60** $\Leftarrow$ sort( c(l.k.sort60, non.donneur60))
**ordre80** $\Leftarrow$ sort( c(l.k.sort80, non.donneur80))
**ordre90** $\Leftarrow$ sort( c(l.k.sort90, non.donneur90))

**F.l60** $\Leftarrow$ summary(as.factor(ordre60))
**F.l80** $\Leftarrow$ summary(as.factor(ordre80))
**F.l90** $\Leftarrow$ summary(as.factor(ordre90))

**F.l60** [ is.na(match(names(F.l60),names(l.k.unique.sort60)))] $\Leftarrow$ 0
**F.l80** [ is.na(match(names(F.l80),names(l.k.unique.sort80)))] $\Leftarrow$ 0
**F.l90** [ is.na(match(names(F.l90),names(l.k.unique.sort90)))] $\Leftarrow$ 0

**o.x.impute60** $\Leftarrow$ o.impute60[, "Rmt85"]
**o.x.impute80** $\Leftarrow$ o.impute80[, "Rmt85"]
**o.x.impute90** $\Leftarrow$ o.impute90[, "Rmt85"]

**o.y.impute60** $\Leftarrow$ o.impute60[, "Rev84"]
**o.y.impute80** $\Leftarrow$ o.impute80[, "Rev84"]
**o.y.impute90** $\Leftarrow$ o.impute90[, "Rev84"]

**o.z.impute60** $\Leftarrow$ o.impute60[, "P75"]
**o.z.impute80** $\Leftarrow$ o.impute80[, "P75"]
**o.z.impute90** $\Leftarrow$ o.impute90[, "P75"]

**s.impute60** $\Leftarrow$ rbind( r60[,-c(4:5)],o.impute60[,-c(4:5)])
**s.impute80** $\Leftarrow$ rbind( r80[,-c(4:5)],o.impute80[,-c(4:5)])
**s.impute90** $\Leftarrow$ rbind( r90[,-c(4:5)],o.impute90[,-c(4:5)])

**s.impute60** $\Leftarrow$ s.impute60[dimnames(s)[[1]],]
**s.impute80** $\Leftarrow$ s.impute80[dimnames(s)[[1]],]
**s.impute90** $\Leftarrow$ s.impute90[dimnames(s)[[1]],]

**s.y.impute60** $\Leftarrow$ s.impute60[, "Rev84"]
**s.y.impute80** $\Leftarrow$ s.impute80[, "Rev84"]
**s.y.impute90** $\Leftarrow$ s.impute90[, "Rev84"]

**s.x.impute60** $\Leftarrow$ s.impute60[, "Rmt85"]
**s.x.impute80** $\Leftarrow$ s.impute80[, "Rmt85"]
**s.x.impute90** $\Leftarrow$ s.impute90[, "Rmt85"]

**s.z.impute60** $\Leftarrow$ s.impute60[, "P75"]
**s.z.impute80** $\Leftarrow$ s.impute80[, "P75"]
**s.z.impute90** $\Leftarrow$ s.impute90[, "P75"]

**R.hat.impute60** $\Leftarrow$ ( sum(r60.y) + sum(o.y.impute60) ) / sum(s.x.impute60)
**R.hat.impute80** $\Leftarrow$ ( sum(r80.y) + sum(o.y.impute80) ) / sum(s.x.impute80)
**R.hat.impute90** $\Leftarrow$ ( sum(r90.y) + sum(o.y.impute90) ) / sum(s.x.impute90)

**y.mean.complete60[i]** $\Leftarrow$ mean(mu284.dat.mod[,"Rmt85"]) * R.hat.impute60
**y.mean.complete80[i]** $\Leftarrow$ mean(mu284.dat.mod[,"Rmt85"]) * R.hat.impute80
**y.mean.complete90[i]** $\Leftarrow$ mean(mu284.dat.mod[,"Rmt85"]) * R.hat.impute90

**estvech[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop)$^2$

**estvimp60[i]** $\Leftarrow$ (y.mean.complete60[i]-y.mean[i])$^2$
**estvimp80[i]** $\Leftarrow$ (y.mean.complete80[i]-y.mean[i])$^2$
**estvimp90[i]** $\Leftarrow$ (y.mean.complete90[i]-y.mean[i])$^2$

**estvmix60[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete60[i]-y.mean[i])
**estvmix80[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete80[i]-y.mean[i])
**estvmix90[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete90[i]-y.mean[i])

**y.variance.complete60[i]** $\Leftarrow$ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[, "Rmt85"])$^2$ / mean(s.x.impute60)$^2$ * 1/(nb.echantillon - 1) * sum( (s.y.impute60 - R.hat.impute60 * s.x.impute60)$^2$ )

**y.variance.complete80[i]** $\Leftarrow$ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[, "Rmt85"])$^2$ / mean(s.x.impute80)$^2$ * 1/(nb.echantillon - 1) * sum( (s.y.impute80 - R.hat.impute80 * s.x.impute80)$^2$ )

**y.variance.complete90[i]** $\Leftarrow$ (1/nb.echantillon - 1/N) * mean(mu284.dat.mod[, "Rmt85"])$^2$ / mean(s.x.impute90)$^2$ * 1/(nb.echantillon - 1) * sum( (s.y.impute90 - R.hat.impute90 * s.x.impute90)$^2$ )

**beta60** $\Leftarrow$ sum (r60.y) / sum (r60.z)

**beta80** $\Leftarrow$ sum (r80.y) / sum (r80.z)

**beta90** $\Leftarrow$ sum (r90.y) / sum (r90.z)

**CV60** $\Leftarrow$ sqrt ( sum( (r60.z - mean(r60.z))$^2$) /(long.r60 - 1) ) / mean(r60.z)

**CV80** $\Leftarrow$ sqrt ( sum( (r80.z - mean(r80.z))$^2$) /(long.r80 - 1) ) / mean(r80.z)

**CV90** $\Leftarrow$ sqrt ( sum( (r90.z - mean(r90.z))$^2$) /(long.r90 - 1) ) / mean(r90.z)

**hatsigma60** $\Leftarrow$ (1/( 1-CV60$^2$/(long.r60))) * (sum((r60.y - beta60 * r60.z)$^2$)/ sum(r60.z) )

**hatsigma80** $\Leftarrow$ (1/( 1-CV80$^2$/(long.r80))) * (sum((r80.y - beta80 * r80.z)$^2$)/ sum(r80.z) )

**hatsigma90** $\Leftarrow$ (1/( 1-CV90$^2$/(long.r90))) * (sum((r90.y - beta90 * r90.z)$^2$)/ sum(r90.z) )

**vmix60[i]** $\Leftarrow$ 2*hatsigma60 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute60))$^2$ * sum((r60[o.impute60[,"l.k60"],"P75"] - o.impute60[,"P75"])) - 2*hatsigma60 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute60) ) * sum((r60[o.impute60[,"l.k60"],"P75"] - o.impute60[,"P75"]))

