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

Techniques d'enquête

Est-ce que la réduction du déséquilibre de la réponse accroît l'exactitude des estimations de l'enquête ?

par Carl-Erik Särndal, Kaur Lumiste et Imbi Traat

Date de diffusion : le 20 décembre 2016

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 sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

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 2016
Vol. 42, nº 2, pp. 233-253
Statistique Canada, nº 12-001-X au catalogue

Est-ce que la réduction du déséquilibre de la réponse accroît l'exactitude des estimations de l'enquête ?

Carl-Erik Särndal, Kaur Lumiste et Imbi Traat[1]

Résumé

Nous présentons des preuves théoriques que les efforts déployés durant la collecte des données en vue d'équilibrer la réponse à l'enquête en ce qui concerne certaines variables auxiliaires augmentera les chances que le biais de non-réponse soit faible dans les estimations qui sont, en fin de compte, produites par pondération calée. Nous montrons que la variance du biais – mesurée ici comme étant l'écart de l'estimateur calé par rapport à l'estimateur sans biais sur échantillon complet (non réalisé) – diminue linéairement en fonction du déséquilibre de la réponse que nous supposons être mesuré et contrôlé continuellement tout au long de la période de collecte des données. Cela offre donc la perspective intéressante d'un plus faible risque de biais si l'on peut gérer la collecte des données de manière à réduire le déséquilibre. Les résultats théoriques sont validés au moyen d'une étude en simulation s'appuyant sur des données réelles provenant d'une enquête-ménages estonienne.

Mots-clés : Non-réponse à l'enquête; biais; plan de collecte adaptatif; estimateur par calage; variables auxiliaires.

1 Introduction

Le problème de l'estimation exacte en dépit d'une non-réponse importante doit être examiné sous deux angles temporellement dépendants, à savoir d'abord les moyens de gérer la collecte des données, puis les moyens de traiter l'estimation à partir des données finalement recueillies. La première activité peut nécessiter d'importantes ressources. Dans une enquête téléphonique, la planification quotidienne des tentatives de prise de contact, l'interaction avec les intervieweurs et les éléments à prendre en considération pour établir leur charge de travail peuvent demander de coûteux efforts. L'étape de l'estimation est administrativement plus simple; elle comporte la recherche des meilleures variables auxiliaires en vue de procéder à une pondération calée pour corriger la non-réponse, après quoi les estimations sont habituellement calculées en se servant de logiciels existants.

La collecte des données est le point de concentration de la littérature sur les plans de collecte dynamique (*responsive designs*); Groves (2006), Groves et Heeringa (2006) ont été parmi les premiers à proposer ces plans. Les plans de collecte adaptatifs (*adaptive survey designs*) sont discutés dans Wagner (2008). L'une des idées qui sous-tend cette approche de recherche est qu'une collecte des données qui s'étend sur une certaine période pourrait être inspectée à des points de décision appropriés, où des mesures peuvent être prises pour obtenir en fin de compte un ensemble bien équilibré de répondants. Schouten, Calinescu et Luiten (2013) expliquent comment les plans de collecte adaptatifs peuvent être taillés sur mesure en vue d'optimiser les taux de réponse et de réduire la sélectivité de la non-réponse, en tenant compte des aspects liés au coût. De nombreux travaux exploratoires ont porté sur les plans de collecte dynamiques (ou adaptatifs). La recherche d'une réponse bien équilibrée ou représentative peut être un objectif en soi. Différentes pistes ont été explorées, à savoir le classement des cas par ordre de priorité, (Peytchev, Riley,

1. Carl-Erik Särndal, Ph.D., professeur émérite, Statistics Sweden. Courriel : carl.sarndal@telia.com; Kaur Lumiste, M.Sc., Institute of Mathematical Statistics, Université de Tartu, Estonie; Imbi Traat, Ph.D., professeur associé, Institute of Mathematical Statistics, Université de Tartu, Estonie.

Rosen, Murphy et Lindblad 2010); les règles d'arrêt pour mettre fin aux tentatives de collecte des données pour des unités particulières de l'échantillon, (Rao, Glickman et Glynn 2008; Wagner et Raghunathan 2010); l'utilisation de paradonnées de manière plus générale pour gérer la réponse à l'enquête, (Couper et Wagner 2011).

La mesure et le contrôle du déséquilibre font partie de la phase de collecte des données. La statistique de déséquilibre (voir la section 3) joue un rôle central dans le présent article; elle a été utilisée par exemple dans Särndal (2011), Lundquist et Särndal (2013), Särndal et Lundquist (2014a, 2014b). Elle est reliée à l'indicateur R (R pour représentativité); voir Schouten, Cobben et Bethlehem (2009), et Bethlehem, Cobben et Schouten (2011).

La seconde étape s'appuie sur la théorie de l'estimation pour résoudre la difficulté que pose la non-réponse, principalement la façon d'obtenir un faible biais dans les estimations. Considérée strictement comme un problème d'estimation, il s'agit d'une activité en soi, après l'achèvement de la collecte des données. L'ensemble des unités répondantes est fixé; la quantité des données sur ces unités est « gelée ». Le choix des variables auxiliaires joue un rôle crucial. Les « meilleures » doivent être sélectionnées. Cet aspect a été traité en profondeur, notamment dans Särndal et Lundström (2005). Deux facteurs sont habituellement mentionnés comme étant importants pour l'exactitude des estimations, à savoir la mesure dans laquelle les variables auxiliaires choisies peuvent expliquer la variable étudiée et la mesure dans laquelle ces variables peuvent expliquer l'indicateur de réponse 0/1 montrant la présence ou l'absence d'une unité dans le jeu de répondants. Ces deux degrés d'explication sont l'un et l'autre partiels au mieux, imparfaits. Les deux rôles des variables auxiliaires interagissent, comme le soulignent par exemple Little et Vartivarian (2005). Une revue détaillée des procédures d'ajustement de la pondération pour corriger la non-réponse est donnée dans Brick (2013).

La disponibilité des variables auxiliaires dépend de l'environnement de l'enquête. En Scandinavie, les enquêtes auprès des particuliers et des ménages peuvent s'appuyer sur les vastes sources de variables auxiliaires que sont les registres administratifs. Et il en est de plus en plus souvent ainsi dans d'autres pays.

D'aucuns pensent que l'estimation est vraiment l'étape importante : toute mesure qui peut être prise à l'étape de la collecte des données, comme l'équilibrage ou l'amélioration de la représentativité, est peut-être superflue; obtenir les estimations les plus exactes possibles est un problème qui peut être traité efficacement à l'étape de l'estimation, grâce à l'usage judicieux des variables auxiliaires dans un processus de pondération pour corriger la non-réponse ou par d'autres moyens. Ce point de vue est défendu, par exemple, dans Beaumont, Bocci et Haziza (2014).

Néanmoins, il est clair que les aspects mesurables de la collecte des données influenceront l'exactitude des estimations qui sont produites en dernière analyse. L'une de ces mesures est la statistique de déséquilibre définie à la section 3. Dans le présent article, les deux activités temporellement dépendantes sont prises en compte. L'équilibrage de la réponse doit être combiné à des méthodes d'estimation efficaces afin d'obtenir en fin de compte les estimations les meilleures (les plus exactes) possibles. Cette façon de penser sous-tend, par exemple, les travaux de Schouten, Cobben, Lundquist et Wagner (2014).

Les considérations qui motivent le présent article sont les suivantes : il existe des méthodes s'appliquant à diverses lignes de conduite – règles d'arrêt, classement des cas par ordre de priorité, et d'autres – durant

la collecte des données, en vue d'obtenir en fin de compte un ensemble de répondants r favorable. Särndal et Lundquist (2014a, 2014b) ont utilisé la statistique de déséquilibre IMB (de l'anglais *imbalance*) donnée à la section 3 comme outil en vue d'arriver à un faible déséquilibre dans l'ensemble de répondants final. Vu que des variables auxiliaires seront également utilisées dans l'estimation, dans quelle mesure une meilleure exactitude des estimations découlera-t-elle d'un faible déséquilibre durant la collecte des données qui précède ? Des signes encourageants, par exemple dans Särndal et Lundquist (2014a), indiquent qu'un déséquilibre plus faible donne lieu à une certaine amélioration de l'exactitude, quoique modeste. Ces travaux étant empiriques, dans le présent article nous présentons un soutien mathématique/analytique menant à une conclusion similaire.

La présentation de l'article est la suivante : le contexte de l'enquête est exposé à la section 2 et la statistique de déséquilibre est présentée à la section 3. L'importance de la relation de régression – celle de la variable étudiée sur le vecteur de variables auxiliaires – est décrite à la section 4, et plus particulièrement pour l'estimateur (appelé CAL) obtenu par repondération calée pour corriger la non-réponse, à la section 5. L'écart de l'estimateur CAL par rapport à l'estimateur (sans biais) nécessitant une réponse complète et est analysé à la section 6, à la section 7 et à la section 8, en montrant comment l'écart dépend du déséquilibre. Deux résultats sont présentés concernant les propriétés statistiques (moyenne et variance) de l'écart de CAL. En particulier, il est montré que la variance de cet écart est, approximativement, une fonction linéaire de la statistique de déséquilibre. D'où, l'écart est vraisemblablement plus petit, et les estimations plus exactes, si le déséquilibre peut être réduit durant la collecte des données. Les résultats théoriques sont validés empiriquement à la section 9 en utilisant des données provenant d'une enquête-ménages estonienne. Le logiciel statistique R est utilisé; R Core Team (2014). Une discussion conclut l'article à la section 10. Trois annexes fournissent les preuves et les dérivations nécessaires.

