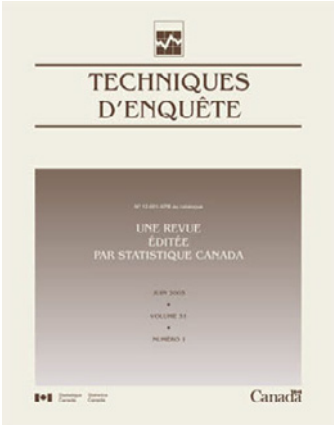
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Une évaluation de l’amélioration de l’exactitude au moyen d’un plan de sondage adaptatif



par Carl-Erik Särndal et Peter Lundquist

Date de diffusion : le 27 juin 2019

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.

Techniques d'enquête, juin 2019
Vol. 45, nº 2, pp. 339-360
Statistique Canada, nº 12-001-X au catalogue

Une évaluation de l'amélioration de l'exactitude au moyen d'un plan de sondage adaptatif

Carl-Erik Särndal et Peter Lundquist[1]

Résumé

De nos jours, il y a une non-réponse élevée dans de nombreuses enquêtes-échantillons, y compris d'importantes enquêtes menées par des organismes statistiques gouvernementaux. Une collecte de données adaptative peut être avantageuse dans cette situation : il est possible de réduire le biais de non-réponse dans les estimations de l'enquête, jusqu'à un certain point, en produisant un ensemble de répondants bien équilibré. Les variables auxiliaires ont un double objectif. Utilisées au cours de la phase d'estimation, elles réduisent le biais, sans toutefois l'éliminer complètement, par une pondération ajustée par calage. Au cours de la phase précédente de collecte de données adaptative, les variables auxiliaires jouent également un rôle important : elles contribuent à réduire le déséquilibre dans l'ensemble final de répondants. Dans le contexte de cette utilisation combinée de variables auxiliaires, le présent article est consacré à un examen de l'écart entre l'estimation par calage et l'estimation sans biais (réponse complète). Nous montrons que cet écart est la somme de deux composantes. La *composante réductible* peut être réduite, par la collecte de données adaptative, jusqu'à zéro si une réponse parfaitement équilibrée est obtenue par rapport à un vecteur auxiliaire choisi. En revanche, la *composante résistante* ne varie pas ou varie peu sous l'effet d'une réponse mieux équilibrée; elle représente une partie de l'écart qu'un plan adaptatif ne permet pas d'éliminer. La taille relative de cette première composante est un indicateur de l'avantage qu'on peut tirer d'un plan de sondage adaptatif.

Mots-clés : Non-réponse; plan de sondage adaptatif; déséquilibre de la réponse; réduction du biais.

1 Introduction

Pour un grand nombre d'importantes enquêtes par échantillonnage probabiliste, la période de collecte des données se clôture par la déclaration d'un taux considérable de non-réponse. À l'étape de l'estimation qui suit, il faut néanmoins établir les meilleures estimations possible en fonction de l'ensemble de répondants produit par une collecte de données qui n'est pas idéale. Quelles que soient les techniques utilisées, il existe un biais de non-réponse, et il faut faire en sorte qu'il demeure le plus faible possible. Un *plan de sondage réactif – ou adaptatif –* pour les enquêtes-ménages et les autres types d'enquêtes a pour objet d'exercer un contrôle actif sur les erreurs et les coûts d'enquête aux phases de la planification et de la collecte des données. L'un des objectifs consiste à obtenir un ensemble final de répondants qui est susceptible d'accroître la probabilité selon laquelle les estimations de l'enquête seront plus exactes. Le concept du plan de sondage réactif a été inventé par Groves et Heeringa (2006).

La collecte de données dans le cadre d'une grande enquête s'étend habituellement sur un certain nombre de jours ou de semaines au cours desquels on tente d'entrer en contact avec les unités de l'échantillon probabiliste choisi. Certaines unités produiront un résultat fructueux au bout de quelques tentatives. D'autres unités seront finalement déclarées non répondantes en raison de leur refus de répondre ou de l'absence de contact malgré des appels répétés. La prémisse du présent article n'est pas qu'on finira par réduire le taux de non-réponse à moins de 10 %. Il est plus réaliste d'affirmer qu'il continuera d'exister un taux élevé de

1. Carl-Erik Särndal, Statistics Sweden, SE-70185 Örebro, Suède. Courriel : carl.sarndal@telia.com; Peter Lundquist, Statistics Sweden, SE-10451 Stockholm, Suède. Courriel : peter.lundquist@scb.se.

non-réponse, de l'ordre de 30 % à 50 %, lorsque la période de collecte des données devra nécessairement se terminer. Il faut néanmoins produire les meilleures estimations possible.

Diverses techniques ont été proposées et mises à l'essai en vue de l'amélioration de la collecte des données : la priorisation des cas, les règles d'arrêt, l'équilibrage, etc. La priorisation des cas est examinée, par exemple, dans Peytchev, Riley, Rosen, Murphy et Lindblad (2010) et dans Beaumont, Haziza et Bocci (2014). Il est question des règles d'arrêt dans Rao, Glickman et Glynn (2008), et dans Wagner et Raghunathan (2010).

Des chercheurs comme Peytcheva et Groves (2009) ont examiné la variation entre les taux de réponse de différents sous-groupes démographiques à la recherche d'indications d'un biais. Andridge et Little (2011) proposent l'analyse indirecte par modèles de mélange de schémas d'observation comme méthode d'évaluation du biais de non-réponse. Le biais de non-réponse ne peut être quantifié, mais des indicateurs du risque de biais peuvent être utilisés, comme il en est question dans Wagner (2012), Kreuter, Olson, Wagner, Yan, Ezzati-Rice, Casas-Cordero, Lemay, Peytchev, Groves et Raghunathan (2010), Lohr, Riddles et Morganstein (2016), Nishimura, Wagner et Elliott (2016).

La *représentativité* et l'*équilibre* sont maintenant des termes fréquemment utilisés en ce qui a trait à l'ensemble des unités répondantes. Des indicateurs de représentativité, dont l'indicateur R fondé sur les probabilités de réponse estimées, ont été élaborés; voir Schouten, Cobben et Bethlehem (2009), Bethlehem, Cobben et Schouten (2011), Schouten, Shlomo et Skinner (2011), et Bianchi, Shlomo, Schouten, da Silva et Skinner (2016).

L'équilibrage est une procédure utilisée en matière de collecte des données pour obtenir au bout du compte un ensemble représentatif de répondants. La clé consiste à faire en sorte que les moyennes des variables auxiliaires pour les répondants soient proches des moyennes correspondantes calculables pour l'échantillon probabiliste, ou connues pour la population. C'est ce qui sous-tend la statistique de *déséquilibre de la réponse* IMB (de l'anglais *imbalance*), utilisée dans Särndal (2011a) et Lundquist et Särndal (2013). Il existe des méthodes de réduction d'IMB dans la collecte de données adaptative; voir par exemple Särndal et Lundquist (2014).

On utilise souvent l'expression « ajustement pour la non-réponse » afin d'évoquer un estimateur idéal qui serait satisfaisant en cas de réponse complète, mais qui devrait être « réparé » pour être un substitut crédible en présence de non-réponse. L'ajustement sert à limiter l'inconvénient, à savoir le biais de non-réponse. La pratique actuelle des organismes statistiques consiste à utiliser des variables auxiliaires dans l'estimation, en calculant les poids d'ajustement de la non-réponse calés. Il est possible d'obtenir une variance réduite et un biais de non-réponse réduit. Plusieurs travaux récents, dont des articles de synthèse, traitent des procédures de pondération de la phase d'estimation, par exemple, Brick (2013), Fattorini, Franceschi et Maffei (2013), Haziza et Lesage (2016), Little et Vartivarian (2005), Tourangeau, Brick, Lohr et Li (2017). Särndal et Lundström (2005, 2008, 2010) et Särndal (2011b) examinent le choix des variables auxiliaires dans les procédures de pondération, mais ce n'est pas l'objet du présent article.

Il est évident qu'un meilleur équilibre de la réponse pourrait considérablement améliorer l'exactitude si l'estimateur utilisé est rudimentaire, comme la moyenne de réponse élargie par la taille de la population. Le point de départ est en fait que les variables auxiliaires sont très utilisées à l'étape d'estimation, et que le fait de les utiliser aussi dans la collecte de données adaptative peut être une caractéristique apportant un avantage supplémentaire. Ce rôle combiné des variables auxiliaires est potentiellement important.

Des travaux ont soulevé la question de savoir si l'équilibrage de la réponse réduira significativement le biais de non-réponse, par exemple dans Schouten, Cobben, Lundquist et Wagner (2014), Lundquist et Särndal (2013), Särndal et Lundquist (2014), Särndal, Lumiste et Traat (2016). Ces études constatent que l'équilibrage apporte une amélioration de l'exactitude, bien qu'elle ne soit pas très prononcée. Les modestes gains obtenus n'étaient ni solides ni convaincants. Quant aux preuves, elles sont pour la plupart empiriques. Il est nécessaire de mieux comprendre les raisons théoriques du « succès limité » d'une collecte de données adaptative. L'article apporte des éclaircissements pour tenter de répondre à la question : l'utilisation de variables auxiliaires dans une collecte de données adaptative peut-elle contribuer à une plus grande amélioration de l'estimation, sachant que ces variables sont de toute façon utilisées dans la pondération calée à l'étape de l'estimation ? La question importe pour les recherches sur les plans de sondage adaptatifs. Ces plans doivent constituer une promesse formelle d'amélioration des estimations. En effet, si l'amélioration est nulle ou faible, dans des conditions assez générales, il semble en partie moins motivant de choisir un plan adaptatif. Nous étudions les raisons théoriques de gain d'exactitude apporté par un meilleur équilibre de la réponse et, plus précisément, la raison pour laquelle on peut attendre « un avantage marginal » d'un plan adaptatif, c'est-à-dire une amélioration supplémentaire des estimations par l'utilisation de variables auxiliaires également à l'étape de la collecte de données. Il est tout à fait légitime de réutiliser les variables auxiliaires à l'étape de l'estimation.

