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

Techniques d'enquête

Estimation des propensions à répondre et indicateurs de représentativité des réponses utilisant l'information au niveau de la population

par Annamaria Bianchi, Natalie Shlomo, Barry Schouten,
Damião N. Da Silva et Chris Skinner

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. 231-264
Statistique Canada, nº 12-001-X au catalogue

Estimation des propensions à répondre et indicateurs de représentativité des réponses utilisant l'information au niveau de la population

Annamaria Bianchi, Natalie Shlomo, Barry Schouten, Damião N. Da Silva et Chris Skinner[1]

Résumé

Ces dernières années, les mesures indirectes du biais de non-réponse dans les enquêtes ou d'autres formes de collecte de données ont suscité un vif intérêt, en raison de la diminution progressive des propensions à répondre aux enquêtes et des pressions exercées sur les budgets d'enquête. Ces changements ont poussé les sondeurs à se concentrer davantage sur la représentativité ou l'équilibre des unités échantillonnées répondantes par rapport à des variables auxiliaires pertinentes. Un exemple de mesure est l'indicateur de représentativité, ou indicateur R. Cet indicateur est basé sur la variation d'échantillon pondérée selon le plan de sondage des propensions à répondre estimées. Cela suppose que l'on dispose de données auxiliaires appariées. L'une des critiques de l'indicateur est qu'il ne peut pas être utilisé si l'information auxiliaire est disponible uniquement au niveau de la population. Dans le présent article, nous proposons une nouvelle méthode d'estimation des propensions à répondre qui ne requiert pas d'information auxiliaire pour les non-répondants à l'enquête et qui est fondée sur de l'information auxiliaire pour la population. Ces propensions à répondre basées sur la population peuvent alors être utilisées pour élaborer des indicateurs R faisant appel à des tableaux de contingence de population ou à des fréquences de population. Nous discutons des propriétés statistiques des indicateurs, et évaluons leur performance au moyen d'une étude portant sur des données réelles de recensement et d'une application à la *Dutch Health Survey*.

Mots-clés : Non-réponse; données manquantes; biais de non-réponse; réponse équilibrée.

1 Introduction

Le biais de non-réponse aux enquêtes est une question de plus en plus préoccupante en raison de la diminution des taux de réponse et de la compression des budgets. Les instituts nationaux de statistique (INS) chargés de réaliser des enquêtes nationales afin de faire part de la situation économique et sociale, et des caractéristiques démographiques de leur pays éprouvent de plus en plus de difficulté à maintenir la qualité de la réponse à leurs enquêtes. Dans le présent article, nous nous concentrons sur la *Dutch Health Survey*, qui est réalisée depuis 1998 par *Statistics Netherlands* et qui, jusqu'en 2010, était une enquête avec interviews sur place. En 2010, la collecte de données en ligne a été ajoutée à titre de mode de collecte séquentiel avant les interviews sur place. Les taux de réponse ont diminué progressivement pour passer de près de 70 % à environ 60 %. D'autres INS et organismes d'enquête ont fait état d'une baisse des taux de réponse, particulièrement après être passés à des modes mixtes de collecte des données afin de réduire les coûts, en orientant les répondants vers les modes de collecte les moins chers. Cependant, à eux seuls, les taux de réponse ne sont pas suffisants pour juger de la qualité de la réponse aux enquêtes, car le biais de non-réponse résulte du contraste entre les personnes qui répondent et celles qui ne répondent pas. La

1. Annamaria Bianchi, University of Bergamo, Italie. Courriel : annamaria.bianchi@unibg.it; Natalie Shlomo, University of Manchester, Royaume-Uni et Social Statistics, School of Social Sciences, Humanities Bridgeford Street Room G17A, University of Manchester M13 9PL Royaume-Uni. Courriel : natalie.shlomo@manchester.ac.uk; Barry Schouten, Statistics Netherlands et Utrecht University. Courriel : jg.schouten@cbs.nl; Damião N. Da Silva, Universidade Federal do Rio Grande do Norte, Brésil. Courriel : damiao@ccet.ufrn.br; Chris Skinner, London School of Economics and Political Science, Royaume-Uni. Courriel : c.j.skinner@lse.ac.uk.

conjecture est que la mauvaise santé, certaines habitudes comme le tabagisme ou les visites peu fréquentes chez le dentiste, et les mauvaises conditions de vie sont des facteurs à l'origine du biais de non-réponse à la *Dutch Health Survey*. L'âge, l'état matrimonial, le revenu et l'origine ethnique sont des variables explicatives importantes.