2 Contexte et notation

Un échantillon probabiliste s est tiré de la population finie $U = \{1, 2, \ldots, k, \ldots, N\}$. L'unité k possède la probabilité d'inclusion connue $\pi_k = \Pr(k \in s)$ et le poids de sondage connu $d_k = 1/\pi_k$. Une non-réponse a lieu. L'ensemble de répondants, désigné r, est un sous-ensemble de s pour lequel la variable étudiée est observée. Nous ne savons pas comment r a été généré à partir de s; les probabilités de réponse sont inconnues (si l'on suppose qu'elles « existent », elles ne sont pas nécessaires dans le présent article). Le taux de réponse (pondéré par les poids de sondage) est

$$P = \sum_r d_k \Big/ \sum_s d_k \,. \tag{2.1}$$

Si A est un ensemble d'unités, $A \subseteq U$, une somme $\sum_{k \in A}$ s'écrira $\sum_A$. L'enquête peut compter de nombreuses variables étudiées. Une variable type, désignée y (continue ou catégorique), possède une valeur y_k qui est enregistrée pour $k \in r$ mais manquante pour $k \in s - r$. Notre objectif est d'estimer le total de population de y, $Y = \sum_U y_k$. L'indicateur de réponse I prend la valeur $I_k = 1$ pour $k \in r$, $I_k = 0$ pour $k \in s - r$. Un objectif dans la pratique est d'obtenir une réponse r qui est bien équilibrée, au sens qui

sera spécifié plus tard. Nous sommes amenés à considérer les différents ensembles r qui peuvent provenir d'un échantillon donné s.

Le vecteur auxiliaire $\mathbf{x}$ de dimension $J \geq 1$ prend la valeur $\mathbf{x}_k$ connue au moins pour toutes les unités $k \in s$. L'information auxiliaire peut être utilisée dans la collecte des données (pour surveiller le flux de données entrantes afin d'arriver à un meilleur équilibre) et/ou dans l'estimation (pour le calcul de poids calés). Le vecteur auxiliaire ne doit pas nécessairement être le même dans les deux cas, mais nous supposons ici que les deux vecteurs concordent et que l'information $\mathbf{x}$ utilisée concerne les unités $k \in s$. Cela englobe le cas important des paradonnées, c'est-à-dire des variables portant sur le processus de collecte des données.

Un type important de vecteur auxiliaire est le *vecteur de groupes*. Il précise l'appartenance de chaque unité k à l'un de J groupes de l'échantillon mutuellement exclusifs et exhaustifs, de sorte que $\mathbf{x}_k = (0,\ldots,1,\ldots 0)'$, où la seule valeur « 1 » indique l'unique groupe (sur les J possibles) auquel k appartient.

Un vecteur de groupes existe lorsque plusieurs variables auxiliaires catégoriques sont entièrement croisées. Par exemple, si $\mathbf{x} = (sexe \times études \times âge)$ représente le croisement de 2 sexes, 3 catégories exhaustives d'études et 4 catégories exhaustives d'âge, alors $\mathbf{x}$ est un vecteur de groupes de dimension $J = 2 \times 3 \times 4 = 24$ et pouvant prendre autant de valeurs possibles $\mathbf{x}_k$. Quand plusieurs variables catégoriques sont utilisées mais qu'elles ne sont pas entièrement croisées – un important aspect pratique dans les organismes statistiques –, la dimension J de $\mathbf{x}_k$ peut être maintenue relativement modeste (disons inférieure à 15) tout en codant un nombre beaucoup plus élevé (disons plus d'une centaine) de propriétés possibles $\mathbf{x}_k$ des unités k. Pour une étude de la Swedish Living Conditions Survey, Särndal et Lundquist (2014a) ont utilisé un vecteur $\mathbf{x}$ de dimension 14 avec 256 valeurs possibles. Le cas du vecteur de groupes et le cas du vecteur sans groupes révèlent d'importantes différences dans les résultats qui suivent.

Tous les vecteurs auxiliaires utilisés ici satisfont une contrainte appliquée pour des raisons de commodité mathématique sans restreindre gravement le choix du vecteur, à savoir qu'il existe un vecteur constant $\boldsymbol{\mu}$ tel que

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ pour tout } k. \tag{2.2}$$

Par exemple, quand $J = 2$ et $\mathbf{x}_k = (1, x_k)'$, alors $\boldsymbol{\mu} = (1,0)'$ satisfait la contrainte. Dans le cas du vecteur de groupes, où $\mathbf{x}_k = (0,\ldots,1,\ldots,0)'$, alors $\boldsymbol{\mu} = (1,\ldots,1,\ldots,1)'$ satisfait la contrainte. Si $\mathbf{x}$ n'est pas un vecteur de groupes, disons un vecteur utilisé pour coder « études » en se servant de trois catégories mutuellement exclusives et exhaustives, et « sexe » au moyen d'une variable univariée égale à 1 ou 0, alors $J = 3 + 1 = 4$ (pas de croisement des variables études et sexe), et $\boldsymbol{\mu} = (1,1,1,0)'$ satisfait la contrainte.

3 Déséquilibre

Le concept d'*équilibre* a souvent été évoqué en littérature statistique en faisant référence à l'égalité des moyennes de variables particulières pour deux ensembles d'unités, dont l'un est un sous-ensemble de l'autre. La méthode du cube de Deville et Tillé (2004) est une méthode permettant d'obtenir un échantillon

probabiliste s tiré de U qui est équilibré par rapport à un vecteur $\mathbf{x}$. Dans le contexte de la non-réponse, nous voulons savoir dans quelle mesure une réponse r est bien équilibrée, comparativement à l'échantillon probabiliste s qui aurait donné des estimations sans biais. Un vecteur auxiliaire donné $\mathbf{x}$ possède des moyennes calculables $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k \Big/ \sum_r d_k$ pour la réponse et $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k \Big/ \sum_s d_k$ pour l'échantillon. Si ces moyennes sont égales, ce qui est peu probable, la réponse est parfaitement équilibrée par rapport à $\mathbf{x}$. Le contraste entre la réponse r et l'échantillon s peut être mesuré par des quantités scalaires

$$Q_s = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right); \quad Q_r = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_r^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right). \tag{3.1}$$

L'échantillon et l'ensemble de répondants ne se distinguent que par la matrice de pondération de dimensions $J \times J$, $\boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_s d_k$ opposée à $\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_r d_k$, toutes deux supposées non singulières. En particulier, la quantité Q_s est importante pour la statistique de *déséquilibre* appelée *IMB* (de l'anglais *imbalance*) de r par rapport au vecteur $\mathbf{x}$ spécifié :

$$IMB(r, \mathbf{x} \mid s) = P^2 \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right) = P^2 Q_s, \tag{3.2}$$

où P est le taux de réponse (2.1); voir, par exemple, Särndal et Lundquist (2014a). La notation complète $IMB(r, \mathbf{x} \mid s)$ souligne le fait que le déséquilibre dépend de la réponse obtenue r et du choix du vecteur $\mathbf{x}$. À moins qu'il ne soit nécessaire de mettre l'accent, nous utilisons la notation plus simple IMB. Nous avons $0 \le IMB \le P(1-P)$ pour n'importe quels ensemble de répondants r et formulation de vecteur $\mathbf{x}$, sachant s. La statistique IMB est une mesure descriptive de la réponse r. Elle est reliée à un cas particulier de l'indicateur R, dont la motivation repose plutôt sur l'estimation des probabilités de réponse (inconnues) pour les unités de population; voir par exemple Bethlehem et coll. (2011).

La statistique IMB (3.2) peut être calculée et surveillée en continu pendant une collecte de données s'étendant sur une certaine période, disons plusieurs jours ou semaines, durant laquelle les tentatives de prise de contact se poursuivent auprès d'une unité de l'échantillon jusqu'à ce que les données souhaitées soient obtenues, ou, en cas d'échec, jusqu'à ce que l'unité soit déclarée être un non-répondant. À mesure que le taux de réponse P augmente, IMB sert d'outil de surveillance et de gestion de la collecte des données en vue d'obtenir en fin de compte un ensemble de répondants r qui, s'il n'est pas parfaitement équilibré pour satisfaire $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, aura au moins une IMB considérablement plus faible que si aucun équilibrage n'avait été tenté durant la collecte des données. Des méthodes d'équilibrage fondées sur la propension à répondre, telle que la méthode du seuil et la méthode des proportions égales, sont décrites dans Särndal et Lundquist (2014a, 2014b).

Nous examinerons plus tard le cas particulier où s est un échantillon autopondéré (comme quand s est un échantillon aléatoire simple), l'ensemble de répondants r possède une taille fixe m, et $\mathbf{x}$ est un vecteur de groupes de dimension J comme il est défini à la section 2. Alors, s ainsi que r sont divisés en J groupes non chevauchants. Pour le groupe de l'échantillon s_j, si nous désignons par n_j la taille et par

$W_{js} = n_j / n$ la taille relative, $\sum_{j=1}^{J} n_j = n.$ Pour le groupe r_j de la réponse, posons que $m_j \le n_j$ est la taille; $\sum_{j=1}^{J} m_j = m.$ Le déséquilibre (3.2) est alors donné par

$$IMB = \sum_{j=1}^{J} W_{js} \left(p_j - p \right)^2, \tag{3.3}$$

où les taux de réponse sont $p_j = m_j / n_j$ dans le groupe j et $p = m/n$ dans l'ensemble. (Le taux de réponse P est défini en (2.1) avec les poids de sondage généraux d_k; pour un échantillon autopondéré, où d_k est constant, nous utilisons p minuscule pour désigner le taux de réponse.) Si $IMB = 0,$ nous avons un équilibre parfait; tous les taux de réponse de groupe p_j sont alors égaux.