L'article est structuré de la manière suivante. La notation est présentée (section 2) pour trois sous-ensembles de population importants, soit l'échantillon probabiliste, l'ensemble de réponses et son complément l'ensemble de non-réponses. Le rôle du vecteur auxiliaire (vecteur $\mathbf{x}$) est mis en évidence, en particulier dans la statistique de déséquilibre de réponse, notée IMB. Pour estimer le total de population y, nous considérons l'estimateur basé sur des poids calés sur le vecteur auxiliaire. Son écart par rapport à l'estimateur sans biais, qui est hypothétique parce qu'il exige une réponse complète, est un indicateur de biais que nous examinons de manière approfondie.

Nous décomposons l'écart de l'estimateur par calage (section 3) en deux termes, le *terme réductible* et le *terme résistant*. Le premier peut être réduit par un plan de sondage adaptatif; il est en fait réductible jusqu'à zéro si un équilibre parfait est atteint. En revanche, la collecte de données adaptative a peu d'effets sur le terme résistant, ce qui laisse penser qu'au mieux, le plan adaptatif est un remède partiel pour éliminer le biais de non-réponse. Le cas particulier où le vecteur $\mathbf{x}$ encode un ensemble de groupes d'échantillons exhaustifs et mutuellement exhaustifs est particulièrement important (section 4). Par sa plus grande résolubilité mathématique, il permet de mieux comprendre les deux termes de la décomposition. La preuve empirique utilisant les données de l'Enquête sur la population active en Suède (section 5) illustre et confirme

les conclusions théoriques dégagées à propos des deux termes. L'article se termine par des remarques finales (section 6). Les démonstrations sont présentées en annexe.

2 Échantillonnage probabiliste suivi par une non-réponse

2.1 Population, échantillon, ensemble de réponses, ensemble de non-réponses

Soit $U = \{1, 2, \ldots, k, \ldots, N\}$ une population finie de taille N dont a été tiré un échantillon probabiliste s, en donnant à l'unité k la probabilité d'inclusion π_k et le poids d'échantillonnage $d_k = 1/\pi_k$. Soit r l'ensemble de réponses, c'est-à-dire l'ensemble d'unités pour lesquelles la valeur y_k de la variable d'enquête y, qu'elle soit catégorique ou continue, est observée. Nous avons donc les observations y_k pour $k \in r \subset s \subset U$, mais elles sont manquantes pour l'ensemble de non-réponses noté $nr = s - r$. Au moyen des valeurs y_k observées, nous voulons estimer le total de population y $Y = \sum_U y_k$. Une sommation $\sum_{k \in A}$ sur un ensemble d'unités $k \in A \subseteq U$ est écrite simplement comme suit : $\sum_A$.

Il est établi que « les non-répondants ne sont pas comme les répondants » et qu'aucun des deux groupes n'est suffisamment semblable à l'échantillon complet. Cela cause un biais dans les estimations du total de Y. Le présent article s'intéresse au contraste entre l'ensemble de réponses et l'ensemble de non-réponses, car cette comparaison a porté ses fruits auparavant, notamment dans des concepts comme celui de fraction d'information manquante.

La collecte de données adaptative est un processus dynamique. Les ensembles de réponses et de non-réponses, ainsi que les quantités connexes comme les moyennes et les coefficients de régression, sont définis temporairement. Une notation plus complète distinguerait les ensembles de réponses $r^{(a)}$, $a = 1, 2, \ldots$, dans l'ordre hiérarchique,

$$r^{(1)} \subseteq r^{(2)} \subseteq \ldots \subseteq r^{(a)} \subseteq \ldots$$

Ici, $r^{(a)}$ est l'ensemble d'unités ayant donné la valeur y_k après a tentatives d'appel (ou, sinon, après a jours de collecte de données), et $nr^{(a)} = s - r^{(a)}$ est l'ensemble de non-réponses correspondant. Pour simplifier, r désignera tout ensemble des ensembles de réponses de plus en plus grands. Des outils de collecte de données, comme le système WinDATI de Statistique Suède, permettent d'enregistrer toutes les tentatives de contact, puis d'intervenir et de rediriger les entrées de données afin d'obtenir un ensemble de réponses final mieux équilibré.

Le *taux de réponse* et le *taux de non-réponse* correspondant (les deux pondérés par d) sont les suivants :

$$P = \sum_r d_k \Big/ \sum_s d_k \,; \quad Q = 1 - P = \sum_{nr} d_k \Big/ \sum_s d_k \,.$$

De plus, des notations de la réponse, la non-réponse et de l'échantillon complet sont données pour le vecteur auxiliaire $\mathbf{x}$ à la section 2.2, pour la variable d'enquête y et pour le vecteur de régression y sur $\mathbf{x}$ à la section 2.3.

2.2 Vecteur auxiliaire et déséquilibre de la réponse

Un vecteur auxiliaire $\mathbf{x}$ est choisi à la suite d'un choix structuré de variables auxiliaires parmi les variables de ce type disponibles. Dans les pays scandinaves, les variables disponibles sont nombreuses et proviennent de sources administratives ainsi que d'enquêtes auprès des particuliers et des ménages. On pourrait s'étendre sur les principes ou les « tentatives d'optimalité » susceptibles de guider ce choix, mais le présent article ne porte pas sur la procédure de sélection.

Soit le vecteur auxiliaire désigné $\mathbf{x}$ de dimension $J \geq 1$. On suppose sa valeur $\mathbf{x}_k$ connue pour toutes les unités $k \in s$. (Le cas où la population totale de $\mathbf{x}_k$ est connue n'est pas pris en compte.) Dans un cas particulier important, $\mathbf{x}$ est un *vecteur de groupes* de dimension J, sous la forme $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$, c'est-à-dire avec $J - 1$ entrées « 0 » et une seule entrée « 1 » pour identifier le groupe contenant l'unité k. Les groupes ne se chevauchent pas et sont exhaustifs, le vecteur de groupes encode J propriétés des unités. Par exemple, deux catégories *Instruction* croisées avec trois catégories *Revenu* donnent $J = 2 \times 3 = 6$ groupes. Mais les variables catégoriques x utilisées dans le vecteur $\mathbf{x}$ n'ont pas toutes besoin d'être entièrement croisées, et elles ne le sont pas dans plusieurs applications.

Nous utilisons des vecteurs $\mathbf{x}$ avec une propriété pratique du point de vue mathématique dans de nombreux calculs : Il existe un vecteur $\boldsymbol{\mu}$ qui ne dépend pas de k de sorte que :

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ pour tous les } k. \tag{2.1}$$

La majorité des vecteurs d'intérêt $\mathbf{x}$ satisfont à cette condition ou peuvent la vérifier. Par exemple, s'il y a une seule variable auxiliaire continue x, et $\mathbf{x}_k = (1, x_k)'$, alors $\boldsymbol{\mu} = (1, 0)'$ satisfait à la contrainte. Dans le cas du vecteur de groupes, la condition est satisfaite par $\boldsymbol{\mu} = (1, \ldots, 1, \ldots, 1)'$. Dans un exemple où $\mathbf{x}$ est catégorique, mais n'est pas un vecteur de groupes, supposons que deux catégories *Études* sont croisées avec trois catégories *Revenu* et que *Sexe* est ajouté au vecteur $\mathbf{x}_k$ comme variable univariée 0/1, alors la dimension est $J = 6 + 1 = 7$, le nombre de valeurs distinctes $\mathbf{x}_k$ est $6 \times 2 = 12$, et $\boldsymbol{\mu} = (1, 1, 1, 1, 1, 1, 0)'$ satisfait à la contrainte (2.1).

Les moyennes − toutes calculables − du vecteur $\mathbf{x}$ pondéré par (d) sont

$$\overline{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k \Big/ \sum_r d_k; \quad \overline{\mathbf{x}}_{nr} = \sum_{nr} d_k \mathbf{x}_k \Big/ \sum_{nr} d_k; \quad \overline{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k \Big/ \sum_s d_k. \tag{2.2}$$

Nous avons besoin des moments du second ordre; les matrices $J \times J$ calculables sont supposées non singulières :

$$\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_r d_k; \quad \boldsymbol{\Sigma}_{nr} = \sum_{nr} d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_{nr} d_k; \quad \boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_s d_k. \tag{2.3}$$

Le *déséquilibre* (IMB) de la réponse est calculé sur les valeurs du vecteur auxiliaire $\mathbf{x}_k$ connues pour $k \in s$, voir Särndal (2011a). Il s'agit d'une mesure du contraste entre la réponse et l'échantillon complet, ou autrement entre la réponse et la non-réponse. Pour un vecteur donné $\mathbf{x}$ et un échantillon s, le déséquilibre est défini comme suit :

$$\text{IMB} = P^2 \left(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\right).$$

Une notation plus explicite serait $\text{IMB}(r, \mathbf{x} \mid s)$ mais pour simplifier, nous utilisons seulement IMB. Étant donné que $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s = (1-P)(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})$, nous pouvons exprimer IMB comme un contraste entre la réponse et la non-réponse :

$$\text{IMB} = \{P(1-P)\}^2 (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})' \, \boldsymbol{\Sigma}_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr}). \tag{2.4}$$

L'un des objectifs d'un plan de collecte de données adaptatif est de créer un ensemble de réponses final r ayant un IMB faible, pour l'échantillon probabiliste donné s.

La dimension et la composition du vecteur $\mathbf{x}$ sont des déterminants importants de la valeur du déséquilibre IMB. IMB a comme propriété, pour les variables fixes r et s, d'augmenter si on étend un vecteur $\mathbf{x}$ donné en y ajoutant d'autres variables x. L'augmentation reflète l'intuition selon laquelle il est plus difficile de rapprocher un nombre plus élevé de moyennes de variables x. Le vecteur $\mathbf{x}$ trivial, $\mathbf{x}_k = 1$ pour tous les k, donne $\text{IMB} = 0$, mais ne présente presque aucun intérêt en pratique. Nous avons $0 \le \text{IMB} \le P(1-P) \le 0{,}25$ pour tout vecteur s, r, et $\mathbf{x}$. Ce sont des conditions générales. Pour la plupart des ensembles de données d'enquête, la valeur calculée IMB est bien inférieure à la borne supérieure, souvent entre 0,01 et 0,05.