**vmix80[i]** $\Leftarrow$ 2*hatsigma80 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute80))$^2$ * sum((r80[o.impute80[,"l.k80"],"P75"]

- o.impute80[,"P75"])) - 2*hatsigma80 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[,
"Rmt85"]) / mean(s.x.impute80) ) * sum((r80[o.impute80[,"l.k80"],"P75"]
- o.impute80[,"P75"]))
**vmix90[i]** $\Leftarrow$ 2*hatsigma90 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[,
"Rmt85"]) / mean(s.x.impute90))$^2$ * sum((r90[o.impute90[,"l.k90"],"P75"]
- o.impute90[,"P75"])) - 2*hatsigma90 * (1 / nb.echantillon)$^2$ * ( mean(mu284.dat.mod[,
"Rmt85"]) / mean(s.x.impute90) ) * sum((r90[o.impute90[,"l.k90"],"P75"]
- o.impute90[,"P75"]))

**variance.impute.complete60[i]** $\Leftarrow$ hatsigma60 * (1 /nb.echantillon)$^2$
* ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute60))$^2$ * ( 2 *
sum(o.z.impute60) + sum( F.l60 * (F.l60 - 1) * r60.z ) )
**variance.impute.complete80[i]** $\Leftarrow$ hatsigma80 * (1 /nb.echantillon)$^2$
* ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute80))$^2$ * ( 2 *
sum(o.z.impute80) + sum( F.l80 * (F.l80- 1) * r80.z ) )
**variance.impute.complete90[i]** $\Leftarrow$ hatsigma90 * (1 /nb.echantillon)$^2$
* ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.x.impute90))$^2$ * ( 2 *
sum(o.z.impute90) + sum( F.l90 * (F.l90 - 1) * r90.z ) )

if (moyenne.pop $\geq$ y.mean.complete60[i] - qnorm(1-alpha/2)*sqrt(y.variance.complete60[i]
+ variance.impute.complete60[i] ) & moyenne.pop $\leq$ y.mean.complete60[i]
+ qnorm(1-alpha/2)*sqrt(y.variance.complete60[i] + variance.impute.complete60[i]))
{**TRMC.complete60** $\Leftarrow$ TRMC.complete60 + 1/nb.monte.carlo}
if (moyenne.pop $\geq$ y.mean.complete80[i] - qnorm(1-alpha/2)*sqrt(y.variance.complete80[i]
+ variance.impute.complete80[i] ) & moyenne.pop $\leq$ y.mean.complete80[i]
+ qnorm(1-alpha/2)*sqrt(y.variance.complete80[i] + variance.impute.complete80[i]))
**TRMC.complete80** $\Leftarrow$ TRMC.complete80 + 1/nb.monte.carlo
if (moyenne.pop $\geq$ y.mean.complete90[i] - qnorm(1-alpha/2)*sqrt(y.variance.complete90[i]
+ variance.impute.complete90[i] ) & moyenne.pop $\leq$ y.mean.complete90[i]

+ qnorm(1-alpha/2)*sqrt(y.variance.complete90[i] + variance.impute.complete90[i]))
**TRMC.complete90** $\Leftarrow$ TRMC.complete90 + 1/nb.monte.carlo

**diffnew60[i]** $\Leftarrow$ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.impute60[, "Rmt85"]))$^2$ * (beta60$^2$ * sum((r60[o.impute60[,"l.k60"],"P75"] - o.impute60[,"P75"])$^2$))
**diffnew80[i]** $\Leftarrow$ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.impute80[, "Rmt85"]))$^2$ * (beta80$^2$ * sum((r80[o.impute80[,"l.k80"],"P75"] - o.impute80[,"P75"])$^2$))
**diffnew90[i]** $\Leftarrow$ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * ( mean(mu284.dat.mod[, "Rmt85"]) / mean(s.impute90[, "Rmt85"]))$^2$ * (beta90$^2$ * sum((r90[o.impute90[,"l.k90"],"P75"] - o.impute90[,"P75"])$^2$))

**if (i%%100 = 0) {cat(i, ") ", date(), "\n")**
**}, grain.size = 15 }**
date()

**EMC.vmix60** $\Leftarrow$ mean(vmix60)
**EMC.vmix80** $\Leftarrow$ mean(vmix80)
**EMC.vmix90** $\Leftarrow$ mean(vmix90)
**EMC.estvmix60** $\Leftarrow$ mean(estvmix60)
**EMC.estvmix80** $\Leftarrow$ mean(estvmix80)
**EMC.estvmix90** $\Leftarrow$ mean(estvmix90)

**EMC.moyenne** $\Leftarrow$ mean(y.mean)
**VMC.moyenne.100.reponse** $\Leftarrow$ var(y.mean)
**EMCV.variance.echantillonnale.100.reponse** $\Leftarrow$ mean(y.var)
**EMC.Vech** $\Leftarrow$ mean(estvech)

**EMC.moyenne.impute60** $\Leftarrow$ mean(y.mean.complete60)
**EMC.moyenne.impute80** $\Leftarrow$ mean(y.mean.complete80)
**EMC.moyenne.impute90** $\Leftarrow$ mean(y.mean.complete90)

**VMC.moyenne.complete60** $\Leftarrow$ var(y.mean.complete60)
**VMC.moyenne.complete80** $\Leftarrow$ var(y.mean.complete80)
**VMC.moyenne.complete90** $\Leftarrow$ var(y.mean.complete90)

**EMCV.variance.echantillonnale.complete60** $\Leftarrow$ mean(y.variance.complete60)
**EMCV.variance.echantillonnale.complete80** $\Leftarrow$ mean(y.variance.complete80)
**EMCV.variance.echantillonnale.complete90** $\Leftarrow$ mean(y.variance.complete90)

**EMCV.variance.impute.complete60** $\Leftarrow$ mean(variance.impute.complete60)
**EMCV.variance.impute.complete80** $\Leftarrow$ mean(variance.impute.complete80)
**EMCV.variance.impute.complete90** $\Leftarrow$ mean(variance.impute.complete90)

**EMC.estvimp60** $\Leftarrow$ mean(estvimp60)
**EMC.estvimp80** $\Leftarrow$ mean(estvimp80)
**EMC.estvimp90** $\Leftarrow$ mean(estvimp90)

**EMCV.diffnew60** $\Leftarrow$ mean(diffnew60)
**EMCV.diffnew80** $\Leftarrow$ mean(diffnew80)
**EMCV.diffnew90** $\Leftarrow$ mean(diffnew90)