4 L'aspect régression

Le déséquilibre IMB est déterminé par le vecteur auxiliaire $\mathbf{x}$ sans tenir compte de la variable étudiée y. Or, la relation entre $\mathbf{x}$ et y est également importante en ce qui concerne le biais des totaux y estimés. Une forte régression de y sur $\mathbf{x}$ donnera vraisemblablement un petit biais, intuitivement parce que les valeurs y prédites par régression peuvent alors donner des valeurs de remplacement approchées pour les valeurs manquantes. Dans le cas de certaines données d'enquête, la force de la régression peut être modeste, mais néanmoins avoir un effet important sur le biais. Les vecteurs des coefficients de régression linaire ordinaire pour l'échantillon complet s et pour la réponse r sont, respectivement,

$$\mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_s d_k \mathbf{x}_k y_k; \quad \mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_r d_k \mathbf{x}_k y_k. \tag{4.1}$$

En présence de non-réponse, $\mathbf{b}_r$ est calculable, mais non $\mathbf{b}_s$. Les matrices de dimensions $J \times J$ qu'il faut inverser sont supposées non singulières. Normalement, $\mathbf{b}_r \ne \mathbf{b}_s,$ la différence étant peut-être considérable (mais inconnue). La régression fondée sur la réponse est incohérente.

Le déséquilibre pour la variable y est $\bar{y}_r - \bar{y}_s,$ où les moyennes sont $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ pour l'échantillon (inconnu) et $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ pour la réponse (calculable). La décomposition

$$\bar{y}_r - \bar{y}_s = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s \right)' \mathbf{b}_r + \left(\mathbf{b}_r - \mathbf{b}_s \right)' \bar{\mathbf{x}}_s, \tag{4.2}$$

met en relief deux différences indésirables, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ (due au déséquilibre dans le vecteur $\mathbf{x}$), et $\mathbf{b}_r - \mathbf{b}_s$ (due à l'incohérence de la régression); pour obtenir (4.2), notons que $\bar{\mathbf{x}}_r' \mathbf{b}_r = \bar{y}_r$ et $\bar{\mathbf{x}}_s' \mathbf{b}_s = \bar{y}_s,$ qui sont des conséquences de la contrainte (2.2) appliquée au vecteur $\mathbf{x}$.

5 Estimation du total de population en présence de non-réponse

L'équation (4.2), quand elle est multipliée par $\hat{N} = \sum_s d_k,$ peut être exprimée en fonction de trois estimateurs courants du total de population $Y = \sum_U y_k.$ Deux d'entre eux sont possibles en présence de non-réponse,

$$\hat{Y}_{EXP} = \hat{N}\frac{\sum_r d_k y_k}{\sum_r d_k}, \quad \hat{Y}_{CAL} = \hat{N}\frac{\sum_r d_k g_k y_k}{\sum_r d_k}, \tag{5.1}$$

avec $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$. De ces estimateurs, $\hat{Y}_{EXP}$ est une simple extension de la moyenne de y dans la réponse et peut être fortement biaisé. L'estimateur par calage $\hat{Y}_{CAL}$ donne à y_k le poids $d_k g_k / P$. La propriété de calage est $\sum_r (d_k g_k / P)\mathbf{x}_k = \sum_s d_k \mathbf{x}_k$, où le deuxième membre de l'équation est sans biais pour le total $\mathbf{x}$ de population $\sum_U \mathbf{x}_k$, ce qui explique pourquoi $\hat{Y}_{CAL}$ peut être considérablement moins biaisé que $\hat{Y}_{EXP}$, quand $\mathbf{x}$ et y sont bien reliées. Si les valeurs y_k sont enregistrées pour l'échantillon complet s, une estimation sans biais sera obtenue au moyen de l'estimateur de Horvitz-Thompson

$$\hat{Y}_{FUL} = \sum_s d_k y_k.$$

Nous désignerons ces trois types d'estimateur par EXP, CAL et FUL (pour *full sample*). Maintenant, l'expression (4.2) multipliée par $\hat{N} = \sum_s d_k$ se lit

$$\hat{Y}_{EXP} - \hat{Y}_{FUL} = \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right). \tag{5.2}$$

Exprimée en mots, Écart de EXP = terme d'ajustement du biais + écart de *CAL*. L'ajustement calculable est $\hat{Y}_{EXP} - \hat{Y}_{CAL} = \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r$. Les deux écarts par rapport à l'estimateur sans biais, $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s$ pour CAL et $\hat{Y}_{EXP} - \hat{Y}_{FUL} = \hat{N}\left(\bar{y}_r - \bar{y}_s\right)$ pour EXP, ne sont pas calculables en présence de non-réponse, parce qu'ils requièrent les valeurs de y pour l'échantillon complet.

Comme nous l'avons mentionné, il existe des méthodes pour réduire le déséquilibre IMB durant la collecte des données. Un faible déséquilibre est intuitivement intéressant, mais aboutit-il à une plus grande exactitude des estimations ? Ou bien, cela suffit-il de faire intervenir les variables auxiliaires à l'étape de l'estimation, par un ajustement de la pondération par calage comme dans l'estimateur CAL ? Le terme d'ajustement $\hat{Y}_{EXP} - \hat{Y}_{CAL} = \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r$ peut clairement être réduit en construisant r de manière que le déséquilibre soit faible; il est égal à zéro pour l'équilibre parfait $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. En pratique, l'estimateur CAL est préféré à l'estimateur EXP, le premier étant habituellement plus précis, en raison de l'information auxiliaire. Mais l'écart $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s$ est-il plus petit si la réponse r a été construite de manière à ce que IMB soit faible ? Autrement dit, cela vaut-t-il l'effort (peut-être coûteux) de gérer la collecte des données de manière que $\bar{\mathbf{x}}_r$ soit plus proche de $\bar{\mathbf{x}}_s$ et, par conséquent, que IMB soit réduit ? La question est essentiellement celle de savoir si cela ferait aussi se rapprocher $\mathbf{b}_r$ et $\mathbf{b}_s$.

6 Propriétés statistiques de l'écart de l'estimateur CAL

Dans la décomposition (5.2), l'écart de CAL par rapport à l'estimateur sans biais FUL est $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$ où $\boldsymbol{\Delta}_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s$. Pour voir si $\boldsymbol{\Delta}_r$ est plus petit, ou susceptible de l'être, lorsqu'on réalise un faible déséquilibre durant la collecte des données, nous cherchons à obtenir des résultats analytiques au sujet

des propriétés statistiques, telles la moyenne et la variance, de $\mathbf{\Delta}_r$ en fonction de la statistique IMB (3.2). Des résultats très généraux de ce type sont difficiles à obtenir. Plusieurs facteurs compliquent l'analyse, dont le plan d'échantillonnage utilisé pour tirer s, la loi de probabilité des ensembles de réponses r sachant s, la composition du vecteur auxiliaire $\mathbf{x}$, et ainsi de suite. Les résultats pour des situations particulières sont obtenus aux sections 7 et 8.

Le résultat 1 à la section 7 donne les propriétés – espérance et variance – de $\mathbf{\Delta}_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \overline{\mathbf{x}}_s$ sur les résultats de réponse r avec une taille fixe m et une moyenne fixe $\overline{\mathbf{x}}_r$ quand $\mathbf{x}$ est un vecteur de groupes, et s est un échantillon aléatoire simple. La moyenne de $\mathbf{\Delta}_r$ sur de tels résultats est nulle. Le déséquilibre apparaît dans la variance de $\mathbf{\Delta}_r$, qui est linéairement croissante en IMB, approximativement. Une raison de choisir $\mathbf{x}$ comme étant un vecteur de groupes est que le conditionnement sur $\overline{\mathbf{x}}_r$ mène à des calculs relativement simples. Une moyenne fixe $\overline{\mathbf{x}}_r$ implique une valeur fixe de IMB. (Mais l'inverse n'est pas vrai; plusieurs $\overline{\mathbf{x}}_r$ peuvent donner le même *IMB*.) Une autre simplification quand $\mathbf{x}$ est un vecteur de groupes tient aux matrices diagonales $\mathbf{\Sigma}_r$ et $\mathbf{\Sigma}_s$. Le test empirique présenté à la section 9.1 concerne le résultat 1.

Les calculs simples obtenus grâce au vecteur de groupes le sont au détriment de la généralité. À Statistics Sweden, par exemple, il est fréquent les vecteurs $\mathbf{x}$ utilisés en production ne soient pas des vecteurs de groupes. L'obtention de résultats mathématiques transparents au sujet de $\mathbf{\Delta}_r$ est alors plus difficile.

Le résultat 2 donné à la section 8 est calculé sous un modèle de régression linéaire entre y et $\mathbf{x}$. Les y_k sont alors considérés comme aléatoires, leurs propriétés étant énoncées par le modèle. Il n'est plus nécessaire que $\mathbf{x}$ soit un vecteur de groupes. Les conclusions sont à certains égards similaires à celles du résultat 1. La situation 2 du test empirique à la section 9.2 se rapporte à la fois au résultat 1 et au résultat 2.