2.3 La variable d'enquête et sa régression sur x

Passons à la variable d'enquête y. Sa valeur y_k est observée pour $k \in r$ de sorte que le modèle de régression linéaire de y sur $\mathbf{x}$ puisse être réalisé pour l'ensemble r. Bien que non faisable en cas d'enquête avec non-réponse, un modèle de régression linéaire de y sur $\mathbf{x}$ existe aussi, théoriquement, pour $nr = s - r$ et pour s. Ajusté à la réponse r, le vecteur de coefficient de régression linéaire (par moindres carrés pondérés par d) est $\mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k \mathbf{x}_k y_k$. De manière analogue, les deux autres modèles de régression donnent les vecteurs de régression $\mathbf{b}_{nr}$ et $\mathbf{b}_s$. Exprimés en fonction des matrices $\mathbf{x}$ de moment du second ordre dans (2.3), les trois vecteurs de régression sont

$$\mathbf{b}_r = \boldsymbol{\Sigma}_r^{-1} \sum_r d_k \mathbf{x}_k y_k \Big/ \sum_r d_k; \;\; \mathbf{b}_{nr} = \boldsymbol{\Sigma}_{nr}^{-1} \sum_{nr} d_k \mathbf{x}_k y_k \Big/ \sum_{nr} d_k; \;\; \mathbf{b}_s = \boldsymbol{\Sigma}_s^{-1} \sum_s d_k \mathbf{x}_k y_k \Big/ \sum_s d_k. \tag{2.5}$$

Les moyennes y sont

$$\bar{y}_r = \sum_r d_k y_k \Big/ \sum_r d_k; \quad \bar{y}_{nr} = \sum_{nr} d_k y_k \Big/ \sum_{nr} d_k; \quad \bar{y}_s = \sum_s d_k y_k \Big/ \sum_s d_k.$$

Les propriétés suivantes sont vérifiées à la suite de (2.1) :

$$\bar{y}_r = \bar{\mathbf{x}}_r' \mathbf{b}_r; \quad \bar{y}_{nr} = \bar{\mathbf{x}}_{nr}' \mathbf{b}_{nr}; \quad \bar{y}_s = \bar{\mathbf{x}}_s' \mathbf{b}_s. \tag{2.6}$$

Un estimateur rudimentaire, en présence de non-réponse, du total de la population $Y = \sum_U y_k = N \bar{y}_U$ utilise l'extension directe (EXP) de la moyenne de la réponse :

$$\hat{Y}_{\text{EXP}} = \hat{N} \sum_r d_k y_k \Big/ \sum_r d_k = \hat{N} \, \bar{y}_r,$$

où $\hat{N} = \sum_s d_k$. Dans $\hat{Y}_{\text{EXP}}$, la pondération est uniforme, sans utilisation d'information auxiliaire. Le biais est susceptible d'être élevé. L'utilisation des valeurs auxiliaires $\mathbf{x}_k$ connues pour $k \in s$ apporte habituellement une amélioration; $\sum_s d_k \mathbf{x}_k$ est un estimateur connu et sans biais (Horvitz-Thompson) du total de la population $\sum_U \mathbf{x}_k$, et nous cherchons donc les poids $d_k w_k$ pour satisfaire l'équation de calage $\sum_r d_k w_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$. Voici une solution (qui n'est pas la seule)

$$w_k = \left(\sum_s d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k = (1/P)\, \overline{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k.$$

L'estimateur par calage linéaire (CAL) de $Y = \sum_U y_k$ est $\hat{Y}_{\text{CAL}} = \sum_r d_k w_k y_k$, ou de façon équivalente

$$\hat{Y}_{\text{CAL}} = \hat{N}\, \overline{\mathbf{x}}_s' \mathbf{b}_r.$$

Une des raisons pour lesquelles nous nous attendons à ce que $\hat{Y}_{\text{CAL}}$ soit moins biaisé que $\hat{Y}_{\text{EXP}}$, surtout quand y et $\mathbf{x}$ sont bien liés, est que les poids $d_k w_k$ doivent respecter la contrainte de quantité calculable non biaisée $\sum_s d_k \mathbf{x}_k$, dans le deuxième membre de l'équation de calage. Cependant, malgré la pondération de l'ajustement, $\hat{Y}_{\text{CAL}}$ a un biais restant non négligeable, peut-être considérable.

Comme repère, nous utilisons l'estimateur sans biais de Y exigeant une réponse complète (FUL), qui est donc hypothétique en présence de non-réponse, à savoir l'estimateur de Horwitz-Thompson

$$Y_{\text{FUL}} = \sum_s d_k y_k = \hat{N}\, \overline{y}_s.$$

Les trois types d'estimateurs sont désignés par EXP, CAL et FUL. En fait, CAL représente une famille d'estimateurs, correspondant à tous les choix possibles de vecteur $\mathbf{x}$. Pour un vecteur $\mathbf{x}$ approprié donné, nous devons examiner de près l'écart de CAL Δ_{CAL}, défini comme étant la différence entre le CAL avec biais et le FUL sans biais, cadrée en divisant par la taille de la population (estimée au besoin) :

$$\left(\hat{Y}_{\text{CAL}} - \hat{Y}_{\text{FUL}}\right) \Big/ \hat{N} = \Delta_{\text{CAL}},$$

où

$$\Delta_{\text{CAL}} = \overline{\mathbf{x}}_s' \mathbf{b}_r - \overline{y}_s = \overline{\mathbf{x}}_s' \left(\mathbf{b}_r - \mathbf{b}_s\right). \tag{2.7}$$

L'écart n'est pas un biais. L'écart Δ_{CAL} est un résultat pour un ensemble de réponses r d'un échantillon particulier s. Par ailleurs, le biais de l'estimateur CAL est la valeur prévue de Δ_{CAL} sur tous les r d'une valeur s, donnée et tous les s de U. Si on peut obtenir une réponse r avec un écart nul Δ_{CAL} quel que soit l'échantillon s, alors l'estimateur CAL a un biais nul, car il est alors toujours égal à l'estimateur de Horvitz-Thompson sans biais. Il n'est cependant pas réaliste de penser qu'on obtiendra en pratique un écart nul par une collecte de données adaptative ou un autre moyen. Il est toutefois judicieux de chercher à réduire l'écart pour l'échantillon donné s, puisque cela permettrait de se rapprocher d'une estimation sans biais. On peut dire que l'objectif ne consiste pas à obtenir une estimation sans biais, parce que cela est impossible en présence de non-réponse, mais qu'il est réaliste de chercher à réduire l'écart par rapport à une estimation sans biais.

3 Analyser l'écart par rapport à une estimation sans biais

3.1 Décomposer l'écart de l'estimateur CAL

Nous décomposons l'écart de CAL, $\Delta_{\text{CAL}} = \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s)$, afin de mieux comprendre son comportement en cas d'amélioration de l'équilibre par une collecte de données adaptative. Nous pouvons prévoir que Δ_{CAL} sera réduit, sinon à zéro au moins dans une certaine mesure, et dans des conditions que nous cherchons à déterminer. Aucun signe apparent immédiat n'indique que $\overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s)$ serait réduit quand $\overline{\mathbf{x}}_r$ se rapprocherait de la valeur fixe $\overline{\mathbf{x}}_s$. Il faut approfondir l'analyse pour expliquer la réduction de Δ_{CAL}, le cas échéant.

Quand le vecteur $\mathbf{x}$ choisi est de dimension relativement élevée, l'équilibre parfait $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$ (et donc $\text{IMB} = 0$) est difficile à atteindre dans une collecte de données, mais on peut s'efforcer de s'en rapprocher. Nous avons néanmoins une idée de la nature de Δ_{CAL} en établissant $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$; au moyen de (2.6) on obtient

$$\Delta_{\text{CAL}} = \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s) = \overline{\mathbf{x}}_r' \mathbf{b}_r - \overline{y}_s = \overline{y}_r - \overline{y}_s = (1 - P)(\overline{y}_r - \overline{y}_{nr}) \neq 0, \qquad (3.1)$$

$$\hat{Y}_{\text{CAL}} - \hat{Y}_{\text{FUL}} = \hat{N}\, \Delta_{\text{CAL}} = \hat{N} (\overline{y}_r - \overline{y}_s) = \hat{Y}_{\text{EXP}} - \hat{Y}_{\text{FUL}}.$$

Par conséquent, l'équilibre parfait, $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$, ne réduit pas Δ_{CAL} à zéro; il donne $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{EXP}} = \hat{N}\, \overline{y}_r$. Cette dernière équation semble d'abord paradoxale, car $\hat{Y}_{\text{CAL}}$ est supposé meilleur que la valeur rudimentaire $\hat{Y}_{\text{EXP}}$. Nous constatons en fait qu'en cas d'équilibre parfait, ils sont égaux. On résout le paradoxe en remarquant que $\overline{y}_r - \overline{y}_s$ sera probablement plus petit, et même considérablement plus petit, quand la réponse r est parfaitement équilibrée sur $\mathbf{x}$ – à cet égard plus proche de s – que quand elle ne l'est pas. Il est faux de dire que l'information auxiliaire n'est pas utilisée. Cette information est précisément ce qu'il faut pour obtenir un équilibre parfait.

Särndal et Lundquist (2014) ont donné une preuve empirique de la relation entre Δ_{CAL} et IMB. Pour les données d'enquête analysées dans le présent article, Δ_{CAL} diminue considérablement quand IMB est réduit, mais ne se rapproche pas de zéro quand IMB tend vers zéro. Ces résultats empiriques laissent penser que Δ_{CAL} se stabilise à une certaine valeur non nulle.

