Université de Montréal

Estimation robuste en population finie

par

Aliou Seydi

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

septembre 2018

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Estimation robuste en population finie

présenté par

Aliou Seydi

a été évalué par un jury composé des personnes suivantes :

Pierre Duchesne

(président-rapporteur)

David Haziza

(directeur de recherche)

Mylène Bédard

(membre du jury)

Mémoire accepté le

28 septembre 2018

SOMMAIRE

Une unité est considérée comme influente lorsque son inclusion ou son exclusion de l'échantillon a un effet important sur l'erreur due à l'échantillonnage. La présence d'unités influentes dans un échantillon rend les estimateurs classiques instables. Beaumont et al. (2013) ont montré que le biais conditionnel est un bon outil qui permet de mesurer l'influence d'une unité. Ils ont développé un estimateur robuste basé sur le biais conditionnel. Cet estimateur dépend d'une constante appelée « seuil de robustesse » déterminée de manière à minimiser le plus grand biais conditionnel estimé de l'estimateur robuste. Le but de ce travail est d'étudier d'autres critères permettant d'obtenir des estimateurs robustes ayant de bonnes propriétés en termes d'erreur quadratique moyenne.

Mots-clés : biais conditionnel ; unité influente ; estimation robuste ; seuil de robustesse.

SUMMARY

A unit is considered influential when its inclusion or exclusion from the sample has a significant effect on the sampling error. The presence of influential units in a sample makes classical estimators unstable. Beaumont et al. (2013) have shown that conditional bias is a good tool for measuring the influence of a unit. They developed a robust estimator based on conditional bias. The proposed estimator depends on a constant, called tuning constant, which is determined by minimizing the largest conditional bias of the robust estimator. The purpose of this work is to study other criteria for obtaining robust estimators with good properties in terms of mean square error.

Keywords : conditional bias ; influential unit ; robust estimation ; tuning constant.

Table des matières

Liste des tableaux

Table des figures

REMERCIEMENTS

Mes sincères remerciements vont tout d'abord au Professeur David Haziza, mon directeur de mémoire. Je le remercie pour tout ce qu'il m'a apporté comme soutien, qui est vraiment inestimable. Je lui dis grand merci pour le soutien financier, qui est arrivé à un moment vraiment critique. Merci Professeur pour ce stage à Statistique Canada, qui m'a permis d'intégrer ce milieu de travail exceptionnel. Je le remercie pour sa générosité, sa patience ainsi que sa disponibilité malgré son emploi du temps chargé.

Je remercie tous les professeurs du département pour la qualité des enseignements qu'ils produisent, contribuant ainsi au rayonnement de l'UDEM en statistique à travers le monde. Je remercie tout le personnel administratif qui fait aussi un excellent travail, facilitant ainsi la vie aux étudiants.

Mes remerciements vont ensuite à mes parents, tout d'abord à ma défunte tante Fafa Seydi et mon défunt père Elh. Daour Seydi (que la terre de Goudomp leur soit légère), pour ce qu'ils m'ont apporté comme éducation.

Je termine mes remerciements à ma femme Fatoumata Barry, mes enfants Aminata et Idrissa pour leur présence et patience pendant la période des cours de cette maîtrise.

INTRODUCTION

Les instituts de statistiques, comme Statistique Canada, effectuent plusieurs dizaines d'enquêtes par an. Parmi celles-ci, il y a l'enquête sur les dépenses des ménages canadiens à Statistique Canada. Dans une telle enquête, on est confronté au problème des unités influentes. En effet, malgré le choix d'un bon plan de sondage pour contrôler leur impact, le problème des unités influentes n'est jamais complètement éliminé. Par exemple, la stratification est un moyen de contrôler l'influence des unités, mais étant donné que seules quelques variables seront utilisées à des fins de stratification, l'influence par rapport aux autres variables n'est pas forcément prise en compte. Il y a également le problème des *stratum jumpers*, qui survient lorsque l'information auxiliaire disponible sur la base de sondage est différente de celle recueillie sur le terrain. Un *stratum jumper* est une unité qui n'appartient pas à la strate à laquelle elle aurait dû appartenir si l'information sur la base de sondage avait été correcte.

En présence d'unités influentes, l'utilisation des estimateurs classiques (Horvitz-Thompson, calage,...) pour par exemple calculer le total ou la moyenne d'un type de dépense, peut conduire à un estimateur instable. En effet, ces estimateurs sont généralement sans biais (ou asymptotiquement sans biais), mais ils sont très sensibles en présence d'unités influentes, d'où la nécessité d'utiliser des estimateurs robustes qui sont résistants aux unités influentes. En l'absence d'unités influentes, ces estimateurs produiront des résultats similaires à ceux des estimateurs non robustes.

Ce mémoire a utilisé l'article de Beaumont, Haziza et Ruiz-Gazen (2013) comme référence principale. Dans cet article, l'estimateur robuste proposé utilise le biais conditionnel comme mesure d'influence et dépend d'une certaine constante qui a été déterminée en minimisant le biais conditionnel de l'estimateur robuste. Un mauvais choix de cette constante peut grandement affecter les propriétés de l'estimateur robuste résultant, en ce sens que cet estimateur peut exhiber une erreur

quadratique moyenne plus grande que celle de l'estimateur non robuste associé.

Dans le premier chapitre, nous allons faire un petit rappel des notions de base de l'échantillonnage dont nous aurons besoin dans ce travail. Le chapitre 2 sera consacré à la notion de biais conditionnel et de son utilisation dans la construction d'un estimateur robuste. Dans le chapitre 3, nous allons chercher d'autres critères pour déterminer le seuil de robustesse et effectuer une étude par simulations. Le chapitre 4 couvrira une application de l'estimation robuste par le biais conditionnel à l'enquête de Statistique Canada sur les dépenses des ménages canadiens de l'année 2015.

Chapitre 1

PLANS DE SONDAGE

1.1. Introduction

Dans ce chapitre, nous allons faire un rappel de certains plans de sondage. La plupart de ces plans ont été rappelés pour les besoins de ce qui sera développé dans la suite.

Considérons une population finie U composée de N unités. En échantillonnage, on cherche à estimer des paramètres de population finie (total, moyenne, proportion,...). Pour ce faire, un échantillon s est tiré aléatoirement de la population U au moyen d'un plan de sondage $p(s)$.

Un plan de sondage est une fonction $p(\cdot)$ qui associe à tout échantillon s possible sa probabilité d'être sélectionné. Parce que $p(s)$ est une distribution de probabilité, elle doit satisfaire aux deux conditions suivantes:

$$\begin{aligned} &1) \quad p(s) \geq 0 \quad \forall s \subset U; \\ &2) \quad \sum_{s \subset U} p(s) = 1. \end{aligned}$$

On définit la variable indicatrice I_i selon:

$$I_i = \begin{cases} 1 & \text{si} \quad i \in s, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, tout échantillon de U est complétement caractérisé par le vecteur de taille N de variables indicatrices $\boldsymbol{I} = (I_1, ..., I_N)^\top$.

Soit $Q = \{s : p(s) > 0\}$ le support du plan de sondage $p(s)$ défini comme l'ensemble de tous les échantillons possibles s tels que $p(s) > 0$. Les probabilités

d'inclusion du premier et du second ordre sont définies respectivement par:

$$\pi_i = Pr(s \ni i) = Pr(I_i = 1) = \sum_{\substack{s \in Q \\ s \ni i}} p(s)$$

et

$$\pi_{ij} = Pr(s \ni i, s \ni j) = Pr(I_i = 1, I_j = 1) = \sum_{\substack{s \in Q \\ s \ni i,j}} p(s), \quad i \neq j.$$

Le poids de sondage associé à l'unité i est défini par $d_i = \pi_i^{-1}$, et peut être interprété comme le nombre d'unités de la population que l'unité i représente dans l'échantillon.

Nous énonçons les résultats suivants:

Proposition 1.1.1. *Pour un plan de sondage donné $p(s)$, nous avons:*

(1) $E_p(I_i) = \pi_i$,

(2) $E_p(I_i I_j) = \pi_{ij}$,

(3) $V_p(I_i) = \pi_i(1 - \pi_i)$,

(4) $Cov_p(I_i, I_j) = \pi_{ij} - \pi_i \pi_j$, *avec* $\pi_{ii} = \pi_i$,

où V_p et Cov_p représentent respectivement la variance et la covariance sous le plan de sondage p.

La démonstration de la proposition 1.1.1 est donnée dans Särndal et al. (1992).

On distingue les plans de sondage à taille fixe des plans de sondage à taille aléatoire. Pour un plan à taille fixe, on note la taille de l'échantillon par n et par n_s pour un plan à taille aléatoire.

1.1.1. Echantillonnage aléatoire simple sans remise

L'échantillonnage aléatoire simple sans remise est un plan de sondage à taille fixe qui attribue la même probabilité à tous les échantillons de même taille n. Comme il existe $\binom{N}{n}$ échantillons possibles de taille n , on a que $p(s) = \binom{N}{n}^{-1}$ pour tout $s \in Q$ tel que $card(s) = n$. Les probabilités d'inclusion du premier et du second ordre sont données par

$$\pi_i = \frac{n}{N}, \qquad \text{pour tout } i \in U \tag{1.1.1}$$

et

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}, \qquad \text{pour tout } i \neq j \in U. \tag{1.1.2}$$

1.1.2. Echantillonnage stratifié

La population U est stratifiée en H strates $U_1, ..., U_H$, de tailles $N_1, ..., N_H$, respectivement. On a $\bigcup_{h=1}^{H} U_h = U$, $\sum_{h=1}^{H} N_h = N$ et $U_h \cap U_l = \emptyset$ si $h \neq l$. De chaque strate h, on tire un échantillon aléatoire s_h de taille n_h selon un plan de sondage $p_h(\cdot)$. La sélection dans une strate est indépendante de la sélection dans n'importe quelle autre strate. L'échantillon résultant s est donné par $s = \bigcup_{h=1}^{H} s_h$ de taille $n = \sum_{h=1}^{H} n_h$.

Un cas particulier d'échantillonnage stratifié est l'échantillonnage stratifié aléatoire simple sans remise, pour lequel s_h est un échantillon aléatoire simple sans remise tiré de U_h. On a:

$$p(s) = \prod_{h=1}^{H} \frac{1}{\binom{N_h}{n_h}},$$

$$\pi_i = \frac{n_h}{N_h} \quad \text{si} \quad i \in U_h,$$

et

$$\pi_{ij} = \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)} & \text{si} \quad i \in U_h \quad \text{et} \quad j \in U_h, \\ \frac{n_h}{N_h}\frac{n_l}{N_l} & \text{si} \quad i \in U_h \quad \text{et} \quad j \in U_l \quad h \neq l. \end{cases}$$

1.1.3. Echantillonnage de Poisson

L'échantillonnage de Poisson consiste à effectuer, pour chaque unité i de la population, une expérience de Bernoulli avec probabilité de succès π_i, où π_i est définie à l'avance. Si le tirage est un succès, alors l'unité i est sélectionnée dans l'échantillon s, sinon elle est rejetée. La probabilité de sélection de l'échantillon s est donc donnée par

$$p(s) = \prod_{i \in s} \pi_i \prod_{i \in U \setminus s} (1 - \pi_i).$$

Le plan de Poisson est un plan à taille aléatoire. Les expériences de Bernoulli étant mutuellement indépendantes, la probabilité d'inclusion d'ordre deux des unités i et j est $\pi_{ij} = \pi_i \pi_j$ si $i \neq j$. Lorsque $\pi_i = \pi$ pour tout i, alors on obtient le plan de Bernoulli.

1.1.4. Echantillonnage par grappes à deux degrés

Supposons que la population U est constituée de N grappes $U_1, ..., U_N$ de tailles $M_1, ..., M_N$, respectivement. On a $U = \bigcup_{g=1}^{N} U_g$ et $U_g \bigcap U_l = \emptyset$ si $g \neq l$. Au premier degré un échantillon s de grappes est tiré selon un plan $p(s)$. Les probabilités d'inclusion d'ordre 1 et 2 sont définies par $\pi_g = Pr(s \ni g)$ et $\pi_{gl} = Pr(s \ni g, s \ni l)$, respectivement.

Pour toute grappe $g \in s$, on tire un échantillon s_g d'éléments selon le plan $p_g(\cdot \mid s)$. Les probabilités d'inclusion d'ordre 1 et 2 sont notées respectivement par $\pi_{i|g} = Pr(s_g \ni i \mid s \ni g)$ et $\pi_{ij|g} = Pr(s_g \ni i, s_g \ni j \mid s \ni g)$.

1.1.5. Echantillonnage à deux phases

Dans le cas d'une base sondage qui est faible en information auxiliaire, il est de coutume d'utiliser un plan à deux phases qui consiste dans un premier temps à recueillir une telle information avec un plan de sondage habituellement rudimentaire (par exemple, un plan aléatoire simple sans remise). De cet échantillon, on utilise un plan plus efficace pour aller chercher de l'information pour la ou les variables d'intérêt. À la première phase, on tire un échantillon s_1 de taille n_1 selon un plan de sondage $p_1(s_1)$. Le vecteur d'indicateur de sélection est donné par $\boldsymbol{I_1} = (I_{11}, ..., I_{1N})^\top$. De l'échantillon s_1 on tire l'échantillon s_2 de la phase deux de taille n_2 à l'aide du plan de sondage $p_2(s_2|s_1)$ et le vecteur d'indicateur de sélection pour cette phase est donné par $\boldsymbol{I_2} = (I_{21}, ..., I_{2N})^\top$.

Soit $\pi_{1i} = Pr(I_{1i} = 1)$, $\pi_{1ij} = Pr(I_{1i} = 1, I_{1j} = 1)$ les probabilités d'inclusion du premier et du second ordre de la première phase et $\pi_{2i}(\boldsymbol{I_1}) = Pr(I_{2i} = 1 \mid \boldsymbol{I_1})$, $\pi_{2ij}(\boldsymbol{I_1}) = Pr(I_{2i} = 1, I_{2j} = 1 \mid \boldsymbol{I_1})$ les probabilités d'inclusion du premier et du second ordre à la deuxième phase. Notons que ces deux dernières probabilités d'inclusion peuvent dépendre de la réalisation de l'échantillon obtenu à la première phase.

Lorsque $Pr(\boldsymbol{I_2} \mid \boldsymbol{I_1}) = Pr(\boldsymbol{I_2})$, alors on dit que le plan à deux phases est fortement invariant. Dans un tel cas, le vecteur $\boldsymbol{I_2}$ peut être obtenu avant le vecteur $\boldsymbol{I_1}$; voir Beaumont et Haziza (2016). Un plan à deux phases est dit faiblement invariant lorsque $\pi_{2i}(\boldsymbol{I_1}) = \pi_{2i}$ et $\pi_{2ij}(\boldsymbol{I_1}) = \pi_{2ij}$ pour tout $i, j \in U$.

1.2. Estimation

Dans cette section, nous présentons l'estimateur Horvitz-Thompson et l'estimateur de calage ainsi que quelques-unes de leurs propriétés respectives.

1.2.1. Estimateur de Horvitz-Thompson

Pour une variable d'intérêt y, on note par t_y son total dans la population, i.e. $t_y = \sum_{i \in U} y_i$. L'estimateur Horvitz-Thompson du total t_y est donné par:

$$\hat{t}_y^{HT} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s} d_i y_i. \tag{1.2.1}$$

Cet estimateur est sans biais pour t_y, c'est-à-dire $E_p\left(\hat{t}_y^{HT}\right) = t_y$.

Proposition 1.2.1. *La variance de l'estimateur* (1.2.1)*, est donnée par*

$$V_p\left(\hat{t}_y^{HT}\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{1.2.2}$$

Démonstration. De la proposition 1.1.1, nous obtenons

$$\begin{aligned} V_p\left(\hat{t}_y^{HT}\right) &= \sum_{i \in U} \frac{y_i^2}{\pi_i^2} V_p\left(I_i\right) + \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} Cov_p\left(I_i, I_j\right) \\ &= \sum_{i \in U} \frac{y_i^2}{\pi_i^2} \pi_i \left(1 - \pi_i\right) + \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \left(\pi_{ij} - \pi_i \pi_j\right) \\ &= \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \end{aligned}$$

□

1.2.2. Estimateurs de calage

Les estimateurs de calage incorporent une certaine quantité d'information auxiliaire à l'étape de l'estimation. On appellera variable auxiliaire toute variable observée pour les unités dans l'échantillon et dont le total dans la population est connu.

Notons par $\boldsymbol{x}_i = (x_{1i}, ..., x_{Ki})^\top$ les valeurs prises par l'unité $i \in s$ sur les K variables auxiliaires et soit $\boldsymbol{t}_x = \sum_{i \in U} \boldsymbol{x}_i$ le vecteur des totaux dans la population.

Un estimateur de calage de t_y est de la forme:

$$\hat{t}_y^{Cal} = \sum_{i \in s} w_i y_i, \tag{1.2.3}$$

où w_i désigne le poids de calage pour l'unité i. Les poids $w_i, i = 1, ..., n$ doivent être aussi proches que possible des poids $d_i = \frac{1}{\pi_i}$ tout en respectant les contraintes de calage

$$\hat{t}_x^{Cal} = \sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{t}_x. \tag{1.2.4}$$

Plus formellement, on cherche un système de pondération $\{w_i, i \in s\}$ tel que

$$\sum_{i \in s} \frac{d_i G(w_i/d_i)}{q_i}$$

soit minimum sous la contrainte (1.2.4), où $G(\cdot)$ désigne une fonction de distance qui permet de mesurer la proximité entre les poids de calage w_i et les poids de sondage d_i, et q_i est un coefficient de pondération qui marque l'importance de l'unité i dans le calcul de la distance; voir Deville et Särndal (1992).

La fonction de distance $G(\cdot)$ doit satisfaire aux trois conditions suivantes:

(i) $G(u) \geq 0$ et $G(1) = 0$,

(ii) $G(u)$ est dérivable par rapport à u et sa dérivée est continue,

(iii) $G(u)$ est strictement convexe.

La technique des multiplicateurs de Lagrange nous permet de résoudre ce problème en minimisant la fonction suivante par rapport à w_i:

$$\phi\left(w_1, ..., w_n, \boldsymbol{\lambda}\right) = \sum_{i \in s} \frac{d_i G(w_i/d_i)}{q_i} - \boldsymbol{\lambda}^\top \left(\sum_{i \in s} w_i \boldsymbol{x}_i - \boldsymbol{t}_x \right), \tag{1.2.5}$$

où $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)^\top$ est un vecteur de multiplicateurs de Lagrange.

En dérivant ϕ par rapport à w_i et en posant la dérivée égale à zéro, on obtient la solution

$$w_i = d_i F\left(q_i \boldsymbol{\lambda}^\top \boldsymbol{x}_i\right), \tag{1.2.6}$$

où $F(\cdot)$ représente l'inverse de la fonction dérivée $G'(\cdot)$. La fonction $F(\cdot)$ est appelée fonction de calage.

Le vecteur $\boldsymbol{\lambda}$ dans (1.2.6) étant inconnu, on substitue w_i de (1.2.6) dans (1.2.4), ce qui conduit à

$$\sum_{i\in s} d_i F\left(q_i \boldsymbol{\lambda}^\top \boldsymbol{x}_i\right) \boldsymbol{x}_i = \boldsymbol{t}_x. \tag{1.2.7}$$

Le système de K équations à K inconnues (1.2.7) peut être résolu en utilisant l'algorithme de Newton-Raphson. Soit $\hat{\boldsymbol{\lambda}}$ la solution de (1.2.7). Le poids de calage w_i est donné par $w_i = d_i F\left(q_i \hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_i\right)$. On remarque que les poids de calage w_i s'écrivent comme le produit des poids de sondage d_i et d'un facteur d'ajustement $F\left(q_i \hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_i\right)$.

Il existe plusieurs fonctions de distance $G(.)$, parmi lesquelles la distance du khi-deux généralisée définie par:

$$G(w_i/d_i) = \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2. \tag{1.2.8}$$

La fonction de calage correspondante est $F\left(q_i \hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_i\right) = 1 + q_i \hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_i$. Le poids de calage w_i dans (1.2.6) devient:

$$w_i = d_i\left(1 + q_i \hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_i\right), \tag{1.2.9}$$

où

$$\hat{\boldsymbol{\lambda}} = \left(\sum_{i\in s} d_i \boldsymbol{x}_i q_i \boldsymbol{x}_i^\top\right)^{-1}\left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^{HT}\right). \tag{1.2.10}$$

On suppose que la matrice $\sum_{i\in s} d_i \boldsymbol{x}_i q_i \boldsymbol{x}_i^\top$ est de rang plein. En insérant (1.2.10) dans (1.2.9), on obtient:

$$\begin{aligned} w_i &= d_i\left(1 + q_i \hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_i\right) \\ &= d_i\left\{1 + q_i\left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^{HT}\right)^\top\left(\sum_{i\in s} d_i q_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \boldsymbol{x}_i\right\}. \end{aligned}$$

Ainsi l'estimateur de calage (1.2.3) peut s'écrire comme

$$\hat{t}_y^{Cal} = \hat{t}_y^{HT} + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^{HT}\right)^\top \widehat{\boldsymbol{B}}, \tag{1.2.11}$$

où

$$\widehat{\boldsymbol{B}} = \left(\sum_{i\in s} d_i \boldsymbol{x}_i q_i \boldsymbol{x}_i^\top\right)^{-1}\left(\sum_{i\in s} d_i \boldsymbol{x}_i q_i y_i\right).$$

L'estimateur (1.2.11) n'est rien d'autre que l'estimateur par la régression généralisée (GREG); voir Särndal et al. (1992).

1.2.2.1. *Propriétés des estimateurs de calage*

Par une série de Taylor du premier ordre, Deville et Särndal (1992) ont montré que:

$$\hat{t}_y^{Cal} - t_y = \left(\sum_{i\in s} d_i E_i - \sum_{i\in U} E_i\right) + O_p\left(\frac{N}{n}\right), \tag{1.2.12}$$

où $E_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{B}$, et

$$\boldsymbol{B} = \left(\sum_{i\in U} \boldsymbol{x}_i q_i \boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i\in U} \boldsymbol{x}_i q_i y_i\right).$$

En ignorant les termes d'ordre supérieur dans (1.2.12), on obtient une expression de la variance approchée de $\hat{t}_y^{Cal}$:

$$V_p\left(\hat{t}_y^{cal}\right) \approx V_p\left(\sum_{i\in s} d_i E_i\right) \tag{1.2.13}$$

$$= \sum_{i\in U}\sum_{j\in U} \left(\pi_{ij} - \pi_i\pi_j\right) \frac{E_i}{\pi_i}\frac{E_j}{\pi_j}. \tag{1.2.14}$$

Remarque: Nous avons $\sum\limits_{i\in U} E_i = 0$ si $q_i^{-1} = \boldsymbol{\lambda}^\top \boldsymbol{x}_i$. En effet,

$$\begin{aligned}
\sum_{i\in U} E_i &= \sum_{i\in U} y_i - \sum_{i\in U} \boldsymbol{x}_i^\top \boldsymbol{B} \\
&= \sum_{i\in U} y_i - \sum_{i\in U} \frac{\boldsymbol{\lambda}^\top \boldsymbol{x}_i}{\boldsymbol{\lambda}^\top \boldsymbol{x}_i} \boldsymbol{x}_i^\top \boldsymbol{B} \\
&= \sum_{i\in U} y_i - \boldsymbol{\lambda}^\top \left(\sum_{i\in U} \boldsymbol{x}_i q_i \boldsymbol{x}_i^\top\right) \left(\sum_{i\in U} \boldsymbol{x}_i q_i \boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i\in U} \boldsymbol{x}_i q_i y_i\right) \\
&= \sum_{i\in U} y_i - \boldsymbol{\lambda}^\top \left(\sum_{i\in U} \boldsymbol{x}_i q_i y_i\right) \\
&= \sum_{i\in U} y_i - \sum_{i\in U} \left(\boldsymbol{\lambda}^\top \boldsymbol{x}_i\right) q_i y_i \\
&= \sum_{i\in U} y_i - \sum_{i\in U} y_i = 0.
\end{aligned} \tag{1.2.15}$$

Chapitre 2

UNE MESURE D'INFLUENCE: LE BIAIS CONDITIONNEL

2.1. Biais conditionnel

2.1.1. Unité influente et configuration

Soit θ un paramètre de la population finie et $\hat{\theta}$ un estimateur de θ. Une unité peut être définie comme influente si son exclusion de la population (et de l'échantillon le cas échéant) a un effet important sur l'erreur d'échantillonnage $(\hat{\theta} - \theta)$; voir Beaumont et Haziza (2016).

Une unité influente est bien un élément de la population, elle n'est pas une erreur de mesure ou de codage, qui est habituellement détectée à l'étape de la vérification et traitée soit manuellement soit par imputation. Les méthodes décrites dans ce document se placent donc à l'étape de l'estimation de l'enquête, soit à la suite des vérifications et de l'imputation des données.

Les estimateurs usuels sont très sensibles à la présence d'unités influentes dans l'échantillon. Autrement dit, en présence d'unités influentes, ces estimateurs tendent à être très instables.

L'objectif de l'estimation robuste est de réduire l'influence des unités dans l'échantillon qui ont un grand impact sur l'erreur d'échantillonnage, ce qui conduit à des estimateurs biaisés mais plus stables. On cherche à obtenir des estimateurs dont l'erreur quadratique moyenne est inférieure à celle des estimateurs classiques.

