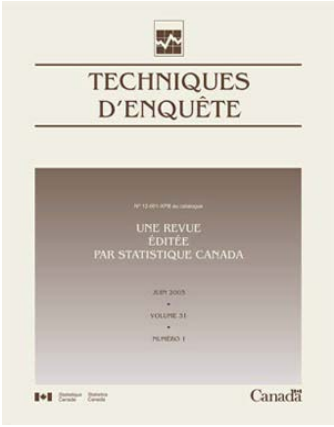
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Estimation de la variance sous non-réponse monotone pour une enquête par panel



par Hélène Juillard et Guillaume Chauvet

Date de diffusion : le 20 décembre 2018

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2018
Vol. 44, nº 2, pp. 295-317
Statistique Canada, nº 12-001-X au catalogue

Estimation de la variance sous non-réponse monotone pour une enquête par panel

Hélène Juillard et Guillaume Chauvet[1]

Résumé

Les enquêtes par panel sont souvent utilisées pour mesurer l'évolution de paramètres au cours du temps. Ces enquêtes peuvent souffrir de différentes formes de non-réponse totale, situation que l'on traite à l'heure actuelle en estimant les probabilités de réponse et en effectuant une nouvelle pondération des répondants. La présente étude porte sur l'estimation, ainsi que l'estimation de la variance en cas de non-réponse totale dans les enquêtes par panel. En étendant les travaux de Kim et Kim (2007) à plusieurs périodes, nous considérons un estimateur ajusté par un score de propension qui tient compte de la non-réponse initiale et de l'attrition, et proposons un estimateur de variance approprié. Nous étendons ensuite cet estimateur afin de couvrir la plupart des estimateurs utilisés dans les enquêtes, y compris les estimateurs calés, les estimateurs de paramètres complexes et les estimateurs longitudinaux. Les propriétés de l'estimateur de variance proposé et d'un estimateur de variance simplifié sont évaluées au moyen d'une étude en simulation. Une illustration de la méthode proposée sur des données provenant de l'enquête ELFE est également présentée.

Mots-clés : Estimateur de variance simplifié; estimation longitudinale; groupes de réponse homogènes; modèle de réponse; plan produit.

1 Introduction

Les enquêtes servent non seulement à produire des estimateurs pour un point particulier dans le temps (estimations transversales), mais aussi à mesurer l'évolution des paramètres (estimations longitudinales), et sont donc répétées au fil du temps. Dans cet article, nous nous intéressons à l'estimation de paramètres, ainsi qu'à l'estimation de la variance pour les enquêtes par panel, dans lesquelles les mesures sont répétées au cours du temps sur les unités d'un même échantillon (Kalton, 2009). Parmi les enquêtes par panel (aussi appelées enquêtes longitudinales, voir Lynn, 2009), les enquêtes de cohortes sont des cas particuliers où les unités de l'échantillon sont liées par un événement originel commun, par exemple être né la même année pour les enfants de l'Enquête longitudinale française depuis l'enfance (enquête ELFE), qui est l'exemple motivant les présents travaux.

L'enquête ELFE, qui est la première étude longitudinale de ce type menée en France, est conçue pour suivre les enfants de la naissance à l'âge adulte (Pirus, Bois, Dufourg, Lanoë, Vandentorren, Leridon et l'équipe ELFE, 2010). Ayant pour champ d'observation la France métropolitaine, elle a été lancée en 2011 et porte sur plus de 18 000 enfants dont les parents ont consenti à ce qu'ils soient inclus dans l'échantillon. L'enquête permettra d'examiner tous les aspects de la vie de ces enfants sous les angles de la santé, des sciences sociales et de la santé environnementale. L'enquête ELFE souffre de non-réponse totale, dont il faut tenir compte en utilisant l'information auxiliaire disponible afin de limiter le biais des estimateurs. L'enquête ELFE servira d'exemple dans le présent article, mais la non-réponse se produit dans presque toutes les enquêtes par panel, de sorte que les méthodes proposées présentent un intérêt général; voir, par

1. Hélène Juillard, INED, 133, boul. Davout, 75020 Paris, France; Guillaume Chauvet, ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France. Courriel : guillaume.chauvet@ensai.fr.

exemple, Laurie, Smith et Scott (1999) pour le traitement de la non-réponse pour la British Household Panel Survey, ou Vandecasteele et Debels (2007) pour le Panel communautaire des ménages d'Eurostat.

Le traitement de la non-réponse totale consiste à modéliser les probabilités de réponse (Kim et Kim, 2007) et à effectuer une nouvelle pondération des répondants par l'inverse de ces probabilités estimées, pour obtenir ce qu'il est convenu d'appeler l'estimateur ajusté par le score de propension. Un échantillon issu d'un panel peut présenter trois types de non-réponse totale (Hawkes et Plewis, 2009), à savoir la non-réponse initiale, qui désigne l'absence originale des unités sélectionnées, la non-réponse à une vague, qui se produit quand certaines unités du panel s'abstiennent de répondre temporairement à un point particulier dans le temps, et l'attrition, ou érosion de l'échantillon, qui a lieu quand certaines unités du panel cessent définitivement de répondre à partir d'un certain point dans le temps. La non-réponse à une vague était assez peu fréquente durant les premières vagues de l'enquête ELFE pour lesquelles nous disposions de données. Par conséquent, nous émettons pour simplifier l'hypothèse d'une non-réponse monotone, comprenant seulement la non-réponse initiale et l'attrition.

La littérature sur le traitement de la non-réponse totale dans les enquêtes longitudinales est abondante, voir Ekholm et Laaksonen (1991), Fuller, Loughin et Baker (1994), Rizzo, Kalton et Brick (1996), Clarke et Tate (2002), Laaksonen et Chambers (2006), Hawkes et Plewis (2009), Rendtel et Harms (2009), Laaksonen (2007), Slud et Bailey (2010), Zhou et Kim (2012). L'estimation de la variance des estimateurs longitudinaux est examinée dans Tam (1984), Laniel (1988), Nordberg (2000), Berger (2004), Skinner et Vieira (2005), Qualité et Tillé (2008) et Chauvet et Goga (2018), mais en se concentrant sur la variance d'échantillonnage uniquement. L'estimation de la variance dans le cas de repondérations pour corriger la non-réponse dans des enquêtes transversales est traitée dans Kim et Kim (2007). Autant que nous sachions, et malgré l'intérêt pour les applications, l'estimation de la variance tenant compte de la non-réponse pour des enquêtes par panel n'a pas été abordée dans la littérature, excepté dans Zhou et Kim (2012).

Zhou et Kim (2012) considèrent l'estimation d'une moyenne pour une enquête par panel sous une non-réponse monotone. Au lieu d'utiliser l'estimateur ajusté par le score de propension, Zhou et Kim (2012) définissent un estimateur à score de propension optimal. Celui-ci est obtenu en notant que, pour toute variable d'intérêt observée avant la période t, l'estimateur produit à la période t diffère de l'estimateur obtenu à la date où la variable a été observée, lequel est fondé sur un échantillon plus grand. L'ajustement pour tenir compte de ces différences au moyen d'une forme de calage mène à l'estimateur proposé par Zhou et Kim (2012). Cet estimateur utilise pleinement l'information recueillie aux périodes précédentes et devrait donc être plus efficace que l'estimateur ajusté par le score de propension. Cependant, une enquête par panel peut comprendre un grand nombre de variables d'intérêt observées à plusieurs périodes, et le calage sur un trop grand nombre de variables risque de donner des estimateurs dont les propriétés sont moins bonnes (Silva et Skinner, 1997). Un exercice de modélisation méticuleux semble donc requis avant d'appliquer l'estimateur optimal de Zhou et Kim (2012). Dans la présente étude, nous nous concentrons sur l'estimateur ajusté par le score de propension, qui est utilisé fréquemment en pratique.

Zhou et Kim (2012) considèrent aussi l'estimation de la variance pour leur estimateur optimal sous le cadre dit inverse (*reverse framework*) de Fay (1992). En considérant l'échantillon obtenu à la période t

comme le résultat d'un processus à deux phases, la première phase étant associée au plan d'échantillonnage original et la seconde, aux étapes de non-réponse successives, il est supposé sous le cadre inverse que ces deux phases peuvent être inversées. Pour cela, le processus à deux phases doit être fortement invariant comme il est défini dans Beaumont et Haziza (2016). Ici, nous proposons pour l'estimateur ajusté par le score de propension un estimateur de variance général qui ne nécessite pas l'hypothèse d'invariance forte. Nous étendons également cet estimateur de variance afin de prendre en compte l'estimation de paramètres complexes, éventuellement avec des poids calés, et pour couvrir les estimateurs longitudinaux. Dans chaque cas, nous proposons aussi un estimateur de variance simplifié et conservatif, plus facilement calculable par les utilisateurs secondaires des données.

La présentation de l'article est la suivante. À la section 2, nous commençons par définir la notation. Puis, nous postulons un modèle paramétrique qui mène à des probabilités de réponse estimées et à un estimateur repondéré. Ensuite, nous établissons un estimateur de variance en suivant l'approche décrite dans Kim et Kim (2007), et proposons également une version simplifiée. Ces estimateurs sont illustrés dans le cas particulier d'un modèle de régression logistique. À la section 3, nous étendons l'estimateur de variance proposé de manière à couvrir les estimateurs calés et les paramètres complexes. À la section 4, nous discutons de l'estimation longitudinale et utilisons l'estimateur de variance proposé pour traiter ce genre de cas. À la section 5, nous comparons les estimateurs de variance au moyen d'une étude par simulations, et à la section 6, nous proposons un exemple d'application aux données de l'enquête ELFE. À la section 7, nous tirons certaines conclusions.