7 Le premier résultat

Le résultat 1 s'applique au contexte d'enquête suivant : un échantillon autopondéré s de taille n est tiré de $U = \{1,\ldots,k,\ldots,N\}$; le poids d_k est le même pour toutes les unités k. Le vecteur auxiliaire $\mathbf{x}$ est un vecteur de groupes de dimension J, de sorte que l'échantillon s et l'ensemble de répondants r, qui sont supposés être de tailles fixes $m < n$, sont divisés en J groupes non chevauchants. La notation pertinente est donnée à la fin de la section 3. Les valeurs y_k sont traitées comme étant fixes, non aléatoires, comme cela se fait habituellement dans l'approche classique fondée sur le plan de sondage. Si les y_k étaient observées pour tout $k \in s$, alors $\hat{Y}_{FUL} = N\,\overline{y}_s$, avec $\overline{y}_s = \sum_s y_k \big/ n$, serait sans biais sous le plan de sondage pour le total de population de y, $Y = \sum_U y_k$. Mais y_k est disponible pour $k \in r$ uniquement; l'estimateur CAL (5.1) devient $\hat{Y}_{CAL} = N \sum_{j=1}^{J} W_{js} \overline{y}_{r_j}$, où $\overline{y}_{r_j}$ est la moyenne des valeurs y_k des répondants dans le groupe j. Les propriétés statistiques – l'espérance et la variance – de $\left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right) \big/ N = \Delta_r$ avec $\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \overline{\mathbf{x}}_s = \sum_{j=1}^{J} W_{js} \overline{y}_{r_j} - \overline{y}_s$ sont données dans le résultat 1 pour les conditions probabilistes suivantes : tous les $\binom{n}{m}$ ensembles de réponses r de taille fixe m sont supposés a priori être également

probables. Sachant s, le déséquilibre IMB est déterminé par $\overline{\mathbf{x}}_r = (1/m)\left(m_1,\ldots,m_j,\ldots,m_J\right)'$. Sachant $\overline{\mathbf{x}}_r$, il nous reste $R = \prod_{j=1}^{J}\binom{n_j}{m_j}$ ensembles r, ayant tous la même probabilité non nulle $1/R$ et le même IMB, donné par (3.3). Les autres ensembles r de taille m ne sont plus dans le champ d'observation. Le conditionnement sur $\overline{\mathbf{x}}_r$ nous permet d'étudier les propriétés de l'estimateur CAL en fonction de l'*IMB*. Le résultat 1 fait intervenir la variance de la variable étudiée y, à l'intérieur des groupes et combinée sur les groupes :

$$S_{yj}^2 = \sum\nolimits_{s_j}\left(y_k - \overline{y}_{s_j}\right)^2 \Big/ \left(n_j - 1\right); \quad S_y^2 = \sum_{j=1}^{J} W_{js} S_{yj}^2. \tag{7.1}$$

Résultat 1. Soit s un échantillon autopondéré de taille n et soit $\mathbf{x}_k$ un vecteur de groupes de dimension J. Supposons que tous les $\binom{n}{m}$ ensembles de réponses r de taille fixe m sont a priori aussi probables les uns que les autres. Alors,

$$\overline{\Delta} = \mathrm{E}\left(\Delta_r \,|\, \overline{\mathbf{x}}_r, m, s\right) = 0 \tag{7.2}$$

$$S_\Delta^2 = \mathrm{E}\left(\left(\Delta_r - \overline{\Delta}\right)^2 \,\middle|\, \overline{\mathbf{x}}_r, m, s\right) = \left(\frac{1}{m} - \frac{1}{n}\right) S_y^2 + \frac{1}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p}{p_j} - 1\right) S_{yj}^2 \tag{7.3}$$

où $W_{js} = n_j/n$ et $p_j = m_j/n_j$ sont la taille relative et le taux de réponse, respectivement, pour le groupe j, $p = m/n$ est le taux de réponse global, et S_y^2 et S_{yj}^2 sont données en (7.1). Si les taux de réponse p_j et les variances S_{yj}^2 ne varient que peu sur les groupes, alors

$$S_\Delta^2 \approx \left(1 - p + \frac{IMB}{p^2}\right)\frac{S_y^2}{m} \tag{7.4}$$

où IMB est donné par (3.3).

En cas de réponse complète, quand $r = s$, les deuxièmes membres des équations (7.3) et (7.4) sont nuls; l'approximation dans (7.4) est exacte : $S_\Delta^2 = 0$. Pour interpréter le résultat 1, notons que le premier terme du deuxième membre de (7.3) est une constante, sachant m. Ce terme donne la variance conditionnelle pour une réponse parfaitement équilibrée, où p_j est la même pour tous les groupes. Le deuxième terme est le *terme de pénalisation*, à savoir la pénalisation du fait de ne pas obtenir l'équilibre parfait durant la collecte des données. La taille de ce terme dépend de la mesure dans laquelle le plan de collecte dynamique réussit à générer des taux de réponse de groupe p_j qui ne varient que faiblement. Ce terme est nul si l'on peut rendre tous les p_j égaux.

La formule (7.4) indique que la variance S_Δ^2 diminue parallèlement à l'IMB de manière plus ou moins linéaire. Donc, un faible déséquilibre accroît les chances d'obtenir un petit écart $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$, ce qui est important en pratique. Par exemple, pour une non-réponse de $1 - p = 40\,\%$, $S_\Delta^2 \approx 0{,}57 S_y^2/m$ si $IMB = 0{,}06$; mais si $IMB = 0$, c'est-à-dire en cas d'équilibre parfait, cette variance est réduite à $S_\Delta^2 \approx 0{,}40 S_y^2/m$. L'amélioration est claire, mais ne peut être qualifiée de très grande. Cela tient au fait que

$IMB/\,p^2$ est souvent faible comparativement à une non-réponse $1-p$ de l'ordre de 30 à 60 %, c'est-à-dire les cas dont nous nous préoccupons principalement ici. Donc, prendre des mesures en vue de réduire le déséquilibre a un effet désirable, mais modeste plutôt que puissant.

Dans (7.2) et (7.3), l'espérance $\mathrm{E}(\cdot)$ est obtenue en prenant la moyenne sur les $R=\prod_{j=1}^{J}\binom{n_j}{m_j}$ ensembles équiprobables r qui restent parmi $\binom{n}{m}$ après avoir fixé $\bar{\mathbf{x}}_r$. Il convient aussi de souligner que plus d'une moyenne $\bar{\mathbf{x}}_r$ peut donner la même valeur IMB. Donc, il peut exister plus d'une valeur S_Δ^2 pour le même IMB. La fonction linéairement croissante de IMB en (7.4) est néanmoins leur approximation commune.

8 Le deuxième résultat

Dans le résultat 1, les valeurs de la variable étudiée y_k sont traitées comme étant fixes, non aléatoires. Dans le résultat 2, elles sont aléatoires et leurs propriétés sont précisées par un modèle de régression linéaire ξ dont les résidus sont $\varepsilon_k = y_k - \mathbf{x}_k'\boldsymbol{\beta}$ pour un certain $\boldsymbol{\beta}$ inconnu :

$$E_\xi\left(y_k \mid \mathbf{x}_k\right)=\mathbf{x}_k'\,\boldsymbol{\beta};\quad E_\xi\left(\varepsilon_k^2 \mid \mathbf{x}_k\right)=\sigma_\varepsilon^2,\ \text{tout}\ k\in s;\quad E_\xi\left(\varepsilon_k\varepsilon_\ell \mid \mathbf{x}_k,\mathbf{x}_\ell\right)=0,\ \text{tout}\ k\neq\ell\in s. \quad (8.1)$$

Les propriétés données en (8.1) s'appliquent également aux unités k et ℓ appartenant à tout sous-ensemble r de s. Le résultat 2 donne l'espérance et la variance approximative de $\boldsymbol{\Delta}_r=\left(\mathbf{b}_r-\mathbf{b}_s\right)'\bar{\mathbf{x}}_s$ conditionnellement à un échantillon autopondéré fixe s et à un ensemble de répondants fixe r ayant respectivement pour taille n et m.

Résultat 2. Soit s un échantillon autopondéré de taille n. Soit $\mathbf{X}$ la matrice des données $\mathbf{x}$ de dimensions $J\times n$ dont les colonnes sont $\mathbf{x}_k$, $k\in s$. Alors, sous le modèle ξ en (8.1),

$$E_\xi\left(\Delta_r \mid \mathbf{X},r,s\right)=0;\quad E_\xi\left(\Delta_r^2 \mid \mathbf{X},r,s\right)\approx\left(1-p+\frac{IMB}{p^2}\right)\frac{\sigma_\varepsilon^2}{m}, \quad (8.2)$$

où m est la taille de l'ensemble de répondants fixe r, $p=m/n$ est le taux de réponse et IMB est donné par (3.2).

Le résultat 2 (pour un vecteur $\mathbf{x}$ arbitraire et les y_k aléatoires) reflète le résultat 1 (pour le vecteur $\mathbf{x}$ de groupes et les y_k non aléatoires) en ce sens que l'un et l'autre donnent une moyenne conditionnelle nulle et la même forme linéairement croissante pour la variance conditionnelle approximative.

Le calcul du résultat 2 donné à l'annexe 3 s'appuie sur une comparaison de deux formes quadratiques en $\bar{\mathbf{x}}_r-\bar{\mathbf{x}}_s$ données dans (3.1), Q_s et Q_r. La première est utilisée dans la statistique de déséquilibre (3.2), $IMB=P^2Q_s$; la seconde détermine les facteurs de pondération g_k pour l'estimateur CAL (5.1). L'approximation $Q_r\approx Q_s$, nécessaire pour le résultat 2, est justifiée à l'annexe 2.

9 Essai empirique

Les résultats 1 et 2 donnent le fondement pour tester empiriquement à la présente section la façon dont la moyenne et la variance de l'écart $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r = \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \overline{\mathbf{x}}_s$ dépendent du déséquilibre IMB. Les deux résultats énoncent que la variance de $\mathbf{\Delta}_r$ augmente de manière à peu près linéaire à mesure que IMB augmente, sans être faible, même si IMB est proche de zéro.

Nous utilisons des données réelles provenant d'une enquête estonienne réalisée auprès de 17 540 ménages. Les variables qui suivent sont disponibles pour chaque ménage, à savoir le revenu net du ménage, utilisé ici comme variable étudiée y, et trois variables catégoriques se rapportant au chef de ménage désigné, utilisées ici comme variables auxiliaires : i) sexe (1 pour masculin, 0 pour féminin), ii) activité économique (1 pour occupé, 0 pour non occupé) et iii) études, avec trois niveaux exhaustifs : faible, moyen et élevé.

Nous calculons la moyenne $\overline{\Delta}$ de $\mathbf{\Delta}_r$ et la variance S_Δ^2 de $\mathbf{\Delta}_r$ en prenant la moyenne sur les ensembles r pour une moyenne $\overline{\mathbf{x}}_r$ fixe, sachant s.