En cas d'équilibre parfait, l'écart (3.1) a la même *expression*, $\Delta_{\text{CAL}} = (1 - P)(\overline{y}_r - \overline{y}_{nr})$, pour tout vecteur $\mathbf{x}$. Cependant, sa *valeur* n'est pas identique, car l'ensemble d'unités dans une réponse parfaitement équilibrée r sera différente d'un vecteur $\mathbf{x}$ au suivant. Pour un vecteur $\mathbf{x}$ donné, l'écart augmente avec le taux de non-réponse $1 - P$ et avec la divergence $\overline{y}_r - \overline{y}_{nr}$ entre les moyennes y de la réponse et de la non-réponse. L'équilibre parfait et par conséquent $\text{IMB} = 0$ se vérifient en particulier pour le vecteur $\mathbf{x}$ trivial, qui ne présente aucun intérêt en pratique, $\mathbf{x}_k = 1$ pour tous les k; dans ce cas, la formule $\Delta_{\text{CAL}} = (1 - P)(\overline{y}_r - \overline{y}_{nr})$ est un rappel d'une note élémentaire mais souvent citée sur l'effet de biais de la non-réponse : Cochran (1977, page 361) donnait déjà une expression analogue pour le biais de la moyenne de l'ensemble de réponses, dans le contexte d'une population modélisée consistant en une strate de réponse (unités répondantes avec une probabilité de un) et une strate de non-réponse (unités répondantes avec une

probabilité de zéro). Le message est : le biais augmente avec le taux de non-réponse et avec la séparation entre les moyennes de la réponse et de la non-réponse.

Nous cherchons à décomposer $\Delta_{\text{CAL}} = \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s)$ en deux composantes de sorte que l'une d'elles soit réductible à zéro en cas d'équilibre parfait $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$. Plusieurs fractionnements peuvent répondre à cette « exigence nulle » pour l'un des deux termes. Särndal et Lundquist (2017) examinent un de ces fractionnements, $\Delta_{\text{CAL}} = \overline{e}_r + u_r$, où $\overline{e}_r = \sum_r d_k e_k \big/ \sum_r d_k = \overline{\mathbf{x}}_r' (\mathbf{b}_r - \mathbf{b}_s)$ est la moyenne, sur la réponse r, des résidus de la régression de y sur $\mathbf{x}$ ajustée à l'échantillon s, $e_k = y_k - \mathbf{x}_k' \mathbf{b}_s$, et u_r est le reste : $\Delta_r - \overline{e}_r = u_r = -(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s)' (\mathbf{b}_r - \mathbf{b}_s)$. Ces auteurs appellent $\overline{e}_r$ *incohérence de la régression*, pour indiquer que le modèle de régression pour l'échantillon complet échoue (ou est incohérent) quand il est appliqué à la réponse : Les résidus e_k ont une moyenne nulle, $\overline{e}_s = 0$, sur l'échantillon s, mais une moyenne non nulle, $\overline{e}_r \neq 0$, sur la réponse r. L'incohérence est un résultat (non surprenant) de « valeurs manquantes non au hasard ».

Nous avons $u_r = 0$ en cas d'équilibre parfait $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$, alors $\Delta_{\text{CAL}} = \overline{e}_r$. Särndal et Lundquist (2017) prouvent empiriquement que quand le déséquilibre IMB est réduit au moyen d'une collecte de données adaptative, Δ_{CAL} et $\overline{e}_r$ sont tous deux réduits, mais aucun n'est réduit à zéro, et que le ratio $\overline{e}_r \big/ \Delta_{\text{CAL}}$ tend lentement vers 1 quand IMB se rapproche de zéro.

3.2 Opposer la réponse et la non-réponse

On comprend mieux en cherchant à décomposer $\Delta_{\text{CAL}} = \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s)$ de sorte que le premier terme demeure à peu près constant quand le déséquilibre est réduit, tandis que le deuxième terme tend vers zéro dans ce processus. Cette décomposition isolerait donc une partie de l'écart non touchée par le plan adaptatif et indiquerait pourquoi le plan adaptatif réussit, au mieux, « partiellement » à éliminer le biais. La décomposition du résultat 3.1 est motivée par la volonté d'opposer l'ensemble de réponses r et l'ensemble de non-réponses $nr = s - r$. La notation est donnée dans (2.2) pour les moyennes $\mathbf{x}$ et dans (2.5) pour les vecteurs de régression.

Résultat 3.1. L'écart du CAL $\Delta_{\text{CAL}} = \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s)$ a la décomposition $\Delta_{\text{CAL}} = D_1 + D_2$ où

$$D_1 = (1 - P)\, \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_{nr}); \quad D_2 = -P(1 - P)(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr})' (\mathbf{b}_r - \mathbf{b}_{nr}).$$

Les expressions suivantes sont équivalentes :

$$\Delta_{\text{CAL}} = (1 - P)\, \overline{\mathbf{x}}_{nr}' (\mathbf{b}_r - \mathbf{b}_{nr}); \quad D_2 = P(\overline{\mathbf{x}}_s - \overline{\mathbf{x}}_r)' (\mathbf{b}_r - \mathbf{b}_{nr}).$$

La partie 1 de l'annexe montre comment les composantes D_1 et D_2 sont calculées à partir de la définition (2.7) de Δ_{CAL}. Le résultat suscite plusieurs commentaires :

1. L'équilibrage de la réponse pendant la collecte des données peut rapprocher $\overline{\mathbf{x}}_r$ de $\overline{\mathbf{x}}_s$ et, par conséquent, réduire IMB et D_2 pour les faire tendre vers zéro. En cas d'équilibre parfait $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$, $D_2 = 0$ mais $D_1 \neq 0$.

2. Pour obtenir $D_1 = 0$, il faut $\mathbf{b}_r = \mathbf{b}_{nr}$. Cela ne se produit pas, y compris en cas d'équilibre parfait. L'échantillon s et sa moyenne $\overline{\mathbf{x}}_s$ sont fixes. L'élimination du terme D_1 semble hors de portée. Mais est-ce qu'une « meilleure composition » de l'ensemble de réponses r – $\overline{\mathbf{x}}_r$ proche de $\overline{\mathbf{x}}_s$ et IMB plus faible – influe favorablement sur la différence $\mathbf{b}_r - \mathbf{b}_{nr}$ et par conséquent sur D_1 ?

Il est difficile d'obtenir une réponse simple à cette question. Intuitivement, le transfert de certaines unités, pendant la collecte des données, de la non-réponse $nr = s - r$ à la réponse r semble indiquer que $\mathbf{b}_r - \mathbf{b}_{nr}$ est peu touché, et que la valeur D_1 reste plutôt constante. Cela est plus explicite quand $\mathbf{x}$ est un vecteur de groupes, c'est-à-dire se présente sous la forme $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$; voir la section 4. Nous appellerons D_2 *le terme réductible* et D_1 *le terme résistant* de l'écart $\Delta_{\text{CAL}} = D_1 + D_2$. Notons que nous considérons ici un choix fixe de vecteur $\mathbf{x}$. Un changement de vecteur aurait nécessairement des effets à la fois sur D_1 et sur D_2.

4 Cas du vecteur de groupes

4.1 Termes de la décomposition

Nous considérons le cas important d'un vecteur de groupes, $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$. En pratique, sa dimension J est souvent considérable, de 30 ou plus, comme quand plusieurs variables catégoriques x sont entièrement croisées. La forme diagonale des matrices nécessitant une inversion dans (2.5) apporte une simplification notable. Pour $j = 1, 2, \ldots, J$, on désigne par s_j le sous-ensemble de l'échantillon s qui se trouve dans le groupe j. Sa proportion de l'échantillon est $W_j = \sum_{s_j} d_k \Big/ \sum_s d_k$. L'ensemble de réponses dans s_j est désigné par r_j; l'ensemble de non-réponses est $nr_j = s_j - r_j$. Le taux de non-réponse du groupe j est $Q_j = \sum_{nr_j} d_k \Big/ \sum_{s_j} d_k = 1 - \sum_{r_j} d_k \Big/ \sum_{s_j} d_k = 1 - P_j$; le taux global est $Q = \sum_{j=1}^{J} W_j Q_j = 1 - P$. Le déséquilibre devient une variance des taux de réponse du groupe, voir par exemple Särndal (2011a),

$$\text{IMB} = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)^2 = \sum_{j=1}^{J} W_j \left(P_j - P\right)^2.$$

Les moyennes y sont $\overline{y}_{r_j} = \sum_{r_j} d_k y_k \Big/ \sum_{r_j} d_k$; $\overline{y}_{nr_j} = \sum_{nr_j} d_k y_k \Big/ \sum_{nr_j} d_k$; $\overline{y}_{s_j} = \sum_{s_j} d_k y_k \Big/ \sum_{s_j} d_k$. Le vecteur-colonne dimensionnel J $\mathbf{b}_r - \mathbf{b}_{nr}$ comporte des éléments $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$, $j = 1, 2, \ldots, J$. Nous appelons δ_j la *divergence du groupe* j, c'est-à-dire entre la moyenne de réponse y et la moyenne de non-réponse y. L'estimateur par calage est $\hat{Y}_{\text{CAL}} = \sum_{j=1}^{J} \hat{N}_j \overline{y}_{r_j}$, où $\hat{N}_j = \sum_{s_j} d_k$, et l'estimateur repère sans biais est $\hat{Y}_{\text{FUL}} = \sum_{j=1}^{J} \hat{N}_j \overline{y}_{s_j}$, donc

$$\Delta_{\text{CAL}} = \left(\hat{Y}_{\text{CAL}} - \hat{Y}_{\text{FUL}}\right) \Big/ \hat{N} = \sum_{j=1}^{J} W_j \left(\overline{y}_{r_j} - \overline{y}_{s_j}\right) = \sum_{j=1}^{J} W_j Q_j \delta_j.$$

L'application directe du résultat 3.1 au cas du vecteur de groupes donne le résultat 4.1 ci-dessous.

Résultat 4.1 : Quand $\mathbf{x}$ est un vecteur de groupes de dimension J, $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$, les termes de la décomposition du résultat 3.1 prennent la forme

$$D_1 = Q\sum_{j=1}^{J} W_j \delta_j; \quad D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)\delta_j; \quad \Delta_{\text{CAL}} = D_1 + D_2 = \sum_{j=1}^{J} W_j Q_j \delta_j,$$

où $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$, et $Q_j = \sum_{nr_j} d_k \Big/ \sum_{s_j} d_k$ et $Q = \sum_{j=1}^{J} W_j Q_j$ sont respectivement le taux de non-réponse du groupe j et le taux global de non-réponse.

Dans quelle mesure une collecte de données adaptative et une réponse mieux équilibrée r peuvent-elles améliorer l'exactitude et réduire l'écart $\Delta_{\text{CAL}} = D_1 + D_2$? Dans ce qui suit, le terme « réduction » désigne une « réduction en valeur absolue », « plus petit » signifie « plus petit en valeur absolue », car Δ_{CAL} et ses composantes peuvent avoir l'un ou l'autre signe.