2 Correction de la non-réponse et de l'attrition

2.1 Notation et hypothèses principales

Nous considérons une population finie U. Un échantillon s_0 est d'abord sélectionné selon un plan d'échantillonnage $p(\cdot)$, et nous supposons que les probabilités d'inclusion de premier ordre π_i sont strictement positives pour tout $i \in U$. Cette première phase d'échantillonnage correspond à l'inclusion originale des unités dans l'échantillon.

Nous considérons le cas d'une enquête par panel dans laquelle seules les unités sélectionnées dans l'échantillon original s_0 sont suivies au cours du temps, sans réentrée ou entrée tardive d'unités à des périodes ultérieures pour représenter d'éventuels nouveau-nés. Nous nous intéressons par conséquent à l'estimation d'un paramètre défini sur la population U, pour une variable étudiée y_t prenant la valeur y_{it} pour l'unité i à la période t. Les unités dans l'échantillon s_0 sont suivies à des périodes subséquentes $\delta = 1, \ldots, t$, et l'échantillon est sujet à la non-réponse totale à chaque période. Nous notons r_i^δ l'indicateur de réponse pour l'unité i à la période δ, et s_δ le sous-ensemble de répondants à la période δ.

Nous supposons une non-réponse monotone donnant lieu à la série emboîtée $s_0 \supset s_1 \supset \ldots \supset s_t$. Pour $\delta = 1, \ldots, t$, nous notons $p_i^\delta = \Pr(i \in s_\delta \mid s_{\delta-1})$ la probabilité de réponse d'une unité i à la période δ.

Nous supposons que les données manquent au hasard (MAR), donc que la probabilité de réponse p_i^δ à la période δ peut être expliquée par les variables observées aux périodes $0, \ldots, \delta-1$, y compris les variables d'intérêt; voir, par exemple, Zhou et Kim (2012). En outre, nous supposons que, à toute période δ, les unités répondent indépendamment les unes des autres, et nous notons $p_{ij}^\delta = p_i^\delta p_j^\delta$ la probabilité que deux unités distinctes i et j répondent conjointement à la période δ.

2.2 Estimateur repondéré

Nous cherchons à estimer le total $Y(t) = \sum_{i\in U} y_{it}$ à la période t. En pratique, les probabilités de réponse à chaque période sont inconnues et doivent être estimées. Nous supposons qu'à la date δ, la probabilité de réponse est modélisée paramétriquement par

$$p_i^\delta = f^\delta\left(z_i^\delta, \alpha^\delta\right) \tag{2.1}$$

pour une fonction connue $f^\delta(\cdot,\cdot)$, où z_i^δ est un vecteur de variables observées pour toutes les unités dans $s_{\delta-1}$, et α^δ désigne un paramètre inconnu. Dans cet article, l'indice supérieur δ sera utilisé quand nous tenons compte de la non-réponse à la période δ, comme pour la probabilité p_i^δ que l'unité i réponde à la période δ. En suivant l'approche de Kim et Kim (2007), nous supposons que le paramètre réel est estimé par $\hat{\alpha}^\delta$, la solution de l'équation estimante

$$\frac{\partial}{\partial\alpha}\sum_{i\in s_{\delta-1}} k_i^\delta \left\{ r_i^\delta \ln\left(p_i^\delta\right) + \left(1-r_i^\delta\right)\ln\left(1-p_i^\delta\right)\right\} = 0, \tag{2.2}$$

avec k_i^δ le poids donné à l'unité i dans l'équation estimante. Les choix habituels de ces poids comprennent $k_i^\delta = 1$ et $k_i^\delta = \pi_i^{-1}$, voir Fuller et An (1998), Beaumont (2005), et Kim et Kim (2007).

La probabilité de réponse estimée à la période δ est $\hat{p}_i^\delta = f^\delta\left(z_i^\delta, \hat{\alpha}^\delta\right)$. L'estimateur ajusté par le score de propension à la période t, que nous appellerons simplement l'estimateur repondéré dans la suite, est défini comme

$$\hat{Y}_t(t) = \sum_{i\in s_t} \frac{y_{it}}{\pi_i \hat{p}_i^{1\to t}} \qquad \text{avec} \qquad \hat{p}_i^{1\to t} = \prod_{\delta=1}^{t} \hat{p}_i^\delta. \tag{2.3}$$

Dans cet article, l'indice inférieur t sera utilisé quand l'échantillon observé à la période t est utilisé pour l'estimation, comme pour $\hat{Y}_t(\cdot)$ qui utilise l'échantillon s_t. Nous simplifions la notation en $\hat{Y}_t(t) \equiv \hat{Y}_t$ quand le total à la période t est estimé en utilisant l'échantillon observé à la période t.

2.3 Calcul de la variance

Sous certaines hypothèses de régularité appliquées aux mécanismes de réponse et certaines conditions de régularité appliquées aux $p^\delta(\cdot,\cdot)$, nous obtenons à partir du théorème 1 dans Kim et Kim (2007) l'estimateur que nous pouvons écrire

$$\hat{Y}_t = \hat{Y}_{\text{lin},t}(t) + O_p\left(Nn^{-1}\right), \tag{2.4}$$

où

$$\hat{Y}_{\text{lin},t}(t) = \sum_{i\in s_{t-1}} \frac{1}{\pi_i \hat{p}_i^{1\to t-1}} \left\{ k_i^t \pi_i \hat{p}_i^{1\to t-1} p_i^t \left(h_i^t\right)^\top \gamma^t + \frac{r_i^t}{p_i^t}\left(y_{it} - k_i^t \pi_i \hat{p}_i^{1\to t-1} p_i^t \left(h_i^t\right)^\top \gamma^t\right)\right\}, \tag{2.5}$$

et où, pour tout $\delta = 1, \ldots, t$, nous désignons par h_i^δ la valeur de $h_i^\delta(\alpha) = \partial \text{logit}\left(p_i^\delta\right)/\partial\alpha$ évaluée à $\alpha = \alpha^\delta$, et

$$\gamma^\delta = \left\{\sum_{i\in s_{\delta-1}} k_i^\delta p_i^\delta \left(1-p_i^\delta\right) h_i^\delta \left(h_i^\delta\right)^\top\right\}^{-1} \sum_{i\in s_{\delta-1}} \frac{1-p_i^\delta}{\hat{p}_i^{1\to\delta-1}} h_i^\delta \frac{y_{it}}{\pi_i}. \tag{2.6}$$

Partant de (2.5), nous obtenons que

$$E\left\{\hat{Y}_{\text{lin},t}(t) \mid s_{t-1}\right\} = \hat{Y}_{t-1}(t), \tag{2.7}$$

avec $\hat{Y}_{t-1}(t)$ l'estimateur de $Y(t)$ calculé sur s_{t-1}. En utilisant une preuve par induction, il découle de (2.4) et de (2.7) que $\hat{Y}_t$ est approximativement sans biais pour $Y(t)$. En outre, la variance de $\hat{Y}_t$ peut être approximée asymptotiquement par

$$V_{\text{app}}\left(\hat{Y}_t\right) = V\left(\sum_{i\in s_0} \frac{y_{it}}{\pi_i}\right) + E\left[\sum_{\delta=1}^{t} V\left\{\hat{Y}_{\text{lin},\delta}(t) \mid s_{\delta-1}\right\}\right]. \tag{2.8}$$

Le premier terme du deuxième membre de (2.8) est la variance due au plan d'échantillonnage, que nous notons $V^p\left(\hat{Y}_t\right)$. Le deuxième terme du deuxième membre de (2.8) est la variance due à la non-réponse, que nous notons $V^{\text{nr}}\left(\hat{Y}_t\right)$. Partant de (2.5), cette variance asymptotique est donnée par

$$V^{\text{nr}}\left(\hat{Y}_t\right) = E\left(\sum_{\delta=1}^{t} V^{\text{nr}\delta}\left(\hat{Y}_t\right)\right), \tag{2.9}$$

où

$$V^{\text{nr}\delta}\left(\hat{Y}_t\right) = \sum_{i\in s_{\delta-1}} p_i^\delta \left(1-p_i^\delta\right)\left(\frac{y_{it}}{\pi_i \hat{p}_i^{1\to\delta-1} p_i^\delta} - k_i^\delta \left(h_i^\delta\right)^\top \gamma^\delta\right)^2. \tag{2.10}$$

Nous notons que, pour chacune de ses composantes $\delta = 1, \ldots, t$, le terme $V^{\text{nr}\delta}\left(\hat{Y}_t\right)$ dans (2.10) inclut un terme de centrage $k_i^\delta \left(h_i^\delta\right)^\top \gamma^\delta$, qui est essentiellement une prédiction de $\left(\pi_i \hat{p}_i^{1\to\delta-1} p_i^\delta\right)^{-1} y_i$ au moyen des variables explicatives h_i^δ. Ce centrage est dû à l'estimation des probabilités de réponse. Si nous supprimions ces termes de centrage, les équations (2.9) et (2.10) mèneraient à la variance de l'estimateur de $Y(t)$ que nous obtiendrions en remplaçant les probabilités estimées par leurs valeurs réelles dans (2.3). La variance de cet estimateur est habituellement plus grande que celle de l'estimateur repondéré en (2.3); voir aussi Beaumont (2005), équation (5.7), et Kim et Kim (2007), équation (17), pour le cas où $t = 1$.