**v.chapeau.completenew60** $\Leftarrow$ EMCV.variance.impute.complete60 +
EMCV.variance.echantillonnale.complete60 - EMCV.diffnew60
**v.chapeau.completenew80** $\Leftarrow$ EMCV.variance.impute.complete80 +
EMCV.variance.echantillonnale.complete80 - EMCV.diffnew80
**v.chapeau.completenew90** $\Leftarrow$ EMCV.variance.impute.complete90 +

EMCV.variance.echantillonnale.complete90 - EMCV.diffnew90

**differencenew60** $\Leftarrow$ v.chapeau.completenew60 - VMC.moyenne.complete60
**differencenew80** $\Leftarrow$ v.chapeau.completenew80 - VMC.moyenne.complete80
**differencenew90** $\Leftarrow$ v.chapeau.completenew90 - VMC.moyenne.complete90

**Birel60** $\Leftarrow$ ( differencenew60 / VMC.moyenne.complete60 )
**Birel80** $\Leftarrow$ ( differencenew80 / VMC.moyenne.complete80 )
**Birel90** $\Leftarrow$ ( differencenew90 / VMC.moyenne.complete90 )

**R.hat.u** $\Leftarrow$ mean( mu284.dat.mod[, "Rev84"] ) / mean( mu284.dat.mod[, "Rmt85"] )
**estimateur.variance.population** $\Leftarrow$ (1/nb.echantillon - 1/N) * 1 / (N-1) * sum( (mu284.dat.mod[, "Rev84"] - R.hat.u * mu284.dat.mod[, "Rmt85"])$^2$ )

**Faire afficher les résultats**

moyenne.pop
EMC.moyenne
EMC.moyenne.impute60
EMC.moyenne.impute80
EMC.moyenne.impute90

TRMC.100.reponse
TRMC.complete60
TRMC.complete80
TRMC.complete90

estimateur.variance.population

VMC.moyenne.100.reponse

EMCV.variance.echantillonnale.100.reponse

EMC.Vech

estimateur.variance.population

VMC.moyenne.complete60

VMC.moyenne.complete80

VMC.moyenne.complete90

EMCV.variance.echantillonnale.complete60

EMCV.variance.echantillonnale.complete80

EMCV.variance.echantillonnale.complete90

EMCV.variance.impute.complete60

EMCV.variance.impute.complete80

EMCV.variance.impute.complete90

EMC.estvimp60

EMC.estvimp80

EMC.estvimp90

EMC.vmix60

EMC.vmix80

EMC.vmix90

EMC.estvmix60

EMC.estvmix80

EMC.estvmix90

EMCV.diffnew60

EMCV.diffnew80
EMCV.diffnew90

v.chapeau.completenew60
v.chapeau.completenew80
v.chapeau.completenew90

differencenew60
differencenew80
differencenew90

APPENDICE C

# Simulations de Monte Carlo pour l'estimateur par la régression

Voici le programme Splus permettant de faire une simulation de Monte Carlo. Nous considérons le plan d'échantillonnage SI et l'estimateur par la régression. Dans cet exemple, nous considérons le taux de réponse de 60 %.

```
options(object.size= 50920016)
nb.echantillon <= 100
nb.monte.carlo <= 1000
N <= 281
alpha <= 0.05
moyenne.pop <= mean(mu284.dat.mod[ ,"Rev84"])
```

**Initialisations**

```
y.var <= NULL
y.mean <= NULL
TRMC.100.reponse <= 0

y.variance.complete60 <= NULL
y.variance.complete80 <= NULL
y.variance.complete90 <= NULL
```

y.mean.complete60 $\Leftarrow$ NULL

y.mean.complete80 $\Leftarrow$ NULL

y.mean.complete90 $\Leftarrow$ NULL

variance.impute.complete60 $\Leftarrow$ NULL

variance.impute.complete80 $\Leftarrow$ NULL

variance.impute.complete90 $\Leftarrow$ NULL

TRMC.complete60 $\Leftarrow$ 0

TRMC.complete80 $\Leftarrow$ 0

TRMC.complete90 $\Leftarrow$ 0

vmix60 $\Leftarrow$ NULL

vmix80 $\Leftarrow$ NULL

vmix90 $\Leftarrow$ NULL

estvmix60 $\Leftarrow$ NULL

estvmix80 $\Leftarrow$ NULL

estvmix90 $\Leftarrow$ NULL

estvimp60 $\Leftarrow$ NULL

estvimp80 $\Leftarrow$ NULL

estvimp90 $\Leftarrow$ NULL

estvech $\Leftarrow$ NULL

prob.inclusion60 $\Leftarrow$ 0.60

prob.inclusion80 $\Leftarrow$ 0.80

prob.inclusion90 $\Leftarrow$ 0.90

```
f ⇐ nb.echantillon/N
diffnew60 ⇐ NULL
diffnew80 ⇐ NULL
diffnew90 ⇐ NULL

date()
For(i= 1:nb.monte.carlo,
{
repeat
{
s ⇐ mu284.dat.mod[(sample(1:281, nb.echantillon, replace=F)), c("Rmt85",
"Rev84", "P75")]
s ⇐ cbind(s,1:length(dimnames(s)[[1]]))dimnames(s) ⇐ list(dimnames(s)[[1]],
c("Rmt85", "Rev84", "P75", "Numero"))
s.y ⇐ s[ ,"Rev84"]
s.x ⇐ s[ ,"Rmt85"]
s.z ⇐ s[ ,"P75"]

estimateur par la régression pour 100 % réponse
Beta.hat ⇐ sum( (s.x - mean(s.x)) * (s.y - mean(s.y))) / ( (length(s.x)-
1) * var(s.x) )
y.mean[i] ⇐ mean(s.y) + Beta.hat * (mean(mu284.dat.mod[, "Rmt85"])-
mean(s.x) )

e.k ⇐ s.y - mean(s.y) - Beta.hat * ( s.x - mean(s.x) )
a.s ⇐ ( mean(mu284.dat.mod[, "Rmt85"]) - mean(s.x) ) / ( ( (length(s.x)-
1) / length(s.x) ) * var(s.x) )
g.k ⇐ (1 + a.s * ( s.x - mean(s.x) ) )
```

**s** $\Leftarrow$ cbind(s, g.k)
**dimnames(s)** $\Leftarrow$ list(dimnames(s)[[1]], c("Rmt85","Rev84", "P75", "Numero", "g.k"))