L'influence d'une unité i dépend de la configuration qui est un ensemble de caractéristiques. Une configuration est un quadruplet, qui consiste en:

(1) une variable d'intérêt y et sa distribution dans la population,

(2) un paramètre θ de la population que l'on cherche à estimer,

(3) un plan de sondage et l'estimateur associé $\hat{\theta}$,

(4) la présence ou non de l'unité i dans l'échantillon.

Une unité peut être très influente selon une certaine configuration et n'avoir aucune influence par rapport à une autre configuration. De ce fait, la modification d'un de ces quadruplets peut avoir un impact très important sur l'influence d'une unité.

2.1.2. Définition du biais conditionnel

Le biais conditionnel de l'unité i par rapport à un estimateur $\hat{\theta}$ du paramètre θ selon un plan de sondage p est défini selon

$$B_i = E_p\left(\hat{\theta}\middle| I_i\right) - \theta. \tag{2.1.1}$$

Comme une unité de la population peut être échantillonnée ou non, on note par B_{1i} le biais conditionnel d'une unité i échantillonnée et par B_{0i} celui d'une unité i non échantillonnée. Nous avons:

$$B_{1i} = E_p\left(\hat{\theta}\middle| I_i = 1\right) - \theta. \tag{2.1.2}$$

Le biais conditionnel d'une unité échantillonnée i peut être vu comme étant la moyenne de l'erreur due à l'échantillonnage prise sur tous les échantillons qui contiennent l'unité i. Autrement dit, on parcourt tous les échantillons qui contiennent i et on prend la moyenne des estimations provenant de ces échantillons. Si cette moyenne s'éloigne beaucoup du vrai paramètre θ, on dira que l'unité i est influente car elle aura un grand biais conditionnel. En pratique, nous n'avons accès qu'à un seul échantillon. Par conséquent, le biais conditionnel est inconnu et devra être estimé.

Pour une unité i non échantillonnée, le biais conditionnel est défini par:

$$B_{0i} = E_p(\hat{\theta}\middle| I_i = 0) - \theta. \tag{2.1.3}$$

Bien qu'une unité non échantillonnée peut avoir de l'influence sur un estimateur, rien ne peut être fait à l'étape de l'estimation afin de réduire son influence car aucune information n'est observée pour une telle unité. On se résignera donc à réduire l'influence des unités échantillonnée sur l'estimateur.

2.2. Biais conditionnel comme mesure d'influence

Nous avons défini une unité influente comme une unité ayant un grand impact sur l'erreur d'échantillonnage. En pratique, comment mesurer cette influence? On peut se demander s'il suffit d'identifier les unités à grand poids de sondage, d_i, ou les unités ayant une grande valeur de la variable d'intérêt, y_i , ou peut-être même une combinaison des deux (c'est-à-dire les unités exhibant une grande contribution $d_i y_i$). En fait, aucun de ces critères n'est complètement satisfaisant, bien qu'ils soient souvent utilisés en pratique. En effet, ces critères ne tiennent compte que de certains aspects de la configuration. Le biais conditionnel a été suggéré comme méthode alternative puisqu'il tiendra compte de tous les éléments de la configuration simultanément. En guise d'illustration, nous donnerons certains exemples en développant la formule du biais conditionnel pour différents plans de sondage et différents estimateurs.

Exemple 2.2.1. *Pour une variable d'intérêt y et un échantillon s tiré d'une population U avec un plan de sondage p, le biais conditionnel de l'unité échantillonnée i par rapport à l'estimateur Horvitz-Thompson* (1.2.1) *du total est donné par:*

$$B_{1i}^{HT} = \sum_{j \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_j. \tag{2.2.1}$$

Pour l'estimateur par calage (1.2.3)*, on utilisera l'expression* (1.2.12) *afin d'approximer le biais conditionnel d'une unité échantillonnée:*

$$B_{1i}^{Cal} \approx \sum_{j \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} E_j, \tag{2.2.2}$$

où $E_j = y_j - \boldsymbol{x}_j^\top \boldsymbol{B}$*. Pour une unité i non échantillonnée, son biais conditionnel par rapport à l'estimateur de Horvitz-Thompson s'obtient comme suit:*

$$B_{0i}^{HT} = -\frac{\pi_i}{1 - \pi_i} B_{1i}^{HT}. \tag{2.2.3}$$

En effet selon les propriétés de l'espérance conditionnelle, nous avons $E_p\left(\hat{t}_y\right) = E_p\left\{E_p\left(\hat{t}_y \middle| I_i\right)\right\}$*. On établit* (2.2.3) *en notant d'une part que:*

$$\begin{aligned} E_p\left(\hat{t}_y^{HT}\right) &= E_p\left\{E_p\left(\hat{t}_y^{HT} \middle| I_i\right)\right\} \\ &= Pr\left(I_i = 1\right) E_p\left(\hat{t}_y^{HT} \middle| I_i = 1\right) + Pr\left(I_i = 0\right) E_p\left(\hat{t}_y^{HT} \middle| I_i = 0\right) \\ &= \pi_i E_p\left(\hat{t}_y^{HT} \middle| I_i = 1\right) + \left(1 - \pi_i\right) E_p\left(\hat{t}_y^{HT} \middle| I_i = 0\right) \end{aligned} \tag{2.2.4}$$

et d'autre part:

$$\begin{aligned} E_p\left(\hat{t}_y^{HT}\right) &= \pi_i E_p\left(\hat{t}_y^{HT}\right) + (1-\pi_i)\, E_p\left(\hat{t}_y^{HT}\right) \\ &= \pi_i t_y + (1-\pi_i)\, t_y. \end{aligned} \tag{2.2.5}$$

En égalisant (2.2.4) *et* (2.2.5) *puis en isolant les deux biais conditionnels, on obtient* (2.2.3).

2.2.0.1. *Lien entre le biais conditionnel et l'erreur due à l'échantillonnage*

Dans cette section, nous allons établir le lien entre le biais conditionnel par rapport à l'estimateur de Horvitz-Thompson et l'erreur due à l'échantillonnage. Cette relation permettra de mettre en évidence que le biais conditionnel est une bonne mesure de l'influence.

Proposition 2.2.2. *Selon Beaumont et al. (2013), nous avons le résultat suivant:*

$$\hat{t}_y^{HT} - t_y = \sum_{i\in s} B_{1i}^{HT} + \sum_{i\in U\setminus s} B_{0i}^{HT} + \left(\sum_{i\in s} \pi_i^{-1} A_i - \sum_{i\in U} A_i\right), \tag{2.2.6}$$

où

$$A_i = -(1-\pi_i)^{-1} \sum_{\substack{j\in U \\ j\neq i}} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_j}\right) y_j.$$

Si $\left(\sum_{i\in s}\pi_i^{-1}A_i - \sum_{i\in U} A_i\right)$ est petit par rapport $\left(\sum_{i\in s} B_{1i}^{HT} + \sum_{i\in U\setminus s} B_{0i}^{HT}\right)$ alors de (2.2.6) on peut écrire:

$$\hat{t}_y^{HT} - t_y \approx \sum_{i\in s} B_{1i}^{HT} + \sum_{i\in U\setminus s} B_{0i}^{HT}. \tag{2.2.7}$$

Ainsi l'erreur due à l'échantillonnage apparaît approximativement comme la somme des biais conditionnels des unités de la population. Pour une unité i donnée, son biais conditionnel par rapport à l'estimateur de Horvitz-Thompson peut donc être interprété comme sa contribution à l'erreur $\hat{t}_y^{HT} - t_y$.
La preuve de la proposition 2.2.7 est donnée en annexe.

Exemple 2.2.3. *Étudions* (2.2.6) *pour deux plans de sondage:*

(1) Pour le plan de Poisson, on a $A_i = 0$ *(puisque* $\pi_{ij} = \pi_i \pi_j$ *pour tout* $j \neq i$*) ce qui implique:*

$$\sum_{i \in s} \pi_i^{-1} A_i - \sum_{i \in U} A_i = 0.$$

On a donc:

$$\hat{t}_y^{HT} - t_y = \sum_{i \in s} B_{1i}^{HT} + \sum_{i \in U \setminus s} B_{0i}^{HT}.$$

(2) Pour le plan aléatoire simple sans remise, on a

$$\sum_{i \in s} \pi_i^{-1} A_i - \sum_{i \in U} A_i = \frac{-1}{N-1} \left(\hat{t}_y^{HT} - t_y \right)$$

et (2.2.6) *peut s'écrire comme:*

$$\frac{N}{N-1} \left(\hat{t}_y^{HT} - t_y \right) = \sum_{i \in s} B_{1i}^{HT} + \sum_{i \in U \setminus s} B_{0i}^{HT}.$$

Si la taille de la population N *est grande, alors* $\frac{N}{N-1} \approx 1$ *et on a* (2.2.7).

En effet, on a

$$\begin{aligned} A_i &= -(1-\pi_i)^{-1} \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j} \right) y_j \\ &= -\frac{N}{N-n} \left(\frac{n-1}{N-1} - \frac{n}{N} \right) \left(\sum_{\substack{j \in U \\ j \neq i}} y_j + y_i - y_i \right) \\ &= \frac{1}{N-1} \left(\sum_{j \in U} y_j - y_i \right) = \frac{-1}{N-1} \left(y_i - t_y \right). \end{aligned}$$

Il s'ensuit que

$$\begin{aligned} \sum_{i \in s} \pi_i^{-1} A_i &= \frac{-1}{N-1} \sum_{i \in s} \frac{N}{n} \left(y_i - t_y \right) \\ &= \frac{-1}{N-1} \left(\hat{t}_y^{HT} - N t_y \right), \end{aligned}$$

et

$$\sum_{i \in U} A_i = \frac{-1}{N-1} \sum_{i \in U} \left(y_i - t_y \right)$$

$$= \frac{-1}{N-1}\left(t_y - Nt_y\right) = t_y,$$

d'où

$$\sum_{i\in s}\pi_i^{-1}A_i - \sum_{i\in U}A_i = \frac{-1}{N-1}\left(\hat{t}_y^{HT} - Nt_y\right) - t_y$$
$$= \frac{-1}{N-1}\left(\hat{t}_y^{HT} - t_y\right).$$

2.2.0.2. *Lien entre le biais conditionnel et la variance de l'estimateur Horvitz-Thompson*

La variance de l'estimateur Horvitz-Thompson est donnée par:

$$\begin{aligned} V_p\left(\hat{t}_y^{HT}\right) &= E_p\left(\hat{t}_y^{HT} - t_y\right)^2 \\ &= \sum_{i\in U}\sum_{j\in U}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \\ &= \sum_{i\in U} y_i \sum_{j\in U}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} y_j \\ &= \sum_{i\in U} y_i B_{1i}^{HT}, \end{aligned} \tag{2.2.8}$$

voir Beaumont et al. (2013).

Il est intéressant de constater qu'il y a un lien direct entre le concept du biais conditionnel (i.e. l'influence) et la variance. Une unité échantillonnée avec un grand biais conditionnel contribuera donc à rendre l'estimateur instable.

2.2.1. Biais conditionnel pour certains plans

Dans l'exemple ci-dessus, dans le cas de l'estimateur de Horvitz-Thompson, on a obtenu la formule (2.2.1) pour le biais conditionnel de l'unité i. Dans cette section, nous appliquerons cette formule à différents plans de sondage. Pour ces plans, nous remplacerons les valeurs de π_i et π_{ij} par les valeurs correspondantes du plan.

2.2.1.1. *Biais conditionnel pour l'échantillonnage aléatoire simple sans remise*

Dans le cas de l'échantillonnage aléatoire simple sans remise, (2.2.1) se simplifie pour donner:

$$B_{1i}^{HT} = \frac{N}{N-1}\left(\frac{N}{n}-1\right)\left(y_i - \overline{Y}\right), \tag{2.2.9}$$

où $\overline{Y} = t_y/N$ est la moyenne de la population.

Dans le cas stratifié aléatoire simple sans remise on obtient:

$$B_{1i}^{HT} = \frac{N_h}{N_h-1}\left(\frac{N_h}{n_h}-1\right)\left(y_i - \overline{Y}_h\right), \quad i \in U_h, \tag{2.2.10}$$

où $\overline{Y}_h$ désigne la moyenne dans la strate U_h.

L'expression (2.2.9) suggère qu'une unité i aura une grande influence si sa valeur sur la variable d'intérêt est loin de la moyenne de la population, $\overline{Y}$. Pour le plan stratifié, (2.2.10), l'unité i sera influente lorsque sa valeur sur la variable d'intérêt est loin de la moyenne de la strate $\overline{Y}_h$ à laquelle elle appartient.

Pour l'estimateur de calage (1.2.3), on obtient

$$B_{1i}^{Cal} \approx \frac{N}{N-1}\left(\frac{N}{n}-1\right)\left(E_i - \overline{E}\right). \tag{2.2.11}$$

Ici, une unité i aura une grande influence si son résidu est loin de la moyenne des résidus $\overline{E}$ au niveau de la population. Lorsque $q_i^{-1} = \boldsymbol{\lambda}^\top \boldsymbol{x_i}$, pour un certain vecteur de constantes $\boldsymbol{\lambda}$, on a $\overline{E} = 0$, voir section 1.2.2.1.

2.2.1.2. *Biais conditionnel pour le plan de Poisson*

Pour le plan de Poisson, le biais conditionnel se réduit à

$$B_{1i}^{HT} = (d_i - 1)y_i, \tag{2.2.12}$$

en notant que $\pi_{ij} = \pi_i\pi_j$, $i \neq j$. Ici, une unité sera influente si elle a un grand poids de sondage et/ou sa valeur sur la variable d'intérêt est grande. Dans ce cas, le biais conditionnel ne dépend que de l'unité échantillonnée et n'a donc pas besoin d'être estimé. Cet exemple démontre que le biais conditionnel tient compte du plan de sondage et qu'une unité influente pour un plan donné pourrait très bien ne pas l'être pour un autre plan. Plus précisément, on voit qu'une unité i ayant une valeur nulle, $y_i = 0$, aura une influence nulle sous un plan Poisson alors que sous un plan aléatoire simple sans remise, ce ne sera pas nécessairement le cas puisque le biais conditionnel, dans ce cas, est plutôt fonction de la valeur $y_i - \overline{Y}$.

De même, on peut constater l'impact du choix de l'estimateur sur le biais conditionnel. En effet, l'utilisation de l'estimateur de calage pour le plan de Poisson donne

$$B_{1i}^{Cal} \approx (d_i - 1)E_i.$$

2.2.1.3. *Biais conditionnel pour le plan par grappes à deux degrés*

Dans un plan à deux degrés, on estime le total $t_y = \sum\limits_{g\in U} \sum_{i\in U_g} y_{gi}$ par:

$$\hat{t}_{2y} = \sum_{g\in s} \sum_{j\in s_g} \pi_g^{-1}\pi_{j|g}^{-1} y_{gj}; \tag{2.2.13}$$

voir section 1.1.4.

Proposition 2.2.4. *Supposons que l'élément i dans l'échantillon provienne de la grappe g. Son biais conditionnel par rapport au plan à deux degrés est donné par:*

$$B_{1i|g}^{HT} = \sum_{l\in U} \left(\frac{\pi_{gl}}{\pi_g\pi_l} - 1\right) \sum_{j\in U_l} y_{lj} + \frac{1}{\pi_g} \sum_{j\in U_g} \left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right) y_{gj}, \quad i \in U_g. \tag{2.2.14}$$

La preuve de la proposition 2.2.4 est donnée en annexe. Remarquons que le premier terme dans le membre de droite de (2.2.14) est similaire à (2.2.1) avec $\sum\limits_{j\in U_l} y_{lj}$ plutôt que y_j.

2.2.1.4. *Biais conditionnel pour le plan à deux phases*

Dans le cas du plan à deux phases, un estimateur sans biais du total t_y est l'estimateur par double dilatation défini comme

$$\hat{t}_y^{DE} = \sum_{j\in s_2} \pi_{1j}^{-1}\pi_{2j}(\boldsymbol{I}_1)^{-1} y_j; \tag{2.2.15}$$

voir section 1.1.5.

En notant par $E_1(\cdot)$ l'espérance mathématique à la première phase et $E_2(\cdot)$ celle à la deuxième phase, nous avons le résultat suivant dont la démonstration est donnée en annexe.

Proposition 2.2.5. *Soit i une unité échantillonnée à la deuxième phase, son biais conditionnel par rapport à l'estimateur* (2.2.15) *est donné par*

$$
\begin{aligned}
B_{1i}^{DE} &= E_p\left(\hat{t}_y^{DE} - t_y \middle| \boldsymbol{I}_1, I_{1i}=1, I_{2i}=1\right) \\
&= E_1\left\{E_2\left(\hat{t}_y^{DE} - t_y \middle| \boldsymbol{I}_1, I_{1i}=1, I_{2i}=1\right)\right\} \\
&= B_{1i}^{1HT} + E_1\left[\pi_{1i}^{-1}\left\{\pi_{2i}\left(\boldsymbol{I}_1\right)^{-1} - 1\right\} y_i \middle| I_{2i}=1\right] \\
&+ E_1\left[\sum_{\substack{j\in s_1 \\ j\neq i}} \frac{1}{\pi_{1j}}\left\{\frac{\pi_{2ij}\left(\boldsymbol{I}_1\right)}{\pi_{2i}\left(\boldsymbol{I}_1\right)\pi_{2j}\left(\boldsymbol{I}_1\right)} - 1\right\} y_j \middle| I_{2i}=1\right],
\end{aligned}
\tag{2.2.16}
$$

où $B_{1i}^{1HT} = E_1\left\{\left(\hat{t}_{1y}^{HT} - t_y\right)\middle| I_{1i}=1\right\}$ *est le biais conditionnel de l'unité i par rapport au plan à une phase; voir Favre Martinoz et al. (2016).*

En supposant l'invariance faible, c'est-à-dire $\pi_{2i}(\boldsymbol{I}_1) = \pi_{2i}$ et $\pi_{2ij}(\boldsymbol{I}_1) = \pi_{2ij}$ pour tout $i \neq j$, nous obtenons

$$
B_{1i}^{DE} = \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1\right) y_j + \sum_{j\in U}\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\left(\frac{\pi_{2ij}}{\pi_{2i}\pi_{2j}} - 1\right) y_j. \tag{2.2.17}
$$

2.2.2. Estimation du biais conditionnel

En général, le biais conditionnel dans (2.2.1) est inconnu car on n'observe les valeurs de la variable d'intérêt que pour les unités dans l'échantillon.

Si $\pi_{ij} > 0$, pour tout $i, j \in U$ alors un estimateur conditionnellement sans biais est donné par

$$
\hat{B}_{1i}^{HT} = \sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}} y_j, \tag{2.2.18}
$$

voir Beaumont et al. (2013). On a

$$
E_p\left(\hat{B}_{1i}^{HT} | I_i = 1\right) = B_{1i}^{HT}.
$$

En effet,

$$
\begin{aligned}
E_p\left(\hat{B}_{1i}^{HT} | I_i = 1\right) &= E_p\left(\sum_{j\in U}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}} y_j I_j \middle| I_i = 1\right) \\
&= \sum_{j\in U}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}} y_j E_p\left(I_j \middle| I_i = 1\right) \\
&= \sum_{j\in U}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}}\frac{\pi_{ij}}{\pi_i} y_j \\
&= \sum_{j\in U}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} y_j \\
&= B_{1i}^{HT}.
\end{aligned}
$$

Exemple 2.2.6. *Dans le cas d'un plan stratifié aléatoire simple sans remise, l'expression* (2.2.18) *devient*

$$\hat{B}_{1i}^{HT} = \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{y}_h), \quad \textit{pour } i \in U_h, \tag{2.2.19}$$

où $\bar{y}_h = \frac{1}{n_h}\sum_{j\in s_h} y_j$.

En effet, partant de l'équation (2.2.18) *et en supposant qu'on a tiré l'échantillon* s_h *de la strate* U_h*, nous avons:*

$$\begin{aligned}
\hat{B}_{1i}^{HT} &= \sum_{j\in s_h} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}} y_j \\
&= \frac{1-\pi_i}{\pi_i} y_i + \sum_{\substack{j\in s_h \\ j\neq i}} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}} y_j \\
&= \frac{N_h - n_h}{n_h} y_i + \sum_{\substack{j\in s_h \\ j\neq i}} \frac{n_h - N_h}{n_h\,(n_h-1)}\,(y_j + y_i - y_i) \\
&= \left\{\frac{N_h - n_h}{n_h} - \frac{n_h - N_h}{n_h\,(n_h-1)}\right\} y_i + \sum_{j\in s_h} \frac{n_h - N_h}{n_h\,(n_h-1)} y_j \\
&= \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right) y_i - \frac{N_h - n_h}{n_h - 1}\bar{y}_h \\
&= \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right) y_i - \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)\bar{y}_h,
\end{aligned}$$

d'où le résultat. Si on utilise l'estimateur de calage (1.2.3)*, le biais conditionnel de l'unité* i *par rapport à cet estimateur sera estimé par:*

$$\hat{B}_{1i}^{Cal} \approx \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(e_i - \overline{e}_h),$$

où $e_i = y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{B}}$ *et* $\overline{e}_h$ *est la moyenne échantillonnale des* e_i *de la strate* h*. En général* $\overline{e}_h = 0$*, de ce fait, on obtient l'approximation*

$$\hat{B}_{1i}^{Cal} \approx \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right) e_i. \tag{2.2.20}$$

Exemple 2.2.7. *Supposons qu'un échantillon de 10 unités a été tiré d'une population de taille 100 selon un plan aléatoire simple sans remise. La moyenne de l'échantillon est* $\bar{y} = 50$ *(données fictives). Le tableau ci-dessous donne une comparaison entre l'estimateur de Horvitz-Thompson et l'estimateur de calage correspondant pour un plan aléatoire simple sans remise. Dans cet exemple, il s'agit d'un modèle de régression linéaire simple sur la variable auxiliaire* $\boldsymbol{x}$ *avec ordonnée à l'origine. On a donc* $\boldsymbol{x}_i = (1, x_i)^\top$.

TABLEAU 2.1. Biais conditionnel pour les estimateurs de Horvitz-Thompson et de calage

i	y_i	x_i	e_i	$\hat{B}_{1i}^{HT}$	$\hat{B}_{1i}^{Cal}$
1	*50*	*70*	*7.8*	*0*	*78*
2	*50*	*85*	*7.2*	*0*	*72*
3	*50*	*90*	*7.1*	*0*	*71*
4	*100*	*1500*	*6.9*	*500*	*69*
5	*50*	*800*	*-18.2*	*0*	*-182*
6	*50*	*75*	*7.6*	*0*	*76*
7	*0*	*0*	*-39.7*	*-500*	*-397*
8	*50*	*100*	*6.7*	*0*	*67*
9	*50*	*80*	*7.4*	*0*	*74*
10	*50*	*85*	*7.2*	*0*	*72*

Le tableau 2.1 montre qu'une unité peut être influente par rapport à l'estimateur de Horvitz-Thompson et moins influente par rapport à l'estimateur de calage et vice versa.

Proposition 2.2.8. *Un estimateur conditionnellement sans biais du biais conditionnel pour le plan à deux degrés* (2.2.14) *est donné par*

$$\hat{B}_{1i|g}^{HT} = \sum_{\substack{l\in s\\ l\neq g}} \frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l} \sum_{j\in s_l} \frac{1}{\pi_{j|l}} y_{lj} + \sum_{j\in s_g} \frac{\pi_{i|g}}{\pi_{ij|g}} \left(\frac{1}{\pi_g} \frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1 \right) y_{gj}, \tag{2.2.21}$$

où $i \in U_g$.
Autrement dit

$$\begin{aligned} E_p\left(\hat{B}_{1i|g}^{HT} \middle| I_{1g}=1, I_{2i}=1\right) &= E_1E_2\left(\hat{B}_{1i|g}^{HT} \middle| I_{1g}=1, I_{2i}=1\right) \\ &= B_{1i|g}^{HT}, \end{aligned}$$

où $E_1(\cdot)$ *est l'espérance mathématique au premier degré et* $E_2(\cdot)$ *celle au second degré.*

La démonstration de la proposition 2.2.8 est donnée en annexe.

2.3. UNE MÉTHODE D'ESTIMATION ROBUSTE UTILISANT LE BIAIS CONDITIONNEL

Nous présentons la méthode d'estimation robuste proposée par Beaumont et al. (2013). Comme l'influence d'une unité sera identifiée par la valeur de son biais

conditionnel, il est donc naturel de se demander comment décider qu'un biais conditionnel est grand? La réponse à cette question dépendra d'un seuil positif (appelé tuning constant, en anglais) que l'on notera c dans la suite.

La recherche de critères d'optimalité pour la détermination du seuil est le coeur de ce mémoire et sera également abordé au chapitre suivant.