2.4 Estimation de la variance

À la période t, un estimateur approximativement sans biais de la variance due au plan d'échantillonnage $V^p\left(\hat{Y}_t\right)$ est

$$\hat{V}_t^p\left(\hat{Y}_t\right) = \sum_{i,j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{y_{it}}{\pi_i} \frac{y_{jt}}{\pi_j}, \tag{2.11}$$

où $\hat{p}_{ij}^{1\to t} \equiv \prod_{\delta=1}^{t} \hat{p}_{ij}^{\delta}$, et où $\hat{p}_{ij}^{\delta} = \hat{p}_i^{\delta}$ si $i = j$, et $\hat{p}_{ij}^{\delta} = \hat{p}_i^{\delta}\hat{p}_j^{\delta}$ autrement. En suivant l'équation (25) dans Kim et Kim (2007), $V^{\mathrm{nr}}\left(\hat{Y}_t\right)$ peut être estimé approximativement sans biais à la période t par

$$\hat{V}_t^{\mathrm{nr}}\left(\hat{Y}_t\right) = \sum_{\delta=1}^{t} \hat{V}_t^{\mathrm{nr}\delta}\left(\hat{Y}_t\right) \tag{2.12}$$

où

$$\hat{V}_t^{\mathrm{nr}\delta}\left(\hat{Y}_t\right) = \sum_{i\in s_t} \frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}} \left(\frac{y_{it}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\hat{\gamma}_t^{\delta}\right)^2, \tag{2.13}$$

$$\hat{h}_i^{\delta} = h\left(z_i, \hat{\alpha}^{\delta}\right), \tag{2.14}$$

$$\hat{\gamma}_t^{\delta} = \left\{\sum_{i\in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}} \hat{h}_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}} \hat{h}_i^{\delta} \frac{y_{it}}{\pi_i}. \tag{2.15}$$

Cela mène à l'estimateur de variance globale à la période t

$$\hat{V}_t\left(\hat{Y}_t\right) = \hat{V}_t^p\left(\hat{Y}_t\right) + \hat{V}_t^{\mathrm{nr}}\left(\hat{Y}_t\right). \tag{2.16}$$

Un estimateur simplifié de la variance due à la non-réponse s'obtient en ignorant les termes de prédiction $k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\hat{\gamma}_t^{\delta}$ pour chacune des $\delta = 1, \ldots, t$ composantes de variance. Après un peu de calcul, nous obtenons l'estimateur de variance simplifié

$$\hat{V}_{t,\,\mathrm{simp}}^{\mathrm{nr}}\left\{\hat{Y}_t(t)\right\} = \sum_{i\in s_t} \frac{1-\hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2}\left(\frac{y_{it}}{\pi_i}\right)^2. \tag{2.17}$$

Le principal avantage de cet estimateur de variance simplifié tient au fait qu'il ne faut connaître que les probabilités de réponse estimées. Par contre, le calcul de l'estimateur de variance en (2.12) requiert la connaissance des modèles de réponse utilisés à toutes les périodes. L'estimateur de variance simplifié est par conséquent particulièrement intéressant pour les utilisateurs secondaires des données d'enquête, pour lesquels les probabilités de réponse estimées pourraient être la seule information disponible liée à la modélisation de la réponse. Cet estimateur de variance simplifié aura tendance à surestimer la variance due à la non-réponse de $\left(\hat{Y}_t\right)$ si le terme de prédiction $k_i^{\delta}\left(h_i^{\delta}\right)^{\top}\gamma^{\delta}$ explique partiellement $\left(\pi_i \hat{p}_i^{1\to\delta-1} p_i^{\delta}\right)^{-1} y_{it}$.

2.5 Application au modèle de régression logistique

Dans le cas particulier où un modèle de régression logistique est utilisé à chaque période δ, le modèle (2.1) peut être réécrit sous la forme

$$\text{logit}\left(p_i^\delta\right) = \left(z_i^\delta\right)^\top \alpha^\delta. \tag{2.18}$$

Nous obtenons $\hat{h}_i^\delta = z_i^\delta$, et l'estimateur de la variance due à la non-réponse est donné par (2.12), avec

$$\hat{V}_t^{\text{nr}\delta}\left(\hat{Y}_t\right) = \sum_{i\in s_t} \frac{\hat{p}_i^\delta\left(1-\hat{p}_i^\delta\right)}{\hat{p}_i^{\delta\to t}}\left(\frac{y_{it}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^\delta\left(z_i^\delta\right)^\top \hat{\gamma}_t^\delta\right)^2, \tag{2.19}$$

$$\hat{\gamma}_t^\delta = \left\{\sum_{i\in s_t} k_i^\delta \frac{\hat{p}_i^\delta\left(1-\hat{p}_i^\delta\right)}{\hat{p}_i^{\delta\to t}} z_i^\delta\left(z_i^\delta\right)^\top\right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^\delta}{\hat{p}_i^{1\to t}} z_i^\delta \frac{y_{it}}{\pi_i}. \tag{2.20}$$

Si l'estimateur repondéré est calculé à la période $t = 1$, l'estimateur en (2.12) de la variance due à la non-réponse peut être réécrit sous la forme

$$\hat{V}_1^{\text{nr}}\left(\hat{Y}_1\right) = \sum_{i\in s_1}\left(1-\hat{p}_i^1\right)\left(\frac{y_{i1}}{\pi_i \hat{p}_i^1} - k_i^1\left(z_i^1\right)^\top \hat{\gamma}_1^1\right)^2. \tag{2.21}$$

Si l'estimateur repondéré est calculé à la période $t = 2$, l'estimateur en (2.12) de la variance due à la non-réponse peut être réécrit sous la forme

$$\begin{aligned}\hat{V}_2^{\text{nr}}\left(\hat{Y}_2\right) &= \sum_{i\in s_2} \frac{\left(1-\hat{p}_i^1\right)}{\hat{p}_i^2}\left(\frac{y_{i2}}{\pi_i \hat{p}_i^1} - k_i^1\left(z_i^1\right)^\top \hat{\gamma}_2^1\right)^2 \\ &\quad + \sum_{i\in s_2}\left(1-\hat{p}_i^2\right)\left(\frac{y_{i2}}{\pi_i \hat{p}_i^1 \hat{p}_i^2} - k_i^2\left(z_i^2\right)^\top \hat{\gamma}_2^2\right)^2.\end{aligned} \tag{2.22}$$

En pratique, on suppose souvent que l'on peut appliquer le modèle des groupes de réponse homogènes (GRH) pour corriger la non-réponse totale. Sous ce modèle, il est posé qu'à chaque période $\delta = 1, \ldots, t$, le sous-échantillon $s_{\delta-1}$ peut être partitionné en $C(\delta-1)$ groupes $s_{\delta-1}^c$, $c = 1, \ldots, C(\delta-1)$, tels que la probabilité de réponse p_i^δ est constante à l'intérieur d'un groupe. Ce modèle est un cas particulier du modèle de régression logistique en (2.18), obtenu avec

$$z_i^\delta = \left[1\left\{i \in s_{\delta-1}^1\right\}, \ldots, 1\left\{i \in s_{\delta-1}^{C(\delta-1)}\right\}\right]^\top, \tag{2.23}$$

et la variance due à la non-réponse est estimée en conséquence. Les formules explicites sont données en annexe.

3 Calage et paramètres complexes

Dans la plupart des enquêtes, une étape de calage est utilisée pour obtenir des poids ajustés qui permettent d'améliorer la précision des estimations du total. Ces estimateurs calés sont examinés à la section 3.1. En

outre, des paramètres plus complexes que les totaux présentent souvent un intérêt, et une étape de linéarisation peut être utilisée pour l'estimation de la variance. Ce sujet est abordé à la section 3.2. L'estimation des paramètres complexes à l'aide de poids calés est traitée à la section 3.3. Dans chaque cas, nous établissons les formules explicites pour l'estimation de la variance et l'estimation simplifiée de la variance, et discutons du biais de l'estimateur de variance simplifié.

3.1 Estimation de la variance pour les estimateurs calés d'un total

Supposons qu'un vecteur x_i de variables auxiliaires existe pour toute unité $i \in s_t$, et que le vecteur de totaux X sur la population U est connu. Alors, une étape de calage additionnelle (Deville et Särndal, 1992) est habituellement appliquée à $\hat{Y}_t$. Elle consiste à modifier les poids $d_{ti} = \pi_i^{-1} \left(\hat{p}_i^{1 \to t} \right)^{-1}$ pour obtenir les poids calés w_{ti} qui permettent de faire correspondre l'estimation au total réel X, en ce sens que

$$\sum_{i \in s_t} w_{ti} x_i = X. \tag{3.1}$$

Les nouveaux poids calés sont choisis de manière à minimiser une fonction de distance par rapport aux poids originaux, tout en satisfaisant (3.1). Cela mène à l'estimateur calé

$$\hat{Y}_{wt} = \sum_{i \in s_t} w_{ti} y_{it}. \tag{3.2}$$

Le résidu estimé pour la régression pondérée de y_{it} sur x_i est donné par

$$e_{it} = y_{it} - \hat{b}_t x_i \tag{3.3}$$

avec

$$\hat{b}_t = \left(\sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1 \to t}} x_i x_i^\top \right)^{-1} \sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1 \to t}} x_i y_{it}. \tag{3.4}$$

Le remplacement de la variable y_{it} par e_{it} dans (2.11) donne l'estimateur de la variance due au plan d'échantillonnage

$$\hat{V}_t^p \left(\hat{Y}_{wt} \right) = \sum_{i,\ j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1 \to t}} \frac{e_{it}}{\pi_i} \frac{e_{jt}}{\pi_j}. \tag{3.5}$$

De même, le remplacement de la variable y_{it} par e_{it} dans en (2.12) donne l'estimateur de la variance due à la non-réponse

$$\hat{V}_t^{\text{nr}} \left(\hat{Y}_{wt} \right) = \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta \left(1 - \hat{p}_i^\delta \right)}{\hat{p}_i^{\delta \to t}} \left(\frac{e_{it}}{\pi_i \hat{p}_i^{1 \to \delta}} - k_i^\delta \left(\hat{h}_i^\delta \right)^\top \hat{\gamma}_{te}^\delta \right)^2 \tag{3.6}$$

$$\hat{\gamma}_{te}^\delta = \left\{ \sum_{i \in s_t} k_i^\delta \frac{\hat{p}_i^\delta \left(1 - \hat{p}_i^\delta \right)}{\hat{p}_i^{\delta \to t}} \hat{h}_i^\delta \left(\hat{h}_i^\delta \right)^\top \right\}^{-1} \sum_{i \in s_t} \frac{1 - \hat{p}_i^\delta}{\hat{p}_i^{1 \to t}} \hat{h}_i^\delta \frac{e_{it}}{\pi_i}. \tag{3.7}$$