Dans une collecte de données adaptative agissant sur un vecteur $\mathbf{x}$ qui encode un assez grand nombre de groupes, il est difficile d'arriver à un équilibre parfait où $Q_j = Q$ pour chaque groupe j, et IMB = 0. Cependant, on peut amener tous les Q_j assez près du taux global de non-réponse Q; les deux IMB et D_2 peuvent alors être proches de zéro. Il faut réfléchir à un scénario dans lequel on compare une collecte de données réelle, suivant un protocole prédéfini qui demeure inchangé jusqu'à la fin, à d'autres collectes de données qui s'efforcent activement, par des interventions, d'aboutir à une réponse présentant un déséquilibre faible. Il s'agit du format de la validation empirique (de la section 5), fondée sur une grande enquête de Statistique Suède : On a l'ensemble de réponses réellement enregistré dans l'enquête, et plusieurs autres ensembles de réponses calculés par des interventions expérimentales sur la réponse réelle visant à réduire successivement le déséquilibre.

Nous traiterons des questions suivantes concernant les termes du résultat 4.1 :

a) Lorsque les différentiels de non-réponse $Q_j - Q$ se rapprochent de zéro, de sorte qu'IMB est réduit, le terme $D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)\delta_j$ tend à diminuer, pour parfois se rapprocher de zéro. La perspective est très différente pour $D_1 = Q\sum_{j=1}^{J} W_j \delta_j$. A priori, il n'est pas exclu que les divergences δ_j soient quelque peu réduites au cours du processus. Si cela se produit dans un certain nombre de groupes, alors il est possible que D_1 soit aussi réduit. Mais dans certaines conditions, un changement important de D_1 est peu probable; D_1 devrait être peu touché par un IMB réduit. Nous le justifions d'abord par un argument théorique assisté par un modèle à la section 4.6, puis nous l'observons de façon empirique sur des données d'enquête réelles à la section 5.

b) Il n'est pas évident que D_1 et D_2 soient toujours du même signe. Quand cela est le cas, une réduction atteignable de D_2 entraîne un écart réduit $\Delta_{\text{CAL}} = D_1 + D_2$, et, par conséquent, une plus grande exactitude. Dans quelles conditions ces valeurs ont-elles le même signe ?

(c) Quand D_1 et D_2 ont le même signe et D_1 reste à peu près constant quand IMB est réduit, alors toute réduction de $\Delta_{\text{CAL}} = D_1 + D_2$ provient d'une réduction de D_2; la taille relative des deux termes détermine alors si le gain obtenu au moyen du plan adaptatif est considérable, modeste ou même insignifiant.

4.2 Corrélation à l'intérieur des groupes entre la variable réponse et la variable d'enquête

La variable réponse et la variable d'enquête y sont corrélées, dans l'échantillon complet s, comme dans chaque groupe de l'échantillon s_j. Cette conséquence inévitable de « valeurs manquantes non au hasard » est essentielle dans l'interprétation de D_1, D_2 et leur somme Δ_{CAL}. Dans le groupe j, soit ρ_j le coefficient de la corrélation entre l'indicateur de réponse i ($i_k = 1$ pour $k \in r_j$, $i_k = 0$ pour $k \in nr_j$) et la variable d'enquête y. Ce ρ_j a une relation simple, au moyen d'un facteur de multiplication, à la divergence du groupe $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$:

$$\rho_j = \sqrt{P_j\left(1 - P_j\right)}\ \delta_j \big/ S_{yj}, \tag{4.1}$$

où $S_{yj}^2 = \left(\sum_{s_j} d_k\right)^{-1} \sum_{s_j} d_k \left(y_k - \overline{y}_{s_j}\right)^2$ est la variance y dans le groupe j. L'expression (4.1) qui suit (voir annexe, partie 2) de la définition habituelle du coefficient de corrélation comme étant la « covariance divisée par le produit des deux écarts types », $\rho_j = S_{iyj} \big/ \left(S_{ij} S_{yj}\right)$.

Dans le cas particulier où y est dichotomique $0/1$ (comme quand $y_k = 1$ si k est « employé », $y_k = 0$ autrement), alors $\overline{y}_{s_j} = \sum_{s_j} d_k y_k \big/ \sum_{s_j} d_k$ est une proportion, à savoir la proportion (pondérée par les poids de sondage) de « 1 » dans le groupe j (le taux d'emploi dans le groupe j), de sorte que l'écart dans le groupe y est $S_{yj}^2 = \overline{y}_{s_j}\left(1 - \overline{y}_{s_j}\right)$.

4.3 Analyse du terme résistant

Nous examinerons maintenant de plus près le comportement de D_1 et D_2 pour le cas de vecteur de groupes du résultat 4.1. Leurs tailles (relatives) dépendent de plusieurs facteurs. Le terme résistant $D_1 = Q\sum_{j=1}^{J} W_j \delta_j$ dépend des tailles et de la répartition sur les groupes J des divergences $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$, ou, sous une autre perspective, des corrélations à l'intérieur du groupe $\rho_j = \sqrt{P_j\left(1 - P_j\right)}\,\delta_j \big/ S_{yj}$.

Dans un contexte d'enquête typique, en observant la répartition de δ_j, $j = 1, \ldots, J$, où J peut être de 30 ou plus, on constate généralement un mélange de valeurs positives et négatives. Il peut y avoir des « groupes critiques » avec une divergence inhabituellement grande de δ_j, du même signe, qu'il soit positif ou négatif. Ces δ_j ont une influence sur la moyenne (pondérée) $\overline{\delta} = \sum_{j=1}^{J} W_j \delta_j$ et, par conséquent, sur $D_1 = Q\overline{\delta}$. La majorité des δ_j peuvent être des valeurs assez petites, mais non nulles.

Exemple 4.1. Soit y = *Employé* ($y_k = 1$ si l'unité k est « employé », $y_k = 0$ sinon). On peut s'attendre à une corrélation positive à l'intérieur du groupe ρ_j entre *Employé* et *Réponse* ($i_k = 1$ pour $k \in r_j$, $i_k = 0$ pour $k \in nr_j$) pour les groupes de personnes ayant une ou plusieurs caractéristiques comme de sexe masculin, jeune, faible niveau d'instruction ou d'origine étrangère. Ces groupes ont tendance à avoir des valeurs faibles (comparativement) à la fois sur *Réponse* et sur *Employé*. Autrement dit, le taux d'emploi est plus élevé chez les répondants que chez les non-répondants, donc δ_j est positif. Par ailleurs, les groupes ayant des caractéristiques comme d'âge moyen, niveau d'instruction élevé ou propriétaires de logement ont tendance à avoir des valeurs élevées sur *Réponse* et *Employé*, de sorte que δ_j est susceptible d'être positif pour ces groupes également. Une autre variable y qui peut avoir une tendance semblable est *Revenu* : dans

certains groupes de l'échantillon, *Revenu* peut être plus élevé pour les répondants que pour les non-répondants, donc δ_j est de nouveau positif. Ainsi, pour les deux variables y, plusieurs variables δ_j nettement positives et influentes peuvent rendre $\overline{\delta} = \sum_{j=1}^{J} W_j \delta_j$ et $D_1 = Q\,\overline{\delta}$ positifs. Cela se confirme à la section 5 pour les données de l'Enquête sur la population active en Suède. Cependant, dans d'autres pays, une variable comme *Revenu* peut se comporter différemment : dans certains groupes, le revenu peut plutôt être plus élevé pour les non-répondants que pour les répondants, ce qui rend δ_j négatif, car les personnes aisées ou au revenu élevé auraient tendance à répondre relativement moins. On voit là qu'il est difficile de prévoir le signe des divergences des groupes pour certaines variables d'enquête.

4.4 Analyse du terme réductible

On peut écrire le terme réductible sous la forme $D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)\left(\delta_j - \overline{\delta}\right)$. Il s'agit de la *covariance* entre la non-réponse $Q_j = 1 - P_j$ et la divergence $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$. Ici D_2 est une *covariance entre groupes*, $j = 1, \ldots, J$, alors que la divergence δ_j donnée dans (4.1) est une *covariance à l'intérieur d'un groupe* j entre l'indicateur de réponse i_k et la variable d'enquête y_k. On peut difficilement tirer des conclusions générales sur la taille de D_2 et sur le fait que la valeur ait le même signe, ou non, que D_1. Cela dépend en effet de plusieurs facteurs.

Exemple 4.1 (suite). Comme on l'a montré dans l'exemple 4.1, un certain nombre de divergences positives δ_j sont susceptibles de se produire pour y = *Employé* ($y_k = 1$ si k est *Employé*, = 0 sinon), par exemple, pour les groupes ayant les caractéristiques de sexe masculin, jeune ou de faible niveau d'instruction. Alors, $D_1 = Q\overline{\delta} = Q\sum_{j=1}^{J} W_j \delta_j$ peut être nettement positif. En revanche, on peut moins évidemment savoir si un taux de non-réponse élevé Q_j tend à aller de pair avec une divergence élevée δ_j, de façon à rendre la covariance D_2 également positive. Cela est susceptible de se produire pour des variables comme *Employé* et *Revenu*, comme cela est le cas pour les données suédoises à la section 5.

4.5 Les deux termes ont-ils le même signe ?

Il semble impossible de tirer des conclusions générales sur les signes des valeurs D_1 et D_2. Le nombre de facteurs est en effet trop élevé : la population, le plan de sondage, la variable d'enquête particulière y, et d'autres facteurs. Quand D_1 et D_2 ont le même signe, une réduction de D_2 vers zéro peut donner un écart réduit (voire considérablement réduit) $\Delta_{\mathrm{CAL}} = D_1 + D_2$. En revanche, si les termes sont de signe opposé, ils se contrebalancent; le fait de s'efforcer à obtenir un déséquilibre faible et une valeur D_2 faible ne réduit pas nécessairement $\Delta_{\mathrm{CAL}} = D_1 + D_2$. Dans cette situation non souhaitable et peut-être rare, les tentatives d'équilibrer la réponse – en réduisant D_2 – pourraient empêcher de parvenir à un écart plus faible Δ_{CAL}; la valeur pourrait en effet être plus élevée et non plus petite. Examinons maintenant la « question du même signe ».