Beaumont et al. (2013) ont considéré la classe d'estimateurs robustes donnée par

$$\hat{t}_y^{RHT}(c) = \hat{t}_y^{HT} + \Delta_c, \tag{2.3.1}$$

où Δ_c est une variable aléatoire qui dépend de c, dont la valeur devra être déterminée. On peut remarquer que l'estimateur (2.3.1) est forcément biaisé puisque

$$Biais\left(\hat{t}_y^{RHT}(c)\right) = E_p\left(\hat{t}_y^{HT} + \Delta_c\right) - t_y = E_p\left(\Delta_c\right).$$

L'estimateur qu'ils ont proposé est le suivant

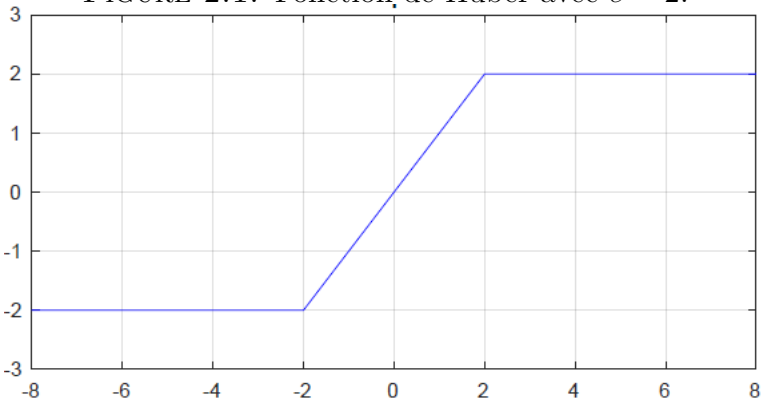
$$\hat{t}_y^{RHT}(c) = \hat{t}_y^{HT} + \sum_{i\in s}\left\{\Psi_c\left(\hat{B}_{1i}^{HT}\right) - \hat{B}_{1i}^{HT}\right\}, \tag{2.3.2}$$

qui appartient à la classe (2.3.1) avec

$$\Delta_c = \sum_{i\in s}\left\{\Psi_c\left(\hat{B}_{1i}^{HT}\right) - \hat{B}_{1i}^{HT}\right\}.$$

Ici, Ψ_c peut être n'importe quelle fonction bornée telle que $\Psi_c(z) \approx 0$ quand z est proche de 0. Dans notre cas, nous prendrons la fonction d'Huber définie par $\Psi_c(z) = sign(z) \times min(|z|, c)$, dont un exemple illustratif est donné ci-dessous avec $c = 2$.

FIGURE 2.1. Fonction de Huber avec $c = 2$.



Le rôle de cette fonction est de réduire l'impact des unités ayant une grande influence. En effet, pour les unités ayant un biais conditionnel supérieur à c, cette fonction va ramener leur biais conditionnel à c réduisant ainsi leur influence comme l'illustre l'exemple ci-dessous.

Exemple 2.3.1. *Considérons les données de l'exemple 2.2.7. Nous allons utiliser différentes valeurs de c pour appliquer la fonction de Huber au biais conditionnel* $\hat{B}_{1i}^{Cal}$.

TABLEAU 2.2. Fonction de Huber en fonction de différentes valeurs de c

i	y_i	w_i	$\hat{B}_{1i}^{Cal}$	$\Psi_{75}\left(\hat{B}_{1i}^{Cal}\right)$	$\Psi_{150}\left(\hat{B}_{1i}^{Cal}\right)$	$\Psi_{397}\left(\hat{B}_{1i}^{Cal}\right)$
1	*50*	*10.1941*	*78*	*75*	*78*	*78*
2	*50*	*10.2357*	*72*	*72*	*72*	*72*
3	*50*	*10.2495*	*71*	*71*	*71*	*71*
4	*100*	*14.1589*	*69*	*69*	*69*	*69*
5	*50*	*12.2181*	*-182*	*-75*	*-150*	*-182*
6	*50*	*10.2079*	*76*	*75*	*76*	*76*
7	*0*	*10.0000*	*-397*	*-75*	*-150*	*-397*
8	*50*	*10.2773*	*67*	*67*	*67*	*67*
9	*50*	*10.2218*	*74*	*74*	*74*	*74*
10	*50*	*10.2357*	*72*	*72*	*72*	*72*

Nous remarquons ici que pour $c = 75$, *la fonction* $\Psi_{75}(\hat{B}_{1i}^{Cal}) = sign(\hat{B}_{1i}^{Cal}) \times min(|\hat{B}_{1i}^{Cal}|, 75)$ *applique le minimum entre* $\hat{B}_{1i}^{Cal}$ *et* 75 *(et multiplie ensuite par le signe de* $\hat{B}_{1i}^{Cal}$*). Dans ce cas, quatre unités ont vu leur biais conditionnel corrigé. Pour* $c = 150$, *la fonction de Huber corrige le biais conditionnel de deux unités et pour* $c = 397 = max\left\{|\hat{B}_{1i}^{Cal}|; i \in s\right\}$, *aucune unité n'a son biais conditionnel corrigé. Il faut donc trouver une valeur optimale de c, selon un certain critère, pour le calcul de l'estimateur robuste. Nous expliquerons ci-dessous comment trouver cette valeur optimale de c. Pour l'instant, prenons* $c = 75$ *comme exemple afin d'illustrer comment l'estimateur robuste est calculé.*

TABLEAU 2.3. Exemple de calcul de l'estimateur robuste avec $c = 75$

i	y_i	w_i	$\hat{B}_{1i}^{Cal}$	$\Psi_{75}\left(\hat{B}_{1i}^{Cal}\right)$
1	*50*	*10.1941*	*78*	*75*
2	*50*	*10.2357*	*72*	*72*
3	*50*	*10.2495*	*71*	*71*
4	*100*	*14.1589*	*69*	*69*
5	*50*	*12.2181*	*-182*	*-75*
6	*50*	*10.2079*	*76*	*75*
7	*0*	*10.0000*	*-397*	*-75*
8	*50*	*10.2773*	*67*	*67*
9	*50*	*10.2218*	*74*	*74*
10	*50*	*10.2357*	*72*	*72*

On a que:

$$\hat{t}_y^{Cal} = \sum_{i \in s} w_i y_i = 5607.89$$

et

$$\hat{t}_y^{RCal} = \hat{t}_y^{Cal} + \Delta_{75} = \hat{t}_y^{Cal} + \sum_{i \in s} \left\{ \Psi_{75}\left(\hat{B}_{1i}^{Cal}\right) - \hat{B}_{1i}^{Cal} \right\} = 6032.89,$$

puisque

$$\Delta_{75} = \sum_{i \in s} \left\{ \Psi_{75}\left(\hat{B}_{1i}^{Cal}\right) - \hat{B}_{1i}^{Cal} \right\} = 425.$$

L'expression de l'estimateur robuste donnée dans (2.3.2) est sous une forme qui n'est pas toujours pratique à utiliser. En effet, les utilisateurs sont habitués d'utiliser des sommes pondérées pour obtenir leurs estimations. Il est donc utile de transformer les valeurs y_i de sorte à ce que leur somme pondérée donne l'estimateur robuste.

Selon Favre Martinoz et al. (2016), l'estimateur (2.3.2) peut s'exprimer comme

$$\hat{t}_y^{RHT} = \sum_{i \in s} w_i \tilde{y}_i, \tag{2.3.3}$$

où

$$\tilde{y}_i = y_i - \phi_i \frac{\hat{B}_{1i}^{HT}}{w_i} \tag{2.3.4}$$

et

$$\phi_i = 1 - \frac{\Psi_c\left(\hat{B}_{1i}^{HT}\right)}{\hat{B}_{1i}^{HT}}. \tag{2.3.5}$$

Comme $0 \leq \phi_i \leq 1$, quand le biais conditionnel de l'unité i est petit (par rapport à c), nous avons $\Psi_c\left(\hat{B}_{1i}^{HT}\right) = \hat{B}_{1i}^{HT}$ et donc $\phi_i = 0$, ce qui implique que $\tilde{y}_i = y_i$. Autrement dit, la valeur d'une unité non influente reste inchangée. Par contre si l'unité est influente, son biais conditionnel va être supérieur à c et sa valeur de y sera modifiée.

Exemple 2.3.2. *Dans cet exemple, qui est la suite de l'exemple 2.3.1, nous allons calculer des valeurs de $\tilde{y}_i$ de l'équation* (2.3.4) *pour différentes valeurs de c.*

TABLEAU 2.4. valeurs de $\tilde{y}_i$ pour différentes valeurs de c

i	y_i	w_i	$\hat{B}_{1i}^{Cal}$	$\tilde{y}_i(c=75)$	$\tilde{y}_i(c=150)$	$\tilde{y}_i(c=397)$
1	*50*	*10.1941*	*78*	*49.7*	*50*	*50*
2	*50*	*10.2357*	*72*	*50*	*50*	*50*
3	*50*	*10.2495*	*71*	*50*	*50*	*50*
4	*100*	*14.1589*	*69*	*100*	*100*	*100*
5	*50*	*12.2181*	*-182*	*58.8*	*52.6*	*50*
6	*50*	*10.2079*	*76*	*49.9*	*50*	*50*
7	*0*	*10.0000*	*-397*	*32.2*	*24.7*	*0*
8	*50*	*10.2773*	*67*	*50*	*50*	*50*
9	*50*	*10.2218*	*74*	*50*	*50*	*50*
10	*50*	*10.2357*	*72*	*50*	*50*	*50*

Dans ce tableau, nous remarquons que plus c est petit, plus on a tendance à corriger les valeurs. On remarque également que les valeurs des unités 5 et 7 ont été corrigées à la hausse.

On peut également exprimer l'estimateur robuste comme $\hat{t}_y^{RHT} = \sum_{i \in s} \tilde{w}_i y_i$, où $\tilde{w}_i = w_i - \phi_i \frac{\hat{B}_{1i}^{HT}}{y_i}$. Lorsque le biais conditionnel de l'unité i est petit, nous avons $\phi_i = 0$ et $\tilde{w}_i = w_i$. Autrement dit le poids d'une unité non influente reste inchangé. Celui d'une unité influente sera modifié.

2.3.1. Critère d'optimatilité selon Beaumont et al. (2013).

Nous avons remarqué dans l'exemple 2.3.1 qu'il est nécessaire de trouver une valeur optimale de c pour décider quand une unité est influente. Pour ce faire, Beaumont et al. (2013) ont utilisé le critère qui consiste à minimiser l'expression suivante par rapport à c

$$max\left\{|\hat{B}_{1i}^{RHT}(c)|; i \in s\right\}, \tag{2.3.6}$$

où

$$\hat{B}_{1i}^{RHT}(c) = \hat{B}_{1i}^{HT} + \sum_{i \in s} \left\{ \Psi_c \left(\hat{B}_{1i}^{HT} \right) - \hat{B}_{1i}^{HT} \right\}$$

représente l'estimateur du biais conditionnel de l'unité i par rapport à l'estimateur robuste.

Autrement dit, on veut trouver un estimateur robuste de la forme (2.3.1) et on veut choisir cet estimateur de sorte à ce que la valeur maximale de son biais conditionnel soit minimisée. C'est ce critère qui nous permet de trouver la valeur optimale de c. Ce critère, appelé min-max, nous permet d'aboutir à

$$\Delta_{C_{opt}} = -\frac{1}{2} \left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT} \right),$$

ce qui conduit à

$$\hat{t}_y^{RHT}(c_{opt}) = \hat{t}_y^{HT} - \frac{1}{2} \left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT} \right), \tag{2.3.7}$$

où

$$\hat{B}_{min}^{HT} = min \left\{ \hat{B}_{1i}^{HT}; i \in s \right\}$$

et

$$\hat{B}_{max}^{HT} = max \left\{ \hat{B}_{1i}^{HT}; i \in s \right\},$$

voir Favre Martinoz (2015).

2.3.2. Calcul de la valeur optimale de c

Comme nous voulons un estimateur robuste de la forme (2.3.2) et que la valeur de c optimale implique que (2.3.7) est satisfaite, il nous faut résoudre l'équation

$$\Delta_c = -\frac{1}{2} \left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT} \right)$$

ou encore

$$\sum_{i \in s} \left\{ \Psi_c \left(\hat{B}_{1i}^{HT} \right) - \hat{B}_{1i}^{HT} \right\} + \frac{1}{2} \left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT} \right) = 0. \tag{2.3.8}$$

En posant $\delta_c = \sum_{i \in s} \left\{ \Psi_c \left(\hat{B}_{1i}^{HT} \right) - \hat{B}_{1i}^{HT} \right\} + \frac{1}{2} \left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT} \right)$, résoudre (2.3.8) revient à résoudre $\delta_c = 0$ par rapport à c.

Techniquement, on cherche c entre 0 et $max \left\{ |\hat{B}_{1i}^{HT}|; i \in s \right\}$. En effet, lorsque $c = 0$, on corrige toutes les unités alors que lorsque $c = max \left\{ |\hat{B}_{1i}^{HT}|; i \in s \right\}$, on ne corrige aucune unité. Pour ce faire, on ordonne $\left\{ |\hat{B}_{1i}^{HT}|; i \in s \right\}$ par ordre décroissant et on calcule δ_c pour chaque valeur de $c = |\hat{B}_{1i}^{HT}|$. On examine à quel moment il y a un changement de signe de δ_c . Lors du premier changement de

signe, on prend les deux valeurs de c donnant lieu au changement de signe, et on fait une interpolation linéaire pour trouver la solution. Ceci sera illustré dans un exemple à la section 2.3.3.

Si cette équation admet plusieurs solutions, on prendra celle obtenue par la plus grande valeur de c. En effet, on veut la plus grande valeur de c possible puisque cela donne lieu à l'estimateur qui corrigera le moins d'unités.

L'objectif cherché dans l'estimation robuste étant de réduire l'influence des unités qui ont une grande influence, cela conduit à des estimateurs biaisés mais stables. En même temps, nous aimerions que quand il n'y a pas d'unité influente, les estimateurs robustes ne soient pas beaucoup moins efficaces que les estimateurs non robustes correspondants. Selon Beaumont et al. (2013), l'équation (2.3.8) admet toujours une solution. Il est donc possible qu'on ajuste la valeur de certaines unités même si elles ne sont pas très influentes. Par conséquent, en cas de solutions multiples à (2.3.8), le fait de choisir la solution qui a la plus grande valeur permet de limiter le nombre de valeurs corrigées et ainsi avoir un estimateur robuste faiblement biaisé.

2.3.3. Un rappel sur l'interpolation linéaire

Supposons que c_{max} est la dernière valeur du biais conditionnel pour laquelle δ_c est positif (ou négatif, respectivement), et notons cette valeur par δ_{max}. De même, supposons que c_{min} est la première valeur du biais conditionnel pour laquelle δ_c est négatif (ou positif, respectivement), et notons cette valeur par δ_{min}. La formule de l'interpolation linéaire est la suivante:

$$c_{opt} = c_{min} - \delta_{min} \times \frac{c_{max} - c_{min}}{\delta_{max} - \delta_{min}}.$$

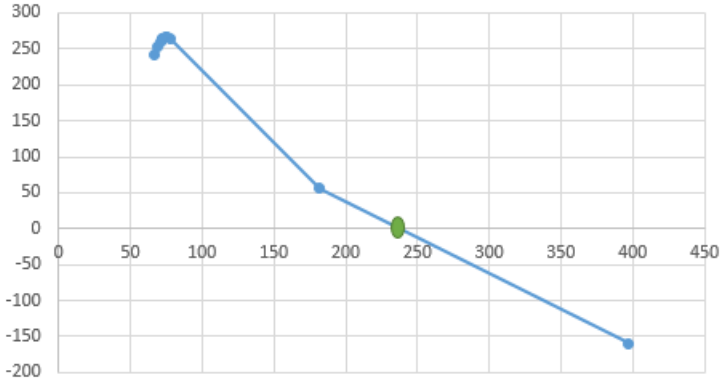
L'intersection de la droite passant par les points (c_{min}, δ_{min}) et (c_{max}, δ_{max}) avec l'axe des abscisses donne la solution $c = c_{opt}$ de l'équation $\delta_c = 0$. Nous allons appliquer l'interpolation linéaire aux données de l'exemple ci-dessus.

TABLEAU 2.5. δ_c en fonction de c

$c(=\lvert\hat{B}_{1i}^{Cal}\rvert)$	δ_c
397	-159.5
182	55.5
78	263.5
76	265.5
74	265.5
72	263.5
72	263.5
71	260.5
69	252.5
67	242.5

Le tableau 2.5 ci-dessus donne les dix valeurs de c triées de façon décroissante en fonction des valeurs correspondantes de δ_c. Pour la plus grande valeur de c, la valeur de δ_c correspondante est négative. C'est à la troisième plus grande valeur de c que le changement de signe a eu lieu. Nous avons obtenu un seul changement de signe, ce qui veut dire que l'équation admet une solution unique.

FIGURE 2.2. Représentation graphique des données du tableau 2.5



On voit que la fonction δ_c change de signe entre $c_{min} = 182$ et $c_{max} = 397$. Le c_{opt} se trouvera alors entre ces deux valeurs. En effectuant les calculs, nous obtenons:

$$c_{opt} = 182 - 55.5 \times \frac{397 - 182}{-159.5 - 55.5} = 237.5.$$

C'est donc cette valeur qu'il faudra utiliser pour déterminer les ajustements (corrections) aux valeurs y , c'est-à-dire les $\tilde{y}$ de l'équation (2.3.4). Nous obtenons le tableau suivant:

TABLEAU 2.6. valeurs de $\tilde{y}_i$ pour la valeur optimale de c

i	y_i	w_i	e_i	$\hat{B}_{1i}^{Cal}$	$\tilde{y}_i$
1	50	10.1941	7.8	78	50
2	50	10.2357	7.2	72	50
3	50	10.2495	7.1	71	50
4	100	14.1589	6.9	69	100
5	50	12.2181	-18.2	-182	50
6	50	10.2079	7.6	76	50
7	0	10.0000	-39.7	-397	15.96
8	50	10.2773	6.7	67	50
9	50	10.2218	7.4	74	50
10	50	10.2357	7.2	72	50

Nous remarquons que la méthode du biais conditionnel a corrigé l'unité ayant le plus grand résidu (unité 7), ce qui est normal selon (2.2.20). Or l'utilisation d'une des méthodes actuelles de détection des unités influentes, c'est-à-dire celles qui consistent à juger l'influence d'une unité selon sa valeur sur la variable d'intérêt, ou son poids ou le produit des deux, n'aurait pas détectée l'unité 7 comme influente. En effet, l'unité 7 a non seulement le poids le plus faible, mais elle a aussi une contribution ($w_7 \times y_7$) nulle. Par contre, l'unité 4 serait détectée sans surprise comme étant influente, puisqu'elle a la plus grande valeur de y et le plus grand poids, par conséquent la plus grande contribution. Ceci marque l'importance de la configuration lorsque l'on cherche à quantifier l'influence d'une unité, puisqu'elle tient compte non seulement du plan de sondage, mais aussi de la variable d'intérêt et du paramètre à estimer.

Chapitre 3

RECHERCHE DE CRITÈRES D'OPTIMALITÉ POUR LA DÉTERMINATION DU SEUIL DE ROBUSTESSE

3.1. Introduction

L'estimateur robuste donné dans Beaumont et al. (2013) est basé sur un critère de type min-max. Dans ce chapitre, on cherche à savoir si on peut trouver des critères différents conduisant à des estimateurs robustes ayant de bonnes propriétés, autrement dit, celles qui minimisent l'erreur quadratique moyenne.

3.2. Un critère général

Rappelons que, comme dans Beaumont et al. (2013), nous cherchons un estimateur robuste de la classe

$$\hat{t}_y^{RHT}(c) = \hat{t}_y^{HT} + \Delta_c,$$

où Δ_c est une variable aléatoire qui dépend de c, dont la valeur devra être déterminée.
Dans cette section, nous considérons le critère général suivant:

$$\sum_{i\in s} \phi_i \left|\hat{B}_{1i}^{RHT}(c)\right|^q = \sum_{i\in s} \phi_i \left|\hat{B}_{1i}^{HT} - \Delta_c\right|^q, \tag{3.2.1}$$

où ϕ_i est un coefficient associé à l'unité i et $q \in \{1, 2, 3, ...\}$.

Ordonnons de façon croissante les valeurs des coefficients ϕ dans l'équation (3.2.1) selon les valeurs du biais conditionnel de l'estimateur robuste. Pour ce faire, on note par $\phi_{(1)}$ le coefficient ϕ_i associé à l'unité exhibant le plus petit biais conditionnel $\hat{B}_{1i}^{RHT}$, $\phi_{(2)}$ le coefficient ϕ_i associé à l'unité exhibant le suivant plus petit

biais conditionnel,..., et $\phi_{(n)}$ le coefficient ϕ_i associé à l'unité exhibant le plus grand biais conditionnel de l'estimateur robuste. Ainsi le critère min-max, défini en (2.3.6), peut être vu comme un cas particulier de (3.2.1) avec $q = 1$, $\phi_{(1)} = ... = \phi_{(n-1)} = 0$ et $\phi_{(n)} = 1$.

3.2.1. Cas particuliers du critère général

Considérons le cas $q = 1$ et $\phi_i = 1$ pour tout $i \in s$. Le critère (3.2.1) se simplifie pour donner

$$\sum_{i\in s} \left| \hat{B}_{1i}^{HT} - \Delta_c \right|. \tag{3.2.2}$$

Il est bien connu que la valeur de Δ_c qui minimise (3.2.2) est

$$\Delta_c = med \left\{ \hat{B}_{1i}^{HT}, i \in s \right\}.$$

Dans ce cas l'estimateur (2.3.2) devient

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - med \left\{ \hat{B}_{1i}^{HT}, i \in s \right\}. \tag{3.2.3}$$

Dans le cas d'un plan aléatoire simple sans remise et en se référant à l'équation (2.2.19), l'estimateur (3.2.3) s'écrit comme

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \frac{n}{n-1}\left(\frac{N}{n} - 1\right) med \left\{ (y_i - \bar{y}), i \in s \right\}. \tag{3.2.4}$$

Pour le plan de Bernoulli, selon l'équation (2.2.12), l'estimateur (3.2.3) s'écrit comme

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \left(\frac{1}{\pi} - 1\right) med \left\{ y_i, i \in s \right\}. \tag{3.2.5}$$

Considérons maintenant le cas $q = 2$ et $\phi_i = 1$ pour tout $i \in s$. Le critère (3.2.1) se simplifie pour donner

$$\sum_{i\in s} \left| \hat{B}_{1i}^{HT} - \Delta_c \right|^2. \tag{3.2.6}$$

La valeur de Δ_c qui minimise (3.2.6) est

$$\Delta_c = \sum_{i\in s} \hat{B}_{1i}^{HT}/n,$$

pour un plan de sondage à taille fixe ou

$$\Delta_c = \sum_{i\in s} \hat{B}_{1i}^{HT}/n_s,$$

pour un plan de sondage à taille aléatoire.

L'estimateur (2.3.2), pour un plan de sondage à taille fixe, se réduit à

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - \frac{1}{n}\sum_{i\in s} \hat{B}_{1i}^{HT}. \tag{3.2.7}$$

Dans le cas d'un plan aléatoire simple sans remise, le deuxième terme du membre de droite de (3.2.7) devient:

$$\begin{aligned}\frac{1}{n}\sum_{i\in s}\hat{B}_{1i}^{HT} &= \frac{1}{n}\sum_{i\in s}\frac{n}{n-1}\left(\frac{N}{n}-1\right)(y_i-\bar{y})\\ &= \frac{1}{n}\frac{n}{n-1}\left(\frac{N}{n}-1\right)\sum_{i\in s}(y_i-\bar{y})\\ &= 0,\end{aligned}$$

et on a

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT}.$$

Autrement dit, ce critère n'a aucun effet sur le traitement des unités influentes lorsqu'on utilise un plan aléatoire simple sans remise.

Pour le plan de Bernoulli, selon les équations (1.2.1) et (2.2.12), l'équation (3.2.7) devient

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT}\left\{1-\frac{1-\pi}{n_s}\right\} \approx \hat{t}_y^{HT}, \tag{3.2.8}$$

pour n_s grand.

Dans le cas $q \geq 3$ et $\phi_i = 1$ pour tout $i \in s$, le critère (3.2.1) s'écrit comme

$$\underset{\Delta_c}{argmin}\left\{\sum_{i\in s}\left|\hat{B}_{1i}^{HT}-\Delta_c\right|^q\right\}, \quad q \geq 3. \tag{3.2.9}$$

Dans ce cas, il n'est pas possible d'obtenir une expression explicite de Δ_c. Il faut donc recourir à des méthodes numériques. De Moliner (2017) a également utilisé le critère (3.2.9) dans le cadre de données fonctionnelles.

On peut également s'intéresser à minimiser la somme des $k \geq 2$ plus grands biais conditionnels robustes en valeur absolue. Pour ce faire, considérons le cas $q = 1$, $\phi_{(i)} = 0$ pour $i = 1, 2, ..., n-k$ et $\phi_{(n-k+1)} = ... = \phi_{(n-1)} = \phi_{(n)} = 1$. Par conséquent, le critère (3.2.1) devient

$$\sum_{i\in s}\phi_i\left|\hat{B}_{1i}^{HT}-\Delta_c\right|^q = \left|\hat{B}_{1(n-k+1)}^{HT}-\Delta_c\right| + ... + \left|\hat{B}_{1(n-1)}^{HT}-\Delta_c\right| + \left|\hat{B}_{1(n)}^{HT}-\Delta_c\right|.$$

Comme

$$\underset{\Delta_c}{argmin}\left\{\sum_{i=n-k+1}^{n}\left|\hat{B}_{1(i)}^{HT}-\Delta_c\right|\right\} \tag{3.2.10}$$

admet pour solution

$$\Delta_c = med\left\{\hat{B}_{1(i)}^{HT}, \quad i = n-k+1, ..., n\right\},$$

l'estmateur robuste (2.3.2) devient dans ce cas

$$\hat{t}_y^{RHT} = \hat{t}_y^{HT} - med\left\{\hat{B}_{1(i)}^{HT}, \quad i = n-k+1, ..., n\right\}. \tag{3.2.11}$$

3.2.2. Critère basé sur les percentiles

Un autre critère que l'on considère consiste à minimiser le *p-ième* percentile de $\left|\hat{B}_{1(i)}^{HT} - \Delta_c\right|$, $i = 1, ..., n$. Nous allons considérer les valeurs $p = 75$, 90, 95, 97.5 et 99%.