L'estimateur de variance globale pour $\hat{Y}_{wt}$ est

$$\hat{V}_t\left(\hat{Y}_{wt}\right) = \hat{V}_t^{\,p}\left(\hat{Y}_{wt}\right) + \hat{V}_t^{\,\mathrm{nr}}\left(\hat{Y}_{wt}\right). \tag{3.8}$$

L'estimateur simplifié de la variance due à la non-réponse est

$$\hat{V}_{t,\,\mathrm{simp}}^{\,\mathrm{nr}}\left(\hat{Y}_{wt}\right) = \sum_{i\in s_t} \frac{1-\hat{p}_i^{\,1\to t}}{\left(\hat{p}_i^{\,1\to t}\right)^2}\left(\frac{e_{it}}{\pi_i}\right)^2. \tag{3.9}$$

De nouveau, cet estimateur de variance simplifié ne tient pas compte des termes de prédiction $k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\hat{\gamma}_{te}^{\delta}$. Si le modèle de calage sous-jacent est approprié, alors le pouvoir explicatif de $\hat{h}_i^{\delta}$ pour e_{it} sera, en principe, faible, de même que le biais de l'estimateur de variance simplifié. Par contre, s'il reste dans e_{it} une part importante de y_{it} qui ne peut pas être expliquée par x_i, le biais de l'estimateur de variance simplifié risque d'être non négligeable. Cette situation peut se produire dans le cas de l'estimation sur domaine, quand les variables de calage ne comprennent aucune information auxiliaire propre au domaine.

3.2 Estimation de la variance pour des paramètres complexes

Nous pourrions vouloir estimer des paramètres plus complexes que les totaux. Supposons que la variable d'intérêt y_{it} est une multivariable q – dimensionnelle, et que le paramètre d'intérêt est $\theta(t) = f\{Y(t)\}$ avec $f(\cdot)$ une fonction connue. À la période t, la substitution de $\hat{Y}_t$ dans $\theta(t)$ donne l'estimateur *plug-in*, ou estimateur par substitution, $\hat{\theta}_t = f\left(\hat{Y}_t\right)$.

La variable linéarisée estimée de $\theta(t)$ est

$$u_{it} = \left\{f'\left(\hat{Y}_t\right)\right\}^{\top} y_{it}, \tag{3.10}$$

avec $f'\left(\hat{Y}_t\right)$ le vecteur q – dimensionnel des dérivées premières de f au point $\hat{Y}_t$. Le remplacement de la variable y_{it} par u_{it} dans (2.11) donne l'estimateur de la variance due au plan d'échantillonnage

$$\hat{V}_t^{\,p}\left(\hat{\theta}_t\right) = \sum_{i,\ j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{\,1\to t}} \frac{u_{it}}{\pi_i} \frac{u_{jt}}{\pi_j}. \tag{3.11}$$

De même, le remplacement de la variable y_{it} par u_{it} dans (2.12) donne l'estimateur de la variance due à la non-réponse

$$\hat{V}_t^{\,\mathrm{nr}}\left(\hat{\theta}_t\right) = \sum_{\delta=1}^{t}\sum_{i\in s_t} \frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\,\delta\to t}}\left(\frac{u_{it}}{\pi_i \hat{p}_i^{\,1\to\delta}} - k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\hat{\gamma}_{t\theta}^{\delta}\right)^2 \tag{3.12}$$

$$\hat{\gamma}_{t\theta}^{\delta} = \left\{\sum_{i\in s_t} k_i^{\delta}\frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\,\delta\to t}}\hat{h}_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\right\}^{-1}\sum_{i\in s_t}\frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{\,1\to t}}\hat{h}_i^{\delta}\frac{u_{it}}{\pi_i}. \tag{3.13}$$

L'estimateur de variance globale pour $\hat{\theta}_t$ est

$$\hat{V}_t\left(\hat{\theta}_t\right) = \hat{V}_t^{p}\left(\hat{\theta}_t\right) + \hat{V}_t^{\mathrm{nr}}\left(\hat{\theta}_t\right). \tag{3.14}$$

L'estimateur simplifié de la variance due à la non-réponse est

$$\hat{V}_{t,\ \mathrm{simp}}^{\mathrm{nr}}\left(\hat{\theta}_t\right) = \sum_{i\in s_t} \frac{1-\hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2}\left(\frac{u_{it}}{\pi_i}\right)^2. \tag{3.15}$$

Le biais de cet estimateur de variance simplifié dépendra du pouvoir explicatif de $\hat{h}_i^{\delta}$ sur la variable linéarisée u_{it}.

3.3 Estimation de la variance pour des paramètres complexes sous calage

Les poids calés w_{ti} peuvent être utilisés pour obtenir un estimateur du paramètre $\theta(t)$. La substitution de $\hat{Y}_{wt}$ dans $\theta(t) = f\{Y(t)\}$ donne l'estimateur *plug-in* calé $\hat{\theta}_{wt} = f\left(\hat{Y}_{wt}\right)$. Pour obtenir un estimateur de variance pour $\hat{\theta}_{wt}$, nous commençons par calculer la variable linéarisée estimée $u_{it} = \left\{f'\left(\hat{Y}_t\right)\right\}^{\top} y_{it}$ et prenons

$$e_{\theta it} = u_{it} - \hat{b}_{\theta t} x_i \tag{3.16}$$

avec

$$\hat{b}_{\theta t} = \left(\sum_{i\in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i x_i^{\top}\right)^{-1} \sum_{i\in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i u_{it}. \tag{3.17}$$

Le remplacement de la variable y_{it} par $e_{\theta it}$ dans (2.11) donne l'estimateur de la variance due au plan d'échantillonnage

$$\hat{V}_t^{p}\left(\hat{\theta}_{wt}\right) = \sum_{i,\ j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{e_{\theta it}}{\pi_i} \frac{e_{\theta jt}}{\pi_j}. \tag{3.18}$$

De même, le remplacement de la variable y_{it} par $e_{\theta it}$ dans (2.12) donne l'estimateur de la variance due à la non-réponse

$$\hat{V}_t^{\mathrm{nr}}\left(\hat{\theta}_{wt}\right) = \sum_{\delta=1}^{t}\sum_{i\in s_t} \frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}}\left(\frac{e_{\theta it}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\hat{\gamma}_{te\theta}^{\delta}\right)^2 \tag{3.19}$$

$$\hat{\gamma}_{te\theta}^{\delta} = \left\{\sum_{i\in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}} \hat{h}_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}} \hat{h}_i^{\delta} \frac{e_{\theta it}}{\pi_i}. \tag{3.20}$$

L'estimateur de variance globale pour $\hat{\theta}_{wt}$ est

$$\hat{V}_t\left(\hat{\theta}_{wt}\right) = \hat{V}_t^{p}\left(\hat{\theta}_{wt}\right) + \hat{V}_t^{\mathrm{nr}}\left(\hat{\theta}_{wt}\right). \tag{3.21}$$

L'estimateur simplifié de la variance due à la non-réponse est

$$\hat{V}_{t,\,\text{simp}}^{\,\text{nr}}\left(\hat{\theta}_{wt}\right)=\sum_{i\in s_t}\frac{1-\hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2}\left(\frac{e_{\theta it}}{\pi_i}\right)^2. \tag{3.22}$$

Puisque la variable $e_{\theta it}$ obtenue correspond au résidu de la régression de la variable linéarisée u_{it} sur les variables de calage x_i, le pouvoir explicatif de $\hat{h}_i^{\delta}$ sur $e_{\theta it}$ sera, en principe, faible en pratique, et le biais de l'estimateur de variance simplifié sera, en principe, faible également.

4 Estimateurs longitudinaux

Nous pourrions nous intéresser à une évolution des paramètres, telle que

$$\Delta\left(u\to t\right)=Y\left(t\right)-Y\left(u\right), \tag{4.1}$$

la différence entre les totaux d'une variable d'intérêt mesurée à deux périodes différentes $u < t$. Puisque la variable y_{iu} est mesurée sur tous les sous-échantillons $s_{u'}$ pour $u' = u, \ldots, t$, il existe plusieurs estimateurs possibles pour $\Delta\left(u\to t\right)$. Pour $u' = u, \ldots, t$, nous désignons par

$$\hat{\Delta}_{u't}\left(u\to t\right)=\sum_{i\in s_t}\frac{y_{it}}{\pi_i\hat{p}_i^{1\to t}}-\sum_{i\in s_{u'}}\frac{y_{iu}}{\pi_i\hat{p}_i^{1\to u'}} \tag{4.2}$$

l'estimateur qui utilise s_t pour l'estimation de $Y\left(t\right)$ et $s_{u'}$ pour l'estimation de $Y\left(u\right)$. Le cas $u' = u$ correspond à l'estimation de $Y\left(u\right)$ sur le plus grand sous-échantillon disponible, s_u. Le cas $u' = t$ correspond à l'estimation de $Y\left(u\right)$ et $Y\left(t\right)$ sur le sous-échantillon commun s_t.