Un certain nombre de mesures indirectes du biais de non-réponse ont été élaborées récemment pour suppléer le taux de réponse classique. Wagner (2012) donne une classification de ces mesures : d'une part les indicateurs qui comprennent uniquement des variables auxiliaires observées et d'autre part, ceux qui incluent en outre des variables d'enquête observées pouvant ou non tenir compte de la pondération pour corriger la non-réponse. Les indicateurs les plus répandus qui utilisent uniquement des variables auxiliaires observées sont les indicateurs R (Schouten, Cobben et Bethlehem, 2009; Schouten, Shlomo et Skinner, 2011) et les indicateurs d'équilibre (Särndal, 2011; Lundquist et Särndal, 2013).

L'élaboration de ces mesures arrive au moment où l'on cherche de plus en plus à adapter la collecte des données (Schouten, Calinescu et Luiten, 2013; Wagner, 2013; Wagner et Hubbard, 2014; Beaumont, Bocci et Haziza, 2014) afin que l'intensité de l'effort visant différents sous-groupes définis par les variables auxiliaires puisse varier au cours du temps, éventuellement grâce à un changement de stratégie, en fonction des profils de réponse (Schouten, Bethlehem, Beulens, Kleven, Loosveldt, Rutar, Shlomo et Skinner, 2012; Särndal et Lundquist, 2014). Tant les indicateurs R que les indicateurs d'équilibre doivent être considérés en association avec les données auxiliaires employées. Les valeurs de ces indicateurs peuvent varier selon les variables auxiliaires choisies.

De plus, Schouten, Cobben, Lundquist et Wagner (2016) donnent des preuves empiriques qu'il est avantageux que les échantillons soient davantage équilibrés par rapport aux variables auxiliaires, même si ces variables sont utilisées par après pour la correction de la non-réponse. En s'appuyant sur 14 ensembles de données d'enquête, ils montrent qu'en moyenne, un plan de sondage produisant une réponse plus représentative possède un biais de non-réponse plus petit, même après des ajustements sur les caractéristiques pour lesquelles la représentativité est évaluée. Särndal et Lundquist (2014) ont également constaté des gains dus à l'équilibrage de l'ensemble de répondants, en sus de ceux obtenus par calage de l'échantillon. De surcroît, il convient de souligner qu'un échantillon plus équilibré mène à une plus faible variabilité des poids d'ajustement, ce qui est une propriété désirable, car une forte variation des poids d'ajustement peut accroître les erreurs-types des estimations. Évidemment, la pondération pour corriger la non-réponse restera nécessaire, car un certain déséquilibre persistera toujours dans l'ensemble final de données de réponse.

Les données auxiliaires utilisées pour mesurer les indicateurs de réponse peuvent provenir de la base de sondage, de données administratives, ainsi que de données sur le processus de collecte des données, appelées paradonnées (Kreuter, 2013). Les indicateurs d'équilibre et les indicateurs R sont très similaires et sont souvent proportionnels en taille. Dans le présent article, nous nous concentrons sur les indicateurs R. Toutefois, la discussion et les résultats peuvent être en grande partie transposés facilement aux indicateurs d'équilibre.

Les indicateurs R reposent sur l'hypothèse que l'on dispose de variables auxiliaires obtenues en appariant des données provenant, par exemple, de la base de sondage ou de registres à l'échantillon de l'enquête. Cette hypothèse d'appariement aux échantillons d'enquête peut être infaisable dans de nombreuses conditions, ce qui entrave l'application. Tandis que les instituts nationaux de statistique ont souvent accès aux registres de l'administration publique, cela n'est habituellement pas le cas des universités et des entreprises d'études de marché. Pour que les indicateurs soient utiles à ces chercheurs, ils doivent être fondés sur d'autres formes d'information auxiliaire. Les ensembles de statistiques produits par les instituts nationaux de statistique sont la seule forme d'information auxiliaire généralement accessible. Ces instituts diffusent des tableaux contenant tout un éventail de statistiques sur la population. Dans le présent article, nous élaborons des indicateurs R qui s'appuient uniquement sur ce genre de statistiques de population et peuvent être calculés sans rien savoir sur les non-répondants. Ainsi, les entreprises d'études de marché comparent les répartitions des réponses pour un ensemble fixé de variables auxiliaires aux statistiques nationales, considérées comme la norme de référence. Les estimateurs des indicateurs R proposés ici permettent de surveiller et d'évaluer les variables servant de norme de référence durant et après la collecte des données.