3.3. ÉTUDE PAR SIMULATIONS

Nous avons effectué une étude par simulations afin de comparer les différents critères présentés dans ce chapitre en terme de biais et d'efficacité relative. Des populations de taille $N = 1\ 000$ ont été générées en utilisant plusieurs distributions asymétriques. Le plan de Bernoulli et le plan aléatoire simple sans remise ont été utilisés pour la sélection des échantillons.

De chaque population, nous avons tiré $R = 5\ 000$ échantillons de taille $n = 25$, 50 ou 100, lorsqu'il s'agit du plan aléatoire sans remise et de taille espérée 25, 50 ou 100, quand on utilise le plan de Bernoulli. Dans chaque échantillon, nous avons calculé l'estimateur Horvitz-Thompson, l'estimateur robuste obtenu en utilisant le critère de Beaumont et al. (2013) et ceux obtenus au moyen des critères décrits ci-dessus.

Comme mesure du biais d'un estimateur $\hat{t}_y$, nous avons utilisé le biais relatif Monte Carlo:

$$RB_{MC}(\hat{t}_y) = 100 \times \frac{E_{MC}(\hat{t}_y) - t_y}{t_y},$$

où

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

et $\hat{t}_y^{(r)}$ désigne la valeur de l'estimateur $\hat{t}_y$ obtenue par le *r-ième* échantillon, $r = 1, ..., R$. Nous avons également calculé l'efficacité relative Monte Carlo par rapport à l'estimateur Horvitz-Thompson, $\hat{t}_y^{HT}$:

$$RE_{MC}(\hat{t}_y) = 100 \times \frac{MSE_{MC}(\hat{t}_y)}{MSE_{MC}(\hat{t}_y^{HT})},$$

où

$$MSE_{MC}(\hat{t}_y) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{t}_y^{(r)} - t_y\right)^2.$$

Dans les tableaux de résultats ci-dessous, l'estimateur de Beaumont et al. (2013) sera noté par $\hat{t}_y^B$, les autres estimateurs par $\hat{t}_y^R$. Pour chaque couple de population et de taille d'échantillon, la première ligne du tableau représente le biais relatif et la deuxième ligne l'efficacité relative mise entre parenthèses.

3.3.1. Mélange de normales

Nous avons utilisé des populations issues d'un mélange de deux normales.
Pour l'unité i issue d'une de ces populations, la valeur y_i est générée selon le modèle

$$Y_i = \alpha Z_{1i} + (1-\alpha)Z_{2i}, \tag{3.3.1}$$

où $Z_{1i} \sim N(100, 30)$, $Z_{2i} \sim N(2\,000, 500)$ et $0 \leq \alpha \leq 1$. Les valeurs de α que nous avons utilisées sont: 0.99, 0.95 et 0.90. Les trois populations ainsi obtenues, seront appelées mixnor1, mixnor2 et mixnor3 lorsque α= 0.99, 0.95 et 0.90, respectivement.

TABLEAU 3.1. Biais et efficacité relative de l'estimateur (3.2.9) avec q=10, 20, 30, 40, 50 pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[10]$	$\hat{t}_y^R[20]$	$\hat{t}_y^R[30]$	$\hat{t}_y^R[40]$	$\hat{t}_y^R[50]$
mixnor1	25	-0.1	-2.5	-2.1	-2.4	-2.4	-2.4	-2.5
		(100)	(45)	(52)	(48)	(46)	(46)	(46)
mixnor1	50	-0.1	-2.4	-2.0	-2.3	-2.3	-2.4	-2.4
		(100)	(51)	(58)	(53)	(52)	(52)	(52)
mixnor1	100	-0.1	-2.1	-1.7	-1.9	-2.0	-2.0	-2.1
		(100)	(63)	(68)	(65)	(64)	(63)	(63)
mixnor2	25	0.2	-7.9	-6.7	-7.5	-7.7	-7.8	-7.9
		(100)	(71)	(74)	(72)	(72)	(72)	(71)
mixnor2	50	0.3	-5.8	-4.8	-5.4	-5.6	-5.7	-5.8
		(100)	(86)	(87)	(87)	(86)	(86)	(86)
mixnor2	100	-0.2	-4.0	-3.7	-3.8	-3.9	-3.9	-3.9
		(100)	(98)	(98)	(98)	(98)	(98)	(98)
mixnor3	25	-0.1	-9.2	-7.8	-8.7	-9.0	-9.1	-9.1
		(100)	(93)	(93)	(93)	(93)	(93)	(93)
mixnor3	50	0.3	-5.5	-4.4	-5.1	-5.3	-5.4	-5.4
		(100)	(98)	(98)	(98)	(98)	(98)	(98)
mixnor3	100	0.1	-3.1	-2.4	-2.8	-3.0	-3.0	-3.1
		(100)	(100)	(100)	(100)	(101)	(101)	(101)

* La valeur de q est indiquée entre crochets.

TABLEAU 3.2. Biais et efficacité relative de l'estimateur (3.2.11) pour le plan aléatoire simple sans remise avec k=2, ..., 7.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[2]$	$\hat{t}_y^{R}[3]$	$\hat{t}_y^{R}[4]$	$\hat{t}_y^{R}[5]$	$\hat{t}_y^{R}[6]$	$\hat{t}_y^{R}[7]$
mixnor1	25	-0.1	-2.5	-0.6	-0.0	0.1	0.1	0.1	0.1
		(100)	(45)	(82)	(99)	(104)	(107)	(107)	(107)
mixnor1	50	-0.1	-2.4	-0.7	-0.2	-0.1	-0.1	-0.0	-0.0
		(100)	(51)	(81)	(92)	(99)	(101)	(102)	(103)
mixnor1	100	-0.1	-2.1	-0.8	-0.4	-0.1	0.1	-0.1	-0.1
		(100)	(63)	(79)	(85)	(95)	(98)	(100)	(100)
mixnor2	25	0.2	-7.9	-3.0	-1.7	-0.3	0.5	0.8	1.1
		(100)	(71)	(78)	(80)	(91)	(97)	(101)	(105)
mixnor2	50	0.3	-5.8	-3.1	-2.7	-1.2	-0.8	-0.1	0.2
		(100)	(86)	(82)	(80)	(85)	(86)	(92)	(94)
mixnor2	100	-0.2	-4.0	-2.8	-2.8	-2.1	-2.0	-1.4	-1.3
		(100)	(98)	(93)	(92)	(89)	(89)	(89)	(89)
mixnor3	25	-0.1	-9.2	-5.2	-4.4	-2.4	-1.4	-0.4	0.3
		(100)	(92)	(87)	(84)	(87)	(88)	(93)	(95)
mixnor3	50	0.3	-5.5	-3.8	-3.8	-2.7	-2.5	-1.6	-1.4
		(100)	(99)	(95)	(94)	(92)	(91)	(90)	(90)
mixnor3	100	0.1	-3.1	-2.4	-2.5	-2.1	-2.1	-1.8	-1.8
		(100)	(101)	(100)	(100)	(99)	(98)	(97)	(97)

* La valeur de k est indiquée entre crochets.

TABLEAU 3.3. Biais et efficacité relative de l'estimateur obtenu en minimisant des percentiles pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[75]$	$\hat{t}_y^{R}[90]$	$\hat{t}_y^{R}[95]$	$\hat{t}_y^{R}[97.5]$	$\hat{t}_y^{R}[99]$
mixnor1	25	-0.1	-2.5	0.1	0.2	0.0	-2.5	-2.5
		(100)	(45)	(108)	(107)	(101)	(45)	(45)
mixnor1	50	-0.1	-2.4	-0.0	0.0	-0.0	-0.2	-2.4
		(100)	(51)	(104)	(103)	(102)	(93)	(51)
mixnor1	100	-0.1	-2.1	-0.0	-0.0	-0.0	-0.1	-0.7
		(100)	(63)	(101)	(101)	(101)	(98)	(79)
mixnor2	25	0.2	-7.9	1.4	0.7	-1.5	-7.9	-7.9
		(100)	(71)	(108)	(98)	(81)	(71)	(71)
mixnor2	50	0.3	-5.8	0.9	0.8	-0.6	-2.5	-5.8
		(100)	(87)	(104)	(102)	(86)	(80)	(86)
mixnor2	100	-0.2	-4.0	-0.0	-0.0	-0.4	-1.9	-2.8
		(100)	(98)	(102)	(101)	(94)	(88)	(93)
mixnor3	25	-0.1	-9.2	1.8	-1.2	-4.2	-9.2	-9.2
		(100)	(93)	(108)	(88)	(84)	(93)	(93)
mixnor3	50	0.3	-5.5	1.2	0.0	-2.4	-3.7	-5.5
		(100)	(99)	(104)	(94)	(90)	(94)	(99)
mixnor3	100	0.1	-3.1	0.5	0.1	-1.2	-2.1	-2.4
		(100)	(101)	(102)	(96)	(95)	(98)	(100)

* La valeur du percentile est indiquée entre crochets.

Les tableaux 3.1, 3.2 et 3.3 donnent les résultats des différentes méthodes examinées pour le plan aléatoire simple sans remise. Pour le tableau 3.1 nous remarquons que lorsque q augmente, le biais augmente et la valeur de l'efficacité relative diminue pour donner des résultats très similaires à ceux du min-max. Dans le tableau 3.2, pour la population mixnor1, le min-max donne de meilleurs résultats en terme d'efficacité relative, même si son biais est plus élevé. Par contre pour les deux autres populations, l'estimateur (3.2.11) donne parfois de meilleurs résultats que le min-max. En effet, lorsque $n = 25$, pour la population mixnor3, le biais et la valeur de l'efficacité relative sont -9.2% et 92% respectivement pour l'estimateur $\hat{t}_y^B$ alors que pour l'estimateur $\hat{t}_y^R[3]$ ils sont de -4.4% et 84% respectivement. Le tableau 3.3 donne les résultats pour la méthode qui consiste à minimiser des percentiles du biais conditionnel de l'estimateur robuste. Nous remarquons dans ce tableau que le 99ème percentile donne des résultats très similaires à ceux du min-max. Pour les plus faibles percentiles, en général les résultats du min-max sont meilleurs. En effet, pour la population mixnor2, avec $n = 50$, le biais et l'efficacité relative sont 0.8% et 102% pour l'estimateur $\hat{t}_y^R[90]$ alors qu'ils sont de -5.8% et 87% respectivement pour l'estimateur de Beaumont.

TABLEAU 3.4. Biais et efficacité relative de l'estimateur (3.2.9) pour le plan de Bernoulli avec q=10, 20, 30, 40, 50.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[10]$	$\hat{t}_y^R[20]$	$\hat{t}_y^R[30]$	$\hat{t}_y^R[40]$	$\hat{t}_y^R[50]$
mixnor1	25	-0.2	-8.6	-7.9	-8.3	-8.5	-8.5	-8.5
		(100)	(60)	(65)	(62)	(61)	(61)	(60)
mixnor1	50	0.5	-6.1	-5.2	-5.7	-5.9	-6.0	-6.0
		(100)	(65)	(69)	(66)	(66)	(65)	(65)
mixnor1	100	0.1	-2.8	-2.4	-2.7	-2.7	-2.8	-2.8
		(100)	(84)	(87)	(85)	(85)	(85)	(85)
mixnor2	25	0.1	-15.9	-13.9	-15.1	-15.5	-15.7	-15.8
		(100)	(82)	(83)	(83)	(83)	(82)	(82)
mixnor2	50	0.1	-10.5	-8.8	-9.8	-10.1	-10.3	-10.4
		(100)	(97)	(96)	(97)	(97)	(97)	(97)
mixnor2	100	0.0	-4.6	-3.8	-4.3	-4.4	-4.5	-4.5
		(100)	(99)	(99)	(99)	(99)	(99)	(99)
mixnor3	25	-0.1	-14.3	-12.5	-13.6	-13.9	-14.0	-14.1
		(100)	(98)	(97)	(98)	(98)	(98)	(98)
mixnor3	50	-0.2	-8.3	-7.0	-7.7	-8.0	-8.1	-8.2
		(100)	(103)	(102)	(102)	(103)	(103)	(103)
mixnor3	100	0.1	-4.0	-3.3	-3.7	-3.8	-3.9	-3.9
		(100)	(101)	(100)	(100)	(101)	(101)	(101)

* La valeur de q est indiquée entre crochets.

TABLEAU 3.5. Biais et efficacité relative de l'estimateur (3.2.11) avec k=2, ..., 7 pour le plan Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[2]$	$\hat{t}_y^{R}[3]$	$\hat{t}_y^{R}[4]$	$\hat{t}_y^{R}[5]$	$\hat{t}_y^{R}[6]$	$\hat{t}_y^{R}[7]$
mixnor1	25	-0.2	-8.6	-3.9	-4.0	-3.6	-3.7	-3.7	-3.6
		(100)	(60)	(93)	(95)	(100)	(101)	(101)	(101)
mixnor1	50	0.5	-6.1	-2.0	-1.8	-1.3	-1.3	-1.2	-1.2
		(100)	(65)	(87)	(90)	(97)	(99)	(100)	(100)
mixnor1	100	0.1	-2.8	-1.1	-1.1	-0.9	-0.8	-0.8	-0.8
		(100)	(84)	(93)	(93)	(98)	(99)	(100)	(100)
mixnor2	25	0.1	-15.9	-7.9	-7.7	-4.3	-3.9	-2.8	-2.6
		(100)	(82)	(77)	(76)	(85)	(87)	(93)	(95)
mixnor2	50	0.1	-10.5	-7.1	-7.0	-4.7	-4.3	-2.8	-2.5
		(100)	(97)	(86)	(85)	(83)	(83)	(87)	(87)
mixnor2	100	0.0	-4.6	-3.4	-3.6	-2.7	-2.6	-2.0	-1.9
		(100)	(99)	(94)	(94)	(91)	(91)	(91)	(91)
mixnor3	25	-0.1	-14.3	-10.1	-10.0	-7.1	-6.7	-4.5	-4.1
		(100)	(98)	(87)	(87)	(83)	(83)	(85)	(86)
mixnor3	50	-0.2	-8.3	-7.1	-7.1	-6.2	-6.1	-5.2	-5.0
		(100)	(103)	(99)	(99)	(95)	(94)	(92)	(91)
mixnor3	100	0.1	-4.0	-3.2	-3.3	-3.0	-3.0	-2.7	-2.7
		(100)	(101)	(101)	(100)	(99)	(99)	(97)	(97)

* La valeur de k est indiquée entre crochets.

TABLEAU 3.6. Biais et efficacité relative de l'estimateur obtenu en minimisant des percentiles pour le plan de Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[75]$	$\hat{t}_y^{R}[90]$	$\hat{t}_y^{R}[95]$	$\hat{t}_y^{R}[97.5]$	$\hat{t}_y^{R}[99]$
mixnor1	25	0.2	-6.0	-3.5	-3.5	-3.8	-5.9	-6.0
		(100)	(77)	(101)	(101)	(99)	(78)	(77)
mixnor1	50	0.5	-3.8	-1.3	-1.3	-1.3	-1.6	-3.8
		(100)	(79)	(100)	(100)	(100)	(97)	(79)
mixnor1	100	0.1	-2.8	-0.8	-0.8	-0.8	-0.8	-1.9
		(100)	(84)	(100)	(100)	(100)	(99)	(97)
mixnor2	25	0.1	-11.4	-2.7	-3.4	-6.1	-11.4	-11.4
		(100)	(80)	(101)	(95)	(86)	(80)	(80)
mixnor2	50	0.1	-7.7	-1.3	-1.4	-2.7	-4.7	-7.7
		(100)	(92)	(100)	(99)	(90)	(87)	(92)
mixnor2	100	0.0	-4.6	-0.6	-0.6	-1.1	-2.5	-4.0
		(100)	(99)	(100)	(100)	(95)	(91)	(100)
mixnor3	25	0.1	-12.4	-2.1	-4.8	-8.3	-12.4	-12.4
		(100)	(93)	(100)	(89)	(88)	(93)	(93)
mixnor3	50	-0.1	-7.6	-1.1	-2.2	-4.5	-5.9	-7.6
		(100)	(100)	(100)	(93)	(93)	(96)	(100)
mixnor3	100	0.1	-4.0	-0.4	-0.9	-2.1	-3.0	-3.6
		(100)	(101)	(100)	(96)	(96)	(99)	(102)

* La valeur du percentile est indiquée entre crochets.

Les trois tableaux ci-dessous donnent les mêmes informations que les trois premiers tableaux, mais pour le plan de Bernoulli. Ce que nous remarquons est que les observations vont dans le même sens que celles obtenues pour le plan aléatoire simple sans remise, ceci quelle que soit la population ou la méthode considérée.

3.3.2. Mélange de lognormales

Nous avons également utilisé des populations issues d'un mélange de deux lognormales. Le principe est le même comme pour le mélange de normales. Autrement dit, pour toute unité issue de ces populations, sa valeur sur la variable d'intérêt s'obtient en utilisant l'équation (3.3.1), avec $Z_1 \sim Lognormale(1, 0.25)$, $Z_2 \sim Lognormale(5, 1)$ et $0 \leq \alpha \leq 1$. Nous avons utilisé les mêmes valeurs de α que pour le mélange de normales, c'est-à-dire 0.99, 0.95 et 0.90. Les trois populations obtenues ont été appelées mixlogn1, mixlogn2 et mixlogn3 lorsque α= 0.99, 0.95 et 0.90, respectivement.

TABLEAU 3.7. Biais et efficacité relative de l'estimateur (3.2.9) avec q=10, 20, 30, 40, 50 pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[10]$	$\hat{t}_y^{R}[20]$	$\hat{t}_y^{R}[30]$	$\hat{t}_y^{R}[40]$	$\hat{t}_y^{R}[50]$
mixlogn1	25	-0.2	-16.3	-13.3	-14.9	-15.4	-15.6	-15.8
		(100)	(35)	(44)	(39)	(37)	(37)	(36)
mixlogn1	50	0.1	-15.5	-12.1	-13.9	-14.5	-14.7	-14.9
		(100)	(39)	(49)	(44)	(42)	(41)	(41)
mixlogn1	100	0.2	-13.8	-10.2	-12.1	-12.7	-13.0	-13.1
		(100)	(50)	(59)	(54)	(53)	(52)	(51)
mixlogn2	25	0.4	-27.2	-22.0	-24.8	-25.6	-26.1	-26.3
		(100)	(48)	(55)	(51)	(50)	(49)	(49)
mixlogn2	50	1.8	-21.7	-16.6	-19.3	-20.2	-20.6	-20.8
		(100)	(55)	(62)	(58)	(57)	(57)	(56)
mixlogn2	100	-0.5	-17.3	-13.0	-15.2	-15.9	-16.3	-16.5
		(100)	(68)	(73)	(70)	(69)	(69)	(68)
mixlogn3	25	-0.6	-24.9	-20.3	-22.7	-23.5	-23.9	-24.1
		(100)	(60)	(65)	(62)	(61)	(61)	(61)
mixlogn3	50	1.5	-17.6	-13.4	-15.6	-16.3	-16.7	-16.9
		(100)	(67)	(71)	(69)	(68)	(68)	(67)
mixlogn3	100	0.5	-12.7	-9.3	-11.1	-11.6	-11.9	-12.1
		(100)	(77)	(80)	(78)	(78)	(78)	(78)

* La valeur de q est indiquée entre crochets.

TABLEAU 3.8. Biais et efficacité relative de l'estimateur (3.2.11) avec k=2, ..., 7 pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[2]$	$\hat{t}_y^{R}[3]$	$\hat{t}_y^{R}[4]$	$\hat{t}_y^{R}[5]$	$\hat{t}_y^{R}[6]$	$\hat{t}_y^{R}[7]$
mixlogn1	25	-0.2	-16.3	-3.9	0.3	0.7	1.1	1.1	1.2
		(100)	(35)	(81)	(102)	(105)	(108)	(108)	(108)
mixlogn1	50	0.1	-15.5	-3.0	-0.6	0.1	0.4	0.5	0.5
		(100)	(39)	(84)	(94)	(99)	(101)	(102)	(102)
mixlogn1	100	0.2	-13.7	-3.1	-1.5	-0.2	0.1	0.2	0.2
		(100)	(50)	(82)	(87)	(96)	(98)	(100)	(100)
mixlogn2	25	0.4	-27.2	-8.5	-3.1	0.0	1.8	2.4	3.0
		(100)	(48)	(82)	(94)	(100)	(103)	(105)	(107)
mixlogn2	50	1.8	-21.7	-7.9	-4.8	-1.1	-0.1	1.3	1.7
		(100)	(55)	(80)	(89)	(95)	(96)	(99)	(100)
mixlogn2	100	-0.5	-17.3	-8.6	-7.2	-4.6	-4.2	-2.7	-2.4
		(100)	(68)	(85)	(91)	(94)	(94)	(96)	(96)
mixlogn3	25	-0.6	-24.9	-10.3	-6.8	-3.1	-1.3	0.0	1.1
		(100)	(60)	(84)	(91)	(97)	(99)	(102)	(104)
mixlogn3	50	1.5	-17.6	-7.8	-5.9	-2.9	-2.3	-0.7	-0.3
		(100)	(67)	(84)	(89)	(94)	(95)	(97)	(97)
mixlogn3	100	0.5	-12.7	-6.5	-5.5	-3.7	-3.4	-2.5	-2.3
		(100)	(77)	(87)	(90)	(95)	(96)	(97)	(97)

* La valeur de k est indiquée entre crochets.

TABLEAU 3.9. Biais et efficacité relative de l'estimateur obtenu en minimisant des percentiles pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[75]$	$\hat{t}_y^{R}[90]$	$\hat{t}_y^{R}[95]$	$\hat{t}_y^{R}[97.5]$	$\hat{t}_y^{R}[99]$
mixlogn1	25	-0.2	-16.3	1.3	1.2	0.5	-16.3	-16.3
		(100)	(35)	(108)	(108)	(103)	(35)	(35)
mixlogn1	50	0.1	-15.5	0.8	0.7	0.6	-0.5	-15.5
		(100)	(39)	(103)	(103)	(103)	(95)	(39)
mixlogn1	100	0.2	-13.7	0.3	0.3	0.3	0.1	-3.1
		(100)	(50)	(100)	(100)	(100)	(99)	(82)
mixlogn2	25	0.4	-27.2	3.5	2.0	-3.0	-27.2	-27.2
		(100)	(48)	(108)	(104)	(94)	(48)	(48)
mixlogn2	50	1.8	-21.7	3.3	3.0	0.1	-4.7	-21.7
		(100)	(55)	(104)	(103)	(96)	(89)	(55)
mixlogn2	100	-0.5	-17.3	-0.2	-0.2	-0.8	-4.2	-8.6
		(100)	(68)	(101)	(101)	(99)	(94)	(85)
mixlogn3	25	-0.6	-24.9	2.9	-1.2	-6.7	-24.9	-24.9
		(100)	(60)	(108)	(100)	(91)	(60)	(60)
mixlogn3	50	1.5	-17.6	3.2	1.7	-2.2	-5.8	-17.6
		(100)	(67)	(104)	(100)	(95)	(89)	(67)
mixlogn3	100	0.5	-12.7	1.1	0.6	-1.0	-3.4	-6.5
		(100)	(77)	(101)	(100)	(100)	(96)	(87)

* La valeur du percentile est indiquée entre crochets.

Le tableau 3.7 donne les mêmes conclusions en terme de comparaison de méthodes que le tableau 3.1. Par contre, avec les populations issues de mélanges de lognormales, le biais est plus élevé et la valeur de l'efficacité relative plus faible par

rapport à celles issues de mélanges de normales. Comme exemple, avec l'estimateur $\hat{t}_y^B$ et quand $n = 25$, le biais et l'efficacité relatives ont donné comme valeur -16.3% et 35% respectivement pour la population mixlogn1, alors que pour la population mixnor1 ils sont de -2.5% et 45% respectivement. Le tableau 3.8, qui est l'équivalent du tableau 3.2 pour des mélanges de lognormales, informe que la méthode du min-max donne toujours les meilleurs résultats quelle que soit la population en terme d'efficacité relative, mais son biais est plus élevé par rapport à l'estimateur qui consiste à minimiser la somme des k plus grands biais conditionnels de l'estimateur robuste. Concernant le tableau 3.9, l'estimateur $\hat{t}_y^R[99]$ est le seul qui donne des résultats proches de ceux du min-max comme observé pour les mélanges de normales.