Dans le contexte d'une réponse complète, plusieurs auteurs ont recommandé l'estimateur $\hat{\Delta}_{tt}\left(u\to t\right)$ qui utilise l'échantillon commun uniquement, si les variables y_{ui} et y_{ti} sont fortement positivement corrélées; voir Caron et Ravalet (2000), Qualité et Tillé (2008), Goga, Deville et Ruiz-Gazen (2009), Chauvet et Goga (2018). Dans notre contexte, ce choix peut être justifié de manière heuristique comme suit. Pour $u' < t$, et en conditionnant sur le sous-échantillon $s_{u'}$, nous obtenons

$$V\left\{\hat{\Delta}_{u't}\left(u\to t\right)\right\}\simeq V\left\{\sum_{i\in s_{u'}}\frac{y_{it}-y_{iu}}{\pi_i\hat{p}_i^{1\to u'}}\right\}+EV\left\{\sum_{i\in s_t}\frac{y_{it}}{\pi_i\hat{p}_i^{1\to t}}\;\middle|\;s_{u'}\right\}, \tag{4.3}$$

$$V\left\{\hat{\Delta}_{tt}\left(u\to t\right)\right\}\simeq V\left\{\sum_{i\in s_{u'}}\frac{y_{it}-y_{iu}}{\pi_i\hat{p}_i^{1\to u'}}\right\}+EV\left\{\sum_{i\in s_t}\frac{y_{it}-y_{iu}}{\pi_i\hat{p}_i^{1\to t}}\;\middle|\;s_{u'}\right\}. \tag{4.4}$$

Dans les équations (4.3) et (4.4), le premier terme du deuxième membre est identique. Puisqu'on s'attend à ce que les variables y_{iu} et y_{it} soient positivement corrélées, on s'attend à ce que la différence $y_{it} - y_{iu}$ soit plus petite que y_{it}. Donc, l'estimateur $\hat{\Delta}_{tt}\left(u\to t\right)$ basé sur l'échantillon commun sera, en principe, plus efficace en ce qui concerne la variance. Les résultats d'une petite étude par simulations présentés à la section 5.2 appuient ce raisonnement heuristique. Par conséquent, nous nous concentrons uniquement ici

sur l'estimateur $\hat{\Delta}_{tt}(u \to t)$ pour l'estimation de $\Delta(u \to t)$. Comme l'a souligné un examinateur, et en suivant l'approche décrite dans Zhou et Kim (2012), nous pourrions réaliser un gain d'efficacité en utilisant toute l'information sur s_u, à savoir en calant les poids $\left(\pi_i \hat{p}_i^{1\to t}\right)^{-1}$ sur l'estimateur $\hat{Y}_u$.

Le remplacement de la variable y_{it} par $y_{it} - y_{iu}$ dans (2.11) donne l'estimateur de la variance due au plan d'échantillonnage

$$\hat{V}_t^p \left\{ \hat{\Delta}_{tt}(u \to t) \right\} = \sum_{i,\ j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{(y_{it} - y_{iu})}{\pi_i} \frac{(y_{jt} - y_{ju})}{\pi_j}. \tag{4.5}$$

De même, le remplacement de la variable y_{it} par $y_{it} - y_{iu}$ dans (2.12) donne l'estimateur de la variance due à la non-réponse

$$\hat{V}_t^{\text{nr}} \left\{ \hat{\Delta}_{tt}(u \to t) \right\} = \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta \left(1 - \hat{p}_i^\delta\right)}{\hat{p}_i^{\delta \to t}} \left(\frac{y_{it} - y_{iu}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^\delta \left(\hat{h}_i^\delta\right)^\top \hat{\gamma}_{t\Delta}^\delta \right)^2 \tag{4.6}$$

avec

$$\hat{\gamma}_{t\Delta}^\delta = \left\{ \sum_{i \in s_t} k_i^\delta \frac{\hat{p}_i^\delta \left(1 - \hat{p}_i^\delta\right)}{\hat{p}_i^{\delta \to t}} \hat{h}_i^\delta \left(\hat{h}_i^\delta\right)^\top \right\}^{-1} \sum_{i \in s_t} \frac{1 - \hat{p}_i^\delta}{\hat{p}_i^{1\to t}} \hat{h}_i^\delta \frac{y_{it} - y_{iu}}{\pi_i}. \tag{4.7}$$

L'estimateur de variance globale pour $\hat{\Delta}_{tt}(u \to t)$ est

$$\hat{V}_t \left\{ \hat{\Delta}_{tt}(u \to t) \right\} = \hat{V}_t^p \left\{ \hat{\Delta}_{tt}(u \to t) \right\} + \hat{V}_t^{\text{nr}} \left\{ \hat{\Delta}_{tt}(u \to t) \right\}. \tag{4.8}$$

L'estimation de la variance pour les mesures de variation est également abordée dans Berger (2004), Qualité et Tillé (2008), Goga et coll. (2009), et Chauvet et Goga (2018), entre autres.

L'estimateur simplifié de la variance due à la non-réponse est

$$\hat{V}_{t,\ \text{simp}}^{\text{nr}} \left\{ \hat{\Delta}_{tt}(u \to t) \right\} = \sum_{i \in s_t} \frac{1 - \hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2} \left(\frac{y_{it} - y_{iu}}{\pi_i} \right)^2. \tag{4.9}$$

Si les variables y_{it} et y_{iu} sont fortement positivement corrélées, le biais de l'estimateur de variance simplifié sera, en principe, faible.

5 Une étude par simulations

À la présente section, plusieurs populations artificielles sont générées conformément au modèle décrit à la section 5.1. À la section 5.2, nous considérons plusieurs estimateurs de la variation d'un total, qui illustrent le raisonnement heuristique de la section 4. À la section 5.3, nous présentons une expérience Monte Carlo et comparons plusieurs estimateurs de variance pour l'estimation d'un total, d'un ratio ou d'un

paramètre d'évolution. Les résultats des tableaux 5.1 et 5.2 peuvent être reproduits facilement en utilisant le code R fourni dans le matériel supplémentaire.

5.1 Configuration de la simulation

Nous considérons sept populations de taille 10 000, contenant chacune trois variables d'intérêt y_{i1}, y_{i2} et y_{i3} observées aux périodes $t = 1,\ 2$ et 3, respectivement. Les variables d'intérêt sont générées selon le modèle de superpopulation

$$y_{i1} = \alpha^0 + \alpha^a x_{ai} + \alpha^b x_{bi} + \sigma u_{i1}, \tag{5.1}$$

$$y_{i2} = \rho y_{i1} + \sigma u_{i2}, \tag{5.2}$$

$$y_{i3} = \rho y_{i2} + \sigma u_{i3}. \tag{5.3}$$

Les variables auxiliaires x_{ai} et x_{bi} sont générées indépendamment à partir d'une loi Gamma de paramètres de forme et d'échelle 2 et 1. Deux variables auxiliaires x_{ci} et x_{di}, non liées aux variables d'intérêt, sont générées de façon similaire. Les variables u_{i1}, u_{i2} et u_{i3} sont générées indépendamment à partir d'une loi normale centrée réduite. Nous utilisons $\alpha^0 = 10$, $\alpha^a = \alpha^b = 5$ et $\sigma = 10$, ce qui donne un coefficient de détermination $\left(R^2\right)$ dans le modèle (5.1) approximativement égal à 0,50. Le paramètre ρ est fixé à 0; 0,2; 0,4; 0,6; 0,8; 1,0 et 1,2 pour les populations 1 à 7, respectivement.

De chaque population est tiré un échantillon aléatoire simple s_0 de taille $n =$ 1 000. Trois phases de non-réponse sont ensuite simulées successivement. À chaque phase $\delta = 1,\ 2,\ 3$, le sous-échantillon de répondants s_δ est obtenu par échantillonnage de Poisson avec probabilité de réponse p_i^δ pour l'unité i, définie comme étant

$$\text{logit}\left(p_i^\delta\right) = \beta^{\delta 0} + \beta^{\delta a} x_{ai} + \beta^{\delta b} x_{bi}. \tag{5.4}$$

Nous utilisons $\beta^{\delta 0} = -1$ à chaque phase $\delta = 1,\ 2,\ 3$. Pour $\delta = 1$, nous utilisons $\beta^{1a} = \beta^{1b} =$ 0,60, ce qui correspond à un taux de réponse moyen de 0,75. Pour $\delta = 2,\ 3$, nous utilisons $\beta^{\delta a} = \beta^{\delta b} =$ 0,75, ce qui correspond à un taux de réponse moyen de 0,81. Dans chaque sous-échantillon s_δ, les probabilités de réponse estimées $\hat{p}_i^\delta$ sont obtenues par une régression logistique non pondérée.

5.2 Comparaison des estimateurs pour une différence de totaux

Dans cette section, nous comparons la précision de deux estimateurs pour une différence de totaux $\Delta\left(u \to t\right)$ pour $u = 1$ et $t = 2$, pour $u = 1$ et $t = 3$, et pour $u = 2$ et $t = 3$. Nous considérons l'estimateur $\hat{\Delta}_{ut}\left(u \to t\right)$, qui utilise l'intégralité des sous-échantillons appropriés pour les variables y_{iu} et y_{it}, et l'estimateur $\hat{\Delta}_{tt}\left(u \to t\right)$, qui utilise le sous-échantillon commun uniquement. Ces deux estimateurs sont comparés en se basant sur la différence relative (DR) de leurs variances, qui est définie comme il suit :

$$\mathrm{DR}\,(u \to t) = 100 \times \frac{V\left\{\hat{\Delta}_{ut}\,(u \to t)\right\} - V\left\{\hat{\Delta}_{tt}\,(u \to t)\right\}}{V\left\{\hat{\Delta}_{tt}\,(u \to t)\right\}}. \qquad (5.5)$$

Les variances réelles sont remplacées par leurs approximations Monte Carlo, obtenues en répétant $B =$ 100 000 fois les phases de sélection de l'échantillon et de non-réponse.