**y.var[i]** $\Leftarrow$ (1-f) / (nb.echantillon *(nb.echantillon-1))* sum(s[, "g.k"]$^2$ * e.k$^2$)

if (moyenne.pop $\geq$ y.mean[i] - qnorm(1-alpha/2)*sqrt(y.var[i]) & moyenne.pop $\leq$ y.mean[i] + qnorm(1-alpha/2)*sqrt(y.var[i]))
{**TRMC.100.reponse** $\Leftarrow$ TRMC.100.reponse + 1/nb.monte.carlo}

estimateur par la régression pour la non-réponse

**unif** $\Leftarrow$ runif(dimnames(s)[[1]], 0, 1)
**r60** $\Leftarrow$ s[unif $\leq$ prob.inclusion60, 1:5]
**r80** $\Leftarrow$ s[unif $\leq$ prob.inclusion80, 1:5]
**r90** $\Leftarrow$ s[unif $\leq$ prob.inclusion90, 1:5]

**long.r60** $\Leftarrow$ length(dimnames(r60)[[1]])
**long.r80** $\Leftarrow$ length(dimnames(r80)[[1]])
**long.r90** $\Leftarrow$ length(dimnames(r90)[[1]])

**r60** $\Leftarrow$ cbind(r60,1:long.r60)
**r80** $\Leftarrow$ cbind(r80,1:long.r80)
**r90** $\Leftarrow$ cbind(r90,1:long.r90)

**dimnames(r60)** $\Leftarrow$ list(dimnames(r60)[[1]], c("Rmt85", "Rev84", "P75", "Numero.de.s", "g.k60", "NO.de.r60"))
**dimnames(r80)** $\Leftarrow$ list(dimnames(r80)[[1]], c("Rmt85", "Rev84",

```
"P75", "Numero.de.s", "g.k80", "NO.de.r80"))
dimnames(r90) ⇐ list(dimnames(r90)[[1]], c("Rmt85", "Rev84",
"P75", "Numero.de.s", "g.k90", "NO.de.r90"))

r60.x ⇐ r60[, "Rmt85"]
r80.x ⇐ r80[, "Rmt85"]
r90.x ⇐ r90[, "Rmt85"]
r60.y ⇐ r60[, "Rev84"]
r80.y ⇐ r80[, "Rev84"]
r90.y ⇐ r90[, "Rev84"]
r60.z ⇐ r60[, "P75"]
r80.z ⇐ r80[, "P75"]
r90.z ⇐ r90[, "P75"]

numero60 ⇐ s[dimnames(r60)[[1]] ,4]
numero80 ⇐ s[dimnames(r80)[[1]] ,4]
numero90 ⇐ s[dimnames(r90)[[1]] ,4]

o60 ⇐ s[-numero60,1:5]
o80 ⇐ s[-numero80,1:5]
o90 ⇐ s[-numero90,1:5]

o60.y ⇐ o60[,"Rev84"]
o80.y ⇐ o80[,"Rev84"]
o90.y ⇐ o90[,"Rev84"]

o60.z ⇐ o60[,"P75"]
o80.z ⇐ o80[,"P75"]
o90.z ⇐ o90[,"P75"]
```

**o60.x** $\Leftarrow$ o60[,"Rmt85"]
**o80.x** $\Leftarrow$ o80[,"Rmt85"]
**o90.x** $\Leftarrow$ o90[,"Rmt85"]

if (length(dimnames(o60)[[1]]) > 1) break
if (length(dimnames(o80)[[1]]) > 1) break
if (length(dimnames(o90)[[1]]) > 1) break
}

**long.o60** $\Leftarrow$ length(dimnames(o60)[[1]])
**long.o80** $\Leftarrow$ length(dimnames(o80)[[1]])
**long.o90** $\Leftarrow$ length(dimnames(o90)[[1]])

**r60.matrix** $\Leftarrow$ matrix( r60[, "P75"], byrow=T, ncol=long.r60, nrow=long.o60)
**r80.matrix** $\Leftarrow$ matrix( r80[, "P75"], byrow=T, ncol=long.r80, nrow=long.o80)
**r90.matrix** $\Leftarrow$ matrix( r90[, "P75"], byrow=T, ncol=long.r90, nrow=long.o90)

**o60.matrix** $\Leftarrow$ matrix( o60[, "P75"], byrow=F, ncol=long.r60, nrow=long.o60)
**o80.matrix** $\Leftarrow$ matrix( o80[, "P75"], byrow=F, ncol=long.r80, nrow=long.o80)
**o90.matrix** $\Leftarrow$ matrix( o90[, "P75"], byrow=F, ncol=long.r90, nrow=long.o90)

**distance.minimiser60** $\Leftarrow$ abs ( o60.matrix - r60.matrix)
**distance.minimiser80** $\Leftarrow$ abs ( o80.matrix - r80.matrix)
**distance.minimiser90** $\Leftarrow$ abs ( o90.matrix - r90.matrix)

**dimnames(distance.minimiser60)** $\Leftarrow$ list(paste("o60", 1:long.o60),
paste("r60", 1:long.r60))
**dimnames(distance.minimiser80)** $\Leftarrow$ list(paste("o80", 1:long.o80),

paste("r80", 1:long.r80))
**dimnames(distance.minimiser90)** ⇐ list(paste("o90", 1:long.o90),
paste("r90", 1:long.r90))

**distance.minimiser60** ⇐ rbind(distance.minimiser60, 1:long.r60)
**distance.minimiser80** ⇐ rbind(distance.minimiser80, 1:long.r80)
**distance.minimiser90** ⇐ rbind(distance.minimiser90, 1:long.r90)

**minimum60** ⇐ t(apply(distance.minimiser60, 1, min))
**minimum80** ⇐ t(apply(distance.minimiser80, 1, min))
**minimum90** ⇐ t(apply(distance.minimiser90, 1, min))