TABLEAU 3.10. Biais et efficacité relative de l'estimateur (3.2.9) avec q=10, 20, 30, 40, 50 pour le plan de Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[10]$	$\hat{t}_y^R[20]$	$\hat{t}_y^R[30]$	$\hat{t}_y^R[40]$	$\hat{t}_y^R[50]$
mixlogn1	25	-1.4	-20.1	-17.3	-18.8	-19.3	-19.5	-19.7
		(100)	(35)	(43)	(38)	(37)	(36)	(36)
mixlogn1	50	0.9	-16.6	-13.2	-15.0	-15.6	-15.9	-16.0
		(100)	(40)	(49)	(44)	(42)	(42)	(41)
mixlogn1	100	-0.1	-14.6	-11.2	-13.0	-13.5	-13.8	-14.0
		(100)	(52)	(60)	(55)	(54)	(53)	(53)
mixlogn2	25	-1.7	-31.0	-26.2	-28.7	-29.5	-29.9	-30.2
		(100)	(47)	(53)	(50)	(49)	(49)	(48)
mixlogn2	50	-1.1	-24.9	-20.1	-22.6	-23.5	-23.8	-24.1
		(100)	(58)	(64)	(61)	(60)	(59)	(59)
mixlogn2	100	-0.6	-17.8	-13.7	-15.9	-16.6	-16.9	-17.1
		(100)	(69)	(73)	(71)	(70)	(70)	(69)
mixlogn3	25	-0.6	-27.7	-23.3	-25.6	-26.4	-26.7	-26.9
		(100)	(59)	(63)	(61)	(60)	(60)	(60)
mixlogn3	50	-0.3	-20.5	-16.4	-18.5	-19.2	-19.6	-19.7
		(100)	(69)	(72)	(70)	(70)	(70)	(69)
mixlogn3	100	-0.0	-13.8	-10.5	-12.2	-12.8	-13.0	-13.2
		(100)	(78)	(81)	(79)	(79)	(79)	(78)

* La valeur de q est indiquée entre crochets.

TABLEAU 3.11. Biais et efficacité relative de l'estimateur (3.2.11) avec k=2, ..., 7 pour le plan de Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[2]$	$\hat{t}_y^R[3]$	$\hat{t}_y^R[4]$	$\hat{t}_y^R[5]$	$\hat{t}_y^R[6]$	$\hat{t}_y^R[7]$
mixlogn1	25	-1.4	-20.1	-5.6	-4.6	-4.0	-4.0	-3.9	-4.0
		(100)	(35)	(91)	(96)	(99)	(100)	(100)	(100)
mixlogn1	50	0.9	-16.6	-2.5	-1.4	-0.5	-0.5	-0.4	-0.4
		(100)	(40)	(89)	(92)	(98)	(99)	(100)	(100)
mixlogn1	100	-0.1	-14.6	-3.1	-2.7	-1.2	-0.9	-0.8	-0.7
		(100)	(52)	(84)	(85)	(95)	(98)	(100)	(100)
mixlogn2	25	-1.7	-31.0	-9.9	-8.7	-4.6	-4.0	-3.0	-3.0
		(100)	(47)	(85)	(87)	(95)	(96)	(99)	(99)
mixlogn2	50	-1.1	-24.9	-10.2	-9.2	-5.1	-4.6	-2.9	-2.7
		(100)	(58)	(83)	(87)	(92)	(93)	(96)	(97)
mixlogn2	100	-0.6	-17.8	-8.6	-8.1	-5.4	-5.2	-3.5	-3.4
		(100)	(69)	(88)	(89)	(92)	(92)	(94)	(94)
mixlogn3	25	-0.6	-27.7	-11.2	-10.3	-5.6	-5.1	-3.0	-2.8
		(100)	(59)	(84)	(85)	(92)	(92)	(96)	(96)
mixlogn3	50	-0.3	-20.5	-10.0	-9.3	-6.1	-5.8	-4.0	-3.8
		(100)	(69)	(86)	(87)	(92)	(92)	(94)	(94)
mixlogn3	100	-0.0	-13.8	-7.1	-6.7	-4.9	-4.8	-3.8	-3.7
		(100)	(78)	(89)	(90)	(94)	(94)	(95)	(95)

* La valeur de k est indiquée entre crochets.

TABLEAU 3.12. Biais et efficacité relative de l'estimateur obtenu en minimisant des percentiles pour le plan de Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[75]$	$\hat{t}_y^R[90]$	$\hat{t}_y^R[95]$	$\hat{t}_y^R[97.5]$	$\hat{t}_y^R[99]$
mixlogn1	25	-1.4	-20.1	-3.8	-3.9	-5.7	-20.0	-20.1
		(100)	(35)	(100)	(100)	(92)	(35)	(35)
mixlogn1	50	0.9	-16.6	-0.3	-0.3	-0.4	-2.3	-16.6
		(100)	(40)	(100)	(100)	(99)	(89)	(40)
mixlogn1	100	-0.1	-14.6	-0.6	-0.7	-0.7	-0.9	-8.1
		(100)	(52)	(100)	(100)	(100)	(98)	(72)
mixlogn2	25	-1.7	-31.0	-2.6	-4.0	-10.9	-30.9	-31.0
		(100)	(47)	(100)	(97)	(84)	(48)	(47)
mixlogn2	50	-1.1	-24.9	-1.6	-1.8	-4.6	-10.2	-24.9
		(100)	(58)	(100)	(99)	(93)	(85)	(58)
mixlogn2	100	-0.6	-17.8	-0.8	-0.8	-1.9	-5.1	-12.7
		(100)	(69)	(100)	(100)	(97)	(92)	(82)
mixlogn3	25	-0.6	-27.7	-1.2	-5.0	-12.1	-27.6	-27.7
		(100)	(59)	(100)	(93)	(84)	(59)	(59)
mixlogn3	50	-0.3	-20.5	-0.5	-2.0	-5.8	-10.0	-20.5
		(100)	(69)	(100)	(97)	(93)	(87)	(69)
mixlogn3	100	-0.0	-13.8	-0.1	-0.7	-2.6	-4.7	-10.0
		(100)	(78)	(100)	(98)	(96)	(94)	(87)

* La valeur du percentile est indiquée entre crochets.

L'analyse des tableaux 3.10, 3.11 et 3.12 s'aligne avec celle faite des tableaux 3.7, 3.8 et 3.9.

3.3.3. Populations issues de différentes distributions

Dans cette sous-section, nous avons utilisé trois populations issues de distributions individuelles, sans mélange. La première population vient de la lognormale de moyenne 2 et de variance 1.5, elle sera notée par « lognormal » dans la suite. La deuxième population est obtenue à partir de la distribution de Weibull de paramètres 0.25 et 2.1, cette population sera identifiée par « Weibull ». Quant à la dernière population, elle est issue d'une distribution de Student à 3 degrés de liberté, nous l'identifierons par « Student ».

TABLEAU 3.13. Biais et efficacité relative de l'estimateur (3.2.9) avec q=10, 20, 30, 40, 50 pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^{B}$	$\hat{t}_y^{R}[10]$	$\hat{t}_y^{R}[20]$	$\hat{t}_y^{R}[30]$	$\hat{t}_y^{R}[40]$	$\hat{t}_y^{R}[50]$
lognormal	25	0.6	-12.3	-10.0	-11.3	-11.7	-11.9	-12.0
		(100)	(60)	(65)	(62)	(61)	(61)	(60)
lognormal	50	-0.3	-9.7	-7.7	-8.8	-9.2	-9.3	-9.4
		(100)	(68)	(73)	(70)	(70)	(69)	(69)
lognormal	100	0.1	-6.5	-4.9	-5.8	-6.0	-6.2	-6.3
		(100)	(72)	(77)	(74)	(74)	(73)	(73)
Weibull	25	0.9	-27.1	-21.8	-24.6	-25.5	-25.9	-26.1
		(100)	(48)	(55)	(51)	(50)	(50)	(49)
Weibull	50	0.4	-22.3	-17.3	-19.9	-20.7	-21.1	-21.4
		(100)	(57)	(63)	(60)	(59)	(58)	(58)
Weibull	100	0.1	-16.4	-12.2	-14.4	-15.1	-15.5	-15.7
		(100)	(68)	(73)	(70)	(69)	(69)	(68)
Student	25	0.0	-36.4	-29.5	-33.1	-34.2	-34.8	-35.1
		(100)	(55)	(61)	(58)	(57)	(57)	(56)
Student	50	0.8	-29.8	-23.0	-26.6	-27.7	-28.3	-28.6
		(100)	(71)	(75)	(73)	(72)	(72)	(72)
Student	100	1.0	-20.7	-15.1	-18.0	-18.9	-19.3	-19.6
		(100)	(88)	(88)	(88)	(88)	(88)	(88)

* La valeur de q est indiquée entre crochets.

TABLEAU 3.14. Biais et efficacité relative de l'estimateur (3.2.11) avec k=2, ..., 7 pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[2]$	$\hat{t}_y^R[3]$	$\hat{t}_y^R[4]$	$\hat{t}_y^R[5]$	$\hat{t}_y^R[6]$	$\hat{t}_y^R[7]$
lognormal	25	0.6	-12.3	-5.4	-3.2	-1.4	-0.6	0.3	0.8
		(100)	(60)	(84)	(96)	(99)	(101)	(102)	(103)
lognormal	50	-0.3	-9.7	-5.0	-4.1	-2.8	-2.5	-1.8	-1.6
		(100)	(68)	(90)	(95)	(97)	(97)	(99)	(99)
lognormal	100	0.1	-6.5	-3.2	-2.8	-2.0	-1.9	-1.5	-1.4
		(100)	(72)	(92)	(95)	(97)	(97)	(98)	(98)
Weibull	25	0.9	-27.1	-12.4	-7.0	-3.7	-2.2	-1.4	-0.8
		(100)	(48)	(84)	(93)	(100)	(104)	(105)	(106)
Weibull	50	0.4	-22.3	-11.5	-8.1	-4.3	-3.4	-2.3	-2.0
		(100)	(57)	(81)	(85)	(95)	(97)	(99)	(100)
Weibull	100	0.1	-16.4	-8.6	-7.2	-3.6	-3.1	-1.6	-1.4
		(100)	(68)	(81)	(82)	(91)	(92)	(96)	(97)
Student	25	0.0	-36.4	-13.4	-4.5	-0.1	2.4	3.0	3.5
		(100)	(55)	(75)	(86)	(97)	(103)	(105)	(107)
Student	50	0.8	-29.8	-11.9	-8.4	-2.5	-1.1	0.6	1.1
		(100)	(71)	(76)	(77)	(89)	(93)	(98)	(100)
Student	100	1.0	-20.7	-11.0	-10.0	-4.8	-3.9	-1.3	-0.9
		(100)	(88)	(80)	(79)	(84)	(85)	(91)	(93)

* La valeur de k est indiquée entre crochets.

TABLEAU 3.15. Biais et efficacité relative de l'estimateur obtenu en minimisant des percentiles pour le plan aléatoire simple sans remise.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[75]$	$\hat{t}_y^R[90]$	$\hat{t}_y^R[95]$	$\hat{t}_y^R[97.5]$	$\hat{t}_y^R[99]$
lognormal	25	0.6	-12.3	2.7	-0.5	-3.2	-12.3	-12.3
		(100)	(60)	(107)	(101)	(96)	(60)	(60)
lognormal	50	-0.3	-9.7	0.7	-0.6	-2.5	-4.1	-9.7
		(100)	(68)	(103)	(100)	(97)	(95)	(68)
lognormal	100	0.1	-6.5	0.5	-0.1	-0.8	-1.9	-3.2
		(100)	(72)	(101)	(100)	(99)	(97)	(91)
Weibull	25	0.9	-27.1	2.7	-1.4	-5.7	-27.1	-27.1
		(100)	(48)	(107)	(100)	(92)	(48)	(48)
Weibull	50	0.4	-22.3	0.9	-0.8	-4.3	-8.2	-22.3
		(100)	(57)	(104)	(102)	(96)	(87)	(57)
Weibull	100	0.1	-16.4	-0.4	-1.1	-2.3	-5.0	-8.7
		(100)	(68)	(101)	(100)	(99)	(93)	(81)
Student	25	0.0	-36.4	4.1	2.7	-4.2	-36.4	-36.4
		(100)	(55)	(108)	(104)	(86)	(55)	(55)
Student	50	0.8	-29.8	2.7	2.4	-0.9	-8.3	-29.8
		(100)	(71)	(104)	(103)	(93)	(77)	(71)
Student	100	1.0	-20.7	1.5	1.4	0.9	-3.9	-11.0
		(100)	(88)	(101)	(101)	(99)	(85)	(80)

* La valeur du percentile est indiquée entre crochets.

TABLEAU 3.16. Biais et efficacité relative de l'estimateur (3.2.9) avec q=10, 20, 30, 40, 50 pour le plan de Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[10]$	$\hat{t}_y^R[20]$	$\hat{t}_y^R[30]$	$\hat{t}_y^R[40]$	$\hat{t}_y^R[50]$
lognormal	25	0.1	-15.8	-13.7	-14.9	-15.3	-15.5	-15.6
		(100)	(64)	(68)	(66)	(65)	(65)	(65)
lognormal	50	0.7	-10.5	-8.5	-9.6	-9.9	-10.1	-10.2
		(100)	(69)	(73)	(71)	(70)	(70)	(70)
lognormal	100	-0.1	-7.5	-5.9	-6.7	-7.0	-7.1	-7.2
		(100)	(77)	(80)	(78)	(77)	(77)	(77)
Weibull	25	0.2	-28.6	-23.9	-26.4	-27.2	-27.6	-27.8
		(100)	(54)	(59)	(56)	(55)	(55)	(55)
Weibull	50	0.2	-21.7	-17.3	-19.6	-20.3	-20.7	-20.9
		(100)	(65)	(69)	(66)	(66)	(65)	(65)
Weibull	100	0.1	-15.3	-11.6	-13.5	-14.1	-14.4	-14.6
		(100)	(76)	(79)	(77)	(77)	(77)	(76)
Student	25	2.2	-37.8	-30.9	-34.5	-35.7	-36.2	-36.5
		(100)	(52)	(57)	(55)	(54)	(53)	(53)
Student	50	1.6	-30.4	-23.7	-27.2	-28.3	-28.8	-29.2
		(100)	(69)	(73)	(71)	(70)	(70)	(70)
Student	100	0.1	-22.1	-16.7	-19.5	-20.4	-20.8	-21.1
		(100)	(88)	(88)	(88)	(88)	(88)	(88)

* La valeur de q est indiquée entre crochets.

TABLEAU 3.17. Biais et efficacité relative de l'estimateur (3.2.11) avec k=2, ..., 7 pour le plan de Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[2]$	$\hat{t}_y^R[3]$	$\hat{t}_y^R[4]$	$\hat{t}_y^R[5]$	$\hat{t}_y^R[6]$	$\hat{t}_y^R[7]$
lognormal	25	0.1	-15.8	-7.9	-7.5	-5.3	-5.0	-3.8	-3.6
		(100)	(64)	(88)	(90)	(94)	(94)	(96)	(96)
lognormal	50	0.7	-10.5	-5.1	-4.9	-3.5	-3.3	-2.5	-2.4
		(100)	(69)	(91)	(92)	(94)	(94)	(95)	(96)
lognormal	100	-0.1	-7.5	-3.9	-3.8	-3.0	-2.9	-2.5	-2.4
		(100)	(77)	(94)	(95)	(96)	(96)	(97)	(97)
Weibull	25	0.3	-28.6	-10.0	-9.0	-4.2	-4.0	-2.2	-2.0
		(100)	(55)	(83)	(84)	(93)	(94)	(97)	(97)
Weibull	50	0.2	-21.3	-8.8	-8.0	-4.4	-4.1	-2.4	-2.3
		(100)	(66)	(82)	(83)	(91)	(92)	(95)	(95)
Weibull	100	0.1	-15.4	-7.3	-7.2	-4.6	-4.4	-3.2	-3.1
		(100)	(79)	(85)	(85)	(91)	(91)	(94)	(95)
Student	25	2.2	-37.8	-7.6	-5.5	0.3	0.6	1.7	1.8
		(100)	(52)	(78)	(80)	(94)	(95)	(99)	(99)
Student	50	1.7	-30.4	-10.4	-9.2	-2.3	-1.8	0.4	0.5
		(100)	(69)	(75)	(75)	(88)	(89)	(96)	(96)
Student	100	0.1	-22.1	-11.8	-11.6	-6.0	-5.6	-2.8	-2.6
		(100)	(88)	(79)	(78)	(84)	(84)	(91)	(91)

* La valeur de k est indiquée entre crochets.

TABLEAU 3.18. Biais et efficacité relative de l'estimateur obtenu en minimisant des percentiles pour le plan de Bernoulli.

population	n	$\hat{t}_y^{HT}$	$\hat{t}_y^B$	$\hat{t}_y^R[75]$	$\hat{t}_y^R[90]$	$\hat{t}_y^R[95]$	$\hat{t}_y^R[97.5]$	$\hat{t}_y^R[99]$
lognormal	25	0.1	-15.8	-1.9	-4.9	-8.3	-15.8	-15.8
		(100)	(64)	(99)	(95)	(90)	(65)	(64)
lognormal	50	0.7	-10.5	-0.2	-1.5	-3.3	-5.2	-10.5
		(100)	(69)	(99)	(97)	(95)	(92)	(69)
lognormal	100	-0.1	-7.5	-0.5	-1.1	-1.9	-2.9	-5.4
		(100)	(77)	(100)	(99)	(98)	(97)	(91)
Weibull	25	0.2	-28.4	0.3	-3.6	-11.1	-28.4	-28.4
		(100)	(55)	(100)	(93)	(81)	(55)	(55)
Weibull	50	0.2	-20.9	0.5	-1.0	-4.5	-9.6	-20.9
		(100)	(66)	(100)	(97)	(90)	(83)	(66)
Weibull	100	0.1	-14.4	0.1	-0.5	-2.0	-4.8	-10.8
		(100)	(80)	(100)	(99)	(96)	(91)	(89)
Student	25	2.2	-37.8	2.2	0.6	-8.8	-37.6	-37.8
		(100)	(52)	(100)	(96)	(78)	(52)	(52)
Student	50	1.7	-30.4	1.7	1.4	-1.8	-10.4	-30.4
		(100)	(69)	(100)	(99)	(90)	(75)	(69)
Student	100	0.1	-22.1	0.1	0.0	-0.8	-5.5	-16.7
		(100)	(88)	(100)	(100)	(97)	(85)	(84)

* La valeur du percentile est indiquée entre crochets.

Nous retenons des tableaux 3.13 à 3.18 que les résultats obtenus vont généralement dans le même sens que ceux des tableaux précédemment analysés. Nous allons donc procéder dans la section 3.4 à une analyse globale des résultats.

3.4. ANALYSE DES RÉSULTATS

La première observation que nous tirons de ces résultats est que, de façon générale, lorsque la taille de l'échantillon augmente, le biais relatif diminue et la valeur de l'efficacité relative augmente.

Lorsque nous comparons les différents critères examinés entre eux, nous remarquons que, de façon générale, le min-max donne les meilleurs résultats en termes d'efficacité relative. Par contre son biais relatif est toujours plus élevé que celui des autres critères. C'est ce que nous observons dans la plupart des tableaux, en particulier dans les tableaux 3.16 et 3.17. L'exception à cette règle s'observe au niveau des tableaux 3.2 et 3.5 pour les populations mixnor2 et mixnor3, autrement dit lorsque la taille de l'échantillon et/ou la proportion d'outliers augmente. En effet, pour ces deux populations, nous remarquons que par moment, des estimateurs issus du critère (3.2.11) ont une valeur de l'efficacité relative plus faible que celle du min-max. Par exemple, dans le tableau 3.5, la valeur obtenue pour l'efficacité relative est 83% alors que celle pour l'estimateur $\hat{t}_y^B$ est 97%, lorsqu'on prend des échantillons de taille 50 tirés de la population mixnor2. De plus pour

cet exemple, malgré que min-max a la plus grande valeur de l'efficacité relative, il a aussi le biais relatif le plus élevé. Cependant, lorsqu'on jette un coup d'oeil aux tableaux 3.8 et 3.11, qui évaluent le même critère pour le même plan que ceux évalués dans les tableaux 3.2 et 3.5, respectivement, mais pour un mélange de lognormales, le min-max a une valeur de l'efficacité relative plus faible. Cette différence peut s'expliquer en partie du fait que les différentes populations issues des mélanges de lognormales ont donné des distributions asymétriques, alors que ceux de normales donnent des distributions « bimodales », en ce sens qu'il y a une séparation entre la majorité des observations et la minorité d'outliers. Donc le fait de minimiser le maximum seulement lorsque la proportion d'outliers est importante, a moins d'effet que lorsque l'on minimise les k plus grands.

Concernant le critère donné par (3.2.9), nous remarquons que quel que soit le plan de sondage, la population ou la taille, la valeur de l'efficacité relative diminue lorsque q augmente, donc ses résultats s'améliorent lorsque q augmente. En effet, lorsque $q = 50$ ce critère donne des résultats très similaires à ceux obtenus par le critère min-max. Le tableau 3.13 et d'autres tableaux relatifs à ce critère confirment cette affirmation. Par contre, au moment où l'efficacité relative diminue avec q, le biais augmente. En effet, dans le tableau 3.16 par exemple, pour l'échantillon de taille 100 issu de la population de Weibull, nous remarquons que le biais est passé de -11.6% lorsque $q = 10$ à -14.6% quand $q = 50$. C'est le compromis à faire entre le biais et l'efficacité relative.

Pour la méthode qui consiste à minimiser les percentiles, le 99ème percentile donne généralement des résultats souvent très proches à ceux du min-max et ceci quel que soit le plan, la population et la taille de l'échantillon. En effet, lorsqu'on regarde le tableau 3.6, on voit que ces deux critères donnent des résultats très similaires. La minimisation du 97.5ème percentile donne des résultats identiques à ceux du min-max lorsque la taille de l'échantillon est égale à 25. Par contre dès que la taille commence à augmenter, ses résultats s'éloignent progressivement de ceux du min-max. Nous pouvons confirmer cette affirmation en se référant au tableau 3.15, dans lequel, pour la population lognormale, lorsque $n = 25$, les deux critères donnent une valeur de l'efficacité relative égale à 60% et un biais relatif de -12.3%. Quand la taille devient 50 ou 100, la valeur de l'efficacité relative pour l'estimateur $\hat{t}_y^R[97.5]$ passe à 95% ou 97% respectivement et celle du biais passe à -4.1% ou -1.9% respectivement, alors que celles du min-max sont à 68% ou 72% respectivement pour l'efficacité et -9.7% ou -6.5% respectivement pour le biais.

Chapitre 4

APPLICATION DU CRITÈRE MIN-MAX À L'ENQUÊTE SUR LES DÉPENSES DES MÉNAGES AU CANADA 2015

4.1. Introduction

Dans ce chapitre, nous allons appliquer le critère min-max à l'enquête sur les dépenses des ménages de Statistique Canada. Ce travail, qui a été effectué sous forme de stage à Statistique Canada, a eu lieu entre le 19 septembre et le 16 décembre 2016. Les données utilisées sont celles collectées par l'agence en 2015 auprès des ménages canadiens.

L'enquête sur les dépenses des ménages (EDM) recueille principalement des renseignements détaillés sur les dépenses des ménages au Canada. L'enquête collecte également le revenu annuel des membres du ménage (provenant de fichiers de données d'impôt des particuliers), les caractéristiques démographiques du ménage, certaines caractéristiques du logement (telles que le type, l'âge et le mode d'occupation du logement) ainsi que certains renseignements sur l'équipement que possède le ménage (par exemple, l'équipement électronique et le matériel de communication). L'enquête est effectuée annuellement dans les 10 provinces et généralement aux deux ans dans les territoires.

L'EDM combine l'utilisation d'un questionnaire et d'un journal de dépenses que les ménages sélectionnés remplissent pendant une période de deux semaines suivant l'entrevue. Les dépenses collectées pendant l'entrevue ont des périodes de références différentes, par exemple 1, 3 et 12 mois ou dernier paiement, etc... Un exemple de dépense que l'EDM collecte est les frais de scolarité des écoles maternelles, primaires et secondaires (ED004).

Les distributions des variables d'intérêts de l'EDM sont asymétriques et des valeurs extrêmes peuvent être observées pour certaines dépenses. Toutes les variables de dépenses de l'entrevue et du journal sont annualisées en multipliant par un facteur approprié selon la période de référence, ce qui peut amplifier l'impact des valeurs extrêmes. Il est donc important d'avoir un mécanisme en place pour identifier les valeurs extrêmes et ensuite les corriger. Dans ce chapitre, nous utiliserons le concept du biais conditionnel pour identifier et traiter les données influentes, ensuite nous expliquerons comment la théorie pour ce concept a été appliquée en pratique aux données de l'EDM. Les résultats seront présentés et comparés aux résultats de la méthode usuelle utilisée en production à Statistique Canada.

Dans cette application, nous avons utilisé l'estimateur robuste obtenu en minimisant le maximum du biais conditionnel de l'estimateur robuste selon Beaumont et al. (2013), présenté à la section 2.3.

4.2. Plan de sondage de l'EDM

L'échantillon de l'enquête sur les dépenses des ménages de 2015 est constitué de 17 603 ménages répartis dans les 10 provinces [1]. Cet échantillon a été sélectionné selon un plan de sondage stratifié à deux degrés; le premier degré est un échantillon d'aires géographiques (appelées grappes). La liste de tous les logements se trouvant dans les grappes sélectionnées est ensuite établie pour permettre, au deuxième degré, la sélection d'un échantillon de logements. Les logements choisis qui sont habités par des individus de la population cible constituent l'échantillon de ménages de l'enquête. L'enquête utilise plusieurs composantes du plan de l'enquête sur la population active (EPA) dans le but de minimiser les coûts d'opération, mais les logements sélectionnés pour l'EDM sont différents de ceux sélectionnés pour l'EPA.