Les résultats sont présentés au tableau 5.1. Une DR positive indique que l'utilisation de l'échantillon commun uniquement produit un estimateur plus précis. Comme on pouvait s'y attendre, la DR augmente dans tous les cas avec ρ, quand la corrélation entre y_{it} et y_{iu} augmente. Pour $u = 1$ et $t = 2$, et pour $u = 2$ et $t = 3$, l'estimateur $\hat{\Delta}_{tt}\,(u \to t)$ est plus précis pour ρ plus grand que 0,6. Pour $u = 1$ et $t = 3$, $\hat{\Delta}_{tt}\,(u \to t)$ est plus précis pour ρ plus grand que 0,8.

Tableau 5.1
Différence relative (DR) entre deux estimateurs pour une différence de totaux

ρ	DR $(1 \to 2)$	DR $(1 \to 3)$	DR $(2 \to 3)$
0,0	-12	-27	-13
0,2	-09	-25	-11
0,4	-04	-20	-03
0,6	05	-09	11
0,8	17	11	39
1,0	30	33	83
1,2	40	46	127

5.3 Performances des estimateurs de variance

Dans cette section, nous considérons la population artificielle 5 $(\rho = 0{,}8)$ générée comme il est décrit à la section 5.1. La sélection de l'échantillon par échantillonnage aléatoire simple de taille $n =$ 1 000 et les trois phases de non-réponse sont appliquées $B =$ 5 000 fois. Nous voulons évaluer les estimateurs de variance et les estimateurs de variance simplifiés, dans le cas de l'estimation d'un total, d'un ratio ou d'une variation d'un total.

Comme pour le total $Y\,(t)$, à chaque période $t = 1, 2, 3$, nous considérons trois estimateurs. L'estimateur $\hat{Y}_t$ utilise les poids $d_{ti} = \pi_i^{-1}\left(\hat{p}_i^{1\to t}\right)^{-1}$. L'estimateur $\hat{Y}_{wt}$ utilise les poids w_i, obtenus par calage des poids d_{ti} sur la taille de la population et les totaux des variables auxiliaires x_{ai} et x_{bi}. L'estimateur $\hat{Y}_{\tilde{w}t}$ utilise les poids $\tilde{w}_i$, obtenus par calage des poids d_{ti} sur la taille de la population et sur les totaux des variables auxiliaires x_{ci} et x_{di}. Le modèle de travail est donc bien spécifié pour $\hat{Y}_{wt}$, mais non pour $\hat{Y}_{\tilde{w}t}$. L'estimateur de variance proposé pour $\hat{Y}_t$ est obtenu à partir de l'équation (2.16), et l'estimateur de variance simplifié est obtenu en insérant dans (2.16) l'estimateur de variance simplifié pour la non-réponse donné en (2.17). Les estimateurs de variance proposés pour $\hat{Y}_{wt}$ et $\hat{Y}_{\tilde{w}t}$ sont obtenus à partir de l'équation (3.8), et les estimateurs de variance simplifiés sont obtenus en insérant dans (3.8) l'estimateur de variance simplifié pour la non-réponse donné en (3.9).

Nous souhaitons aussi estimer le ratio $R(t) = Y(t)/Y(1)$ pour $t = 2, 3$. À chaque période t, nous considérons trois estimateurs. L'estimateur $\hat{R}_t$ utilise les poids d_i. L'estimateur de variance proposé est obtenu à partir de l'équation (3.14), en utilisant la variable linéarisée estimée $u_{it} = \left(\hat{Y}_1\right)^{-1}\left(y_{ti} - \hat{R}_t y_{1i}\right)$. L'estimateur de variance simplifié est obtenu en insérant dans (3.14) l'estimateur de variance simplifié pour la non-réponse donné en (3.15). Les estimateurs $\hat{R}_{wt}$ et $\hat{R}_{\tilde{w}t}$ utilisent les poids calés w_i et $\tilde{w}_i$. Les estimateurs de variance proposés sont obtenus à partir de l'équation (3.21). Les estimateurs de variance simplifiés sont obtenus en insérant dans (3.21) l'estimateur de variance simplifié pour la non-réponse donné en (3.22).

Enfin, nous souhaitons estimer la variation des totaux $\Delta(1 \to t)$ pour $t = 2, 3$. À chaque période t, nous considérons trois estimateurs. L'estimateur $\hat{\Delta}_{tt}(1 \to t)$ utilise les poids d_i. L'estimateur de variance proposé est obtenu à partir de l'équation (4.8), et l'estimateur de variance simplifié est obtenu en insérant dans (4.8) l'estimateur de variance simplifié pour la non-réponse donné en (4.9). Les estimateurs $\hat{\Delta}_{tt,\,w}(1 \to t)$ et $\hat{\Delta}_{tt,\,\tilde{w}}(1 \to t)$ utilisent les poids calés w_i et $\tilde{w}_i$. Les estimateurs de variance proposés sont obtenus à partir de l'équation (4.8), en remplaçant $y_{it} - y_{iu}$ par le résidu estimé pour la régression pondérée de $y_{it} - y_{iu}$ sur les variables de calage. Les estimateurs de variance simplifiés sont obtenus en insérant dans (4.8) l'estimateur de variance simplifié pour la non-réponse donné en (4.9).

Pour un estimateur de variance proposé $\hat{V}$, nous calculons le biais relatif Monte Carlo en pourcentage

$$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}\right) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V}$$

où la variance globale V a été approximée par exécution d'un ensemble indépendant de 100 000 simulations. Pour évaluer la contribution d'une composante $\hat{V}_a$ à l'estimateur de variance $\hat{V}$, nous avons calculé la contribution (en pourcentage)

$$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_a\right) = 100 \times \frac{\frac{1}{B}\sum_{b=1}^{B}\hat{V}_a^{(b)}}{\frac{1}{B}\sum_{b=1}^{B}\hat{V}^{(b)}}.$$

Pour évaluer l'estimateur de variance simplifié pour la non-réponse $\hat{V}_{\mathrm{simp}}^{\mathrm{nr}}$, nous avons calculé le biais relatif Monte Carlo en pourcentage

$$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}_{\mathrm{simp}}^{\mathrm{nr}}\right) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}_{\mathrm{simp}}^{(b)} - V^{\mathrm{nr}}}{V^{\mathrm{nr}}},$$

où la variance V^{nr} due à la non-réponse a été approximée en exécutant un ensemble indépendant de 100 000 simulations.

Les résultats des simulations sont présentés au tableau 5.2. L'estimateur de variance proposé est presque sans biais dans tous les cas. Comme on pouvait s'y attendre, la contribution de la variance due au plan d'échantillonnage diminue avec le temps, à mesure que le nombre de répondants diminue et que la variance

due à la non-réponse devient plus grande. L'estimateur de variance simplifié est fortement biaisé pour la variance due à la non-réponse dans le cas de $\hat{Y}_t$. Le biais diminue rapidement avec le temps, mais demeure grand à la période $t = 3$. L'estimateur de variance simplifié est presque sans biais pour un estimateur calé quand le modèle de travail est spécifié adéquatement, mais est sévèrement biaisé autrement. Ces observations sont cohérentes avec notre raisonnement de la section 3.1. L'estimateur de variance simplifié est presque sans biais pour les trois estimateurs du ratio, et pour les estimateurs calés de la variation des totaux. Dans le cas de l'estimateur non calé pour la variation des totaux, le biais peut atteindre 30 %.

Tableau 5.2
Biais relatif d'un estimateur de variance globale, contribution relative des estimateurs des composantes de la variance et biais relatif d'un estimateur de variance simplifié pour la variance due à la non-réponse pour l'estimation d'un total, d'un ratio ou d'une variation des totaux avec trois ensembles de poids

	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$
	$\hat{Y}_t$			$\hat{Y}_{wt}$			$\hat{Y}_{\tilde{w}t}$		
$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}\right)$	-0	-1	-2	-1	-1	-2	-1	-1	-3
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{p}\right)$	81	57	35	69	49	32	80	56	35
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}1}\right)$	19	19	13	31	22	15	20	18	13
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}2}\right)$	-	25	18	-	28	19	-	25	17
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}3}\right)$	-	-	34	-	-	34	-	-	34
$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}_{t,\,\mathrm{simp}}^{\mathrm{nr}}\right)$	559	188	80	0	-1	-2	83	34	15
	$\hat{R}_t$			$\hat{R}_{wt}$			$\hat{R}_{\tilde{w}t}$		
$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}\right)$	-	-0	-2	-	-1	-2	-	-1	-2
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{p}\right)$	-	49	32	-	49	32	-	50	33
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}1}\right)$	-	22	15	-	22	15	-	22	15
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}2}\right)$	-	28	19	-	28	19	-	28	19
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}3}\right)$	-	-	34	-	-	34	-	-	34
$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}_{t,\,\mathrm{simp}}^{\mathrm{nr}}\right)$	-	0	0	-	-1	-2	-	-1	-1
	$\hat{\Delta}_{tt}\ (1 \to t)$			$\hat{\Delta}_{tt,\,w}\ (1 \to t)$			$\hat{\Delta}_{tt,\,\tilde{w}}\ (1 \to t)$		
$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}\right)$	-	-0	-2	-	-0	-2	-	-1	-3
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{p}\right)$	-	50	33	-	49	32	-	50	33
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}1}\right)$	-	22	14	-	22	15	-	22	14
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}2}\right)$	-	28	18	-	28	19	-	28	18
$\mathrm{CONTR}_{\mathrm{mc}}\left(\hat{V}_t^{\mathrm{nr}3}\right)$	-	-	34	-	-	34	-	-	34
$\mathrm{BR}_{\mathrm{mc}}\left(\hat{V}_{t,\,\mathrm{simp}}^{\mathrm{nr}}\right)$	-	19	30	-	-1	-2	-	3	5

6 Illustration

Dans cette section, nous voulons illustrer nos résultats sur un ensemble de données réelles provenant de l'enquête ELFE. La population d'inférence comprend les enfants nés dans l'une des 544 maternités en France durant 2011, à l'exception des très grands prématurés. Notre illustration vise à reproduire aussi rigoureusement que possible la méthodologie de l'enquête ELFE. En particulier, la modélisation de l'attrition à chaque période est effectuée uniquement avec les variables disponibles comme variables explicatives à la période de référence. Comme l'a souligné l'éditeur associé, sous l'hypothèse de données manquant au hasard (hypothèse MAR), les variables d'intérêt mesurées à toute période $\delta < t$ peuvent aussi avoir été utilisées pour modéliser l'attrition entre les périodes $t-1$ et t.