L'exemple 4.1 et sa suite, à la section 4.4, décrivaient une situation, pour les variables y *Employé* et *Revenu*, où la moyenne $\overline{\delta} = \sum_{j=1}^{J} W_j \delta_j$ et la covariance $\sum_{j=1}^{J} W_j \left(Q_j - Q\right)\delta_j = D_2$ ont probablement le

même signe, de sorte que $D_1 = Q\bar{\delta}$ et D_2 ont le même signe. Mais de façon plus générale, les signes opposés ne sont pas exclus. Donnons l'exemple d'une non-réponse inhabituellement faible, $Q_j - Q < 0$, qui se produirait dans des groupes influents affichant une divergence nettement positive δ_j. Il peut s'ensuivre que D_2 soit négatif, alors que D_1 serait positif. Les deux termes se contrebalanceraient alors. Il reste toutefois la perspective positive où les deux termes sont de signe identique : Nous pouvons réduire D_2 à de faibles niveaux en rendant tous les taux de non-réponse des groupes presque égaux à Q_j. Ensuite, en supposant que D_1 varie peu, comme dans les conditions proposées ci-dessous à la section 4.6, une réduction considérable de l'écart Δ_{CAL} peut se produire. Voici une synthèse de ces conclusions.

Proposition 4.1. Dans le cas du vecteur de groupes, supposons que la divergence moyenne $\bar{\delta} = \sum_{j=1}^{J} W_j \delta_j$ et la covariance $D_2 = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)\delta_j$ ont un signe identique, disons positif. Alors, les termes $D_1 = Q\bar{\delta}$ et D_2 sont positifs dans l'écart $\Delta_{\text{CAL}} = D_1 + D_2 = \sum_{j=1}^{J} W_j Q_j \delta_j$. Une collecte de données adaptative qui réduit le déséquilibre $\text{IMB} = \sum_{j=1}^{J} W_j \left(Q_j - Q\right)^2$ réduira D_2 (jusqu'à zéro si IMB = 0), alors que le terme résistant D_1 variera probablement peu.

Le choix de vecteur $\mathbf{x}$, à savoir le choix des variables entrant dans le vecteur et sa dimension, peut avoir des effets importants sur D_1 et D_2. Supposons que nous doublons le nombre de groupes codés par le vecteur de groupes, comme quand le nombre de groupes préexistants $J = 8 = 2^3$ (obtenus par croisement de trois variables x dichotomiques) est doublé pour être $J = 16 = 2^4$, par un croisement complet sur une quatrième variable x dichotomique. Une certaine asymétrie dans la répartition des divergences $\delta_j = \bar{y}_{r_j} - \bar{y}_{nr_j}$, $j = 1, \ldots, J$, persistera probablement. Pour une réponse fixe r, on peut réduire D_1 et D_2 en prolongeant le vecteur $\mathbf{x}$. La preuve empirique présentée à la section 5 éclaire en partie cette question.

4.6 Insensibilité des divergences de groupes

Dans le cas du vecteur de groupes du résultat 4.1, le terme résistant est $D_1 = Q\bar{\delta}$, où $\bar{\delta} = \sum_{j=1}^{J} W_j \delta_j$ est une moyenne pondérée des divergences δ_j, $j = 1, \ldots, J$. Nous aimerions en savoir plus sur la façon dont la valeur δ_j évolue dans une collecte de données susceptible d'être étendue, dans une grande enquête, sur une certaine période.

Pendant cette période, on surveille une collecte de données adaptative. On peut évaluer et modifier le protocole et l'accent mis, à un ou plusieurs points d'intervention. Par exemple, on peut, à un point donné, décider de cibler particulièrement les unités dont la propension à répondre calculée est perçue comme faible, afin d'accroître leur propension à répondre pour atteindre une propension moyenne globale. Certaines unités qui étaient non répondantes jusqu'à maintenant sont converties en unités répondantes; leur *statut de réponse* change.

Dans le cas du vecteur de groupes, le groupe de l'échantillon s_j consiste, à un moment donné de la période de collecte des données, en un certain ensemble de réponses et à l'ensemble de non-réponses correspondant. Les tentatives de contact se poursuivent, auprès des unités n'ayant pas encore répondu. Une variation du statut de réponse se produit pour certaines unités dans s_j. Quelques-unes de plus deviennent

des répondants, ce qui entraîne une variation de la divergence δ_j. Cependant, dans certaines conditions, il se peut que δ_j soit peu touché. Si cela devait se vérifier pour la plupart des groupes ou tous les groupes, D_1 varierait peu. Examinons cette possibilité.

On ne peut pas observer seul le fait que δ_j varie peu dans une enquête réelle, parce que y_k est une valeur manquante pour la non-réponse, mais cela peut être plausible au moyen d'un modèle pour la variation du statut de réponse. Nous considérons des modèles dans lesquels les unités d'un seul et même groupe sont *échangeables* ou *substituables* les unes aux autres, par rapport au statut de réponse : à un moment donné de la collecte des données, toutes les unités sont considérées comme identiques en ce qui concerne la variation du statut de réponse; le passage de la non-réponse à la réponse, ou vice versa, est tout aussi probable pour toutes les unités.

On peut justifier un tel modèle en cas de répartition suffisamment détaillée de l'échantillon s en petits groupes homogènes s_j, de sorte que les unités d'un seul et même groupe aient des caractéristiques importantes communes. Le modèle serait en revanche irréaliste dans un grand groupe hétérogène avec un mélange d'unités très différentes en ce qui concerne l'âge, le sexe, le revenu, le niveau d'instruction et d'autres aspects importants. Dans un groupe hétérogène, la probabilité de conversion est très grande pour certaines unités, alors que d'autres sont peu susceptibles de se convertir. Par exemple, si un groupe contient à la fois des personnes plus âgées au niveau d'instruction élevé et des personnes plus jeunes et moins instruites, la probabilité de conversion peut être élevée pour le premier sous-groupe et considérablement plus faible pour le deuxième.

À un certain point de la collecte des données, nous examinerons deux modèles pour la variation du statut de réponse dans un groupe de l'échantillon s_j. Par simplicité, nous supposons que s est un échantillon aléatoire simple de U, de taille N, de sorte que les poids d'échantillonnage sont constants, $d_k = N/n$ pour tous les k.

Modèle 4.1. Supposons qu'un transfert ait lieu, dans le groupe s_j, de l'ensemble de non-réponses $nr_j = s_j - r_j$ à l'ensemble de réponses r_j (une conversion) de sorte que tous les ensembles de transfert $tr_j \subset nr_j$ de taille fixe q_j ont une probabilité égale de se produire.

Dans ce modèle, nr_j et r_j sont des ensembles fixes de taille respective $n_j - m_j$ et m_j, tandis que tr_j est une sélection aléatoire simple de q_j non-répondants. Tout ensemble de transfert particulier tr_j de q_j unités peut être converti, et avec une probabilité égale. Chaque unité de l'ensemble de non-réponse nr_j a la même probabilité de conversion, $q_j/(n_j - m_j)$.

Avant le transfert, la divergence dans le groupe j est $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$. Après le transfert, la nouvelle divergence, indiquée par une étoile * dans les indices, est $\delta_{*j} = \overline{y}_{*r_j} - \overline{y}_{*nr_j}$. Sa valeur prévue dans le modèle 1 satisfait

$$P_{*j}E(\delta_{*j}) = P_j\delta_j, \tag{4.2}$$

où le taux de réponse du nouveau groupe est $P_{*j} = (m_j + q_j)/n_j > m_j/n_j = P_j$. La démonstration est fournie dans la partie 3 de l'annexe. Si la variation du taux de réponse du groupe est petite, comme quand

q_j est petit par rapport à n_j, alors $P_{*j} \approx P_j$, et il se peut que le transfert fasse peu varier la divergence. Sa nouvelle valeur δ_{*j} est légèrement inférieure, en espérance, à la valeur avant transfert δ_j.

Dans la section 5 empirique, nous essayons le modèle avec des ensembles de réponses qui deviennent plus petits plutôt que plus grands. Le modèle 4.2 qui traite d'un transfert dans ce sens aboutit à une conclusion semblable.

Modèle 4.2. Supposons qu'un transfert ait lieu à l'intérieur du groupe j de l'ensemble de réponses r_j à l'ensemble de non-réponses $nr_j = s_j - r_j$ de telle façon que tous les ensembles de transfert $tr_j \subset r_j$ de taille fixe q_j aient une probabilité égale de se produire.

Un calcul analogue à celui qui a donné (4.2) montre que

$$Q_{*j}\, E\left(\delta_{*j}\right) = Q_j \delta_j. \tag{4.3}$$

Le nouveau taux de non-réponse (plus élevé) est $Q_{*j} = 1 - \left(m_j - q_j\right)/n_j > 1 - m_j/n_j = Q_j$. La nouvelle divergence δ_{*j} est légèrement inférieure en valeur espérée, quand q_j est petit par rapport à n_j. La démonstration est analogue à celle de (4.2).

Les résultats (4.2) et (4.3) indiquent que si peu importe quelles sont les unités du groupe j dont le statut de réponse varie, alors la divergence δ_j demeure à peu près identique, en espérance. Cependant, l'hypothèse des modèles 4.1 et 4.2 d'ensembles de transferts de probabilité égale est difficile ou impossible à corroborer en pratique. Il serait difficile de spécifier un vecteur $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ tel que les groupes codés par le vecteur obéissent aux modèles. Nous ne pouvons pas affirmer la « vérité » de ces modèles, mais seulement suggérer qu'ils peuvent être plausibles pour un groupe suffisamment diversifié de l'échantillon.

5 Preuve empirique

Les travaux empiriques présentés dans cette section illustrent une partie de la théorie avancée dans les sections précédentes. Nous avons utilisé des données d'enquête du cycle de 2012 de l'Enquête sur la population active (EPA) de Statistique Suède. Les résultats sont présentés dans les tableaux 5.1 à 5.4.