La moitié des ménages sélectionnés doit aussi remplir un journal de dépenses. Ainsi, dans chacune des grappes sélectionnées, un sous-échantillon des logements préalablement choisis est également sélectionné en vue d'identifier les logements pour lesquels les ménages auront à remplir le journal. L'évaluation décrite dans ce document ne porte que sur la portion de l'entrevue, c'est-à-dire que les estimateurs robustes ne seront calculés que sur des dépenses mesurées à l'entrevue.

1. L'EDM 2015 couvre également les trois capitales des territoires. Les résultats de cette évaluation ne porteront que sur l'échantillon dans les provinces.

4.2.1. Plan de sondage au premier et au deuxième degré

Au premier degré, l'échantillon de grappes de l'EDM est sélectionné selon un plan de Rao-Hartley-Cochran; voir Rao et al. (1962). Ce plan est un plan de sondage à taille fixe. Supposons que r représente la taille de l'échantillon que nous aimerions tirer au premier degré. Dans un premier temps, on subdivise de façon aléatoire la population en r groupes et, dans un second temps, on tire une et une seule unité dans chaque groupe.

Soit p_g la probabilité de sélection de l'unité g de la population avant la subdivision; par exemple, en utilisant un plan proportionnel à la taille, p_g sera de la forme

$$p_g = \frac{x_g}{\sum\limits_{l \in U} x_l},$$

où x est la mesure de taille et U représente la population de grappes.

Si g appartient au groupe G, sa probabilité d'inclusion sera donnée par

$$\pi_g = \frac{p_g}{\sum\limits_{l \in G} p_l}.$$

Pour le plan de l'EPA, il y a six groupes dans chaque strate, et ces groupes sont appelés groupes de rotation. En effet, les logements restent dans l'échantillon pendant six mois consécutifs, et chaque mois un sixième de l'échantillon est remplacé.
Nous avons approximé les plans de sondage au premier et au deuxième degré pour simplifier le calcul des probabilités d'inclusion doubles. Au premier degré, nous avons approximé le plan de Rao-Hartley-Cochran par le plan de Poisson.
Au second degré, les logements sont sélectionnés selon un plan systématique. Comme pour le plan au premier degré, ce plan a été approximé par un plan aléatoire simple sans remise.

Il faut noter que ces approximations ont été faites pour des raisons de simplicité.

4.2.2. Calcul des probabilités d'inclusion

Dans cette section, nous expliquerons comment les probabilités d'inclusion simple et double ont été calculées et les hypothèses qui ont été nécessaires à ces calculs.

4.2.2.1. *Calcul des probabilités d'inclusion simple*

La formule du biais conditionnel requiert le calcul des probabilités d'inclusion simple et double, aux deux degrés de l'enquête ainsi qu'à la phase de non-réponse. Cette section décrira comment nous avons procédé au calcul des probabilités d'inclusion simple.

La probabilité d'inclusion d'un ménage dans l'échantillon est calculée selon le plan d'échantillonnage de l'EPA. Ce poids, appelé poids théorique (*theor_wt*), doit être ajusté pour tenir compte des modifications au plan pour les fins de l'EDM. Ces modifications découlent principalement du fait que l'échantillon de l'EDM est beaucoup plus petit que celui de l'EPA ($\approx$ 20 000 -vs- $\approx$ 54 000 logements). En effet, l'échantillon fourni par l'EPA suppose que le plan de l'EDM est le même que celui de l'EPA ce qui n'est pas le cas.

On a donc besoin de réduire la taille initiale de l'échantillon fourni par l'EPA pour satisfaire aux besoins de l'EDM. De plus, étant donné que le taux d'échantillonnage est fixe au niveau des grappes, deux méthodes sont utilisées afin de contrôler la taille d'échantillon: le sous-échantillonnage de grappes et la stabilisation.
Le sous-échantillonnage de grappes est effectué lorsque le listage des logements dans une grappe donnée, donne un nombre de logements largement supérieur à ce qui était prévu. Dans un tel cas, la grappe en question est divisée en sous-grappes et on sélectionne un échantillon de ces sous-grappes, ce qui crée un degré supplémentaire de sélection, induisant ainsi un facteur (*cluster_wt*) dans le calcul du poids initial (*init_wt*) dans l'échantillon de l'EDM.
Pour réduire la taille obtenue de l'EPA, l'EDM choisit aussi moins de rotations dans une strate que le nombre de rotations choisies par l'EPA. Malgré ce fait, il arrive que trop de logements soient sélectionnés, on procède donc à un sous-échantillonnage de logements sélectionnés. Ce sous-échantillonnage est appelé la stabilisation et induit le facteur d'ajustement *stab_wt* dans le calcul du poids initial.
Le poids initial (*init_wt*) correspond donc au produit du poids théorique et de tous les poids tenant compte des changements au plan de sondage de l'EPA

$$init\_wt = theor\_wt \times cluster\_wt \times rot\_wt \times stab\_wt,$$

où *rot_wt* est un facteur d'ajustement pour tenir compte de la sélection d'un sous-ensemble de rotations.

Ce poids théorique tient compte des deux premiers degrés de l'enquête. Pour les besoins de cette évaluation, nous avons demandé à l'EPA de séparer le terme $theor\_wt$ en deux composantes, pour chacun des deux degrés. Pour chaque grappe, nous avons alors obtenu la probabilité d'inclusion au premier degré, psu_{prob}, et la probabilité d'inclusion au deuxième degré pour les logements de la grappe, $stg2_{prob}$. On a alors que

$$theor\_wt = \frac{1}{psu_{prob}} \times \frac{1}{stg2_{prob}}.$$

Ensuite, il a fallu ajuster ces probabilités d'inclusion pour tenir compte du plan de l'EDM. Nous avons alors défini la probabilité d'inclusion au premier et deuxième degré, respectivement, comme étant

$$\pi_g = psu_{prob} \times \frac{1}{cluster\_wt} \times \frac{1}{rot\_wt},$$

pour la grappe g, et

$$\pi_{i|g} = stg2_{prob} \times \frac{1}{stab\_wt},$$

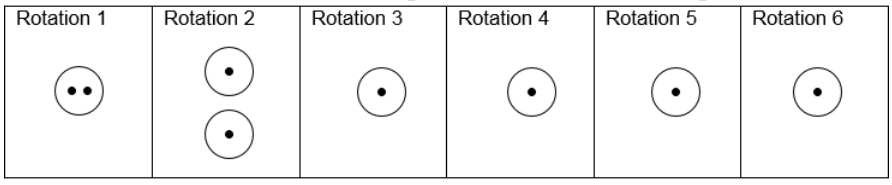
pour l'unité i dans la grappe g.

Le poids initial de l'unité i dans la grappe g, $init\_wt$, est donc donné par

$$init\_wt = \pi_g^{-1} \times \pi_{i|g}^{-1}.$$

Il peut arriver que deux départs aléatoires (c'est-à-dire deux échantillons de deuxième degré) soient sélectionnés dans la même grappe. En 2015, 27 strates sur 1 038 contiennent au moins une grappe dans cette situation. Il s'agit de la situation illustrée dans le premier groupe de rotation dans la figure ci-dessous, où le cercle représente une grappe et le point dans le cercle est un départ aléatoire sélectionné.

FIGURE 4.1. Exemple de sélection de départ



Pour ces cas, le poids de la rotation, rot_wt, sera de $\frac{1}{2}$. Le terme rot_wt apparaît normalement dans la probabilité d'inclusion du premier degré. Par contre, pour les fins de cette évaluation, nous avons jugé préférable de tenir compte de ce facteur dans la probabilité d'inclusion du deuxième degré puisque ce sont les logements dans la grappe qui ont une plus grande probabilité d'être sélectionnés (et

non pas la grappe). Après avoir identifié tous les logements dans cette situation, nous avons transféré le facteur de $\frac{1}{2}$ de π_g à $\pi_{i|g}$, afin que ces deux probabilités reflètent davantage le plan de sondage aux deux degrés.

4.2.2.2. *Calcul des probabilités d'inclusion double*

La taille d'échantillon voulue pour l'EDM est moins grande que celle de l'EPA. Alors que l'EPA utilise en général six groupes de rotation par strate, l'EDM en utilise généralement entre un et trois. Un ajustement est appliqué au poids théorique pour tenir compte de ce fait. Sous un tel plan, il est très complexe de calculer les probabilités d'inclusion double au premier degré. Nous avons donc supposé que les grappes sont choisies indépendamment à l'intérieur d'une strate. Ceci nous semblait une hypothèse acceptable puisque suite à l'assignation aléatoire des grappes aux groupes de rotation, une seule grappe est sélectionnée dans chaque groupe de rotation (et la sélection dans un groupe de rotation est faite de façon indépendante d'un groupe de rotation à l'autre). Cette hypothèse d'indépendance revient donc à supposer un plan de sondage de Poisson au premier degré.

Cependant, cette indépendance n'est pas entièrement satisfaite pour deux raisons: d'une part, les grappes sont assignées de façon aléatoire aux groupes de rotation. D'autre part, il arrive que l'on sélectionne deux grappes dans un même groupe de rotation au lieu d'une. Ainsi:

— l'assignation aléatoire des grappes au groupe de rotation fait que l'hypothèse d'indépendance n'est pas entièrement satisfaite. En effet, la création des groupes de rotation étant aléatoire, deux grappes peuvent se retrouver dans un même groupe de rotation. Étant donné que l'on en sélectionne une seule par groupe de rotation, la probabilité de sélection de l'une dépendra de celle de l'autre. Conditionnellement aux groupes de rotation, la sélection des grappes est indépendante mais pas inconditionnellement;

— la sélection de deux grappes par groupe de rotation n'est pas chose fréquente. En effet, pour l'année 2015, nous avons dénombré 12 grappes dans de tels cas pour un total de 2601 grappes.

Après la sélection des grappes au premier degré, on sélectionne les logements selon un plan systématique. Dans l'estimation du biais conditionnel, nous avons besoin des probabilités d'inclusion double au deuxième degré. Or l'échantillonnage systématique est tel que certaines probabilités d'inclusion double sont nulles. Pour remédier à cette situation, nous avons supposé un échantillonnage aléatoire simple

sans remise au deuxième degré. Cette hypothèse n'a pas été suffisante pour régler le problème pour toutes les grappes. En effet, de certaines grappes (qui sont au nombre de 2), seul un logement a été sélectionné, ce qui fait que la probabilité d'inclusion double au deuxième degré est égale à zéro ($\pi_{ij|g} = \frac{n_g}{N_g}\frac{n_g-1}{N_g-1} = 0$). Dans un tel cas, nous avons fait une autre approximation en posant

$$\pi_{ij|g} \approx \left(\frac{n_g}{N_g}\right)^2.$$

Les hypothèses ci-dessus nous ont permis de calculer plus facilement les probabilités d'inclusion d'ordre deux.

Pour le premier degré, étant donné deux grappes g et l, on a:

$$\pi_{gl} = \begin{cases} \pi_g, & \text{si} \quad l = g \\ \pi_g\pi_l, & \text{sinon.} \end{cases}$$

Au second degré, pour deux logements i et j dans la grappe g, avec l'hypothèse d'échantillonnage aléatoire simple sans remise, nous avons:

$$\pi_{ij|g} = \begin{cases} \frac{n_g}{N_g}, & \text{si} \quad j = i \\ \left(\frac{n_g}{N_g}\right)^2, & \text{si} \quad n_g = 1 \\ \frac{n_g}{N_g}\frac{n_g-1}{N_g-1}, & \text{si} \quad n_g \geq 2. \end{cases}$$

4.3. Estimation

4.3.1. Ajustement de la probabilité de réponse

Comme dans toute enquête, celle-ci a son lot de non-répondants. Le traitement de la non-réponse dans le plan de sondage correspond à une phase additionnelle à partir du deuxième degré, et la probabilité d'inclusion à cette phase est donnée par la probabilité de réponse. On suppose que la probabilité de réponse d'un ménage est indépendante de celle d'un autre, autrement dit, l'ensemble des répondants peut être vu comme un échantillon qui a été sélectionné selon un plan de Poisson avec des probabilités d'inclusion connues.

La méthode courante pour estimer les probabilités de réponse est la méthode du score. Cette méthode consiste dans un premier temps à faire une régression

logistique pour estimer la probabilité de réponse de toute unité et dans un second temps de former les classes d'unités avec probabilités de réponse estimées semblables. On ajuste ensuite les poids des répondants par le taux de réponse pondéré dans la classe à laquelle ils appartiennent.

Par souci de simplicité, dans cette étude, nous avons supposé que les probabilités de réponse sont connues.
La probabilité de réponse de l'unité i appartenant à la grappe g est donnée par

$$\hat{p}_{i|g} = \frac{1}{balfact},$$

où $balfact$ est l'ajustement de non-réponse appliqué au poids initial du ménage. Cet ajustement est égal à l'inverse du taux de réponse pondéré dans la classe de non-réponse. Nous faisons ainsi deux hypothèses:

(1) La probabilité estimée par le modèle de non-réponse, $\hat{p}_{i|g}$ est égale à la vraie probabilité de réponse, $p_{i|g}$.

(2) Les probabilités de réponse de deux ménages sont indépendantes l'une de l'autre: $p_{ij|g} = p_{i|g}p_{j|g}$.

4.3.2. Calcul des résidus pour les variables de dépenses à l'entrevue

Lorsqu'un estimateur de calage est utilisé, tel qu'à l'EDM, la formule d'estimation du biais conditionnel requiert les résidus de calage (4.4.7). Ces résidus doivent être calculés pour toutes les variables de dépense. Ceci est possible en effectuant la régression linéaire entre la variable de dépense (variable dépendante) et les variables de calage (les variables indépendantes). Nous présenterons dans cette section comment cette étape a été effectuée pour l'EDM. Le calage utilisé pour l'enquête est intégré, au sens où les personnes du même ménage ont le même poids de calage et ce poids est le même que le poids du ménage; voir Lemaître et al. (1987).

Les totaux de contrôle des poids de calage sont soit au niveau de la personne (comme par exemple, le groupe d'âge) ou soit au niveau du ménage (comme par exemple, la taille du ménage). Une liste complète des totaux de contrôle utilisés peut être retrouvée dans le guide de l'utilisateur de l'enquête. Nous illustrons le calcul des résidus en utilisant un exemple de totaux de contrôle au niveau personne (groupe d'âge) et au niveau ménage (taille du ménage). D'abord, dans une province donnée, neuf groupes d'âge sont utilisés dans le calage (0-6 ans, 7-17

ans, ..., 65-74 ans, 75+) et trois catégories de tailles de ménage sont utilisées (1, 2, 3+). Pour une personne donnée, nous calculons les variables indicatrices $x_1, x_2, ..., x_{12}$ comme suit:

$x_1 = 1$ si la personne est âgée de 0 à 6 ans et 0 sinon,

$x_2 = 1$ si la personne est âgée de 7 à 17 ans et 0 sinon,

...,

$x_9 = 1$ si la personne est âgée de 75 ans et plus et 0 sinon,

$x_{10} = 1$ si la personne habite dans un ménage de taille 1 et 0 sinon,

$x_{11} = 1$ si la personne habite dans un ménage de taille 2 et 0 sinon,

$x_{12} = 1$ si la personne habite dans un ménage de taille 3 ou plus et 0 sinon.

Nous définissons ensuite les variables z_k, $k = 1, \ldots, 12$, en prenant la somme des x_k de toutes les personnes d'un même ménage et en divisant ensuite par la taille du ménage. Les z_k peuvent donc être interprétés comme une moyenne des x_k dans le ménage, lorsqu'il s'agit de totaux de contrôle au niveau personne. Par contre, pour les totaux de contrôle au niveau des ménages, tels que la taille du ménage, l'interprétation est un peu différente. Ceci est dû au fait que pour ces totaux, les x_k ne sont que définis pour la première personne du ménage et est posée à 0 pour les autres.

Illustrons ceci à l'aide d'un exemple. Considérons un ménage de taille 4, avec deux individus dans le 2e groupe d'âge (7-17 ans) et deux individus dans le 6e groupe d'âge (45-54 ans). Pour ce ménage, les deux enfants auront une valeur de $x_2 = 1$ et $x_k = 0$ pour tous les autres groupes d'âge (c'est à dire, k=1, 3, 4, ..., 9). Les deux parents auront une valeur de $x_6 = 1$ et $x_k = 0$ pour tous les autres groupes d'âge (c'est à dire, k=1, 2, 3, 4, 5, 7, 8, 9). En ce qui concerne les x_k reliés à la taille du ménage (pour k=10, 11, 12), ils seront calculés pour la première personne du ménage dans le fichier. Pour cette personne, on aura $x_{10} = 0, x_{11} = 0$ et $x_{12} = 1$ puisqu'elle habite dans un ménage de taille 4 (qui correspond à la catégorie 3+). Les autres personnes du ménage se feront toutes assigner les valeurs $x_{10} = 0, x_{11} = 0$ et $x_{12} = 0$, par définition. Les z_k sont calculés en prenant la moyenne des x_k dans le ménage. Le tableau ci-dessous résume les résultats des calculs de cet exemple.

TABLEAU 4.1. Exemple de calcul de certaines variables auxiliaires

Personne	x_2 $(7-17)$	x_6 $(45-54)$	x_{12} (taille 3+)	z_2 $(7-17)$	z_6 $(45-54)$	z_{12} (taille 3+)
1	1	0	1	0.5	0.5	0.25
2	1	0	0	0.5	0.5	0.25
3	0	1	0	0.5	0.5	0.25
4	0	1	0	0.5	0.5	0.25

Lorsque les z_k sont calculés, nous ajustons la régression linéaire entre la variable de dépense y et les z_k. Cette régression doit être pondérée par les poids d'entrée au calage (les poids ajustés pour la non-réponse, dans ce cas-ci).

4.4. Biais conditionnels

Le plan de sondage de l'EDM étant un plan relativement complexe, nous avons formulé certaines hypothèses dans le but d'obtenir les expressions du biais conditionnel. Dans cette section, nous allons considérer trois scénarios allant du plus simple au plus complexe pour donner l'expression du biais conditionnel:

(1) le plan à deux degrés avec un taux de réponse de cent pour cent en utilisant l'estimateur Horvitz-Thompson (2.2.13). Dans ce contexte, le résultat est celui de la proposition 2.2.4, c'est-à-dire l'équation (2.2.14);

(2) le plan à deux degrés suivi de la non-réponse, avec probabilité de réponse connue, par rapport à l'estimateur Horvitz-Thompson ajusté pour la non-réponse (4.4.1);

(3) le scénario précédent mais pour lequel l'estimateur de Horvitz-Thompson est remplacé par l'estimateur de calage (4.4.4).

Pour chacun des scénarios (2) et (3), nous obtenons l'expression du biais conditionnel. Celle pour le scénario (1) a déjà été obtenue au chapitre 2 à travers l'équation (2.2.14). La section sera terminée par quelques exemples.

Pour alléger les notations, nous allons considérer le cas d'une seule strate.

4.4.1. Pour le plan à deux degrés suivi de la non-réponse

Supposons que la probabilité de réponse de chaque ménage est connue. L'estimateur ajusté pour la non-réponse est donné par

$$\hat{t}_y^{NR} = \sum_{l\in s} \sum_{j\in s_{lr}} \frac{1}{\pi_l \pi_{j|l} p_{j|l}} y_{lj}, \tag{4.4.1}$$

où s_{lr} est l'ensemble des répondants de l'échantillon tiré de la grappe l au deuxième degré.

Le biais conditionnel de tout ménage i provenant de la grappe g par rapport à l'estimateur (4.4.1) et ayant répondu à l'enquête est donné par

$$B_{1i|g}^{NR} = E_p\left(\hat{t}_y^{NR} - t_y | I_{1g} = 1, I_{2i} = 1, I_{3i} = 1\right),$$

où I_{3i} est la variable indicatrice de réponse, telle que $I_{3i} = 1$ si i est un répondant et $I_{3i} = 0$, sinon.

Proposition 4.4.1. *En tenant compte de l'hypothèse formulée ci-dessus, nous obtenons le résultat suivant dont la démonstration est donnée en annexe*

$$B_{1i|g}^{NR} = \sum_{l\in U}\left(\frac{\pi_{gl}}{\pi_g\pi_l} - 1\right)\sum_{j\in U_l} y_{lj} + \frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right)y_{gj} + \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}} - 1\right)y_{gi}. \tag{4.4.2}$$

On remarquera que l'expression de $B_{1i|g}^{NR}$ s'obtient à partir de celle de l'équation (2.2.13) en y ajoutant un troisième terme qui tient compte de la phase additionnelle de non-réponse.

Proposition 4.4.2. *Un estimateur conditionnellement sans biais de $B_{1i|g}^{NR}$ est:*

$$\hat{B}_{1i|g}^{NR} = \sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_g}{\pi_{gl}}\left(\frac{\pi_{gl}}{\pi_g\pi_l} - 1\right)\sum_{j\in s_{lr}}\frac{y_{lj}}{p_{j|l}\pi_{j|l}} + \sum_{\substack{j\in s_{gr}\\ j\neq i}}\frac{1}{p_{j|g}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{1}{\pi_g}\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right)y_{gj} + \left(\frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\frac{1}{p_{i|g}} - 1\right)y_{gi}. \tag{4.4.3}$$

La démonstration de la proposition 4.4.2 est donnée en annexe.

Dans le cadre de cette évaluation, du fait que le plan de Rao-Hartley-Cochran a été approximé par le plan de Poisson, nous avons $\pi_{gl} = \pi_g\pi_l$, pour tout $l \neq g$. Ainsi, l'estimateur (4.4.3) se réduit à:

$$\hat{B}_{1i|g}^{NR} = \sum_{\substack{j\in s_{gr}\\ j\neq i}}\frac{1}{p_{j|g}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{1}{\pi_g}\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right)y_{gj} + \left(\frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\frac{1}{p_{i|g}} - 1\right)y_{gi}.$$

4.4.2. Pour le plan à deux degrés suivi de la non-réponse utilisant l'estimateur de calage

L'estimateur de calage s'écrit comme

$$\hat{t}_y^{cal} = \hat{t}_y^{NR} + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^{NR}\right)^\top \widehat{\boldsymbol{B}}, \tag{4.4.4}$$

où

$$\widehat{\boldsymbol{B}} = \left(\sum_{l\in s}\sum_{j\in s_{lr}}\frac{1}{\pi_l\pi_{j|l}p_{j|l}}\boldsymbol{x}_j q_j \boldsymbol{x}_j^\top\right)^{-1}\left(\sum_{l\in s}\sum_{j\in s_{lr}}\frac{1}{\pi_l\pi_{j|l}p_{j|l}}\boldsymbol{x}_j q_j y_j\right).$$

Le biais conditionnel dans le cas du calage s'obtient à partir de celui de l'estimateur non-calé en remplaçant les valeurs y par les « résidus » E, de sorte que le biais conditionnel par rapport à $\hat{t}_y^{cal}$ s'écrit comme

$$B_{1i|g}^{cal} = \sum_{l\in U}\left(\frac{\pi_{gl}}{\pi_g\pi_l}-1\right)\sum_{j\in U_l}E_{lj}+\frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)E_{gj}+\frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}}-1\right)E_{gi}, \tag{4.4.5}$$

où

$$E_{gi} = y_{gi} - \boldsymbol{x}_{gi}^\top \boldsymbol{B},$$

et

$$\boldsymbol{B} = \left(\sum_{l\in U}\sum_{j\in U_l}\boldsymbol{x}_j q_j \boldsymbol{x}_j^\top\right)^{-1}\left(\sum_{l\in U}\sum_{j\in U_l}\boldsymbol{x}_j q_j y_j\right).$$

On estime $B_{1i|g}^{cal}$ par

$$\hat{B}_{1i|g}^{cal} = \sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_g}{\pi_{gl}}\left(\frac{\pi_{gl}}{\pi_g\pi_l}-1\right)\sum_{j\in s_{lr}}\frac{e_{lj}}{p_{j|l}\pi_{j|l}}+\sum_{\substack{j\in s_{gr}\\ j\neq i}}\frac{1}{p_{j|g}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{1}{\pi_g}\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)e_{gj} + \left(\frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\frac{1}{p_{i|g}}-1\right)e_{gi}, \tag{4.4.6}$$

où

$$e_{gi} = y_{gi} - \boldsymbol{x}_{gi}^\top \widehat{\boldsymbol{B}}. \tag{4.4.7}$$

Exemple 4.4.3.
Supposons qu'au premier degré nous avons un plan de Poisson et au second degré un plan aléatoire simple sans remise, nous obtenons une expression du biais conditionnel pour chacun des trois scénarios:

(1) Scénario 1:

$$B_{1i|g}^{HT} = \left(\frac{1}{\pi_g}-1\right)Y_g + \frac{1}{\pi_g}\frac{N_g}{N_g-1}\left(\frac{N_g}{n_g}-1\right)\left(y_{gi}-\bar{Y}_g\right),$$

où $\overline{Y}_g$ est la moyenne de la grappe g, N_g la taille de la grappe g et n_g celle de l'échantillon tiré de g, s_g. Ici, on remarque que l'influence d'une unité provenant d'une grappe donnée dépend de l'écart entre sa valeur sur la variable d'intérêt et la moyenne de la grappe. Son estimateur devient

$$\hat{B}_{1i|g}^{HT} = \left(\frac{1}{\pi_g}-1\right)\hat{Y}_g + \frac{1}{\pi_g}\frac{n_g}{n_g-1}\left(\frac{N_g}{n_g}-1\right)\left(y_{gi}-\bar{y}_g\right),$$

où

$$\hat{Y}_g = \sum_{j\in s_g}\frac{1}{\pi_{j|g}}y_{gj},$$

et

$$\bar{y}_g = \frac{1}{n_g} \sum_{j \in s_g} y_{gj}.$$

(2) Scénario 2:

$$B_{1i|g}^{NR} = \left(\frac{1}{\pi_g} - 1\right) Y_g + \frac{1}{\pi_g} \frac{N_g}{N_g - 1} \left(\frac{N_g}{n_g} - 1\right) \left(y_{gi} - \bar{Y}_g\right) + \frac{1}{\pi_g} \frac{N_g}{n_g} \left(\frac{1}{p_{i|g}} - 1\right) y_{gi}.$$

Son estimateur s'écrit alors comme

$$\hat{B}_{1i|g}^{NR} = \left(\frac{1}{\pi_g} - 1\right) \hat{Y}_{gr} + \frac{1}{\pi_g} \frac{n_g}{n_g - 1} \left(\frac{N_g}{n_g} - 1\right) \left(y_{gi} - \hat{\bar{y}}_{gr}\right) + \frac{1}{\pi_g} \frac{N_g}{n_g} \left(\frac{1}{p_{i|g}} - 1\right) y_{gi},$$

où

$$\hat{Y}_{gr} = \sum_{j \in s_{gr}} \frac{1}{p_{j|g} \pi_{j|g}} y_{gj}$$

et

$$\hat{\bar{y}}_{gr} = \frac{1}{n_g} \hat{Y}_{gr}.$$

(3) Scénario 3:

$$B_{1i|g}^{Cal} = \left(\frac{1}{\pi_g} - 1\right) E_g + \frac{1}{\pi_g} \frac{N_g}{N_g - 1} \left(\frac{N_g}{n_g} - 1\right) \left(E_{gi} - \bar{E}_g\right) + \frac{1}{\pi_g} \frac{N_g}{n_g} \left(\frac{1}{p_{i|g}} - 1\right) E_{gj}.$$

Ce biais conditionnel est estimé par

$$\hat{B}_{1i|g}^{Cal} = \left(\frac{1}{\pi_g} - 1\right) e_{gr} + \frac{1}{\pi_g} \frac{n_g}{n_g - 1} \left(\frac{N_g}{n_g} - 1\right) (e_{gi} - \bar{e}_{gr}) + \frac{1}{\pi_g} \frac{N_g}{n_g} \left(\frac{1}{p_{i|g}} - 1\right) e_{gi},$$

où

$$e_{gr} = \sum_{j \in s_{gr}} \frac{1}{p_{j|g} \pi_{j|g}} e_{gj}$$

et

$$\bar{e}_{gr} = \frac{1}{n_g} e_{gr}.$$

4.5. Estimation robuste et résultats

4.5.1. Méthode utilisée en production pour détecter les valeurs influentes

L'EDM utilise actuellement une méthode pour détecter et corriger les valeurs influentes. Nous la décrivons dans cette section afin de pouvoir ensuite la comparer avec la méthode du biais conditionnel.