Bien que les indicateurs R fondés sur de l'information auxiliaire au niveau de la population décrits dans le présent article soient motivés par les pratiques de collecte de données d'enquête, ils peuvent être appliqués à toute situation où des données manquent pour les variables d'intérêt et où l'on dispose de données auxiliaires (presque) complètes. Ces estimateurs peuvent être utilisés, par exemple, pour surveiller et évaluer la complétude des données administratives, exercice utile si les données sont transmises et accumulées progressivement au cours du temps. Dans ce cas, les indicateurs R basés sur la population fourniraient une évaluation de la représentativité des données administratives transmises. Une autre application utile de ces indicateurs est pour l'évaluation de la représentativité des enregistrements de données appariés. Van der Laan et Bakker (2015) ont proposé un indicateur de représentativité d'appariement (indicateur LR pour *linkage representativeness*) qui examine la similarité des enregistrements appariés avec la population cible étudiée.

Les indicateurs R et leurs propriétés statistiques, comme il est exposé dans Shlomo, Skinner et Schouten (2012), se rapportent au cas où nous disposons d'information auxiliaire appariée au niveau de l'échantillon pour les non-répondants. Pour élaborer les indicateurs R basés sur des statistiques de population, nous proposons une nouvelle méthode d'estimation des propensions à répondre qui ne requiert pas d'information auxiliaire pour les non-répondants à l'enquête. Nous les appellerons propensions à répondre basées sur la population. Autant que nous sachions, aucun modèle pour les propensions à répondre faisant appel à de l'information sur la population uniquement n'est mentionné dans la littérature. À cet égard, le présent article est novateur et pourrait également être utile et pertinent dans d'autres domaines statistiques. Dans cet article, nous nous concentrons sur l'utilisation des propensions à répondre basées sur la population pour calculer les indicateurs R.

En ce qui concerne l'adaptation de la collecte de données, il est évident que les conditions qui requièrent l'emploi d'indicateurs R basés sur la population rendent plus difficile la mise en œuvre des types de plans

de collecte adaptatifs mentionnés plus haut, car nous ne connaissons pas les valeurs des covariables pour les non-répondants. Toutefois, en utilisant ces indicateurs R basés sur l'information auxiliaire au niveau de la population, nous pouvons donner au plan de collecte des caractéristiques qui sont plus pertinentes pour les personnes qui tardent à répondre. Ainsi, si les taux de réponse sont plus faibles chez les jeunes, nous pouvons envoyer un rappel général axé davantage sur les jeunes ou bien donner aux intervieweurs l'instruction de surveiller plus méticuleusement les adresses où ils s'attendent à trouver des personnes jeunes.

L'information auxiliaire pour le calcul des propensions à répondre basées sur la population est tirée de tableaux de données sur la population et de chiffres de population. Pour cela, nous proposons d'abord d'estimer les propensions à répondre basées sur les valeurs de population, en remplaçant les matrices de covariance d'échantillon et les moyennes d'échantillon par les covariances de population et les moyennes de population connues. Puis, en utilisant les propensions à répondre basées sur la population, nous calculons les estimations de l'indicateur R. Nous appelons l'indicateur résultant indicateur R basé sur la population, et nous appelons l'indicateur R classique l'indicateur R basé sur l'échantillon. Nous nous intéressons à trois questions de recherche :

- Comment étendre les propensions à répondre et les indicateurs R basés sur l'échantillon aux propensions à répondre et aux indicateurs R basés sur la population ?
- Quelles sont les propriétés statistiques des indicateurs R basés sur la population ?
- Les indicateurs R basés sur la population sont-ils applicables dans des conditions d'enquête réelles ?