9.1 Situation de test 1

En harmonie avec le résultat 1, nous voulons prendre en considération les ensembles de réponses r de taille fixe m issus de l'échantillon s de taille n. Le volume des calculs est prohibitif, même pour une valeur assez faible de n. Nous avons tiré s comme un échantillon aléatoire simple de taille $n = 20$ parmi une population de 17 540. Les d_k sont alors constants. La moyenne d'échantillon de la variable y (revenu du ménage) était $\overline{y}_s = 10\,386{,}65$. Nous définissons $\mathbf{x}_k$ comme étant le vecteur de groupes de dimension $J = 3$ qui précise les trois niveaux d'études exhaustifs, à savoir faible, moyen et élevé. Pour l'échantillon réalisé s, nous avons $n\overline{\mathbf{x}}_s = (5,8,7)'$.

Nous avons fixé la taille des ensembles de réponses r à $m = 12$. Le taux de réponse est de 60 % pour chacun des $\binom{20}{12} \approx 1{,}26 \times 10^5$ ensembles de réponses r possibles. De ceux-ci, nous avons exclu tous ceux pour lesquels le vecteur des nombres de réponses $m\overline{\mathbf{x}}_r$ contenait un zéro, pour éviter une matrice $\mathbf{\Sigma}_r$ singulière. Cela nous a laissé 31 configurations (m_1, m_2, m_3) telles que $m_1 + m_2 + m_3 = 12$ et les trois nombres $m_j \geq 1$. Pour chacune des 31 possibilités, nous avons calculé $\overline{\Delta}$ et S_Δ^2 en prenant la moyenne sur les ensembles de réponses r satisfaisant la configuration fixe. Par exemple, $(m_1, m_2, m_3) = (3,4,5)$ est satisfaite par 14 700 ensembles de réponses r, de sorte que la moyenne et la variance de Δ_r sont calculées sur ces ensembles. D'autres configurations donnent un nombre beaucoup plus faible d'ensembles de réponses, par exemple, seulement 70 pour la configuration (3, 8, 1); quelques-unes de ces configurations peuvent être très influentes dans les calculs. Pour chacun des 31 cas, $\overline{\Delta}$ est théoriquement nulle, en vertu du résultat 1. Les calculs l'ont confirmé; une représentation graphique de $\overline{\Delta}$ en fonction de IMB n'est pas nécessaire. La figure 9.1 montre le tracé des 31 points de S_Δ^2 en fonction de IMB. En raison du caractère non unique de IMB mentionné plus haut, il arrive plusieurs fois que plus d'une valeur de S_Δ^2 existe pour une même valeur IMB. La figure 9.1 montre que S_Δ^2 suit clairement une tendance à la hausse quand IMB

augmente. La figure 9.1 montre aussi l'approximation $S_\Delta^2 \approx S_{\Delta approx}^2 = \left(S_y^2 / m\right)(1 - p + IMB/p^2)$ issue du résultat 1. Nous avons $p = 0{,}6$; $m = 12$ et $S_y^2 = 26{,}3 \times 10^6$, de sorte que l'approximation calculée, linéaire en IMB, est $S_{\Delta approx}^2 = a + b\, IMB$ avec $a = 0{,}879 \times 10^6$ et $b = 6{,}102 \times 10^6$. Pour les points associés à une faible valeur IMB, S_Δ^2 concorde étroitement avec $S_{\Delta approx}^2$ linéairement croissante. L'une des raisons est que, si IMB est faible, les taux de réponse de groupe p_j varient peu, ce qui est une des conditions pour une approximation proche, comme l'explique le calcul du résultat 1 à l'annexe 1. Pour des valeurs plus élevées de IMB, la tendance à la hausse en S_Δ^2 demeure évidente, mais la dispersion autour de la ligne droite théorique est plus prononcée. Cinq points aberrants dans la figure 9.1 ont une très grande valeur de S_Δ^2; trois d'entre eux sont obtenus quand une composante de (m_1, m_2, m_3) est égale au nombre maximal (5 ou 8 ou 7). Pour ces points, l'on s'attend à une approximation linéaire moins précise, les p_j étant loin d'être égaux.

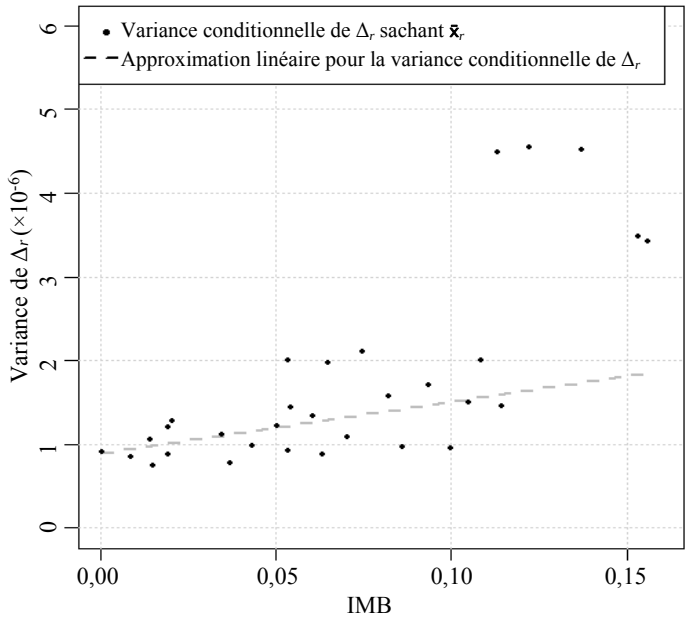

Figure 9.1 Variance conditionnelle de Δ_r en fonction du déséquilibre IMB; $\mathbf{x}_k$ est un vecteur de groupes de dimension 3; ensembles de réponses r de taille fixe 12 provenant d'un échantillon fixe s de taille 20.

9.2 Situation de test 2

La configuration et les étapes de calcul sont similaires à celles de la situation de test 1, mais $\mathbf{x}_k$ n'est plus un vecteur de groupes; certains résultats changent considérablement comparativement à la situation de test 1.

Un nouvel échantillon aléatoire simple s de taille $n = 20$ a été tiré parmi les 17 540 ménages. Pour cet échantillon, $\bar{y}_s = 9\,618{,}4$. Posons que $\mathbf{x}_k$ englobe les trois variables auxiliaires (i), (ii) et (iii), mais que

celles-ci ne sont pas entièrement croisées : sexe (variable univariée codée 0 ou 1), activité économique (variable univariée codée 0 ou 1) et niveau d'études (trois catégories exhaustives codées (1,0,0) ou (0,1,0) ou (0,0,1)). $\mathbf{x}_k$ n'est pas un vecteur de groupes; sa dimension est de $1+1+3=5,$ et $\mathbf{x}_r$ possède $2\times 2\times 3=12$ valeurs possibles $\boldsymbol{\Sigma}_r$ et $\boldsymbol{\Sigma}_s$ ne sont pas des matrices diagonales. Nous avons $n\overline{\mathbf{x}}_s = (9, 11, 4, 7, 9)'$. Pour cet échantillon $s,$ nous avons considéré les ensembles de réponses r de taille fixe $m=12,$ sauf ceux pour lesquels une ou plusieurs des cinq composantes du vecteur des nombres de réponses $m\overline{\mathbf{x}}_r$ sont nulles. Cela a laissé 658 vecteurs différents $m\overline{\mathbf{x}}_r,$ chacun composé de cinq nombres non nuls et satisfait par un certain nombre d'ensembles de réponses r sur lesquels nous avons calculé, en prenant la moyenne, le terme $\overline{\Delta}$ et la variance $S^2_\Delta.$ Il s'agit donc de moments conditionnés sur $\overline{\mathbf{x}}_r.$

La figure 9.2 montre le tracé des 658 points pour $\overline{\Delta}$ en fonction de $IMB.$ Dans la situation de test 1, $\overline{\Delta}$ était nulle pour chaque point parce que $\mathbf{x}_k$ était un vecteur de groupes. Ce n'est pas le cas dans la figure 9.2, où les moyennes $\overline{\Delta}$ s'étalent en éventail quand IMB augmente. Elles sont nettement plus concentrées autour de zéro pour les faibles valeurs que pour les grandes valeurs de $IMB.$ Plusieurs points (c'est-à-dire plusieurs moyennes $\overline{\mathbf{x}}_r$) peuvent donner la même ou quasi la même valeur de $IMB.$ La figure 9.2 montre que dans un voisinage restreint d'une valeur fixe IMB_0 sur l'axe $IMB,$ la moyenne des moyennes $\overline{\Delta}$ est approximativement nulle. En référence au résultat 2, nous pouvons nous attendre à ce que la moyenne de $\overline{\Delta}$ pour une valeur fixe de IMB soit quasi nulle : sous le modèle (8.1) pour $y_k,$ le résultat 2 dit que $E_\xi(\Delta_r \mid \mathbf{X}, r, s)=0.$ Quand $\mathbf{X}$ et r sont fixes, il en est de même de $IMB.$ Si le modèle est une raisonnablement bonne représentation, la moyenne de Δ_r pour un IMB fixe devrait être proche de zéro, comme l'indique la figure 9.2.