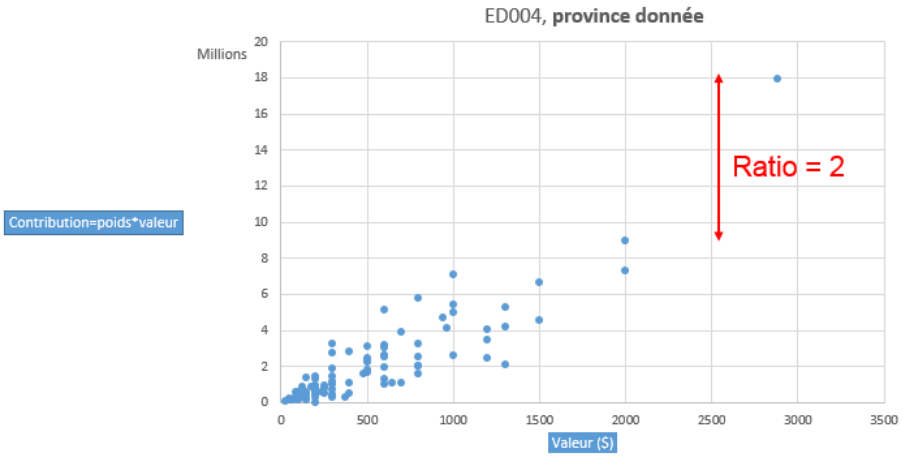
La méthode présentement utilisée en production de l'EDM pour détecter et corriger les valeurs influentes est basée sur la contribution, $w_i \times y_i$, de l'unité i à

l'estimation du total t_y. La détection se fait d'abord pour chaque variable se trouvant au niveau le plus détaillé. Nous appelons parfois ces variables, les variables «enfants», puisqu'elles sont au plus bas niveau possible et sont utilisées pour calculer les variables agrégées.

Illustrons ceci à l'aide d'un exemple. La variable de frais de scolarité ED003 est la somme des frais de scolarité des écoles maternelles, primaires et secondaires (ED004), des frais de scolarité universitaires (ED005), des frais de scolarité des autres programmes post-secondaires (ED006), des frais en service d'éducation (ED007) et des frais d'autres cours et leçons, excluant les cours de conduite (ED008). Du point de vue des spécifications de l'enquête, on aura alors que ED003=ED004 + ED005 + ED006 + ED007 + ED008. Les codes ED004 à ED008 sont les variables «enfants» et ont ED003 comme variable «parent».

La méthode de détection de valeurs influentes utilisée en production est appliquée aux codes enfants d'abord. Cette détection consiste à calculer la contribution $w_i \times y_i$ de chaque unité dans une province donnée, où y_i est la valeur annualisée de la variable enfant pour l'unité i et w_i est son poids final. On ordonne ensuite ces contributions de la plus petite valeur à la plus grande, et calcule le ratio entre deux contributions consécutives. Lorsqu'une unité a un ratio supérieur à un certain seuil prédéterminé (souvent égal à 1.85 en pratique), l'unité sera identifiée comme influente. Toutes les autres unités ayant une contribution supérieure à une telle unité seront aussi jugées influentes. Le graphique suivant, qui utilise des données fictives, contient sur l'axe des y la valeur de la contribution et sur l'axe des x la valeur de la variable de dépense (ED004, par exemple). L'unité ayant la plus grande contribution se démarque des autres. Son ratio, par rapport à l'unité précédente est de 2 et serait donc détectée comme influente.

FIGURE 4.2. Méthode de correction de l'EDM: avant correction



On fait ensuite le même exercice avec les variables enfants mais au niveau national. Finalement, la même détection est effectuée pour le code parent aux mêmes deux niveaux (provincial et national). La détection est donc effectuée à 4 niveaux distincts:

— Variable enfant dans la province,
— Variable enfant au Canada,
— Variable parent dans la province,
— Variable parent au Canada.

Si une unité est détectée à au moins 2 niveaux parmi 4, sa valeur sera corrigée. Ce dernier critère est utilisé afin de ne pas corriger trop d'unités. De plus, si dans un niveau donné, il y a moins que 25 ménages déclarants, aucune correction ne sera apportée à la variable. Nous jugeons que la méthode est moins valide lorsqu'il y a peu d'unités déclarantes et c'est pourquoi nous utilisons un minimum de 25 ménages déclarants. De plus, seulement les 4% plus grandes unités (en termes de contribution) seront éligibles à être corrigées. Encore une fois, ce seuil est pour éviter de corriger trop d'unités.

Lorsque les unités à corriger sont identifiées, il faut ensuite calculer le facteur d'ajustement pour corriger la variable d'intérêt. Pour ce faire, il faut d'abord identifier la première unité non-influente qui a la contribution la plus proche de l'unité à corriger. Dénotons cette unité par j. Le facteur d'ajustement utilisé est défini comme étant le ratio de l'unité j, multipliée par sa contribution et divisée par la contribution de l'unité i à corriger. La nouvelle valeur de la variable

d'intérêt de l'unité i sera donc:

$$\tilde{y}_i = ajust_j \times y_i = \frac{ratio_j w_j y_j}{w_i y_i} \times y_i.$$

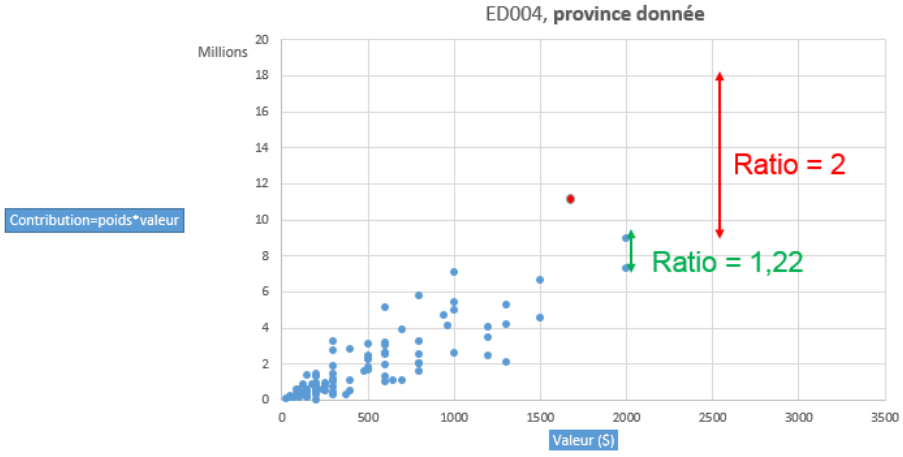
La nouvelle contribution de l'unité i sera donc:

$$w_i \tilde{y}_i = w_i \times ajust_i \times y_i = ratio_j w_j y_j,$$

ce qui revient à prendre la contribution de la première unité non-influente j et la multiplier par son ratio (qui, par définition, sera inférieur au paramètre utilisé puisque l'unité j est non-influente). Ceci fait en sorte que la nouvelle contribution de l'unité i est encore la plus grande dans la province mais son nouveau ratio ne dépasse plus le seuil limite fixé par la méthode.

Désignons l'unité à contribution maximale du graphique précédent par l'indice i. L'unité à contribution juste en dessous de l'unité i a un ratio de 1.22 qui est inférieur au seuil de 1.85. Cette unité est donc la première unité après i à être non-influente. Elle sera l'unité j utilisée dans l'ajustement. La nouvelle contribution de i après la correction sera de 1.22 fois celle de l'unité j. Cette nouvelle contribution est illustrée par un point rouge sur le graphique ci-dessous.

FIGURE 4.3. Méthode de correction de l'EDM: après correction



4.5.2. Résultats de la méthode du biais conditionnel

Comme nous avons supposé que les probabilités de réponse sont connues, nous avons utilisé l'estimateur de calage (4.4.4) pour calculer le biais conditionnel, c'est-à-dire celui de l'équation (4.4.6). Un exemple de calcul des résidus de cette équation est donné à la section 4.3.2. Donc le biais conditionnel dans le calcul de l'estimateur robuste de la section 2.3 a été remplacé par celui dans (4.4.6).

Le principal domaine d'intérêt dans cette enquête est la province. Nous avons ainsi déterminé la valeur optimale du seuil c, ou bien $\hat{B}_{max}^{Cal}$ et $\hat{B}_{min}^{Cal}$, pour chaque province et chaque variable de dépense. Autrement dit, le biais conditionnel décrit à la section 4.4.2 a été calculé pour chaque variable de dépense de plus petit niveau dans chacune des dix provinces séparément, pour obtenir les estimations provinciales et nationales.

4.5.2.1. *Quelques modifications non dictées par la méthode*

Avant tout, il faut noter qu'avec la méthode du biais conditionnel qui consiste à faire l'ajustement des valeurs pour chaque répondant, il peut arriver qu'un non-déclarant devienne un déclarant, ou qu'une dépense positive devienne négative. De la formule utilisée pour faire cet ajustement, de la forme (2.3.4), rien n'empêche qu'une telle situation ne se produise.

En particulier, lorsque l'estimateur de calage est utilisé, le biais conditionnel dépend alors du résidu de calage et non plus directement de la valeur y. On peut ainsi avoir un non-déclarant (i.e. $y = 0$) avec un grand résidu et ce ménage peut alors être influent et voir sa valeur de y changer.
Pour être cohérent avec la manière dont les données de l'EDM sont diffusées, et pour ne pas créer de confusion chez les utilisateurs, nous avons contourné ces situations, sans pour autant affecter l'estimation donnée par la méthode. En effet, les ajustements ont été faits indépendamment d'une province à l'autre. Ainsi, l'estimation donnée par le modèle doit demeurer inchangée dans chaque province quelle que soit la méthode utilisée pour contourner les situations non voulues.

Pour contourner ces situations, nous avons d'abord calculé l'estimation provinciale à partir de l'estimateur robuste. Nous voulons maintenir cette estimation tout en évitant de modifier la valeur y_i d'une unité i non-déclarante. Autrement dit, si $y_i = 0$ on exige que $\tilde{y}_i = 0$. Les autres unités de la province, qui sont déclarantes, devront être modifiées différemment pour compenser cette contrainte de

sorte que l'estimation provinciale reste inchangée.

Pour ce faire, nous avons calculé un facteur d'ajustement interne pour chaque province. Ce facteur est donné par la somme pondérée (par le poids de calage) des valeurs initialement ajustées divisée par la somme pondérée des vraies valeurs. Ainsi pour obtenir l'ajustement d'un ménage, on prend sa valeur initiale y_i et on la multiplie par ce facteur. C'est-à-dire, si nous appelons par $\tilde{y}_{i_{final}}$ la valeur finale ajustée de l'unité i alors elle se calcule comme suit:

$$\tilde{y}_{i_{final}} = y_i \times \frac{\sum_j w_j \tilde{y}_j}{\sum_j w_j y_j},$$

où w_j est le poids de calage du ménage j.

Pour éviter un double ajustement de toutes les valeurs dans une province, l'idéal serait de calculer le facteur d'ajustement uniquement dans la grappe où se trouve l'unité à problème. Et si l'ajustement dans cette grappe est insuffisant pour régler le problème, on peut la combiner avec d'autres grappes jusqu'à ce que le problème soit résolu. Or dans notre cas, pour des fins de simplicité, nous avons utilisé toutes les grappes de la province, qui se trouve être la solution extrême. On a que dans une province donnée, $\hat{t}_y^R = \sum_s w_i \tilde{y}_i = \sum_s w_i \tilde{y}_{i_{final}}$ avec la propriété que si $y_i = 0$ alors $\tilde{y}_{i_{final}} = 0$, mais aussi $\tilde{y}_{i_{final}}$ reste positive.

4.5.2.2. *Comparaison des deux méthodes*

Pour comparer la méthode utilisée actuellement en production à celle du biais conditionnel, nous avons calculé la différence relative des estimations en utilisant les données ajustées par chaque méthode par rapport aux vraies données observées. Les résultats sont donnés ci-dessous sous forme de graphiques et la définition des variables utilisées est donnée en annexe. La méthode utilisée en production a été testée avec deux poids de sondage différents: le poids après le premier calage (méthode la plus comparable à celle du biais conditionnel qui utilise également ce poids) et le poids après le deuxième calage. L'EDM effectue deux calages puisqu'à la suite du premier calage, certains poids calés sont ajustés parce qu'ils sont trop extrêmes. On doit donc effectuer un deuxième calage pour retomber sur les totaux de contrôle.

À première vue, la méthode du biais conditionnel semble davantage corriger les estimations puisque les différences relatives sont de plus grande magnitude. En effet, cela est beaucoup plus visible lorsqu'on regarde les petites provinces. Il faut

rappeler que la méthode actuelle utilisée en production ne corrige pas toujours en raison de certaines restrictions. Par exemple, la méthode utilisée en production n'applique aucune correction s'il y a moins de 25 déclarants dans la province pour une dépense donnée. Cette restriction n'a pas été appliquée avec la méthode du biais conditionnel et rend la comparaison entre les deux méthodes plus difficile. Or les petites provinces ont tendance à avoir moins de déclarants que les grandes provinces. De plus, la méthode actuelle utilise le même seuil de ratio pour détecter les valeurs influentes, peu importe la province. Ceci pourrait être une autre raison expliquant pourquoi la méthode actuelle ne corrige pas autant dans les petites provinces. En regardant les plus grandes provinces comme le Québec ou l'Ontario, on voit que pour certaines variables, c'est la méthode actuelle qui corrige plus, alors que pour d'autres, c'est la méthode du biais conditionnel qui corrige davantage.

FIGURE 4.4. Résultats des provinces T.-N.-L (10), Î.-P.-E (11), N.-E (12) et N.-B (13)

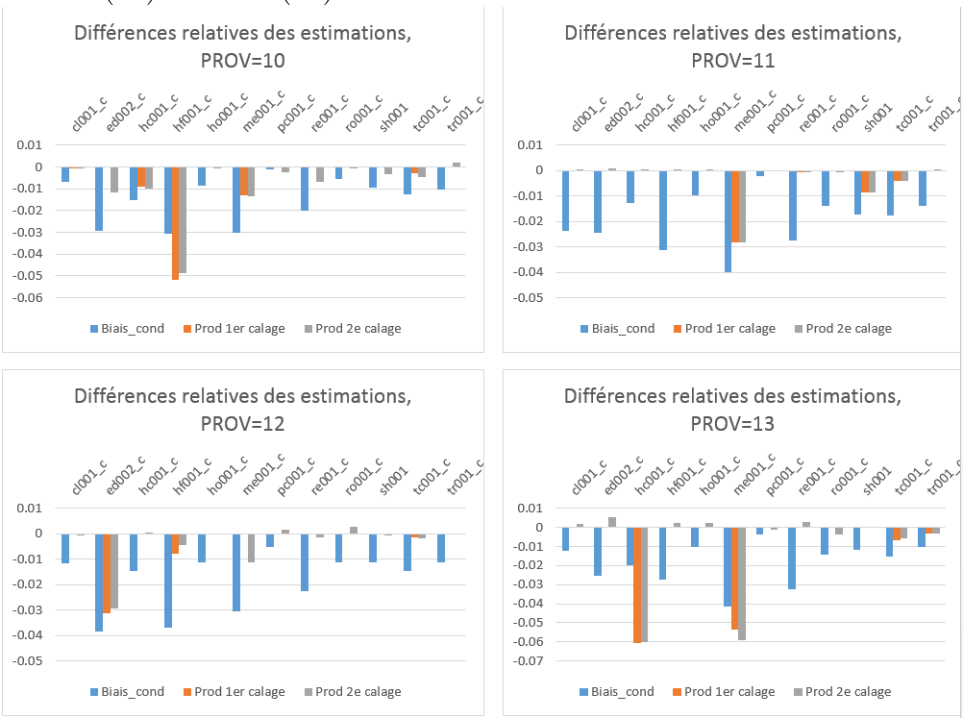


FIGURE 4.5. Résultats des provinces QC (24), ON (35), MB (46) et SK (47)

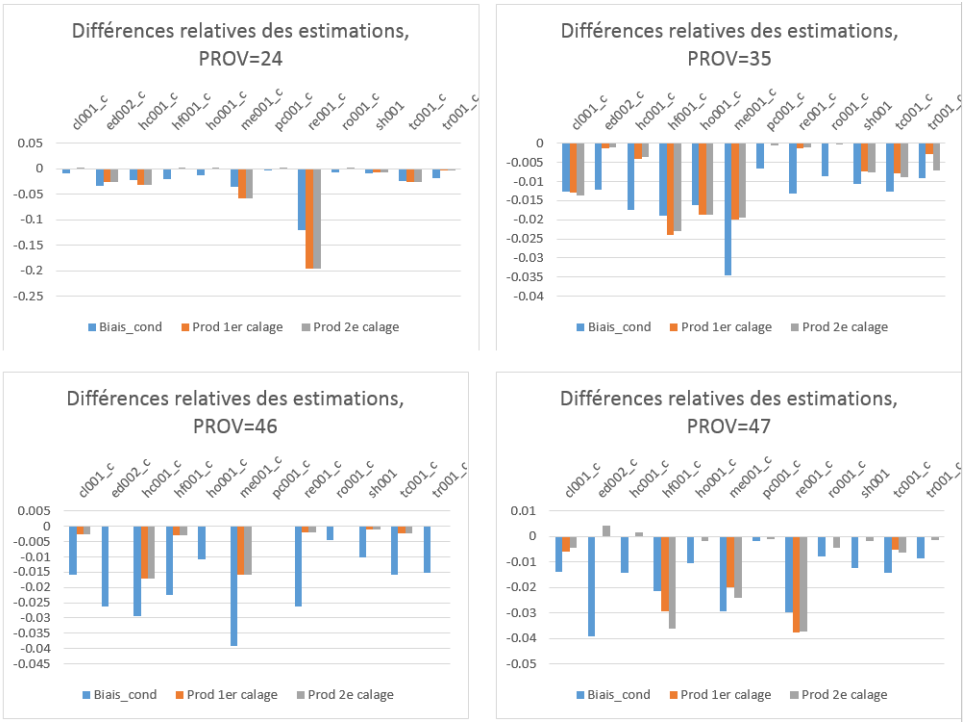
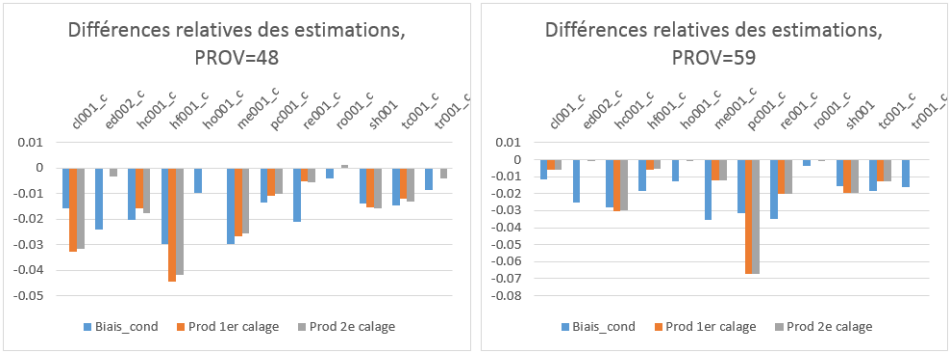


FIGURE 4.6. Résultats des provinces AB (48) et CB (59)



Chapitre 5

CONCLUSION

Dans ce mémoire, nous avons cherché à trouver des critères permettant de trouver des estimateurs robustes avec de bonnes propriétés. L'ensemble des critères investigués, y compris celui utilisé par Beaumont et al (2013), sauf celui qui consiste à minimiser des percentiles, ont été regroupés dans un critère général. Le biais conditionnel et son estimateur pour certains plans de sondage ont été établis. La méthode min-max a été évaluée sur des données réelles issues de l'enquête sur les dépenses des ménages canadiens, une enquête de Statistique Canada.

Une étude par simulations nous a permis de comparer différents critères avec la méthode min-max. Pour le critère qui consiste à minimiser la somme des biais conditionnels robustes en valeur absolue élevés à une certaine puissance q, nous remarquons que pour les faibles puissances, le critère min-max donne de meilleurs résultats en termes d'efficacité relative. Lorsque la valeur de q augmente, ce critère a tendance à converger vers la méthode min-max et ceci quelle que soit la population, le plan de sondage ou la taille de l'échantillon. En même temps, à mesure que l'efficacité relative s'améliore avec la puissance, le biais relatif augmente. Pour ce critère, lorsque nous augmentons la proportion d'unités ayant de grande valeurs dans une distribution, nous remarquons une augmentation de la valeur de l'efficacité relative, ce qui est également le cas pour le min-max.

Le critère du min-max donne également de meilleurs résultats en ce qui concerne l'efficacité relative par rapport au critère qui consiste à minimiser les percentiles, lorsqu'on considère les 95ème percentile et moins, indépendamment du plan de sondage. Par contre, lorsque la quantité d'unités ayant de grandes valeurs est importante dans la population, certains de ces percentiles donnent parfois de

meilleurs résultats que le min-max, toujours en termes d'efficacité relative. Le 99ème percentile donne généralement des résultats très similaires à ceux obtenus par la méthode du min-max. Lorsqu'on minimise le 97.5ème percentile, on obtient en général le même résultat que Beaumont et al. (2013) lorsque la taille de l'échantillon est égale à 25, et ceci peu importe le plan de sondage et la population. En dehors de ce contexte, les résultats du 97.5ème percentile sont similaires à ceux des autres percentiles sauf le 99ème.

Concernant le critère qui consiste à minimiser la somme des $k \geq 2$ plus grands biais conditionnels robuste en valeur absolue, le résultat semble dépendre de la population. En effet, lorsqu'on regarde les différents mélanges de lognormales, la valeur de l'efficacité relative augemente avec k au moment où le biais relatif diminue; et ceci quel que soit le plan de sondage utilisé. Par contre, en regardant les mélanges de normales, la première population (mixnor1) donne un résultat similaire aux mélanges de lognormales. Le troisième mélange, lui, donne une valeur de l'efficacité relative décroissante et un biais relatif décroissant avec k. Pour la deuxième population de mélange de normales, nous avons une croissance de la valeur de l'efficacité relative avec k lorsque $n = 25$ et une décroissance lorsque $n = 100$. Pour $n = 50$, il est plus difficile de voir ce qui se passe. Lorsqu'on compare ce critère au min-max, on peut dire que le min-max est plus consistant et plus prévisible. Mais il arrive que ce critère donne de meilleurs résultats que le min-max. En effet lorsqu'on regarde les populations 2 et 3 issues de mélanges de normales, nous remarquons que le min-max donne parfois des résultats moins bons en termes d'efficacité relative par rapport à ce critère.