À la section 2, nous proposons une nouvelle méthode d'estimation des propensions à répondre basées sur la population. À la section 3, nous examinons brièvement les définitions et la méthodologie qui sous-tendent les indicateurs R, puis nous considérons leur estimation dans les conditions de population. À la section 4, nous présentons une étude d'évaluation portant sur des échantillons tirés de données de recensement réelles sous des hypothèses réalistes au sujet de la non-réponse dans les enquêtes sociales, et nous évaluons les propriétés des indicateurs R basés sur la population. À la section 5, nous illustrons les indicateurs R proposés à l'aide d'une application à la *Dutch Health Survey*, l'enquête sur la santé réalisée aux Pays-Bas. À la section 6, nous concluons par une discussion et formulons certaines mises en garde concernant les indicateurs proposés et les futurs travaux.

2 Propensions à répondre basées sur la population

2.1 Notation générale

Nous supposons qu'on entreprend un sondage dans lequel un échantillon s est tiré d'une population finie U. Les tailles de s et de U sont notées n et N, respectivement. Les unités dans U sont étiquetées

$i = 1, 2, \ldots, N$. L'échantillon est présumé être tiré selon un plan d'échantillonnage probabiliste $p(.)$, où l'échantillon s est sélectionné avec la probabilité $p(s)$. La probabilité d'inclusion d'ordre un de l'unité i est notée π_i et $d_i = \pi_i^{-1}$ est le poids de sondage. L'étude d'évaluation est fondée sur un échantillonnage aléatoire simple sans remise. Bien que les enquêtes nationales à grande échelle puissent utiliser des plans de sondage à deux degrés plus complexes, nombre d'entre elles sont généralement planifiées de manière que toutes les unités d'enquête aient la même probabilité d'inclusion. Nous donnons aussi les expressions théoriques sous des plans de sondage plus complexes généraux.

Nous supposons que l'enquête est sujette à la non-réponse totale. L'ensemble d'unités répondantes est noté r, de sorte que $r \subset s \subset U$. Nous notons la sommation sur les répondants, l'échantillon et la population par Σ_r, Σ_s et Σ_U, respectivement. Soit r_i la variable indicatrice de réponse telle que $r_i = 1$ si l'unité i répond et $r_i = 0$, autrement. Donc, $r = \{i \in s;\ r_i = 1\}$. Nous supposons que la cible type de l'inférence est la moyenne de population $\bar{Y} = N^{-1}\sum_U y_i$ d'une variable étudiée, qui prend la valeur y_i pour l'unité i.

D'abord, nous supposons que les données disponibles pour l'estimation comprennent les valeurs $\{y_i;\ i \in r\}$ de la variable d'enquête, observées pour les répondants uniquement. Puis, nous supposons que l'information est disponible sur les valeurs $\mathbf{x}_i = \left(x_{1,i}, x_{2,i}, \ldots, x_{K,i}\right)^T$ d'un vecteur de variables auxiliaires $\mathbf{X}$. Nous supposerons habituellement que chaque $x_{k,i}$ est une variable indicatrice binaire, où $\mathbf{x}_i$ représente une ou plusieurs variables catégoriques, puisque ce sera le cas dans l'application que nous considérons, mais notre exposé permet des valeurs $x_{k,i}$ générales. Nous supposons que les valeurs de $\mathbf{x}_i$ sont observées pour tous les répondants, de sorte que $\{y_i, x_i;\ i \in r\}$ est observé.

Nous distinguons deux situations : l'une dans laquelle $\mathbf{x}_i$ est connue pour toutes les unités de l'échantillon, c'est-à-dire pour les répondants ainsi que les non-répondants, et l'autre dans laquelle $\mathbf{x}_i$ est connue uniquement au niveau agrégé, c'est-à-dire le total de population $\sum_U \mathbf{x}_i$ et/ou les produits croisés de population $\sum_U \mathbf{x}_i \mathbf{x}_i^T$. Nous désignons les deux types d'information comme étant l'*information auxiliaire d'échantillon* et l'*information auxiliaire de population agrégée*. La première situation est pertinente si les variables qui constituent $\mathbf{X}$ sont disponibles dans un registre. Cependant, comme il est mentionné dans l'introduction, dans nombre de pays et d'enquêtes, l'information auxiliaire disponible sur les non-répondants peut être limitée et la deuxième situation où l'on utilise l'information auxiliaire de population pourrait être plus utile.