Un échantillon original s_0 d'environ 35 600 enfants a été sélectionné au départ quand les bébés n'avaient que quelques jours et étaient encore à la maternité. L'échantillon a été tiré selon un plan d'échantillonnage produit (Skinner, 2015; Juillard, Chauvet et Ruiz-Gazen, 2016). Un échantillon de jours et un échantillon de maternités ont été sélectionnés indépendamment, la sélection pouvant être approximée dans les deux cas par un échantillonnage aléatoire simple stratifié (EASS). L'échantillon était constitué de tous les enfants nés durant l'un des 25 jours sélectionnés dans l'une des 320 maternités sélectionnées.

Parmi les 35 600 enfants sélectionnés au départ, un total de 18 329 entretiens en face à face ont été réalisés auprès de leurs familles, ce qui représente un taux de réponse de 51 %. Cela a abouti au sous-échantillon s_1 après avoir tenu compte de la non-réponse. Les poids à la période $t = 1$ ont été calculés en se basant sur les poids d'échantillonnage originaux, ajustés en deux étapes. Premièrement, les probabilités de réponse ont été estimées au moyen d'un modèle de groupes de réponse homogènes (GRH), avec 20 GRH définis en se servant d'un modèle de régression logistique avec les variables explicatives *Âge de la mère*, *Identité gémellaire* et *Saison de naissance*. Puis, un calage par la méthode du raking ratio a été effectué sur les variables binaires *Né dans le mariage*, *Mère immigrante* et *Identité gémellaire*.

Quand les enfants ont atteint l'âge de deux mois, le premier entretien téléphonique avec les parents a eu lieu avec un taux de réponse de 87 %. Cela a mené au sous-échantillon s_2. Les poids à la période $t = 2$ ont été calculés en se basant sur le poids obtenu à la période $t = 1$, avec un ajustement en deux étapes. Premièrement, les probabilités de réponse ont été estimées au moyen de 20 GRH, définis en utilisant une régression logistique avec les variables explicatives *Âge de la mère*, *Nationalité de la mère* et *Père présent à l'accouchement*. Ensuite, un calage par la méthode du raking ratio a été effectué sur les mêmes variables de calage qu'à la période $t = 1$.

Quand les enfants ont eu un an, les parents ont été contactés par téléphone avec un taux de réponse de 77 %. Cela a mené au sous-échantillon s_3. Les poids à la période $t = 3$ ont été calculés en se basant sur les poids obtenus à la période $t = 2$, avec un ajustement en deux étapes similaire à celui réalisé à la période $t = 2$.

Nous avons considéré trois variables d'intérêt : *Allaitement maternel exclusif à la naissance, à deux mois, à un an*. Pour chacune de ces variables, nous avons calculé l'estimateur $\hat{R}_t$ et l'estimateur calé $\hat{R}_{wt}$

pour le pourcentage $R(t)$ d'allaitement maternel parmi tous les enfants à la période t, et les estimateurs de variance associés. Nous avons également calculé le coefficient de variation estimé (en pourcentage), défini par

$$\widehat{\mathrm{CV}}_t\left(\hat{Y}_t\right) = 100 \times \frac{\sqrt{\hat{V}_t\left(\hat{Y}_t\right)}}{\hat{Y}_t}. \tag{6.1}$$

Pour chaque composante $\hat{V}_{ta}$ de la variance estimée $\hat{V}_t$, nous avons calculé la contribution (en pourcentage) définie par

$$\mathrm{CONTR}\left(\hat{V}_{ta}\right) = 100 \times \frac{\hat{V}_{ta} - \hat{V}_t}{\hat{V}_t}. \tag{6.2}$$

Nous avons également calculé l'estimateur de variance simplifié pour la non-réponse $\hat{V}^{\mathrm{nr}}_{t,\,\mathrm{simp}}$, et la différence relative (en pourcentage) par rapport à l'estimateur de variance approximativement sans biais $\hat{V}^{\mathrm{nr}}$ définie par

$$\mathrm{DR}\left(\hat{V}^{\mathrm{nr}}_{t,\,\mathrm{simp}}\right) = 100 \times \frac{\hat{V}^{\mathrm{nr}}_{t,\,\mathrm{simp}} - \hat{V}^{\mathrm{nr}}_t}{\hat{V}^{\mathrm{nr}}_t}. \tag{6.3}$$

Les résultats sont donnés au tableau 6.1. Comme nous l'avons observé dans l'étude en simulation, la DR de l'estimateur de variance simplifié pour la non-réponse est négligeable dans tous les cas.

Tableau 6.1
Estimations d'un ratio, estimations de la variance, coefficients de variation, contributions relatives des composantes de la variance et différence relative d'un estimateur de variance simplifié pour une variable de l'enquête ELFE

Allaitement maternel exclusif	$t = 1$ Maternité	$t = 2$ 2 mois	$t = 3$ 1 an	$t = 1$ Maternité	$t = 2$ 2 mois	$t = 3$ 1 an
	sans calage			avec calage		
$\hat{R}_t$ (%)	59,0	30,6	3,3	59,4	31,0	3,4
$\hat{V}\left(\hat{R}_t\right)$	1,34E-05	1,50E-05	2,58E-06	1,28E-05	1,48E-05	2,60E-06
$\hat{\mathrm{CV}}\left(\hat{Y}_t\right)$ (%)	0,6	1,3	4,8	0,6	1,2	4,7
$\mathrm{CONTR}\left(\hat{V}^{p}_t\right)$	31	34	24	28	34	25
$\mathrm{CONTR}\left(\hat{V}^{\mathrm{nr}1}_t\right)$	69	51	42	72	51	41
$\mathrm{CONTR}\left(\hat{V}^{\mathrm{nr}2}_t\right)$	-	15	13	-	15	13
$\mathrm{CONTR}\left(\hat{V}^{\mathrm{nr}3}_t\right)$	-	-	21	-	-	21
$\mathrm{DR}\left(\hat{V}^{\mathrm{nr}}_{t,\,\mathrm{simp}}\right)$	2	2	0	1	2	0

7 Conclusion

Dans cet article, nous avons examiné l'estimation de la variance en tenant compte des repondérations dans les enquêtes par panel. Nous avons proposé un estimateur de variance approximativement sans biais, ainsi qu'un estimateur de variance simplifié pour les estimateurs de totaux, de paramètres complexes et de mesures de variation, qui couvrent la plupart des cas qui peuvent se présenter dans la pratique. Nos résultats de simulation indiquent que l'estimateur de variance proposé donne de bons résultats dans tous les cas pris en considération. L'estimateur de variance simplifié a tendance à surestimer la variance de l'estimateur par dilatation des totaux, et à surestimer la variance des estimateurs calés des totaux quand les variables de calage manquent de pouvoir explicatif pour la variable d'intérêt. Cependant, la performance de l'estimateur de variance simplifié est bonne pour l'estimation de ratios et de variations de totaux en se servant de poids calés, même si le modèle de calage n'est pas approprié à la variable étudiée.

L'hypothèse d'un comportement de réponse indépendant n'est habituellement pas défendable dans le cas des enquêtes à plusieurs degrés, puisque les unités à l'intérieur des grappes ont tendance à être corrélées en ce qui concerne le comportement de réponse. L'estimation des probabilités de réponse fondée sur la régression logistique conditionnelle dans le contexte de réponses corrélées a été étudiée par Skinner et D'Arrigo (2011), voir aussi Kim, Kwon et Park (2016). L'extension des présents travaux dans le contexte du comportement de réponse corrélé serait un problème intéressant à étudier dans le cadre de futurs travaux de recherche.

Remerciements

Nous remercions les éditeurs, un éditeur associé et les examinateurs de leurs commentaires et suggestions utiles qui nous ont permis d'améliorer l'article.

Annexe

Estimation de la variance due à la non-réponse pour les groupes de réponse homogènes

Nous considérons le modèle des groupes de réponse homogènes présenté à la section 2.5. Rappelons que ce modèle peut se résumer comme il suit : à chaque période $\delta = 1, \ldots, t,$ le sous-échantillon $s_{\delta-1}$ est partitionné en $C(\delta-1)$ groupes $s_{\delta-1}^{c}$, $c = 1, \ldots, C(\delta-1)$. Les probabilités de réponse sont supposées constantes à l'intérieur des groupes.