**l.k60** ⇐ NULL
**l.k80** ⇐ NULL
**l.k90** ⇐ NULL

**z.l60** ⇐ NULL
**z.l80** ⇐ NULL
**z.l90** ⇐ NULL

**o.impute60** ⇐ o60
**o.impute80** ⇐ o80
**o.impute90** ⇐ o90

for(j in 1:long.o60)
{ **tempo60** ⇐ r60 [distance.minimiser60[long.o60 + 1, distance.minimiser60[j,]
== minimum60[j]] , c("Rev84", "NO.de.r60", "P75")]
if( length( tempo60[, "Rev84"] ) > 1 ) {tempo260 ⇐ tempo60[sample(1:dim(tempo60)[1],
size=1), ]

```
} else
{tempo260 ⇐ tempo60
}
o.impute60[j,2] ⇐ tempo260[,"Rev84"]
l.k60[j] ⇐ tempo60[as.numeric(dimnames(tempo60)[[1]]) ==
as.numeric(dimnames(tempo260)[[1]]), "NO.de.r60"]
z.l60[j] ⇐ tempo60[as.numeric(dimnames(tempo60)[[1]]) ==
as.numeric(dimnames(tempo260)[[1]]), "P75"]
}

for(j in 1:long.o80)
{ tempo80 ⇐ r80 [distance.minimiser80[long.o80 + 1, distance.minimiser80[j,]==
minimum80[j]] , c("Rev84", "NO.de.r80", "P75")]
if( length( tempo80[, "Rev84"] ) > 1 ) {tempo280 ⇐ tempo80[sample(1:dim(tempo80)[1],
size=1), ]
} else
{tempo280 ⇐ tempo80
}
o.impute80[j,2] ⇐ tempo280[,"Rev84"]
l.k80[j] ⇐ tempo80[as.numeric(dimnames(tempo80)[[1]]) ==
as.numeric(dimnames(tempo280)[[1]]), "NO.de.r80"]
z.l80[j] ⇐ tempo80[as.numeric(dimnames(tempo80)[[1]]) ==
as.numeric(dimnames(tempo280)[[1]]), "P75"]
}

for(j in 1:long.o90)
{ tempo90 ⇐ r90 [distance.minimiser90[long.o90 + 1, distance.minimiser90[j,]
== minimum90[j]] , c("Rev84", "NO.de.r90", "P75")]
if( length( tempo90[, "Rev84"] ) > 1 ) {tempo290 ⇐ tempo90[sample(1:dim(tempo90)[1],
```

```
size=1), ]
} else
{tempo290 <= tempo90
}
o.impute90[j,2] <= tempo290[,"Rev84"]
l.k90[j] <= tempo90[as.numeric(dimnames(tempo90)[[1]]) ==
as.numeric(dimnames(tempo290)[[1]]), "NO.de.r90"]
z.l90[j] <= tempo90[as.numeric(dimnames(tempo90)[[1]]) ==
as.numeric(dimnames(tempo290)[[1]]), "P75"]
}

o.impute60 <= cbind(o.impute60,l.k60)
o.impute80 <= cbind(o.impute80,l.k80)
o.impute90 <= cbind(o.impute90,l.k90)

names(o.impute60) <= c("Rmt85", "Rev84", "P75", "No.de.o60","g.k60","l.k60")
names(o.impute80) <= c("Rmt85", "Rev84", "P75", "No.de.o80","g.k80","l.k80")
names(o.impute90) <= c("Rmt85", "Rev84", "P75", "No.de.o90","g.k90","l.k90")

l.k.sort60 <= sort(l.k60)
l.k.sort80 <= sort(l.k80)
l.k.sort90 <= sort(l.k90)

names(l.k.sort60) <= c(as.character(l.k.sort60))
names(l.k.sort80) <= c(as.character(l.k.sort80))
names(l.k.sort90) <= c(as.character(l.k.sort90))

l.k.unique60 <= unique(l.k60)
l.k.unique80 <= unique(l.k80)
```

**l.k.unique90** $\Leftarrow$ unique(l.k90)

**l.k.unique.sort60** $\Leftarrow$ sort(l.k.unique60)
**l.k.unique.sort80** $\Leftarrow$ sort(l.k.unique80)
**l.k.unique.sort90** $\Leftarrow$ sort(l.k.unique90)

**names(l.k.unique.sort60)** $\Leftarrow$ c(as.character(l.k.unique.sort60))
**names(l.k.unique.sort80)** $\Leftarrow$ c(as.character(l.k.unique.sort80))
**names(l.k.unique.sort90)** $\Leftarrow$ c(as.character(l.k.unique.sort90))

**non.donneur60** $\Leftarrow$ r60[ -l.k.unique60, "NO.de.r60"]
**non.donneur80** $\Leftarrow$ r80[ -l.k.unique80, "NO.de.r80"]
**non.donneur90** $\Leftarrow$ r90[ -l.k.unique90, "NO.de.r90"]

**names(non.donneur60)** $\Leftarrow$ c(as.character(non.donneur60))
**names(non.donneur80)** $\Leftarrow$ c(as.character(non.donneur80))
**names(non.donneur90)** $\Leftarrow$ c(as.character(non.donneur90))

**ordre60** $\Leftarrow$ sort( c(l.k.sort60, non.donneur60))
**ordre80** $\Leftarrow$ sort( c(l.k.sort80, non.donneur80))
**ordre90** $\Leftarrow$ sort( c(l.k.sort90, non.donneur90))

**F.l60** $\Leftarrow$ summary(as.factor(ordre60))
**F.l80** $\Leftarrow$ summary(as.factor(ordre80))
**F.l90** $\Leftarrow$ summary(as.factor(ordre90))