2.2 Définition de la propension à répondre

La théorie des scores de propension a été introduite par Rosenbaum et Rubin (1983) et discutée dans le contexte de la non-réponse aux enquêtes par Little (1986; 1988). La propension à répondre est définie comme étant l'espérance conditionnelle de la variable indicatrice de réponse r_i sachant les valeurs de variables spécifiées et les conditions d'enquête : $\rho_X(\mathbf{x}_i) = E_m(r_i \mid \mathbf{x}_i)$, où le vecteur de variables auxiliaires est défini comme à la section 2.1. Pour simplifier, nous écrirons $\rho_i = \rho_X(\mathbf{x}_i)$ et notons donc la propension à répondre simplement par ρ_i. $E_m(.)$ désigne l'espérance par rapport au modèle qui sous-tend le mécanisme de réponse. Une discussion détaillée des propensions à répondre et de leurs propriétés est

présentée dans Shlomo et coll. (2012). Ces auteurs soutiennent qu'il est souhaitable de choisir les variables auxiliaires constituant $\mathbf{x}_i$ de manière que l'hypothèse de données manquant au hasard, notée MAR pour *missing at random* (Little et Rubin, 2002), soit vérifiée aussi étroitement que possible.

2.3 Estimation des propensions à répondre en utilisant l'information au niveau de la population

Dans le cas de l'information auxiliaire d'échantillon, il est possible d'estimer les propensions à répondre pour toutes les unités échantillonnées au moyen de modèles de régression $g(\rho_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, où $g(.)$ est une fonction de lien, r_i est la variable dépendante, et $\mathbf{x}_i$ est un vecteur de variables explicatives. En général, les propensions à répondre sont modélisées au moyen de modèles linéaires généralisés. Shlomo et coll. (2012) utilisent une fonction de lien logistique.

Dans le cadre fondé sur la population, il est commode de considérer la fonction de lien identité. Cette dernière est une bonne approximation de la fonction de lien logistique d'usage plus répandu quand les taux de réponse sont de niveau intermédiaire, entre 30 % et 70 %, ce qui est le niveau de taux de réponse habituellement obtenu pour les enquêtes nationales et d'autres. Nous démontrons ce fait dans l'étude d'évaluation présentée à la section 4, où trois gammes de taux de réponse sont examinées : faible, moyen et élevé. La fonction de lien identité constitue aussi le fondement d'autres indicateurs de représentativité dans la littérature, comme les indicateurs de déséquilibre et de distance proposés par Särndal (2011), dont certains sont similaires aux poids g calculés dans les estimateurs par la régression généralisée (GREG).

Sous la fonction de lien identité, nous supposons que les propensions à répondre réelles satisfont le « modèle probabiliste linéaire »

$$\rho_i = \mathbf{x}_i^T \boldsymbol{\beta}, \quad i \in U. \tag{2.1}$$

Le modèle probabiliste linéaire en (2.1) peut être estimé par les moindres carrés pondérés, où d_i est le poids de sondage. L'estimateur implicite de ρ_i est donné par

$$\hat{\rho}_i^{\text{MCO}} = \mathbf{x}_i^T \left(\sum_s d_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_s d_i \mathbf{x}_i r_i, \quad i \in s. \tag{2.2}$$

Dans le cas de l'information auxiliaire de population, nous notons d'abord que $\sum_s d_i \mathbf{x}_i$ et $\sum_s d_i \mathbf{x}_i \mathbf{x}_i^T$ sont sans biais pour $\sum_U \mathbf{x}_i$ et $\sum_U \mathbf{x}_i \mathbf{x}_i^T$, respectivement et que, dans les grands échantillons, nous pouvons nous attendre à ce que $\sum_s d_i \mathbf{x}_i \approx \sum_U \mathbf{x}_i$ et $\sum_s d_i \mathbf{x}_i \mathbf{x}_i^T \approx \sum_U \mathbf{x}_i \mathbf{x}_i^T$. Il découle de (2.2) que, dans le cadre fondé sur la population, nous pouvons approximer $\hat{\rho}_i^{\text{MCO}}$ par

$$\tilde{\rho}_{i,T1} = \mathbf{x}_i^T \mathbf{T}_1^{-1} \sum_r d_k \mathbf{x}_k, \quad i \in r \tag{2.3}$$

où $\mathbf{T}_1 = \sum_U \mathbf{x}_j \mathbf{x}_j^T$. Nous notons que $\tilde{\rho}_{i,T1}$ est calculé uniquement sur l'ensemble d'unités répondantes.