Ce modèle est équivalent au modèle de régression logistique en (2.18), avec

$$z_i^{\delta} = \left[1\left\{i \in s_{\delta-1}^{1}\right\}, \ldots, 1\left\{i \in s_{\delta-1}^{C(\delta-1)}\right\}\right]^{\top}. \tag{A.1}$$

L'équation (2.2) mène aux probabilités de réponse estimées

$$\hat{p}_i^{\delta} = \frac{\sum_{i \in s_{\delta-1}^c} k_i^{\delta} r_i^{\delta}}{\sum_{i \in s_{\delta-1}^c} k_i^{\delta}} \qquad \text{pour} \qquad i \in s_{\delta-1}^c. \tag{A.2}$$

Nous considérons d'abord le cas où l'estimateur repondéré est calculé à la période $t = 1$. Dans l'estimateur de la variance due à la non-réponse donné en (2.21), le vecteur $\hat{\gamma}_1^1$ se simplifie en

$$\hat{\gamma}_1^1 = \left(\frac{\sum_{i \in s_1 \cap s_0^1} \frac{y_{i1}}{\pi_i}}{\hat{p}_1^1 \sum_{i \in s_1 \cap s_0^1} k_i^1}, \ \ldots, \ \frac{\sum_{i \in s_1 \cap s_0^{C(0)}} \frac{y_{i1}}{\pi_i}}{\hat{p}_{C(0)}^1 \sum_{i \in s_1 \cap s_0^{C(0)}} k_i^1} \right)^{\top}. \tag{A.3}$$

Après un peu de calcul, l'estimateur de variance en (2.21) peut être réécrit sous la forme

$$\hat{V}_1^{\text{nr}}\left(\hat{Y}_1\right) = \sum_{c=1}^{C(0)} \frac{\left(1 - \hat{p}_c^1\right)}{\left(\hat{p}_c^1\right)^2} \sum_{i \in s_1 \cap s_0^c} \left(\frac{y_{i1}}{\pi_i} - k_i^1 \frac{\sum_{j \in s_1 \cap s_0^c} \frac{y_{j1}}{\pi_j}}{\sum_{j \in s_1 \cap s_0^c} k_j^1} \right)^2. \tag{A.4}$$

Nous considérons maintenant le cas où l'estimateur repondéré est calculé à la période $t = 2$. Nous nous concentrons sur le cas plus simple où le même système de GRH est conservé au fil du temps. Dans l'estimateur de la variance due à la non-réponse donné en (2.22), les vecteurs $\hat{\gamma}_2^1$ et $\hat{\gamma}_2^2$ se simplifient en

$$\hat{\gamma}_2^1 = \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_{i2}}{\pi_i}}{\hat{p}_1^1 \sum_{i \in s_2 \cap s_1^1} k_i^1}, \ \ldots, \ \frac{\sum_{i \in s_2 \cap s_1^{C(0)}} \frac{y_{i2}}{\pi_i}}{\hat{p}_{C(0)}^1 \sum_{i \in s_2 \cap s_1^{C(0)}} k_i^1} \right)^{\top}, \tag{A.5}$$

$$\hat{\gamma}_2^2 = \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_{i2}}{\pi_i}}{\hat{p}_1^1 \hat{p}_1^2 \sum_{i \in s_2 \cap s_1^1} k_i^2}, \ \ldots, \ \frac{\sum_{i \in s_2 \cap s_1^{C(0)}} \frac{y_{i2}}{\pi_i}}{\hat{p}_{C(0)}^1 \hat{p}_{C(0)}^2 \sum_{i \in s_2 \cap s_1^{C(0)}} k_i^2} \right)^{\top}. \tag{A.6}$$

Après un peu de calcul, l'estimateur de variance en (2.22) peut être réécrit sous la forme

$$\begin{aligned} \hat{V}_2^{\text{nr}}\left(\hat{Y}_2\right) = & \sum_{c=1}^{C(0)} \frac{\left(1 - \hat{p}_c^1\right)}{\hat{p}_c^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_{i2}}{\pi_i \hat{p}_c^1} - k_i^1 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_{j2}}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^1} \right)^2 \\ & + \sum_{c=1}^{C(0)} \left(1 - \hat{p}_c^2\right) \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_{i2}}{\pi_i \hat{p}_c^1 \hat{p}_c^2} - k_i^2 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_{j2}}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^2} \right)^2. \end{aligned} \tag{A.7}$$

Si nous supposons en outre que k_i^{δ} est constant sur les périodes $\delta = 1, 2,$ et peut donc être réécrit comme k_i, l'expression en (A.7) se simplifie en

$$\hat{V}_2^{\text{nr}}\left(\hat{Y}_2\right) = \sum_{c=1}^{C(0)} \frac{\left(1 - \hat{p}_c^{1 \to 2}\right)}{\left(\hat{p}_c^{1 \to 2}\right)^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_{i2}}{\pi_i} - k_i \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_{j2}}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j} \right)^2. \tag{A.8}$$

avec $\hat{p}_c^{1\to 2} = \prod_{\delta=1}^{2} \hat{p}_c^{\delta}$ pour $c = 1, \ldots, C(0)$. Cette simplification de l'estimateur de variance peut être étendue à l'estimateur repondéré à la période t. En supposant que les GRH sont conservés au fil du temps, et que $k_i^{\delta} = k_i$ pour tout $\delta = 1, \ldots, t$, l'estimateur de variance en (2.12) peut se réécrire sous la forme

$$\hat{V}_t^{\mathrm{nr}}\left(\hat{Y}_t\right) = \sum_{c=1}^{C(0)} \frac{\left(1-\hat{p}_c^{1\to t}\right)}{\left(\hat{p}_c^{1\to t}\right)^2} \sum_{i\in s_t\cap s_{t-1}^c} \left(\frac{y_{it}}{\pi_i} - k_i \frac{\sum_{j\in s_t\cap s_{t-1}^c} \frac{y_{jt}}{\pi_j}}{\sum_{j\in s_t\cap s_{t-1}^c} k_j} \right)^2 \tag{A.9}$$

avec $\hat{p}_c^{1\to t} = \prod_{\delta=1}^{t} \hat{p}_c^{\delta}$ pour $c = 1, \ldots, C(0)$.

Bibliographie

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasimodel-assisted approach. *Journal of the Royal Statistical Society, Series B*, 67, 445-458.

Beaumont, J.-F., et Haziza, D. (2016). Une note sur le concept d'invariance dans les plans d'échantillonnage à deux phases. *Techniques d'enquête*, 42, 2, 337-342. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2016002/article/14662-fra.pdf.

Berger, Y. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 4, 451-467.

Caron, N., et Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête emploi en continu. Rapport technique de l'INSEE, Paris.

Chauvet, G., et Goga, C. (2018). Linéarisation contre Bootstrap pour estimer la variance de l'évolution de l'indice de Gini. *Techniques d'enquête*, 44, 1, 19-44. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2018001/article/54926-fra.pdf.

Clarke, P., et Tate, P. (2002). An application of non-ignorable non-response models for gross flows estimation in the British Labour Force Survey. *Australian & New Zealand Journal of Statistics*, 4, 413-425.

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

Ekholm, A., et Laaksonen, S. (1991). Weighting via response modeling in the finnish household budget survey. *Journal of Official Statistics*, 7, 325-327.

Fay, R. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 81, 1, 227-232.

Fuller, W., et An, A. (1998). Regression adjustment for non-response. *Journal of the Indian Society of Agricultural Statistics*, 51, 331-342.

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la *Nationwide Food Consumption Survey* de 1987-1988. *Techniques d'enquête*, 20, 1, 79-89. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1994001/article/14429-fra.pdf.

Goga, C., Deville, J.-C. et Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. *Biometrika*, 96, 691-709.

Hawkes, D., et Plewis, I. (2009). Modelling nonresponse in the national child development study. *Journal of the Royal Statistical Society, Series A*, 169, 479-491.

Juillard, H., Chauvet, G. et Ruiz-Gazen, A. (2017). Estimation under cross-classified sampling with application to a childhood survey. *Journal of the American Statistical Association*, 112, 850-858.

Kalton, G. (2009). Design for surveys over time. *Handbook of Statistics*, 29, 89-108.

Kim, J.K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35, 501-514.

Kim, J.K., Kwon, Y. et Park, M. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling. *Biometrika*, 103, 461-473.

Laaksonen, S. (2007). Pondération de données d'enquête recueillies en deux phases. *Techniques d'enquête*, 33, 2, 137-147. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2007002/article/10489-fra.pdf.

Laaksonen, S., et Chambers, R.L. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22, 81-95.

Laniel, N. (1988). Variances for a rotating sample from a changing population. *Proceedings of the Business and Economics Statistics Section*, American Statistical Association, 246-250.

Laurie, H., Smith, R. et Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15, 269-282.

Lynn, P. (2009). Methods for longitudinal surveys. *Methodology of Longitudinal Surveys*, 1-19.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Pirus, C., Bois, C., Dufourg, M., Lanoë, J., Vandentorren, S., Leridon, H. et the Elfe team (2010). Constructing a cohort: Experience with the French Elfe project. *Population*, 65, 637-670.

Qualité, L., et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée. *Techniques d'enquête*, 34, 2, 193-201. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2008002/article/10758-fra.pdf.

Rendtel, U., et Harms, T. (2009). Weighting and calibration for household panels. *Methodology of Longitudinal Surveys*, 265-286.

Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 1, 43-53. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1996001/article/14386-fra.pdf.

Silva, P., et Skinner, C. (1997). Cross-classiffed sampling: Some estimation theory. *Variable Selection for Regression Estimation in Finite Populations*, 23, 23-32.

Skinner, C. (2015). Cross-classiffed sampling: Some estimation theory. *Statistics & Probability Letters*, 104, 163-168.

Skinner, C., et D'Arrigo, J. (2011). Inverse probability weighting for clustered non-response. *Biometrika*, 98, 953-966.

Skinner, C., et Vieira, M. (2005). Design effects in the analysis of longitudinal survey data. S3RI Methdology Working Papers, M05/13. Southampton, UK: Southampton Statistical Sciences Research Institute.

Slud, E.V., et Bailey, L. (2010). Evaluation and selection of models for attrition nonresponse adjustment. *Journal of Official Statistics*, 26, 1-18.

Tam, S. (1984). On covariance from overlapping samples. *The American Statistician*, 38, 1-18.

Vandecasteele, L., et Debels, A. (2007). Attrition in panel data: The effectiveness of weighting. *European Sociological Review*, 23, 1, 81-97.

Zhou, M., et Kim, J. (2012). An efficient method of estimation for longitudinal surveys with monotone missing data. *Biometrika*, 99, 631-648.