La figure 9.3 montre le tracé de la variance conditionnelle S^2_Δ en fonction de $IMB.$ La tendance d'une variance S^2_Δ linéairement croissante en IMB prédomine, même si $\mathbf{x}_k$ n'est pas un vecteur de groupes ici. La figure 9.3 montre la droite approximative calculée $S^2_{\Delta approx} = (\hat{\sigma}^2_\varepsilon / m)(1 - p + IMB/p^2)$ dérivée du résultat 2, en utilisant $\hat{\sigma}^2_\varepsilon = \sum_s (y_k - \mathbf{x}'_k\mathbf{b}_s)^2 / (n-J)$ pour estimer $\sigma^2_\varepsilon.$ Nous avons $J=5;\ p=0{,}6;$ $m=12$ et $\hat{\sigma}^2_\varepsilon = 33{,}6\times 10^6,$ de sorte que la droite dans la figure 9.3 est $S^2_{\Delta approx} = a + b\ IMB$ avec $a = 1{,}12\times 10^6$ et $b = 7{,}78\times 10^6.$ L'approximation linéaire tient particulièrement bien pour une faible valeur de $IMB,$ disons inférieure à 0,1. Pour les grandes valeurs de $IMB,$ la dispersion est importante; S^2_Δ prend certaines valeurs très grandes, et certaines valeurs très faibles également. La figure 9.4 montre le comportement conjoint de $\overline{\Delta}$ et S^2_Δ pour les 658 points. La taille d'un point est proportionnelle à IMB^2; l'élévation au carré a pour objet de mieux contraster les grandes et les petites valeurs de $IMB.$ Nous constatons que les ensembles de réponses r pour lesquels IMB est petit donnent de faibles valeurs de $\overline{\Delta}$ et $S^2_\Delta,$ signe favorable car les estimateurs CAL et FUL sont alors proches. Par exemple, pour les points satisfaisant $IMB \le 0{,}1;$ $\overline{\Delta}$ est dans l'intervalle (-1 390; 1 447) et S^2_Δ est dans ($0{,}846\times 10^6$; $4{,}86\times 10^6$). Ces intervalles sont étroits; cela est encore plus prononcé pour $IMB \le 0{,}05.$ Quand IMB est grand, cette situation avantageuse disparaît. Par exemple, $\overline{\Delta}$ peut être très petite et, simultanément, $S^2_\Delta,$ très grande (points au milieu et du côté droit de la figure). Par ailleurs, S^2_Δ peut être quasi nulle tandis que $\overline{\Delta}$ est très

grande en valeur absolue (points dans les parties supérieure et inférieure gauche de la figure.) La situation de test 2 illustre le fait qu'un vecteur sans groupes $\mathbf{x}_k$ peut donner à la fois une moyenne distinctement non nulle de Δ_r et une variance élevée de Δ_r, et que ces tendances sont accentuées par un grand déséquilibre.

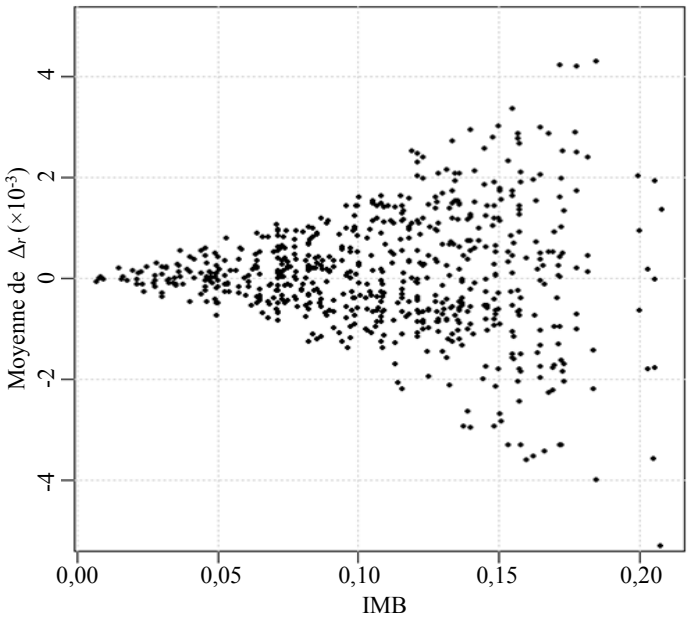

Figure 9.2 Moyenne conditionnelle de Δ_r en fonction du déséquilibre *IMB*; $\mathbf{x}_k$ est un vecteur sans groupes de dimension 5; ensembles de réponses r de taille fixe 12 tirés d'un échantillon fixe de taille 20.

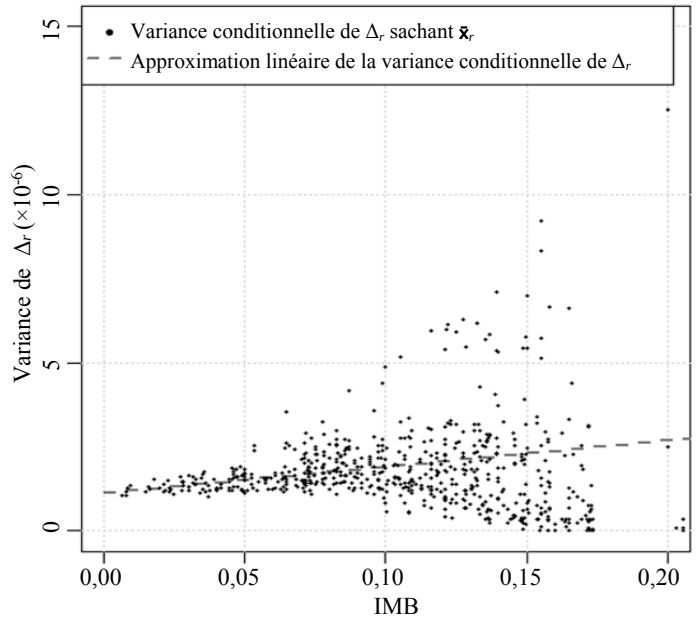

Figure 9.3 Variance conditionnelle de Δ_r en fonction du déséquilibre *IMB*; $\mathbf{x}_k$ est un vecteur sans groupes de dimension 5; ensembles de réponses r de taille fixe 12 tirés d'un échantillon fixe de taille 20.

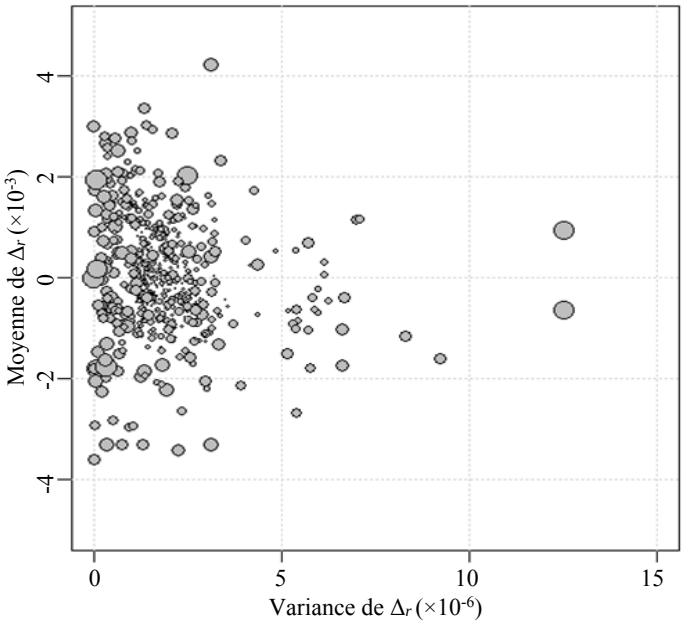

Figure 9.4 Tracé de la moyenne conditionnelle de Δ_r en fonction de la variance conditionnelle de Δ_r; $\mathbf{x}_k$ est un vecteur sans groupes de dimension 5; ensembles de réponses r de taille fixe 12 tirés d'un échantillon fixe de taille 20. Taille des points proportionnelle au carré du déséquilibre.

10 Discussion

Nous commentons ici plusieurs questions qui se posent et indiquons les limites de notre étude.

1. Choix des variables pour le vecteur auxiliaire. Les variables auxiliaires pour le vecteur $\mathbf{x}$ sont traitées comme un choix fixe dans le présent article. Ce choix est important quand une éventuellement grande quantité de ces variables est disponible. Lesquelles faut-il choisir pour atteindre l'objectif ultime, qui est la production d'estimations les plus exactes possibles ? Le résultat 1 montre que, dans le cas du vecteur de groupes, deux facteurs sont importants pour S_Δ^2 (qui détermine la variance conditionnelle de CAL) : le déséquilibre de la réponse IMB et la variance S_y^2 de la variable étudiée y. Le fait que S_Δ^2 soit (approximativement) linéairement décroissante en IMB incite à essayer de réduire IMB durant la collecte des données. Mais inclure plus de variables dans $\mathbf{x}$ augmente IMB (parce que l'on recherche une concordance sur un plus grand nombre de moyennes $\mathbf{x}$). Dans le cas de la variance de y, S_y^2, c'est le contraire qui se produit. En vertu de (7.1), S_y^2 est une variance résiduelle moyenne autour des moyennes de groupe; l'introduction de variables supplémentaires dans $\mathbf{x}$ réduira S_y^2, surtout si elles expliquent bien y. Les deux facteurs travaillent en sens opposé : un plus grand nombre de variables auxiliaires donne un plus grand déséquilibre IMB, mais une plus faible variance de y. Cela suggère un compromis possible, question qui n'est pas examinée dans le présent article. L'une des particularités d'un vecteur de groupes $\mathbf{x}$ joue un rôle, à savoir que, si un plus grand nombre de variables catégoriques entre dans le vecteur, la dimension de celui-ci augmente de façon multiplicative. Le risque que des cellules soient petites ou vides limite

l'extension. En guise d'illustration, si $\mathbf{x} = (sexe \times études \times âge)$ de dimension $J = 2 \times 3 \times 4 = 24$ est étendu afin d'inclure également la *profession* avec 4 catégories, de manière entièrement croisée, la nouvelle dimension (égale au nouveau nombre de groupes) est $J = 24 \times 4 = 96$. En principe, S_y^2 diminue, mais le risque d'obtenir de petites cellules est une bonne raison de s'abstenir de croiser entièrement toutes les variables et de les faire plutôt intervenir dans un vecteur $\mathbf{x}$ sans groupes. Ce cas est abordé dans le résultat 2, selon lequel, si $\mathbf{x}$ explique bien y, alors σ_ε^2 est petite et donnera une faible variance souhaitée pour Δ_r.