L'estimateur en (2.3) requiert que l'on connaisse les sommes des carrés et des produits croisés $\sum_U \mathbf{x}_i \mathbf{x}_i^T$ des éléments de $\mathbf{x}_i$ pour la population. Cependant, les produits croisés pourraient être inconnus. Le cas échéant, nous pouvons estimer $\sum_s d_i \mathbf{x}_i \mathbf{x}_i^T$ dans (2.2) en réécrivant ce terme sous la forme

$$\sum_s d_i \, \mathbf{x}_i \, \mathbf{x}_i^T = \sum_s d_i \, (\mathbf{x}_i - \overline{\mathbf{x}}_s)(\mathbf{x}_i - \overline{\mathbf{x}}_s)^T + N \overline{\mathbf{x}}_s \overline{\mathbf{x}}_s^T, \tag{2.4}$$

où $\overline{\mathbf{x}}_s = \sum_s d_i \mathbf{x}_i / N$. $\overline{\mathbf{x}}_s$ peut être remplacé par $\overline{\mathbf{x}}_U$ et la matrice de covariance

$$\mathbf{S}_{xx} = N^{-1} \sum_s d_i \, (\mathbf{x}_i - \overline{\mathbf{x}}_s)(\mathbf{x}_i - \overline{\mathbf{x}}_s)^T \tag{2.5}$$

peut être remplacée par son estimation en utilisant l'ensemble de réponses

$$\hat{\mathbf{S}}_{xx} = \left(\sum_s d_j r_j \right)^{-1} \sum_s d_i r_i \, (\mathbf{x}_i - \overline{\mathbf{x}}_U)(\mathbf{x}_i - \overline{\mathbf{x}}_U)^T. \tag{2.6}$$

Nous pouvons également estimer (2.6) en utilisant la pondération des propensions par $\tilde{\rho}_i^{-1}$ pour corriger le biais de non-réponse dans la variance des propensions à répondre relatives à un ensemble de variables X.

En combinant (2.3), (2.4) et (2.6), nous obtenons l'estimateur suivant :

$$\tilde{\rho}_{i,T2} = \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1} \sum_r d_k \mathbf{x}_k, \quad i \in r, \tag{2.7}$$

où $\hat{\mathbf{T}}_2 = N \hat{\mathbf{S}}_{xx} + N \overline{\mathbf{x}}_U \overline{\mathbf{x}}_U^T$.

Nous faisons donc la distinction entre deux types d'information auxiliaire de population agrégée, qui sont désignés par les indices « T_1 » dans (2.3) et « T_2 » dans (2.7) :

TYPE 1 *Information auxiliaire complète de population agrégée* : les produits croisés de population sont disponibles, c'est-à-dire $\sum_U \mathbf{x}_i \mathbf{x}_i^T$ ou $\sum_U (\mathbf{x}_i - \overline{\mathbf{x}}_U)(\mathbf{x}_i - \overline{\mathbf{x}}_U)^T$, où $\overline{\mathbf{x}}_U = \sum_U \mathbf{x}_i / N$;

TYPE 2 *Information auxiliaire de marge de population agrégée* : seuls les chiffres de marge de la population sont disponibles, c'est-à-dire $\sum_U \mathbf{x}_i$.

Le premier type implique que nous disposons de tous les tableaux de contingence 2×2, par exemple, âge par sexe, âge par état matrimonial, et sexe par état matrimonial. Cette information pourrait être disponible pour un institut national de statistique qui a accès aux registres de population ou à des données démographiques détaillées et qui souhaite utiliser de l'information au niveau de la population pour surveiller la collecte des données, en raison d'un manque d'information d'échantillon dans les bases de sondage. Le deuxième type est plus contraignant, car nous disposons uniquement des fréquences, par exemple, âge, sexe, état matrimonial, sans aucune connaissance des interactions entre ces variables. Cette information serait disponible ordinairement sur les sites Web des instituts nationaux de statistique et pourrait par conséquent être utilisée par les entreprises de marketing et d'autres organismes de collecte des données pour surveiller leurs opérations de collecte.