Pour créer l'ensemble de données de l'EPA de 2012, nous avons combiné les 12 échantillons de la première vague de 2012. La taille de la première vague mensuelle compte environ 2 650 unités (personnes). L'ensemble de données de l'EPA de 2012 est traité comme un échantillon aléatoire simple s de taille $n =$ 32 265. La réponse ou la non-réponse dans la collecte des données réelle est enregistrée et disponible pour toutes les unités. Dans la collecte de données réelle, le taux de réponse était de 70,6 %. La description détaillée des données et la construction des ensembles de réponses expérimentaux des tableaux sont décrites plus en détail dans Särndal et Lundquist (2017).

Dans l'analyse de l'ensemble de données de l'EPA de 2012, nous avons utilisé différents vecteurs $\mathbf{x}$ obtenus par croisement de variables x binaires : *Instruc*, égale à 1 en cas de personne à niveau d'instruction élevée et à 0 sinon; *Propriétaire*, égale à 1 pour une personne propriétaire de son logement et à 0 sinon; *Origine*, égale à 1 pour une personne née en Suède et à 0 sinon; *État civil*, égale à 1 pour une personne

mariée ou veuve et à 0 sinon; *Sexe*, égale à 1 pour une personne de sexe masculin et à 0 sinon. Les résultats sont présentés ici pour deux vecteurs $\mathbf{x}$.

Le vecteur $\mathbf{x}$ dans les tableaux 5.1 et 5.3 représente le croisement des trois premières variables binaires : $\mathbf{x} = \mathbf{x}_1 =$ (*Instruc* × *Propriétaire* × *Origine*); sa dimension, égale au nombre de groupes, est $J = 2^3 = 8$. On a obtenu le vecteur $\mathbf{x}$ dans les tableaux 5.2 et 5.4 en croisant également la valeur binaire *État civil* : $\mathbf{x} = \mathbf{x}_2 =$ (*Instruc* × *Propriétaire* × *Origine* × *État civil*), de dimension $J = 2^4 = 16$.

Dans cette étude expérimentale, nous avons utilisé deux variables y, *Employé* et *Revenu*. Ce sont deux variables de registre; avec les valeurs y_k disponibles pour toutes les unités $k \in s$. Elles sont donc des « pseudo-variables y » plutôt que des variables y d'enquête réelles. Le fait de connaître y_k pour $k \in s$ nous permet de calculer les coefficients de régression et les moyennes y à la fois pour la réponse et la non-réponse. La variable y est *Employé* dans le tableau 5.1 (avec $\mathbf{x} = \mathbf{x}_1$) et dans le tableau 5.2 (avec $\mathbf{x} = \mathbf{x}_2$). *Employé* est binaire, avec une valeur $y_k = 1$ si k est une personne employée, et zéro sinon. La variable y est *Revenu* dans le tableau 5.3 (avec $\mathbf{x} = \mathbf{x}_1$) et dans le tableau 5.4 (avec $\mathbf{x} = \mathbf{x}_2$). *Revenu* est une variable continue de registre, disponible dans le registre fiscal suédois. Nous avons normalisé *Revenu* pour qu'il y ait une variance moyenne et unitaire nulle sur s. Si *Revenu* a une variabilité et une asymétrie considérables, les résultats sont quelque peu instables. Une valeur y_k peut avoir une incidence considérable dans un petit groupe, par rapport à la plus grande stabilité de la variable y 0/1 *Employé*.

Les quatre lignes des tableaux 5.1 à 5.4 renvoient à une série de quatre ensembles de réponses différents. Leur caractéristique importante est qu'ils sont, par construction, des ensembles avec des IMB de plus en plus faibles. La première ligne, *Réel*, est l'ensemble de réponses consigné dans la collecte des données de l'Enquête sur la population active de 2012 pour les 32 265 unités (personnes). Les trois derniers ensembles de réponses, *A65*, *A63*, *A60*, tirés de Särndal et Lundquist (2017), sont construits à partir de *Réel* au moyen de la méthode du seuil de façon à obtenir une réduction successive du déséquilibre IMB.

Ainsi, on a créé l'ensemble de réponses *A65* à partir de *Réel* en diminuant, à chacun des points d'intervention, les unités répondantes de *Réel* dont la propension à répondre calculée dépasse le seuil de 0,65. Cela tend à aplanir les différences de propension à répondre, de sorte que cette construction réduit IMB, et que le taux de réponse global P diminue quelque peu. On a obtenu les ensembles de réponses désignés par *A63* et *A60* en fixant leur seuil respectif à 0,63 et 0,60; encore une fois, IMB et P sont réduits.

Les colonnes du tableau montrent : le taux de réponse P, le déséquilibre IMB (multiplié par 10^2), les composantes D_1 et D_2 de l'écart $\Delta_{\text{CAL}} = D_1 + D_2$ (toutes les trois multipliées par 10^2), la proportion D_2 de Δ_{CAL}, $\text{Prop}D_2 = 100 \times D_2/\Delta_{\text{CAL}}$, et enfin le rapport de taille D_1, $\text{Rel}\,D_1 = D_1/$ la moyenne(D_1), où la moyenne (D_1) est la moyenne arithmétique des quatre valeurs de D_1. du tableau. Nous utilisons $\text{Rel}\,D_1$ pour voir si sa valeur est proche de « un » pour toutes les lignes, conformément à la théorie selon laquelle D_1 est peu touché par la réduction du déséquilibre IMB.

Les résultats des tableaux 5.1 à 5.4 donnent lieu aux observations suivantes. La deuxième et la troisième sont particulièrement intéressantes, car elles confirment ce que la théorie des sections précédentes suggère, à savoir que quand IMB est réduit, $\text{Prop}D_2$ diminue assez nettement, tandis que $\text{Rel}\,D_1$ reste très proche de 1.

1. Dans les quatre tableaux, D_1 et D_2 ont le même signe. Les deux sont positifs, et le déséquilibre IMB réduit (de la première à la quatrième ligne) entraîne une réduction de $\Delta_{\text{CAL}} = D_1 + D_2$, presque entièrement attribuable à la diminution de D_2.
2. Dans chaque tableau, les rapports $\text{Rel}\, D_1$ ne sont pas éloignés de 1. La valeur de D_1 est ainsi remarquablement constante dans les quatre lignes (ensembles de réponses), et par conséquent insensible à la réduction d'IMB.
3. Dans chaque tableau, les valeurs D_2 et $\text{Prop}D_2$ diminuent sur les quatre lignes, comme le laissait prévoir la théorie. En fait, $\text{Prop}D_2$ tend vers zéro avec IMB. L'effet de la variable y est important; $\text{Prop}D_2$ est beaucoup plus grand pour *Revenu* que pour *Employé*.
4. La variation du vecteur $\mathbf{x}$ a une influence importante sur D_1, pour les deux variables y. Le passage de $\mathbf{x} = \mathbf{x}_1$, qui est plus petit, (tableaux 5.1 et 5.3) à $\mathbf{x} = \mathbf{x}_2$, plus vaste, (tableaux 5.2 et 5.4) entraîne une réduction considérable de D_1, tandis que D_2 varie très peu.

Nous avons examiné la répartition des divergences du groupe J, soit δ_j, $j = 1, \ldots, J$, pour les vecteurs $\mathbf{x}_1$ (avec $J = 8$), $\mathbf{x}_2$ (avec $J = 16$) et $\mathbf{x}_3$ = (*Instruc* × *Propriétaire* × *Origine* × *État civil* × *Sexe*) (avec $J = 32$). Pour *Employé* et *Revenu*, et pour les quatre ensembles de réponses, on constate sans surprise quelques grandes valeurs δ_j positives et une certaine asymétrie dans la répartition. Pour les deux variables, $\bar{\delta} = \sum_{j=1}^{J} W_j \delta_j$ est nettement positif. Cela signifie qu'en moyenne dans les groupes, les moyennes y sont, pour ces données, plus élevées pour les répondants que pour les non-répondants. Il s'agit d'une caractéristique de ces variables y particulières.

Pour $\mathbf{x}_3$, le graphique des 32 divergences δ_j par rapport au différentiel de non-réponse $Q_j - Q$ montre une majorité de points près de zéro sur les deux axes, et des valeurs dispersées dans les quatre quadrants du graphique. Le graphique indique une corrélation positive, mais peu prononcée, entre δ_j et $Q_j - Q$, ce qui rend le terme de covariance D_2 positif.

Tableau 5.1
Variable d'enquête y = Employé; vecteur x : (*Instruc* × *Propriétaire* × *Origine*). Lignes : quatre ensembles de réponses. Colonnes : taux de réponse P, déséquilibre IMB (multiplié par 10^2), composantes D_1 et D_2 de $\Delta_{\text{CAL}} = D_1 + D_2$ (toutes trois multipliées par 10^2), Prop D_2 et Rel D_1

Ensemble rép.	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Réel	0,706	0,608	0,558	0,151	0,709	21,3	0,96
A65	0,659	0,135	0,586	0,098	0,684	14,2	1,01
A63	0,648	0,113	0,596	0,086	0,682	12,6	1,03
A60	0,625	0,062	0,579	0,058	0,637	9,3	1,00

Tableau 5.2
Variable d'enquête y = Employé; vecteur x : (*Instruc* × *Propriétaire* × *Origine* × *État civil*). Lignes : quatre ensembles de réponses. Colonnes : taux de réponse P, déséquilibre IMB (multiplié par 10^2), composantes D_1 et D_2 de $\Delta_{CAL} = D_1 + D_2$ (toutes trois multipliées par 10^2), Prop D_2 et Rel D_1

Ensemble rép.	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Réel	0,706	0,672	0,459	0,153	0,612	25,0	0,92
A65	0,659	0,165	0,515	0,101	0,616	16,4	1,03
A63	0,648	0,142	0,524	0,083	0,607	13,7	1,05
A60	0,625	0,088	0,493	0,067	0,560	12,0	0,99

Tableau 5.3
Variable d'enquête y = Revenu; vecteur x : (*Instruc* × *Propriétaire* × *Origine*). Lignes : quatre ensembles de réponses. Colonnes : taux de réponse P, déséquilibre IMB (multiplié par 10^2), composantes D_1 et D_2 de $\Delta_{CAL} = D_1 + D_2$ (toutes trois multipliées par 10^2), Prop D_2 et Rel D_1

Ensemble rép.	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Réel	0,706	0,608	0,668	0,648	1,316	49,2	1,26
A65	0,659	0,135	0,479	0,261	0,740	35,3	0,90
A63	0,648	0,113	0,449	0,250	0,699	35,8	0,84
A60	0,625	0,062	0,530	0,169	0,699	24,2	1,00