**F.l60** [ is.na(match(names(F.l60),names(l.k.unique.sort60)))] $\Leftarrow$ 0
**F.l80** [ is.na(match(names(F.l80),names(l.k.unique.sort80)))] $\Leftarrow$ 0
**F.l90** [ is.na(match(names(F.l90),names(l.k.unique.sort90)))] $\Leftarrow$ 0

```
Somme.gk.ol60 <= NULL
Somme.gk.ol80 <= NULL
Somme.gk.ol90 <= NULL

for(i in 1:length(F.l60))
{ if(length(l.k60[l.k60 ==as.numeric(names( F.l60[i] ))]) == 0 ) {Somme.gk.ol60[i]
<= 0 } else
{if(length( l.k60[l.k60 == as.numeric(names(F.l60[i]))]) == 1 |
apply(matrix(o.impute60[o.impute60[,6] == as.numeric(names(F.l60[i])),"Rmt85"],
nrow=length(l.k60[l.k60 == as.numeric(names(F.l60[i]))]),ncol=1),2,var)==0)
{Somme.gk.ol60[i] <= 1 } else
{if(length( l.k60[l.k60 == as.numeric(names(F.l60[i])) ] ) != 0 | length(l.k60[l.k60
== as.numeric(names(F.l60[i])) ] ) != 1 | apply(matrix( o.impute60[o.impute60[,6]==
as.numeric(names(F.l60[i])), "Rmt85"], nrow=length( l.k60[l.k60 ==
as.numeric(names(F.l60[i])) ] ), ncol=1) ,2, var) !=0 ) { Somme.gk.ol60[i]
<= sum ( 1 + (mean(mu284.dat.mod[, "Rmt85"]) - apply(matrix(
o.impute60[ o.impute60[,6] == as.numeric(names(F.l60[i])), "Rmt85"],
nrow=length( l.k60[l.k60 == as.numeric(names(F.l60[i])) ] ), ncol=1)
,2,mean ) ) / (( (F.l60[i]-1) / F.l60[i]) * apply(matrix(o.impute60[o.impute60[,6]==
as.numeric(names(F.l60[i])), "Rmt85"], nrow=length( l.k60[l.k60 ==
as.numeric(names(F.l60[i])) ] ), ncol=1) ,2, var) ) * ( o.impute60[o.impute60[,6]==
as.numeric(names(F.l60[i])), "Rmt85"] - apply(matrix( o.impute60[o.impute60[,6]==
as.numeric(names(F.l60[i])), "Rmt85"], nrow=length( l.k60[l.k60 ==
as.numeric(names(F.l60[i])) ] ), ncol=1) ,2,mean )) )
} } } }

for(i in 1:length(F.l80))
{ if(length(l.k80[l.k80 ==as.numeric(names( F.l80[i] ))]) == 0 ) {Somme.gk.ol80[i]
```

```
<= 0 } else
{if(length( l.k80[l.k80 == as.numeric(names(F.l80[i]))]) == 1 |
apply(matrix(o.impute80[o.impute80[,6] == as.numeric(names(F.l80[i])),"Rmt85"],
nrow=length(l.k80[l.k80 == as.numeric(names(F.l80[i]))]),ncol=1),2,var)==0)
{Somme.gk.ol80[i] <= 1 } else
if(length( l.k80[l.k80 == as.numeric(names(F.l80[i])) ] ) != 0 | length(l.k80[l.k80
== as.numeric(names(F.l80[i])) ] ) != 1 | apply(matrix( o.impute80[o.impute80[,6]==
as.numeric(names(F.l80[i])), "Rmt85"], nrow=length( l.k80[l.k80 ==
as.numeric(names(F.l80[i])) ] ), ncol=1) ,2, var) !=0 ) {Somme.gk.ol80[i]
<= sum ( 1 + ( mean(mu284.dat.mod[, "Rmt85"]) - apply(matrix(
o.impute80[ o.impute80[,6] == as.numeric(names(F.l80[i])), "Rmt85"],
nrow=length( l.k80[l.k80 == as.numeric(names(F.l80[i])) ] ), ncol=1)
,2,mean ) ) / (( (F.l80[i]-1) / F.l80[i]) * apply(matrix(o.impute80[o.impute80[,6]==
as.numeric(names(F.l80[i])), "Rmt85"], nrow=length( l.k80[l.k80 ==
as.numeric(names(F.l80[i])) ] ), ncol=1) ,2, var) ) * ( o.impute80[o.impute80[,6]==
as.numeric(names(F.l80[i])), "Rmt85"] - apply(matrix( o.impute80[o.impute80[,6]==
as.numeric(names(F.l80[i])), "Rmt85"], nrow=length( l.k80[l.k80 ==
as.numeric(names(F.l80[i])) ] ), ncol=1) ,2,mean )) )
} } } }

for(i in 1:length(F.l90))
{ if(length(l.k90[l.k90 ==as.numeric(names( F.l90[i] ))]) == 0){Somme.gk.ol90[i]
<= 0 } else
{if(length( l.k90[l.k90 == as.numeric(names(F.l90[i]))]) == 1 |
apply(matrix(o.impute90[o.impute90[,6] == as.numeric(names(F.l90[i])),"Rmt85"],
nrow=length(l.k90[l.k90 == as.numeric(names(F.l90[i]))]),ncol=1),2,var)==0)
{Somme.gk.ol90[i] <= 1 } else
{if(length( l.k90[l.k90 == as.numeric(names(F.l90[i])) ] ) != 0 | length(l.k90[l.k90
== as.numeric(names(F.l90[i])) ] ) != 1 | apply(matrix( o.impute90[o.impute90[,6]==
```

```
as.numeric(names(F.l90[i])), "Rmt85"], nrow=length( l.k90[l.k90 ==
as.numeric(names(F.l90[i])) ] ), ncol=1) ,2, var) !=0 ) {Somme.gk.ol90[i]
<= sum ( 1 + (mean(mu284.dat.mod[, "Rmt85"]) - apply(matrix(
o.impute90[o.impute90[,6]== as.numeric(names(F.l90[i])), "Rmt85"],
nrow=length( l.k90[l.k90 == as.numeric(names(F.l90[i])) ] ), ncol=1)
,2,mean ) ) / (( (F.l90[i]-1) / F.l90[i]) * apply(matrix(o.impute90[o.impute90[,6]==
as.numeric(names(F.l90[i])), "Rmt85"], nrow=length( l.k90[l.k90 ==
as.numeric(names(F.l90[i])) ] ), ncol=1) ,2, var) ) * ( o.impute90[o.impute90[,6]==
as.numeric(names(F.l90[i])), "Rmt85"] - apply(matrix( o.impute90[o.impute90[,6]==
as.numeric(names(F.l90[i])), "Rmt85"], nrow=length( l.k90[l.k90 ==
as.numeric(names(F.l90[i])) ] ), ncol=1) ,2,mean )) )
} } } }

o.x.impute60 <= o.impute60[, "Rmt85"]
o.x.impute80 <= o.impute80[, "Rmt85"]
o.x.impute90 <= o.impute90[, "Rmt85"]

o.y.impute60 <= o.impute60[, "Rev84"]
o.y.impute80 <= o.impute80[, "Rev84"]
o.y.impute90 <= o.impute90[, "Rev84"]

o.z.impute60 <= o.impute60[, "P75"]
o.z.impute80 <= o.impute80[, "P75"]
o.z.impute90 <= o.impute90[, "P75"]

s.impute60 <= rbind( r60[,-c(4,6)], o.impute60[,-c(4,6)])
s.impute80 <= rbind( r80[,-c(4,6)], o.impute80[,-c(4,6)])
s.impute90 <= rbind( r90[,-c(4,6)], o.impute90[,-c(4,6)])
```

**s.impute60** $\Leftarrow$ s.impute60[dimnames(s)[[1]],]
**s.impute80** $\Leftarrow$ s.impute80[dimnames(s)[[1]],]
**s.impute90** $\Leftarrow$ s.impute90[dimnames(s)[[1]],]

**s.y.impute60** $\Leftarrow$ s.impute60[, "Rev84"]
**s.y.impute80** $\Leftarrow$ s.impute80[, "Rev84"]
**s.y.impute90** $\Leftarrow$ s.impute90[, "Rev84"]

**s.x.impute60** $\Leftarrow$ s.impute60[, "Rmt85"]
**s.x.impute80** $\Leftarrow$ s.impute80[, "Rmt85"]
**s.x.impute90** $\Leftarrow$ s.impute90[, "Rmt85"]

**s.z.impute60** $\Leftarrow$ s.impute60[, "P75"]
**s.z.impute80** $\Leftarrow$ s.impute80[, "P75"]
**s.z.impute90** $\Leftarrow$ s.impute90[, "P75"]

**beta60** $\Leftarrow$ sum (r60.y) / sum (r60.z)
**beta80** $\Leftarrow$ sum (r80.y) / sum (r80.z)
**beta90** $\Leftarrow$ sum (r90.y) / sum (r90.z)