3 Estimation des indicateurs R basée sur les totaux de population

À la présente section, nous examinons d'abord brièvement la définition et les concepts des indicateurs R, et leur estimation basée sur l'information auxiliaire d'échantillon. Des renseignements détaillés peuvent être consultés dans Shlomo et coll. (2012). Ensuite, en appliquant la théorie présentée à la section 2.3, nous

adaptons l'indicateur R basé sur l'échantillon au cas où l'information auxiliaire est tirée de tableaux de données de population et de fréquences de population. En outre, nous étudions les propriétés statistiques de cet estimateur.

3.1 Indicateurs R

Schouten et coll. (2009) présentent le concept de la réponse représentative. Une réponse à une enquête est dite *représentative par rapport à* $\mathbf{X}$ quand les propensions à répondre sont constantes pour $\mathbf{X}$, c'est-à-dire

$$\rho_i = \rho_X(\mathbf{x}_i) = \bar{\rho}, \quad \forall \mathbf{x}_i,$$

où $\bar{\rho}$ désigne la propension à répondre moyenne dans la population.

La mesure globale de la réponse représentative est l'indicateur R. L'indicateur R associé à un ensemble de propensions à répondre dans la population $\{\rho_i : i \in U\}$ est défini comme étant

$$R_\rho = 1 - 2S_\rho, \tag{3.1}$$

où S_ρ désigne l'écart-type des propensions à répondre individuelles

$$S_\rho^2 = \frac{1}{N-1}\sum_U (\rho_i - \bar{\rho}_U)^2 = \frac{N}{N-1}\left\{\frac{1}{N}\sum_U \rho_i^2 - \left[\frac{1}{N}\sum_U \rho_i\right]^2\right\}, \tag{3.2}$$

où $\bar{\rho}_U = \sum_U \rho_i / N$.

L'*indicateur R* prend des valeurs dans l'intervalle $\left[1 - \sqrt{\frac{N}{N-1}}, 1\right]$ dont la borne supérieure 1 indique la réponse la plus représentative, où les ρ_i ne présentent aucune variation, et la borne inférieure $1 - \sqrt{\frac{N}{N-1}}$ (qui est proche de 0 pour les grandes enquêtes) indique la réponse la moins représentative, où les ρ_i présentent une variation maximale.

Une importante mesure apparentée à la représentativité est le coefficient de variation des propensions à répondre

$$\mathrm{CV}_\rho = \frac{S_\rho}{\bar{\rho}_U}. \tag{3.3}$$

Cette mesure est pertinente si les moyennes ou les totaux de population sont des paramètres d'intérêt. Dans ces cas, elle peut être utilisée à la place de l'indicateur R. Pour d'autres types de paramètres d'intérêt, tels que la médiane ou un ratio, d'autres indicateurs peuvent être utilisés (Brick et Jones, 2008).

Le coefficient de variation en (3.3) borne le biais de non-réponse absolu des moyennes de réponse non ajustées pour une variable Y divisée par son écart-type. Schouten et coll. (2016) utilisent aussi le coefficient de variation pour évaluer les « pires cas » d'intervalles de biais de non-réponse pour des estimateurs classiques ajustés pour la non-réponse après l'enquête, tels que l'estimateur par la régression généralisée (GREG) (Deville et Särndal, 1992) et la pondération par l'inverse de la propension (IPW pour *inverse propensity weighting*) (Little, 1988).