2. Information auxiliaire à différents niveaux. Dans le présent article, le déséquilibre IMB et l'estimateur par calage $\hat{Y}_{CAL}$ utilisent le même vecteur $\mathbf{x}$, et plus particulièrement un vecteur qui contient des données auxiliaires pour les unités de l'échantillon uniquement. Il s'agit d'un cas réaliste. Cependant, dans des formulations plus générales, la collecte des données ferait usage d'un vecteur de surveillance $\mathbf{x}_{MV}$ éventuellement différent du vecteur de calage $\mathbf{x}_{CAL}$ utilisé plus tard dans l'estimation. Le premier est un instrument destiné à obtenir un faible déséquilibre IMB dans la réponse, tandis que le second sert à obtenir de bons poids calés pour $\hat{Y}_{CAL}$. Une raison pour laquelle $\mathbf{x}_{MV}$ et $\mathbf{x}_{CAL}$ pourraient différer en pratique est que les variables auxiliaires pour l'estimation peuvent être des versions mises à jour des mêmes variables disponibles au moment de la collecte des données. Il peut y avoir d'autres raisons de choisir $\mathbf{x}_{MV}$ et $\mathbf{x}_{CAL}$ de sorte qu'ils soient différents. En outre, ils peuvent contenir de l'information (si elle est disponible) au niveau de la population. Des extensions de notre approche à ce genre de situations sont possibles.

3. Avantage incertain d'une réduction du déséquilibre. Schouten et coll. (2014) dégagent des preuves que l'équilibrage de la réponse réduit le biais. Nous constatons aussi qu'il existe une motivation à s'efforcer d'obtenir, durant la collecte des données, un ensemble ultime de réponses dont le déséquilibre IMB est faible. Comme le montrent théoriquement les résultats 1 et 2, et comme le confirment empiriquement les situations de test 1 et 2, un faible déséquilibre donne un écart $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$ dont l'espérance est nulle ou quasi nulle et dont la variance est petite. Il s'agit de notre protection contre un biais important. Si IMB augmentait, la variance aurait tendance à augmenter. L'espérance nulle de l'écart $\hat{Y}_{CAL} - \hat{Y}_{FUL}$ est une propriété moyenne. Il n'y a aucune garantie que l'écart soit petit pour tout ensemble de répondants particulier r présentant un faible IMB.

4. L'équilibre parfait n'élimine pas le biais. Un déséquilibre nul, $IMB = 0$, implique une égalité des moyennes pour l'ensemble de répondants et l'échantillon complet, $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. Si cet équilibre parfait était atteint, le terme d'ajustement du biais dans (5.2) serait nul; l'estimateur par calage (CAL) et l'estimateur par facteur d'extension (EXP) sont identiquement égaux. On peut dire que, si un équilibre parfait est atteint, la puissance du vecteur auxiliaire est épuisée, non pas en ce qui concerne sa possibilité d'expliquer la variable étudiée, mais en ce qui concerne sa possibilité de se distinguer de l'estimateur EXP rudimentaire, qui, même s'il n'utilise aucune information auxiliaire, est aussi bon que l'estimateur CAL qui, autrement, est meilleur. Cependant, $\text{CAL} \equiv \text{EXP}$ n'est toujours pas une situation idéale. Comme le montre le

résultat 1, la variance de l'écart de CAL n'est pas proche de zéro même si le déséquilibre IMB est quasi nul. L'équilibre parfait n'élimine pas l'écart de CAL, mais un petit IMB protège contre un grand écart.

5. Implications pratiques. Dans le présent article, nous nous intéressons principalement aux enquêtes présentant une « non-réponse importante et inévitable » qui ne peut pas raisonnablement (compte tenu des contraintes de temps et de budget de l'enquête) être réduite à un pourcentage à un chiffre, même en engageant d'importantes ressources. Les enquêtes où la non-réponse est égale ou supérieure à 30 % sont fréquentes aujourd'hui. On est loin de la situation idéale où la réponse est quasi totale, et où le déséquilibre et la non-réponse cesseraient essentiellement de poser problème, vu que les estimateurs EXP, CAL et FUL seraient proches les uns des autres.

6. Indications pour la généralisation. Les résultats 1 et 2 montrent les propriétés de l'écart de l'estimateur CAL parmi les ensembles de réponses sous une formulation donnée du vecteur auxiliaire. Il serait souhaitable de généraliser les résultats à d'autres situations. Nos preuves reposent sur l'hypothèse qu'il existe certaines matrices inverses. Les extensions à d'autres cas seraient possibles à l'aide de l'inverse généralisée de Moore-Penrose.

Remerciements

Les présents travaux ont été financés par la subvention 9127 de l'*Estonian Science Foundation* et par l'*Institutional Research Funding IUT34-5* de l'Estonie. Les auteurs tiennent à remercier un rédacteur associé et un examinateur, tous deux anonymes, de leurs commentaires constructifs.

Annexe 1

Obtention du résultat 1

Nous obtenons les expressions (7.2) à (7.4) sous les conditions et la notation exposées à la section 7. L'échantillon s est autopondéré, de taille n, et $\mathbf{x}$ est un vecteur de groupes de dimension J. Nous supposons la probabilité $\binom{n}{m}^{-1}$ pour chaque ensemble de répondants r de taille fixe m. Nous calculons l'espérance et la variance de $\Delta_r = (\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s = \sum_{j=1}^{J} W_{js} \bar{y}_{r_j} - \bar{y}_s$, où $W_{js} = n_j / n$, conditionnellement à m fixe et à la moyenne $\bar{\mathbf{x}}_r = (1/m)(m_1, \ldots, m_j, \ldots, m_J)$; $\sum_{j=1}^{J} m_j = m$. Sous ce conditionnement, $R = \prod_{j=1}^{J} \binom{n_j}{m_j}$ ensembles r ont la même probabilité, où n_j est la taille du groupe d'échantillon s_j; $\sum_{j=1}^{J} n_j = n$. Cela est identique à la structure de probabilité pour l'échantillonnage aléatoire simple stratifié de m_j à partir de n_j dans la strate $j; j = 1, \ldots, J$. Sachant m et $\bar{\mathbf{x}}_r$, l'espérance et la variance de $\bar{y}_{r_j}$ sont, respectivement, $\bar{y}_{s_j} = \sum_{s_j} y_k / n_j$ et $(1/m_j - 1/n_j) S_{yj}^2$ avec S_{yj}^2 donné en (7.1). Donc, $\bar{\Delta} = \sum_{j=1}^{J} W_{js} \bar{y}_{s_j} - \bar{y}_s = 0$, ce qui prouve (7.2), et $S_{\Delta}^2 = \sum_{j=1}^{J} W_{js}^2 (1/m_j - 1/n_j) S_{yj}^2$. En substituant $p_j = m_j / n_j$ et $p = m/n$, et en utilisant $S_y^2 = \sum_{j=1}^{J} W_{js} S_{yj}^2$ donné en (7.1), nous obtenons

$$S_{\Delta}^2 = \frac{1}{n}\sum_{j=1}^{J} W_{js}\left(\frac{1}{p_j}-1\right)S_{yj}^2 = \left(\frac{1}{m}-\frac{1}{n}\right)S_y^2 + \frac{1}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p}{p_j}-1\right)S_{yj}^2. \qquad \text{(A.1)}$$

Cela prouve (7.3). Pour analyser le terme de pénalisation (deuxième terme du deuxième membre) dans (A.1), supposons que les p_j ne varient que peu autour du taux global p. Alors, $\delta_j = p_j / p - 1$, $j = 1,\ldots,J$, sont de petites quantités, et $1/p_j = 1/p\left(1+\delta_j\right) = (1/p)\left(1-\delta_j+\delta_j^2-\delta_j^3+\ldots\right)$. En gardant les termes jusqu'à l'ordre deux, $p/p_j - 1 \approx -\delta_j + \delta_j^2$. Le terme de pénalisation est alors approximé par

$$\frac{1}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p}{p_j}-1\right)S_{yj}^2 \approx -\frac{1}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p_j}{p}-1\right)S_{yj}^2 + \frac{1}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p_j}{p}-1\right)^2 S_{yj}^2. \qquad \text{(A.2)}$$

Supposons en outre que les variances de groupe S_{yj}^2, $j = 1,\ldots,J$ ne varient que peu autour de leur moyenne pondérée S_y^2. En utilisant l'approximation $S_{yj}^2 \approx S_y^2$ dans (A.2), nous obtenons

$$\frac{S_y^2}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p}{p_j}-1\right) \approx -\frac{S_y^2}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p_j}{p}-1\right) + \frac{S_y^2}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p_j}{p}-1\right)^2.$$

Ici, le premier terme du deuxième membre est nul. Le deuxième terme, égal à $(IMB/p^2)\left(S_y^2/m\right)$ avec IMB donné en (3.3), devient une deuxième approximation pour le terme de pénalisation en (A.1). Par conséquent, $S_{\Delta}^2 \approx (1/m - 1/n)S_y^2 + (IMB/p^2)\left(S_y^2/m\right)$. Cela donne le résultat souhaité (7.4).

Annexe 2

Comparaison de deux formes quadratiques

Nous comparons les deux formes quadratiques en $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$, Q_r et Q_s définies en (3.1), et justifions l'approximation $Q_r \approx Q_s$ nécessaire dans la preuve à l'annexe 3 du résultat 2. Les matrices de pondération respectives, $\boldsymbol{\Sigma}_r$ et $\boldsymbol{\Sigma}_s$, sont définies positives. Par conséquent, Q_r (ou Q_s) ne peut être égal à zéro que sous l'équilibre parfait $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. Puisque $Q_r = Q_s$ pour l'équilibre parfait, l'argument de continuité implique que $Q_r \approx Q_s$ pour des ensembles de réponses quasi équilibrés. Dans quelle mesure sont-ils proches plus généralement ?