**CV60** $\Leftarrow$ sqrt ( sum( (r60.z - mean(r60.z))$^2$ ) / (long.r60 -1) ) / mean(r60.z)
**CV80** $\Leftarrow$ sqrt ( sum( (r80.z - mean(r80.z))$^2$ ) / (long.r80 -1) ) / mean(r80.z)
**CV90** $\Leftarrow$ sqrt ( sum( (r90.z - mean(r90.z))$^2$ ) / (long.r90 -1) ) / mean(r90.z)

**hatsigma60** $\Leftarrow$ (1/(1-CV60$^2$/(long.r60))) * (sum((r60.y - beta60 * r60.z)$^2$ )/ sum(r60.z) )

**hatsigma80** $\Leftarrow$ (1/(1-CV80$^2$/(long.r80))) * (sum((r80.y - beta80 * r80.z)$^2$ )/ sum(r80.z) )
**hatsigma90** $\Leftarrow$ (1/(1-CV90$^2$/(long.r90))) * (sum((r90.y - beta90 * r90.z)$^2$ )/ sum(r90.z) )

**Beta.hat.impute60** $\Leftarrow$ sum( (s.x.impute60 - mean(s.x.impute60)) * (s.y.impute60 - mean(s.y.impute60))) / ( (length(s.x.impute60)-1) * var(s.x.impute60) )
**Beta.hat.impute80** $\Leftarrow$ sum( (s.x.impute80 - mean(s.x.impute80)) * (s.y.impute80 - mean(s.y.impute80))) / ( (length(s.x.impute80)-1) * var(s.x.impute80) )
**Beta.hat.impute90** $\Leftarrow$ sum( (s.x.impute90 - mean(s.x.impute90)) * (s.y.impute90 - mean(s.y.impute90))) / ( (length(s.x.impute90)-1) * var(s.x.impute90) )

**y.mean.complete60[i]** $\Leftarrow$ mean(s.y.impute60) + Beta.hat.impute60 * ( mean(mu284.dat.mod[, "Rmt85"]) - mean(s.x.impute60) )
**y.mean.complete80[i]** $\Leftarrow$ mean(s.y.impute80) + Beta.hat.impute80 * ( mean(mu284.dat.mod[, "Rmt85"]) - mean(s.x.impute80) )
**y.mean.complete90[i]** $\Leftarrow$ mean(s.y.impute90) + Beta.hat.impute90 * ( mean(mu284.dat.mod[, "Rmt85"]) - mean(s.x.impute90) )

**estvech[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop)$^2$

**estvimp60[i]** $\Leftarrow$ (y.mean.complete60[i]-y.mean[i])$^2$
**estvimp80[i]** $\Leftarrow$ (y.mean.complete80[i]-y.mean[i])$^2$
**estvimp90[i]** $\Leftarrow$ (y.mean.complete90[i]-y.mean[i])$^2$

**estvmix60[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete60[i]-y.mean[i])

**estvmix80[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete80[i]-y.mean[i])
**estvmix90[i]** $\Leftarrow$ (y.mean[i]-moyenne.pop) * (y.mean.complete90[i]-y.mean[i])

**e.k.complete60** $\Leftarrow$ s.y.impute60 - mean(s.y.impute60) - Beta.hat.impute60
* ( s.x.impute60 - mean(s.x.impute60) )
**e.k.complete80** $\Leftarrow$ s.y.impute80 - mean(s.y.impute80) - Beta.hat.impute80
* ( s.x.impute80 - mean(s.x.impute80) )
**e.k.complete90** $\Leftarrow$ s.y.impute90 - mean(s.y.impute90) - Beta.hat.impute90
* ( s.x.impute90 - mean(s.x.impute90) )

**y.variance.complete60[i]** $\Leftarrow$ (1-f) / (nb.echantillon * (nb.echantillon-
1) ) * sum(s.impute60[, "g.k60"]$^2$ * e.k.complete60$^2$)
**y.variance.complete80[i]** $\Leftarrow$ (1-f) / (nb.echantillon * (nb.echantillon-
1) ) * sum(s.impute80[, "g.k80"]$^2$ * e.k.complete80$^2$)
**y.variance.complete90[i]** $\Leftarrow$ (1-f) / (nb.echantillon * (nb.echantillon-
1) ) * sum(s.impute90[, "g.k90"]$^2$ * e.k.complete90$^2$)
**variance.impute.complete60[i]** $\Leftarrow$ hatsigma60 * (1 / nb.echantillon)$^2$
* ( sum(Somme.gk.ol60$^2$ * r60.z) + sum(o60[,"g.k"]$^2$ * o60.z) )
**variance.impute.complete80[i]** $\Leftarrow$ hatsigma80 * (1 / nb.echantillon)$^2$
* ( sum(Somme.gk.ol80$^2$ * r80.z) + sum(o80[,"g.k"]$^2$ * o80.z) )

**variance.impute.complete90[i]** $\Leftarrow$ hatsigma90 * (1 / nb.echantillon)$^2$
* ( sum(Somme.gk.ol90$^2$ * r90.z) + sum(o90[,"g.k"]$^2$ * o90.z) )

if (moyenne.pop $\geq$ y.mean.complete60[i] - qnorm(1-alpha/2) * sqrt(
y.variance.complete60[i] + variance.impute.complete60[i] ) & moyenne.pop
$\leq$ y.mean.complete60[i] + qnorm(1-alpha/2) * sqrt(
y.variance.complete60[i] + variance.impute.complete60[i]))
{**TRMC.complete60** $\Leftarrow$ TRMC.complete60 + 1/nb.monte.carlo}

```
if (moyenne.pop ≥ y.mean.complete80[i] - qnorm(1-alpha/2) * sqrt(
y.variance.complete80[i] + variance.impute.complete80[i] ) & moyenne.pop
≤ y.mean.complete80[i] + qnorm(1-alpha/2) * sqrt(
y.variance.complete80[i] + variance.impute.complete80[i]))
{TRMC.complete80 ⇐ TRMC.complete80 + 1/nb.monte.carlo}

if (moyenne.pop ≥ y.mean.complete90[i] - qnorm(1-alpha/2) * sqrt(
y.variance.complete90[i] + variance.impute.complete90[i] ) & moyenne.pop
≤ y.mean.complete90[i] + qnorm(1-alpha/2) * sqrt(y.variance.complete90[i]
+ variance.impute.complete90[i]))
{TRMC.complete90 ⇐ TRMC.complete90 + 1/nb.monte.carlo}

diffnew60[i] ⇐ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * (beta60^2
* sum(o60[,"g.k"]^2 * (r60[o.impute60[,"l.k60"],"P75"] - o.impute60[,"P75"])^2))
diffnew80[i] ⇐ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * (beta80^2
* sum(o80[,"g.k"]^2 * (r80[o.impute80[,"l.k80"],"P75"] - o.impute80[,"P75"])^2))
diffnew90[i] ⇐ ( (1-f)/ (nb.echantillon * (nb.echantillon-1))) * (beta90^2
* sum(o90[,"g.k"]^2 * (r90[o.impute90[,"l.k90"],"P75"] - o.impute90[,"P75"])^2))

if (i%%100 = 0) {cat(i, ") ", date(), "\n") }
}, grain.size=15)
date()

EMC.vmix60 ⇐ mean(vmix60)
EMC.vmix80 ⇐ mean(vmix80)
EMC.vmix90 ⇐ mean(vmix90)
EMC.estvmix60 ⇐ mean(estvmix60)
EMC.estvmix80 ⇐ mean(estvmix80)
```