Pour avoir une meilleure compréhension par rapport à ce qui se passe réellement pour ce dernier critère, il serait intéressant d'investiguer d'autres types de populations. L'autre critère qui peut être intéressant d'investiguer, est de minimiser l'erreur quadratique moyenne, qui sera estimée en utilisant le bootstrap.

Concernant l'application de la méthode du min-max aux données de Statistique Canada, en plus des estimations provenant de cette méthode, les estimations de variance ont également été calculées et comparées aux estimations de variance de la méthode actuelle. Puisque la méthode du biais conditionnel corrige davantage les unités, elle est davantage biaisée. Par contre, on s'attend alors à un gain du

point de vue de la variance estimée. Ce que nous avons observé, dans les faits, est que ce gain est plus faible qu'attendu. Ceci est dû au fait que la méthode du biais n'est pas suivie à la lettre, en raison de l'ajustement additionnel nécessaire pour maintenir les valeurs finales des non-déclarants à zéro. Lorsque les estimations ont été calculées avec les valeurs $\tilde{y}_i$ au lieu de $\tilde{y}_{i_{final}}$, le gain espéré au niveau de l'estimation de la variance est alors observé et la méthode du biais conditionnel a une erreur quadratique moyenne plus basse que la méthode actuelle utilisée en production. Par contre, tel que mentionné plus tôt, les valeurs $\tilde{y}_i$ ne sont pas pratiques à utiliser et créent certaines incohérences dans les données du fait que des non-déclarants peuvent devenir déclarants et des dépenses positives peuvent devenir négatives.

Comme travail futur, il pourrait être intéressant d'étudier la possibilité d'obtenir des valeurs de $\tilde{y}_{i_{final}}$ différemment, en réallouant au niveau de certaines grappes seulement, et non pas à toutes les grappes de la province. Pour l'estimation de la variance, les ajustements aux valeurs ont simplement été appliqués aux mêmes valeurs correspondantes de chaque réplique bootstrap (le même principe étant utilisé en production). Ceci peut mener à une sous-estimation de la variance et à ce moment-ci, nous sommes à la recherche d'une meilleure façon de procéder pour estimer la variance.

Il serait également souhaitable de développer les formules du biais conditionnel pour le plan de sondage de l'EDM en considérant les probabilités de réponse comme étant estimées et non pas connues. Ceci impliquera une étape de linéarisation.

La présente évaluation portait sur les dépenses de l'entrevue seulement et pour l'échantillon dans les provinces. Il serait intéressant de faire le même exercice pour les dépenses du journal et il faudrait alors tenir compte de la phase additionnelle du journal puisque seulement 50% des ménages sélectionnés pour l'entrevue sont également choisis pour remplir le journal. De même, on pourrait appliquer la méthode du biais conditionnel à l'échantillon dans les territoires, où un plan de sondage différent est utilisé.

Bibliographie

[1] Beaumont, J.-F., Haziza, D. et Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika* **100**, 555–569.

[2] Beaumont, J.-F. et Haziza, D. (2016). A note on the concept of invariance in two-phase sampling designs. *Survey Methodology* **42**, 319–323.

[3] Favre Martinoz, C., Haziza, D. et Beaumont, J.-F. (2016). Robust inference in two-phase sampling designs with application to unit nonresponse. *Scandinavian Journal of Statistics* **43**, 1019–1034.

[4] Favre Martinoz, C. (2015). Estimation robuste en population finie et infinie. Thèse soutenue à l'Ensai le 13 octobre 2015, 176–177.

[5] Deville, J.-C., Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87** (1992), 376–382.

[6] Estevao, V. M. et Särndal, C.-E. (2003). A new perspective on calibration estimators. *JSM-Section on Survey Research Methods, American Statistical Association*, 1346–1356.

[7] Estevao, V. M. et Särndal, C.-E. (2009). A new face on two-phase sampling with calibration estimators. *Survey Methodology* **35**, 3–14.

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

[9] Rao, J. N. K., Hartley, H. O. et Cochran, W. G. (1962). One simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*. Series B, **24**, 482–491.

[10] Lemaître, G.E. et Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology* **13**, 199–207.

[11] De Moliner, A. (2017). Estimation robuste de courbes de consommation électrique moyennes par sondage pour de petits domaines en présence de valeurs manquantes. PhD thesis, Université Bourgogne Franche-Comté.

Annexe A

DÉMONSTRATION DES RÉSULTATS

A.1. Proposition 2.2.2

En effet en partant de la droite de l'équation (2.2.6), on a

$$\sum_{i\in s} B_{1i}^{HT} + \sum_{i\in U-s} B_{0i}^{HT} + \left(\sum_{i\in s}\pi_i^{-1}A_i - \sum_{i\in U} A_i\right) = \sum_{i\in s}\left\{B_{1i}^{HT} + \left(\pi_i^{-1}-1\right)A_i\right\} + \sum_{i\in U\setminus s}\left(B_{0i}^{HT} - A_i\right),$$

or

$$\begin{aligned} B_{1i}^{HT} &= \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)y_j \\ &= \left(\pi_i^{-1}-1\right)y_i + \sum_{\substack{j\in U\\ j\neq i}}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}y_j \\ &= \left(\pi_i^{-1}-1\right)y_i - \pi_i^{-1}\left(1-\pi_i\right)A_i \\ &= \left(\pi_i^{-1}-1\right)\left(y_i - A_i\right), \end{aligned}$$

de même on a

$$B_{0i}^{HT} = -\left(y_i - A_i\right).$$

Ainsi la droite de l'équation (2.2.6) devient:

$$\begin{aligned} &\sum_{i\in s}\left\{\left(\pi_i^{-1}-1\right)\left(y_i-A_i\right) + \left(\pi_i^{-1}-1\right)A_i\right\} \\ &\qquad + \sum_{i\in U-s}\left(A_i - y_i - A_i\right) \\ &\qquad = \sum_{i\in s}\left(\pi_i^{-1}-1\right)y_i - \sum_{i\in U\setminus s} y_i \\ &\qquad = \sum_{i\in s}\pi_i^{-1}y_i - \sum_{i\in U} y_i = \hat{t}_y^{HT} - t_y. \end{aligned}$$

A.2. Proposition 2.2.4

Le biais conditionnel de l'unité $i \in U_g$ par rapport à l'estimateur $\hat{t}_{2y}$ ci-dessus est défini par:

$$\begin{aligned} B_{1i|g}^{HT} &= E_p\left(\hat{t}_{2y} - t_y \middle| I_{1g} = 1, I_{2i} = 1\right) \\ &= E_p\left\{\left(\hat{t}_{1y} - t_y\right) + \left(\hat{t}_{2y} - \hat{t}_{1y}\right) \middle| I_{1g} = 1, I_{2i} = 1\right\} \\ &= E_p\left\{\left(\hat{t}_{1y} - t_y\right) \middle| I_{1g} = 1\right\} + E_p\left\{\left(\hat{t}_{2y} - \hat{t}_{1y}\right) \middle| I_{1g} = 1, I_{2i} = 1\right\}. \end{aligned}$$

Nous avons

$$\begin{aligned} E_p\left\{\left(\hat{t}_{1y} - t_y\right) \middle| I_{1g} = 1\right\} &= E_1\left\{\left(\hat{t}_{1y} - t_y\right) \middle| I_{1g} = 1\right\} \\ &= E_1\left\{\sum_{l \in U}\left(\frac{I_{1l}}{\pi_l} - 1\right)\sum_{j \in U_l} y_{lj} \middle| I_{1g} = 1\right\} \\ &= \sum_{l \in U}\left\{\frac{E_1\left(I_{1l} \middle| I_{1g} = 1\right)}{\pi_l} - 1\right\}\sum_{j \in U_l} y_{lj} \\ &= \sum_{l \in U}\left(\frac{\pi_{gl}}{\pi_g \pi_l} - 1\right)\sum_{j \in U_l} y_{lj}. \end{aligned}$$

Et

$$\begin{aligned} &E_p\left\{\left(\hat{t}_{2y} - \hat{t}_{1y}\right) \middle| I_{1g} = 1, I_{2i} = 1\right\} \\ &\qquad = E_1 E_2\left(\hat{t}_{2y} - \hat{t}_{1y} \middle| s, I_{1g} = 1, I_{2i} = 1\right) \\ &= E_1\left\{E_2\left(\sum_{l \in s}\sum_{j \in s_l} \frac{1}{\pi_l \pi_{j|l}} y_{lj} \middle| s, I_{1g} = 1, I_{2i} = 1\right) - \hat{t}_{1y} \middle| I_{1g} = 1\right\} \\ &= E_1\left\{\sum_{l \in s}\sum_{j \in U_l} \frac{1}{\pi_l \pi_{j|l}} E_2\left(I_{2j} \middle| I_{1g} = 1, I_{2i} = 1\right) y_{lj} - \hat{t}_{1y} \middle| I_{1g} = 1\right\} \end{aligned}$$

Or

$$E_2\left(I_{2j}|I_{1g}=1,I_{2i}=1\right)=\begin{cases}1, & \text{si} \quad j=i\\ \frac{\pi_{ij|g}}{\pi_{i|g}}, & \text{si} \quad j\neq i, \quad j\in U_g\\ \pi_{j|l}, & \text{si} \quad j\in U_l, \quad l\neq g\end{cases}$$

Ainsi

$$E_2\left(\sum_{l\in s}\frac{1}{\pi_l}\sum_{j\in s_l}\frac{1}{\pi_{j|l}}y_{lj}\Big|s,I_{1g}=1,I_{2i}=1\right)-\hat{t}_{1y}=\frac{y_{gi}}{\pi_g\pi_{i|g}}+\frac{1}{\pi_g}\sum_{\substack{j\in U_g\\ j\neq i}}\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}y_{gj}-\frac{1}{\pi_g}\sum_{j\in U_g}y_{gj}$$

D'où,

$$E_p\left\{\left(\hat{t}_{2y}-\hat{t}_{1y}\right)\Big|I_{1g}=1,I_{2i}=1\right\}=\frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}.$$

Par conséquent

$$B_{1i|g}^{HT}=\sum_{l\in U}\left(\frac{\pi_{gl}}{\pi_g\pi_l}-1\right)\sum_{j\in U_l}y_{lj}+\frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}.$$

A.3. Proposition 2.2.5

Le biais conditionnel par rapport à l'estimateur $\hat{t}_y^{DE}$ est donné par:

$$\begin{aligned}B_{1i}^{DE}&=E_p\left\{\left(\hat{t}_y^{DE}-t_y\right)\Big|I_{1i}=1,I_{2i}=1\right\}\\&=E_p\left\{\left(\hat{t}_y^{DE}-\hat{t}_{1y}\right)+\left(\hat{t}_{1y}-t_y\right)\Big|I_{1i}=1,I_{2i}=1\right\}\\&=E_p\left\{\left(\hat{t}_{1y}-t_y\right)\Big|I_{1i}=1,I_{2i}=1\right\}+E_p\left\{\left(\hat{t}_y^{DE}-\hat{t}_{1y}\right)\Big|I_{1i}=1,I_{2i}=1\right\}\end{aligned}$$

En posant $B_{1i}^{1HT}=E_p\left\{\left(\hat{t}_{1y}-t_y\right)\Big|I_{1i}=1,I_{2i}=1\right\}=E_1\left\{\left(\hat{t}_{1y}-t_y\right)\Big|I_{1i}=1\right\}$ le biais conditionnel de l'unité i par rapport au plan à un degré
On obtient:

$$\begin{aligned}B_{1i}^{DE}&=B_{1i}^{1HT}+E_p\left\{\left(\hat{t}_y^{DE}-\hat{t}_{1y}\right)\Big|I_{1i}=1,I_{2i}=1\right\}\\&=B_{1i}^{1HT}+E_1E_2\left[\left\{\left(\hat{t}_y^{DE}-\hat{t}_{1y}\right)\Big|I_{1i}=1,I_{2i}=1\right\}\Big|s_1\right]\\&=B_{1i}^{1HT}+E_1\left[E_2\left\{\left(\hat{t}_y^{DE}\Big|I_{2i}=1\right)\Big|s_1\right\}-\hat{t}_{1y}\Big|I_{1i}=1\right]\end{aligned}$$

A-iv

$$= B_{1i}^{1HT} + E_1\left[E_2\left\{\sum_{j\in s_1}\pi_{1j}^{-1}\pi_{2j}(I_1)^{-1}y_jI_{2j}\Big|I_{2i}=1\Big|s_1\right\} - \sum_{j\in s_1}\pi_{1j}^{-1}y_j\Big|I_{1i}=1\right]$$

$$= B_{1i}^{1HT} + E_1\left[\left\{\sum_{j\in s_1}\pi_{1j}^{-1}\pi_{2j}(I_1)^{-1}y_jE_2\left(I_{2j}\Big|I_{2i}=1\right) - \sum_{j\in s_1}\pi_{1j}^{-1}y_j\right\}\Big|I_{2i}=1\right].$$

Or

$$E_2\left(I_{2j}\Big|I_{2i}=1\right) = Pr(I_{2j}=1\Big|I_{2i}=1) = \begin{cases}1, & \text{si} \quad j=i\\ \frac{\pi_{2ij}(I_1)}{\pi_{2i}(I_1)}, & \text{sinon}\end{cases}$$

donc

$$\begin{aligned}B_{1i}^{DE} &= B_{1i}^{1HT} + E_1\left\{\pi_{1i}^{-1}\pi_{2i}(I_1)^{-1}y_iE_2\left(I_{2j}\Big|I_{2i}=1\right) - \pi_{1i}^{-1}y_i\Big|I_{2i}=1\right\}\\ &\qquad + E_1\left\{\sum_{\substack{j\in s_1\\ j\neq i}}\pi_{1j}^{-1}\pi_{2j}(I_1)^{-1}y_jE_2\left(I_{2j}\Big|I_{2i}=1\right) - \sum_{\substack{j\in s_1\\ j\neq i}}\pi_{1j}^{-1}y_j\Big|I_{2i}=1\right\}\\ &= B_{1i}^{1HT} + E_1\left[\frac{1}{\pi_{1i}}\left\{\pi_{2i}(I_1)^{-1}-1\right\}y_i\Big|I_{2i}=1\right]\\ &+ E_1\left[\sum_{\substack{j\in s_1\\ j\neq i}}\frac{1}{\pi_{1j}}\left\{\frac{\pi_{2ij}(I_1)}{\pi_{2i}(I_1)\pi_{2j}(I_1)}-1\right\}y_j\Big|I_{2i}=1\right].\end{aligned}$$

A.4. Proposition 2.2.8

En effet, $\hat{B}_{1i|g}^{HT}$ peut s'écrire:

$$\hat{B}_{1i|g}^{HT} = \sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in s_l}\frac{1}{\pi_{j|l}}y_{lj} + \left(\frac{1}{\pi_g}-1\right)\sum_{j\in s_g}\frac{\pi_{i|g}}{\pi_{ij|g}}y_{gj} + \frac{1}{\pi_g}\sum_{j\in s_g}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj},$$

donc

$$E_p\left(\hat{B}_{1i|g}^{HT}\Big|I_{1g}=1, I_{2i}=1\right) =$$

$$E_p\left[\sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in s_l}\frac{1}{\pi_{j|l}}y_{lj} + \left(\frac{1}{\pi_g}-1\right)\sum_{j\in s_g}\frac{\pi_{i|g}}{\pi_{ij|g}}y_{gj}\Big|I_{1g}=1, I_{2i}=1\right]$$

$$+ E_p\left[\frac{1}{\pi_g}\sum_{j\in s_g}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}\middle| I_{1g}=1, I_{2i}=1\right]$$

$$= E_1E_2\left[\sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in s_l}\frac{1}{\pi_{j|l}}y_{lj}+\left(\frac{1}{\pi_g}-1\right)\sum_{j\in s_g}\frac{\pi_{i|g}}{\pi_{ij|g}}y_{gj}\middle| s, I_{1g}=1, I_{2i}=1\right]$$

$$+ E_1E_2\left[\frac{1}{\pi_g}\sum_{j\in s_g}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}\middle| s, I_{1g}=1, I_{2i}=1\right]$$

$$= E_1\left[\sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in U_l}y_{lj}\middle| I_{1g}=1\right]+\left(\frac{1}{\pi_g}-1\right)\sum_{j\in U_g}y_{gj}$$

$$+\frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}$$

$$=\sum_{\substack{l\in U\\ l\neq g}}\left(\frac{\pi_{gl}}{\pi_g\pi_l}-1\right)\sum_{j\in U_l}y_{lj}+\left(\frac{1}{\pi_g}-1\right)\sum_{j\in U_g}y_{gj}$$

$$+\frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}$$

$$=\sum_{l\in U}\left(\frac{\pi_{gl}}{\pi_g\pi_l}-1\right)\sum_{j\in U_l}y_{lj}+\frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}$$

$$= B^{HT}_{1i|g}.$$

A.5. Proposition 4.4.1

L'erreur totale $\hat{t}_y^{NR}-t_y$ se décompse en trois termes:
$(\hat{t}_{1y}-t_y)+(\hat{t}_{2y}-\hat{t}_{1y})+(\hat{t}_y^{NR}-\hat{t}_{2y})$, où $\hat{t}_{1y}$ et $\hat{t}_{2y}$ sont définis au chapitre 2.

De ce fait $B^{NR}_{1i|g}$ s'écrit:

$$B^{NR}_{1i|g}=E_p\left(\hat{t}_{1y}-t_y\middle| I_{1g}=1\right)+E_p\left(\hat{t}_{2y}-\hat{t}_{1y}\middle| I_{1g}=1, I_{2i}=1\right)$$
$$+E_p\left(\hat{t}_y^{NR}-\hat{t}_{2y}\middle| I_{1g}=1, I_{2i}=1, I_{3i}=1\right).$$

Les deux premiers termes de cette expression sont déjà connus d'après le calcul du biais conditionnel pour le plan à deux degrés, calculons donc le dernier terme.

A-vi

Posons $A := E_p\left(\hat{t}_y^{NR} - \hat{t}_{2y}\middle| I_{1g}=1, I_{2i}=1, I_{3i}=1\right)$, nous avons:

$$A = E_1E_2E_3\left\{\sum_{l\in s}\frac{1}{\pi_l}\left(\sum_{j\in s_{lr}}\frac{y_{lj}}{\pi_{j|l}p_{j|l}} - \sum_{j\in s_l}\frac{y_{lj}}{\pi_{j|l}}\right)\middle| s, s_l, I_{1g}=1, I_{2i}=1, I_{3i}=1\right\}$$

$$= E_1E_2\left[\sum_{l\in s}\frac{1}{\pi_l}\sum_{j\in s_l}\frac{1}{\pi_{j|l}}\left\{\frac{E_3\left(I_{3j}\middle| I_{1g}=1, I_{2i}=1, I_{3i}=1\right)}{p_{j|l}} - 1\right\} y_{lj}\middle| I_{1g}=1, I_{2i}=1\right]$$

Comme, $$E_3\left(I_{3j}\middle| I_{1g}=1, I_{2i}=1, I_{3i}=1\right) = \begin{cases} 1, & \text{si} \quad j=i \\ p_{j|g}, & \text{si} \quad j\neq i, \quad j\in U_g \\ p_{j|l}, & \text{si} \quad j\in U_l, \quad l\neq g \end{cases}$$

$$E_3\left\{\sum_{l\in s}\frac{1}{\pi_l}\left(\sum_{j\in s_{lr}}\frac{y_{lj}}{\pi_{j|l}p_{j|l}} - \sum_{j\in s_l}\frac{y_{lj}}{\pi_{j|l}}\right)\middle| s, s_l, I_{1g}=1, I_{2i}=1, I_{3i}=1\right\} = \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}} - 1\right) y_{gi},$$

qui est une constante.

Donc $$E_p\left(\hat{t}_y^{NR} - \hat{t}_{2y}\middle| I_{1g}=1, I_{2i}=1, I_{3i}=1\right) = \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}} - 1\right) y_{gi}.$$

Par conséquent, on obtient:

$$B_{1i|g}^{NR} = \sum_{l\in U}\left(\frac{\pi_{gl}}{\pi_g\pi_l} - 1\right)\sum_{j\in U_l} y_{lj} + \frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right) y_{gj} + \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}} - 1\right) y_{gi}$$

A.6. Proposition 4.4.2

En effet, nous avons

$$\hat{B}_{1i|g}^{NR} = \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}} - 1\right) y_{gi} + \frac{1}{\pi_g}\left(\frac{1}{\pi_{i|g}} - 1\right) y_{gi} + \left(\frac{1}{\pi_g} - 1\right) y_{gi}$$
$$+ \frac{1}{\pi_g}\sum_{\substack{j\in s_{gr} \\ j\neq i}}\frac{1}{p_{j|g}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right) y_{gj} + \sum_{\substack{j\in s_{gr} \\ j\neq i}}\frac{1}{p_{j|g}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{1}{\pi_g} - 1\right) y_{gj}$$

$$+\sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in s_{lr}}\frac{1}{p_{j|l}}\frac{1}{\pi_{j|l}}y_{lj}$$

Donc

$$\begin{aligned}
E_p\left(\hat{B}^{NR}_{1i|g}\middle|I_{1g}=1,I_{2i}=1,I_{3i}=1\right) &= \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}}-1\right)y_{gi}\\
&\quad+\frac{1}{\pi_g}\left(\frac{1}{\pi_{i|g}}-1\right)y_{gi}+\left(\frac{1}{\pi_g}-1\right)y_{gi}\\
&\quad+\frac{1}{\pi_g}\sum_{\substack{j\in s_g\\ j\neq i}}\frac{1}{p_{j|g}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}E_2E_3\left(I_{3j}\middle|s_g,I_{1g}=1,I_{2i}=1,I_{3i}=1\right)\\
&\quad+\sum_{\substack{j\in s_g\\ j\neq i}}\frac{1}{p_{j|g}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{1}{\pi_g}-1\right)y_{gj}E_2E_3\left(I_{3j}\middle|s_g,I_{1g}=1,I_{2i}=1,I_{3i}=1\right)\\
&\quad+\sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in s_l}\frac{1}{p_{j|l}}\frac{1}{\pi_{j|l}}y_{lj}E_1E_2E_3\left(I_{3j}\middle|s,s_l,I_{1g}=1,I_{2i}=1,I_{3i}=1\right)\\
&= \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}}-1\right)y_{gi}+\frac{1}{\pi_g}\left(\frac{1}{\pi_{i|g}}-1\right)y_{gi}+\left(\frac{1}{\pi_g}-1\right)y_{gi}\\
&\quad+\frac{1}{\pi_g}\sum_{\substack{j\in U_g\\ j\neq i}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}E_2\left(I_{2j}\middle|I_{1g}=1,I_{2i}=1\right)\\
&\quad+\sum_{\substack{j\in U_g\\ j\neq i}}\frac{\pi_{i|g}}{\pi_{ij|g}}\left(\frac{1}{\pi_g}-1\right)y_{gj}E_2\left(I_{2j}\middle|I_{1g}=1,I_{2i}=1\right)\\
&\quad+\sum_{\substack{l\in s\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in U_l}\frac{1}{\pi_{j|l}}y_{lj}E_1E_2\left(I_{2j}\middle|s,I_{1g}=1,I_{2i}=1\right)\\
&= \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}}-1\right)y_{gi}+\frac{1}{\pi_g}\left(\frac{1}{\pi_{i|g}}-1\right)y_{gi}+\left(\frac{1}{\pi_g}-1\right)y_{gi}\\
&\quad+\frac{1}{\pi_g}\sum_{\substack{j\in U_g\\ j\neq i}}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}+\sum_{\substack{j\in U_g\\ j\neq i}}\left(\frac{1}{\pi_g}-1\right)y_{gj}\\
&\quad+\sum_{\substack{l\in U\\ l\neq g}}\frac{\pi_{gl}-\pi_g\pi_l}{\pi_{gl}\pi_l}\sum_{j\in U_l}y_{lj}E_1\left(I_{1g}\middle|I_{1g}=1\right)\\
&= \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}}-1\right)y_{gi}+\frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)y_{gj}
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{j \in U_g} \left(\frac{1}{\pi_g} - 1 \right) y_{gj} + \sum_{\substack{l \in U \\ l \neq g}} \left(\frac{\pi_{gl}}{\pi_g \pi_l} - 1 \right) \sum_{j \in U_l} y_{lj} \\
&= \frac{1}{\pi_g} \frac{1}{\pi_{i|g}} \left(\frac{1}{p_{i|g}} - 1 \right) y_{gi} + \frac{1}{\pi_g} \sum_{j \in U_g} \left(\frac{\pi_{ij|g}}{\pi_{i|g} \pi_{j|g}} - 1 \right) y_{gj} \\
&\qquad + \sum_{l \in U} \left(\frac{\pi_{gl}}{\pi_g \pi_l} - 1 \right) \sum_{j \in U_l} y_{lj} = B_{1i|g}^{NR}.
\end{aligned}
$$

Annexe B

DÉFINITION DES VARIABLES UTILISÉES

Variable de dépense	**Description**
TC001_C	Consommation courante totale - Questionnaire
SH001	Logement
HO001_C	Entretien ménager
HF001_C	Ameublement et équipement ménagers - Questionnaire
CL001_C	Vêtements et accessoires - Questionnaire
TR001_C	Transport - Questionnaire
HC001_C	Soins de santé - Questionnaire
PC001_C	Soins personnels - Questionnaire
RE001_C	Loisirs - Questionnaire
ED002_C	Éducation - Questionnaire
RO001_C	Matériel de lecture et autres imprimés - Questionnaire
ME001_C	Dépenses diverses - Questionnaire