L'estimateur CAL (5.1) utilise les facteurs de pondération $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1}\mathbf{x}_k$, définis pour tout $k \in s$. Leur lien avec Q_r est montré dans les deuxième et troisième expressions en (A.3) ci-dessous. Considérons également les facteurs $f_k = \bar{\mathbf{x}}_r' \boldsymbol{\Sigma}_s^{-1}\mathbf{x}_k$ pour $k \in s$. Ils jouent un rôle déterminant pour Q_s, et pour $IMB = P^2 Q_s$, comme le montrent les deux dernières expressions dans (A.3). Les moments qui suivent de g_k et f_k sont vérifiés à l'aide de la condition (2.2) appliquée au vecteur $\mathbf{x}$:

$$\bar{g}_r = 1, \mathrm{var}_r(g) = Q_r,\ \bar{g}_s = 1 + Q_r; \quad \bar{f}_s = 1, \mathrm{var}_s(f) = Q_s,\ \bar{f}_r = 1 + Q_s. \qquad \text{(A.3)}$$

Pour g_k, les moyennes sont définies comme étant $\bar{g}_s = \sum_s d_k g_k \big/ \sum_s d_k$, $\bar{g}_r = \sum_r d_k g_k \big/ \sum_r d_k$, et les variances sont $\text{var}_s(g) = \sum_s d_k (g_k - \bar{g}_s)^2 \big/ \sum_s d_k$, $\text{var}_r(g) = \sum_r d_k (g_k - \bar{g}_r)^2 \big/ \sum_r d_k$. Pour les moments correspondants de f_k, remplaçons g_k par f_k. Les variances $\text{var}_s(g)$ et $\text{var}_r(f)$ n'ont pas une forme aussi transparente et seront approximées. Une autre propriété importante découlant de (2.2) est $\sum_s d_k f_k g_k \big/ \sum_s d_k = \sum_r d_k f_k g_k \big/ \sum_r d_k = 1$. Ces équations et expressions appropriées dans (A.3) donnent

$$\text{cov}_s(f,g) = \sum_s d_k (f_k - \bar{f}_s)(g_k - \bar{g}_s) \big/ \sum_s d_k = 1 - \bar{f}_s \bar{g}_s = -Q_r,$$
$$\text{cov}_r(f,g) = \sum_r d_k (f_k - \bar{f}_r)(g_k - \bar{g}_r) \big/ \sum_r d_k = 1 - \bar{f}_r \bar{g}_r = -Q_s.$$

Maintenant, utilisons $\text{cov}_s^2(f,g) \le \text{var}_s(f)\,\text{var}_s(g)$ et l'inégalité analogue où r remplace s. En utilisant aussi $\text{var}_s(f) = Q_s$ et $\text{var}_r(g) = Q_r$ provenant de (A.3), nous obtenons les bornes pour le ratio Q_r / Q_s :

$$\frac{Q_s}{\text{var}_r(f)} \le \frac{Q_r}{Q_s} \le \frac{\text{var}_s(g)}{Q_r}. \tag{A.4}$$

Afin d'obtenir des bornes supérieure et inférieure plus transparentes, approximons les deux variances en (A.4) en supposant que le coefficient de variation (écart-type divisé par la moyenne) est approximativement le même pour la réponse r que pour l'échantillon s, et cela pour f ainsi que g. Cela suppose une certaine stabilité du coefficient de variation. Alors, $\text{var}_s(g) \approx (\bar{g}_s)^2 \text{var}_r(g) \big/ (\bar{g}_r)^2 = (1+Q_r)^2 Q_r$, de sorte que la borne supérieure en (A.4) est approximativement $(1+Q_r)^2 > 1$. De même, $\text{var}_r(f) \approx (\bar{f}_r)^2 \text{var}_s(f) \big/ (\bar{f}_s)^2 = (1+Q_s)^2 Q_s$, ce qui donne $(1+Q_s)^{-2} < 1$ comme borne inférieure approximative en (A.4). L'intervalle approximatif pour le ratio Q_r / Q_s est donc

$$Q_r / Q_s \in \left((1+Q_s)^{-2}, (1+Q_r)^2 \right).$$

Cela illustre le fait que le ratio n'est pas très éloigné de 1, car pour la plupart des données, les valeurs de Q_s et Q_r sont toutes deux faibles comparativement à 1, Q_r étant habituellement la valeur quelque peu plus grande. Des travaux empiriques donnent néanmoins à penser que la borne supérieure approximative $(1+Q_r)^2$ peut souvent être trop faible.

Annexe 3

Obtention du résultat 2

Nous calculons les expressions en (8.2) sous les conditions énoncées. Les tailles de r et s sont m et n, respectivement; le taux de réponse est $p = m/n$. L'écart de l'estimateur CAL par rapport à l'estimateur sans biais FUL est $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$ où

$$\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \overline{\mathbf{x}}_s = \Sigma_r d_k g_k y_k \,/ \Sigma_r d_k - \Sigma_s d_k y_k \,/ \Sigma_s d_k$$

avec $\mathbf{b}_r$ et $\mathbf{b}_s$ donnés par (4.1), et $g_k = \overline{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$. Notons que $\mathbf{b}_s' \overline{\mathbf{x}}_s = \overline{y}_s$ en vertu de (2.2). Or, $\Sigma_r d_k g_k \mathbf{x}_k' \,/ \Sigma_r d_k = \Sigma_s d_k \mathbf{x}_k' \,/ \Sigma_s d_k = \overline{\mathbf{x}}_s'$. En utilisant le résultat de la multiplication à droite de cette équation par $\boldsymbol{\beta}$, nous obtenons $\Delta_r = \Sigma_r d_k g_k \left(y_k - \mathbf{x}_k' \boldsymbol{\beta}\right) / \Sigma_r d_k - \Sigma_s d_k \left(y_k - \mathbf{x}_k' \boldsymbol{\beta}\right) / \Sigma_s d_k$, qui exprime Δ_r en fonction des résidus $\varepsilon_k = y_k - \mathbf{x}_k' \boldsymbol{\beta}$ du modèle (8.1) :

$$\Delta_r = \frac{\sum_r d_k g_k \varepsilon_k}{\sum_r d_k} - \frac{\sum_s d_k \varepsilon_k}{\sum_s d_k}.$$

Puis, nous utilisons les propriétés sous le modèle de ε_k dans (8.1). De $E_\xi\left(\varepsilon_k \mid \mathbf{x}_k\right) = 0$ pour tout k, il découle que $E_\xi\left(\Delta_r \mid \mathbf{X}, r, s\right) = 0$. Pour évaluer la variance, nous utilisons $E_\xi\left(\varepsilon_k^2 \mid \mathbf{x}_k\right) = \sigma_\varepsilon^2$, pour tout $k \in s$, et $E_\xi\left(\varepsilon_k \varepsilon_\ell \mid \mathbf{x}_k, \mathbf{x}_\ell\right) = 0$ pour tout $k \neq \ell \in s$. Cela donne

$$E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) = \sigma_\varepsilon^2 \frac{\sum_r d_k^2 g_k^2}{\left(\sum_r d_k\right)^2} + \sigma_\varepsilon^2 \frac{\sum_s d_k^2}{\left(\sum_s d_k\right)^2} - 2\sigma_\varepsilon^2 \frac{\sum_r d_k^2 g_k}{\left(\sum_r d_k\right)\left(\sum_s d_k\right)}.$$

Ici, d_k s'élimine, parce que sa valeur est constante. Les première et deuxième expressions dans (A.3) sont vérifiées pour tout d_k, en particulier d_k constant, de sorte que nous obtenons $\sum_r g_k \,/ m = 1$ pour la moyenne et $\sum_r g_k^2 \,/ m = Q_r + 1$ pour la variance plus le carré de la moyenne. Par conséquent,

$$E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) = \left(\frac{1}{m}\left(1 + Q_r\right) + \frac{1}{n} - 2\frac{1}{n}\right)\sigma_\varepsilon^2 = \left(\frac{1}{m} - \frac{1}{n} + \frac{Q_r}{m}\right)\sigma_\varepsilon^2.$$

En guise d'étape finale, nous utilisons l'approximation $Q_r \approx Q_s$ justifiée à l'annexe 2, et $IMB = p^2 Q_s$. Alors, comme il est affirmé dans le résultat 2, $E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) \approx \left(1 - p + IMB / p^2\right)\left(\sigma_\varepsilon^2 \,/ m\right)$.

Bibliographie

Beaumont, J.-F., Bocci, C. et Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-622.

Bethlehem, J., Cobben, F. et Schouten, B. (2011). *Handbook of nonresponse in households surveys*. New York: John Wiley & Sons, Inc.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

Couper, M.P., et Wagner, J. (2011). Using paradata and responsive design to manage survey nonresponse. Proceedings, 58th World Statistics Congress, International Statistical Institute.

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

Groves, R. (2006). Research synthesis: Nonresponse rates and nonresponse error in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Groves, R.M., et Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

Little, R.J.A., et Vartivarian, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage ? *Techniques d'enquête*, 31, 2, 175-183. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2005002/article/9046-fra.pdf.

Lundquist, P., et Särndal, C.-E. (2013). Aspects of responsive design. With applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29, 557-582.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. et Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rao, R.S., Glickman, M.E. et Glynn, R.J. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, 27, 2196-2213.

Särndal, C.-E. (2011). The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics,* 27, 1-21.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: John Wiley & Sons, Inc.

Särndal, C.-E., et Lundquist, P. (2014a). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology,* 2, 361-387.

Särndal, C.-E., et Lundquist, P. (2014b). Balancing the response and adjusting estimates for nonresponse bias: Complementary activities. *Journal de la Société Française de Statistique*, 155(4), 28-50.

Schouten, B., Calinescu, M. et Luiten, A. (2013). Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs. *Techniques d'enquête,* 39, 1, 33-66. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/12-001-x2013001-fra.pdf.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête,* 35, 1, 107-121. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2009001/article/10887-fra.pdf.

Schouten, B., Cobben, F., Lundquist, P. et Wagner, J. (2016). Does more balanced survey response imply less non-response bias? *Journal of the Royal Statistical Society, Series A*, 179, 727-748.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias. Thèse de doctorat, University of Michigan, Ann Arbor.

Wagner, J., et Raghunathan, T.E. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29, 1014-1024.