**EMC.estvmix90** $\Leftarrow$ mean(estvmix90)

**EMC.moyenne** $\Leftarrow$ mean(y.mean)
**VMC.moyenne.100.reponse** $\Leftarrow$ var(y.mean)
**EMCV.variance.echantillonnale.100.reponse** $\Leftarrow$ mean(y.var)
**EMC.Vech** $\Leftarrow$ mean(estvech)

**EMC.moyenne.impute60** $\Leftarrow$ mean(y.mean.complete60)
**EMC.moyenne.impute80** $\Leftarrow$ mean(y.mean.complete80)
**EMC.moyenne.impute90** $\Leftarrow$ mean(y.mean.complete90)

**VMC.moyenne.complete60** $\Leftarrow$ var(y.mean.complete60)
**VMC.moyenne.complete80** $\Leftarrow$ var(y.mean.complete80)
**VMC.moyenne.complete90** $\Leftarrow$ var(y.mean.complete90)

**EMCV.variance.echantillonnale.complete60** $\Leftarrow$ mean(y.variance.complete60)
**EMCV.variance.echantillonnale.complete80** $\Leftarrow$ mean(y.variance.complete80)
**EMCV.variance.echantillonnale.complete90** $\Leftarrow$ mean(y.variance.complete90)

**EMCV.variance.impute.complete60** $\Leftarrow$ mean(variance.impute.complete60)
**EMCV.variance.impute.complete80** $\Leftarrow$ mean(variance.impute.complete80)
**EMCV.variance.impute.complete90** $\Leftarrow$ mean(variance.impute.complete90)

**EMC.estvimp60** $\Leftarrow$ mean(estvimp60)
**EMC.estvimp80** $\Leftarrow$ mean(estvimp80)
**EMC.estvimp90** $\Leftarrow$ mean(estvimp90)

**EMCV.diffnew60** $\Leftarrow$ mean(diffnew60)
**EMCV.diffnew80** $\Leftarrow$ mean(diffnew80)

**EMCV.diffnew90** $\Leftarrow$ mean(diffnew90)

**v.chapeau.completenew60** $\Leftarrow$ EMCV.variance.impute.complete60
+ EMCV.variance.echantillonnale.complete60 - EMCV.diffnew60

**v.chapeau.completenew80** $\Leftarrow$ EMCV.variance.impute.complete80
+ EMCV.variance.echantillonnale.complete80 - EMCV.diffnew80
**v.chapeau.completenew90** $\Leftarrow$ EMCV.variance.impute.complete90
+ EMCV.variance.echantillonnale.complete90 - EMCV.diffnew90

**differencenew60** $\Leftarrow$ v.chapeau.completenew60 - VMC.moyenne.complete60
**differencenew80** $\Leftarrow$ v.chapeau.completenew80 - VMC.moyenne.complete80
**differencenew90** $\Leftarrow$ v.chapeau.completenew90 - VMC.moyenne.complete90

**Birel60** $\Leftarrow$ ( differencenew60 / VMC.moyenne.complete60 )
**Birel80** $\Leftarrow$ ( differencenew80 / VMC.moyenne.complete80 )
**Birel90** $\Leftarrow$ ( differencenew90 / VMC.moyenne.complete90 )

**estimateur.variance.population** $\Leftarrow$ (1-f)/(nb.echantillon * (N-1)) *
sum( (mu284.dat.mod[, "Rev84"] - mean(mu284.dat.mod[, "Rev84"])
- Beta.hat * ( mu284.dat.mod[, "Rmt85"] - mean(mu284.dat.mod[,
"Rmt85"] ) ))$^2$ )

<u>**Faire afficher les résultats**</u>
moyenne.pop
EMC.moyenne
EMC.moyenne.impute60
EMC.moyenne.impute80
EMC.moyenne.impute90

TRMC.100.reponse
TRMC.complete60
TRMC.complete80
TRMC.complete90

estimateur.variance.population
VMC.moyenne.100.reponse
EMCV.variance.echantillonnale.100.reponse
EMC.Vech

estimateur.variance.population
VMC.moyenne.complete60
VMC.moyenne.complete80
VMC.moyenne.complete90
EMCV.variance.echantillonnale.complete60
EMCV.variance.echantillonnale.complete80
EMCV.variance.echantillonnale.complete90

EMCV.variance.impute.complete60
EMCV.variance.impute.complete80
EMCV.variance.impute.complete90

EMC.estvimp60
EMC.estvimp80
EMC.estvimp90

EMC.vmix60
EMC.vmix80

EMC.vmix90

EMC.estvmix60
EMC.estvmix80
EMC.estvmix90

EMCV.diffnew60
EMCV.diffnew80
EMCV.diffnew90

v.chapeau.completenew60
v.chapeau.completenew80
v.chapeau.completenew90

differencenew60
differencenew80
differencenew90

# RÉFÉRENCES


Deville, J.-C. et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Gagnon, F., Lee, H., Provost, M., Rancourt et E., Särndal, C.-E. (1997). Estimation de la variance en présence d'imputation », *Recueil du Symposium 97 : Nouvelles orientations pour les enquêtes et les recensements.* À paraître, Statistique Canada, Ottawa.

Provost, M. (1995). *Estimation de la variance dans les sondages utilisant l'imputation hot-deck.* Mémoire de maîtrise, Université de Montréal.

Rancourt, E. (1996). L'estimation de variance en présence d'imputation : Concepts et méthodes. *Recueil des résumés des communications des XXVIII*$^{ième}$ *journées de statistique*, ASU, 628-634.

Rancourt, E., Lee, H. et Särndal, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 374-379.

Rancourt, E., Särndal C.-E. et Lee, H. (1994). Estimation of the variance in the presence of nearest neighbour imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 888-892.

Rao, J.N.K. (1991). Jackknife variance estimation under imputation for missing data. Rapport technique, Statistique Canada, Ottawa.

Rubin, D.B. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, 72, pp. 538-543.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*, New York: John Wiley et Sons.

Särndal, C.-E. (1990). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Recueil du Symposium '90: Mesure et amélioration de la qualité des données*, 369-380. Statistique Canada, Ottawa.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

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