Tableau 5.4
Variable d'enquête y = Revenu; vecteur x : (*Instruc* × *Propriétaire* × *Origine* × *État civil*). Lignes : quatre ensembles de réponses. Colonnes : taux de réponse P, déséquilibre IMB (multiplié par 10^2), composantes D_1 et D_2 de $\Delta_{CAL} = D_1 + D_2$ (toutes trois multipliées par 10^2), Prop D_2 et Rel D_1

Ensemble rép.	P	IMB	D_1	D_2	Δ_{CAL}	Prop D_2	Rel D_1
Réel	0,706	0,672	0,324	0,639	0,963	66,4	0,98
A65	0,659	0,165	0,327	0,247	0,574	43,0	0,99
A63	0,648	0,142	0,313	0,232	0,545	42,6	0,95
A60	0,625	0,088	0,355	0,166	0,521	31,9	1,08

6 Remarques finales

Le présent article repose sur la question suivante : si l'ajustement de la pondération calé à l'étape de l'estimation élimine *partiellement* le biais de non-réponse dans les estimations, pourquoi le fait d'utiliser des variables auxiliaires également dans la collecte de données adaptative qui précède ne peut-il pas éliminer *le reste* du biais ? Les motifs portant à le croire seraient qu'après une collecte de données adaptative, on peut obtenir un ensemble final de répondants qui, à bien des égards, est une copie de l'échantillon probabiliste (mais plein de non-réponses) sélectionné et qu'il ne devrait par conséquent pas rester de biais appréciable. Nous avons examiné l'estimateur de la pondération calée et son écart Δ_{CAL} par rapport à l'estimateur sans biais exigeant une réponse complète. L'examen reste théorique, car dans une enquête réelle en présence de non-réponse, l'estimateur sans biais (de Horvitz-Thompson) n'est pas disponible.

En général, les répondants diffèrent systématiquement des non-répondants. En gardant cette différence à l'esprit, nous avons pu écrire Δ_{CAL} comme étant la somme d'un *terme résistant* D_1 et d'un *terme réductible* D_2. En cas d'échantillon divisé en sous-groupes, le terme réductible D_2 est déterminé par la covariance (sur les groupes) entre le taux de non-réponse de groupe et la corrélation à l'intérieur du groupe entre *Réponse* et variable y. On peut alors réduire D_2 à zéro si les taux de non-réponse de tous les groupes peuvent être égaux dans une collecte de données adaptative. Cependant, la collecte de données adaptative n'élimine pas le terme résistant D_1. Cela est en quelque sorte un message qui donne à réfléchir : l'écart par rapport à l'estimation sans biais n'est pas éliminé. Il reste que la collecte de données adaptative peut promettre un meilleur point de départ pour la phase d'estimation qui suit la fin de la collecte des données.

Annexe

Partie 1. Calcul de la décomposition $\Delta_{\text{CAL}} = D_1 + D_2$ dans le résultat 3.1. Par définition $\Delta_{\text{CAL}} = \overline{\mathbf{x}}_s' (\mathbf{b}_r - \mathbf{b}_s) = \mathbf{x}_s' \mathbf{b}_r - \overline{y}_s$ par l'utilisation de (2.6). Substituons $\overline{\mathbf{x}}_s = P\overline{\mathbf{x}}_r + (1-P)\overline{\mathbf{x}}_{nr}$ et $\overline{y}_s = P\overline{y}_r + (1-P)\overline{y}_{nr} = P\overline{\mathbf{x}}_r'\mathbf{b}_r + (1-P)\overline{\mathbf{x}}_{nr}'\mathbf{b}_{nr}$. Cela donne $\Delta_{\text{CAL}} = (1-P)\overline{\mathbf{x}}_{nr}'(\mathbf{b}_r - \mathbf{b}_{nr})$. Enfin, substituons $\overline{\mathbf{x}}_{nr} = \overline{\mathbf{x}}_s - P(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr})$ pour aboutir aux termes D_1 et D_2 du résultat 3.1. Le fait que les deux expressions de D_2 sont équivalentes provient de $(1-P)(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{nr}) = \overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s$.

Partie 2. Calcul du coefficient de corrélation à l'intérieur du groupe (4.1) entre l'indicateur de réponse i et la variable d'enquête y. Par définition, la corrélation est $\rho_j = S_{iyj} / S_{ij} S_{yj}$, où la covariance est $S_{iyj} = \left(\sum_{s_j} d_k\right)^{-1} \sum_{s_j} d_k \left(i_k - \overline{i}_{s_j}\right)\left(y_k - \overline{y}_{s_j}\right)$ avec $\overline{i}_{s_j} = \sum_{s_j} d_k i_k \Big/ \sum_{s_j} d_k = P_j$. Un développement donne $S_{iyj} = P_j(1-P_j)\delta_j$ avec $\delta_j = \overline{y}_{r_j} - \overline{y}_{nr_j}$. La variance y est $S_{yj}^2 = \left(\sum_{s_j} d_k\right)^{-1} \sum_{s_j} d_k \left(y_k - \overline{y}_{s_j}\right)^2$, et S_{ij}^2 est analogue avec $\left(i_k - \overline{i}_{s_j}\right)^2$ remplaçant $\left(y_k - \overline{y}_{s_j}\right)^2$, de sorte que $S_{ij}^2 = P_j(1-P_j)$. Le résultat $\rho_j = \sqrt{P_j(1-P_j)}\,\delta_j / S_{yj}$ suit.

Partie 3. Démonstration de (4.2) : Selon le modèle 4.1, r_j et nr_j sont des ensembles fixes, de taille respective m_j et $n_j - m_j$ et de moyennes fixes $\overline{y}_{r_j}$ et $\overline{y}_{nr_j}$. L'ensemble de transfert tr_j de taille fixe q_j est aléatoire, retiré par échantillonnage aléatoire simple à partir de la non-réponse $nr_j = s_j - r_j$ et transféré à la réponse r_j. Les nouvelles moyennes y, pour la réponse et la non-réponse, sont

$$\overline{y}_{*r_j} = \left(\sum_{r_j} y_k + \sum_{tr_j} y_k\right) \Big/ (m_j + q_j); \quad \overline{y}_{*nr_j} = \left(\sum_{nr_j} y_k - \sum_{tr_j} y_k\right) \Big/ (n_j - m_j - q_j).$$

Parce que tr_j est un échantillon aléatoire simple provenant de l'ensemble fixe nr_j, la moyenne de l'ensemble de transfert $\overline{y}_{tr_j} = \sum_{tr_j} y_k / q_j$ a une valeur prévue de $\overline{y}_{nr_j}$. Les valeurs prévues des nouvelles moyennes y sont alors les valeurs suivantes :

$$E\left(\overline{y}_{*r_j}\right) = \left(m_j \overline{y}_{r_j} + q_j \overline{y}_{nr_j}\right) \Big/ (m_j + q_j); \quad E\left(\overline{y}_{*nr_j}\right) = \left((n_j - m_j)\overline{y}_{nr_j} - q_j \overline{y}_{nr_j}\right) \Big/ (n_j - m_j - q_j) = \overline{y}_{nr_j}.$$

L'expression (4.2) pour $E\left(\delta_{*j}\right) = E\left(\overline{y}_{*r_j}\right) - E\left(\overline{y}_{*nr_j}\right)$ suit.

Bibliographie

Andridge, R.R., et Little, R.J.A. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.

Beaumont, J.-F., Haziza, D. et Bocci, C. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

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

Bianchi, A., Shlomo, N., Schouten, B., da Silva, D. et Skinner, C. (2016). Estimation of response propensities and R-indicators using population-level information. Document de discussion 2016/21, CBS, Voorburg, Pays-Bas.

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

Cochran, W.G. (1977). *Sampling Techniques, Third edition*. New York: John Wiley & Sons, Inc.

Fattorini, L., Franceschi, S. et Maffei, D. (2013). Design-based treatment of unit nonresponse in environmental surveys. *Biometrical Journal*, 55, 925-943.

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.

Haziza, D., et Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32, 129-145.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M. et Raghunathan, T.E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response; examples from multiple surveys. *Journal of the Royal Statistical Society A*, 173, 389-407.

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 https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9046-fra.pdf.

Lohr, S.L., Riddles, M.K. et Morganstein, D. (2016). Tests pour évaluer le biais de non-réponse dans les enquêtes. *Techniques d'enquête*, 42, 2, 207-232. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016002/article/14677-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.

Nishimura, R., Wagner, J. et Elliott, M. (2016). Alterative indicators for the risk of non-response bias. *Revue Internationale de Statistique*, 84, 43-62.

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.

Peytcheva, E., et Groves, R.M. (2009). Using variation in response rates of demographic subgroups as evidence on nonresponse bias in survey estimates. *Journal of Official Statistics*, 25, 193-201.

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. (2011a). Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1-21.

Särndal, C.-E. (2011b). Three factors to signal nonresponse bias, with applications to categorical auxiliary variables. *Revue Internationale de Statistique*, 79, 233-254.

Särndal, C.-E., Lumiste, K. et Traat, I. (2016). Est-ce que la réduction du déséquilibre de la réponse accroît l'exactitude des estimations de l'enquête ? *Techniques d'enquête*, 42, 2, 233-253. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016002/article/14663-fra.pdf.

Särndal, C.-E., et Lundquist, P. (2014). 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. (2017). Inconsistent regression and nonresponse bias: Exploring their relationship as a function of response imbalance. *Journal of Official Statistics*, 33(3), 1-27.

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

Särndal, C.-E., et Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24, 251-260.

Särndal, C.-E., et Lundström, S. (2010). Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse. *Techniques d'enquête*, 36, 2, 141-156. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010002/article/11376-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 https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-fra.pdf.

Schouten, B., Cobben, F., Lundquist, P. et Wagner, J. (2014). Theoretical and empirical evidence for balancing of survey response by design. Document de discussion 201415, Statistics Netherlands.

Schouten, B., Shlomo, N. et Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 231-253.

Tourangeau, R., Brick, J.M., Lohr, S. et Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society, Series A*, 180, 201-223.

Wagner, J. (2012). Research synthesis: A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, 76, 555-575.

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