3.2 Indicateurs R basés sur l'échantillon

Dans le cas de l'information auxiliaire d'échantillon, il est possible d'estimer les propensions à répondre pour toutes les unités échantillonnées. Soit $\hat{\rho}_i$ un estimateur de ρ_i. L'estimateur basé sur l'échantillon de l'indicateur R est

$$\hat{R}_{\hat{\rho}} = 1 - 2\hat{S}_{\hat{\rho}}, \tag{3.4}$$

où $\hat{S}^2_{\hat{\rho}}$ est la variance d'échantillon pondérée selon le plan de sondage des propensions à répondre estimées, calculée en utilisant la première expression en (3.2)

$$\hat{S}^2_{\hat{\rho}} = \frac{1}{N-1}\sum_s d_i \left(\hat{\rho}_i - \hat{\bar{\rho}}_U\right)^2,$$

où $\hat{\bar{\rho}}_U = \left(\sum_s d_i \hat{\rho}_i\right)\Big/ N$.

L'indicateur R basé sur l'échantillon défini par (3.4) est une statistique présentant une certaine précision et un certain biais. Shlomo et coll. (2012) discutent des corrections du biais et des intervalles de confiance pour $\hat{R}_{\hat{\rho}}$. Ceux-ci peuvent être obtenus en SAS et en code R à l'adresse www.risq-project.eu, et un manuel est fourni par De Heij, Schouten et Shlomo (2015). Nous reviendrons sur les propriétés statistiques à la section 3.4.

3.3 Indicateurs R basés sur la population

Nous montrons à la section 4 que les indicateurs R ne dépendent que légèrement du type de fonction de lien quand on estime les propensions à répondre si les taux de réponse ne sont pas dans les queues de la distribution, c'est-à-dire très élevés ou très faibles. En outre, nous obtenons une estimation similaire des indicateurs R quand les propensions à répondre dans la population sont estimées en se servant d'information de type 1 ou de type 2.

Dans le cadre fondé sur la population, un estimateur de l'indicateur R est alors donné par

$$\tilde{R}_{\tilde{\rho}} = 1 - 2\tilde{S}_{\tilde{\rho}}, \tag{3.5}$$

où

$$\tilde{S}^2_{\tilde{\rho}} = \frac{N}{N-1}\left\{\frac{1}{N}\sum_r d_i \tilde{\rho}_i - \left(\frac{1}{N}\sum_r d_i\right)^2\right\}, \tag{3.6}$$

et $\tilde{\rho}_i$ désigne les propensions à répondre calculées sous l'information de type 1 $\left(\tilde{\rho}_{i,T1}\right)$ ou les propensions à répondre estimées sous l'information de type 2 $\left(\tilde{\rho}_{i,T2}\right)$.

Il convient de souligner que l'estimation de l'indicateur R est basée sur la deuxième expression pour S^2_ρ dans (3.2). Ce choix rend en effet l'estimateur $\tilde{S}^2_{\tilde{\rho}}$ linéaire en $\tilde{\rho}_i$, ce qui est avantageux pour les calculs du biais, comme il est décrit à la section 3.4. L'étude d'évaluation présentée à la section 4 démontre

empiriquement que les deux expressions pour S_ρ^2 sont similaires pour les types d'enquêtes nationales à grande échelle prises en considération. En outre, nous utilisons la pondération des propensions par $\tilde{\rho}_i^{-1}$ à titre d'ajustement pour le biais de non-réponse. Comme dans la pondération classique pour tenir compte de la non-réponse, la validité de cet ajustement dépend de la validité des estimations $\tilde{\rho}_i$.

Il convient de souligner que toute technique d'ajustement pour tenir compte de la non-réponse peut être appliquée pour construire les estimateurs pour R_ρ, par exemple, des estimateurs par calage faisant appel à la pondération linéaire ou multiplicative (Särndal et Lundström, 2005) ou à des ajustements par classes de pondération (Little, 1986). Il est généralement reconnu que la pondération par les propensions à répondre peut mener à de plus grandes erreurs-types. Donc, il pourrait être plus efficace d'utiliser des modèles parcimonieux pour estimer l'indicateur R, par exemple, en stratifiant sur les classes de propensions à répondre. Cependant, nous n'avons pas étudié ce genre d'estimateurs et nous nous sommes limités à l'estimateur pondéré par les propensions (3.5). Il s'agit d'un sujet pour de futurs travaux de recherche.

L'estimation du coefficient de variation (3.3) basée sur la population est simple

$$\mathrm{CV}_{\tilde{\rho}} = \frac{\tilde{S}_{\tilde{\rho}}}{\tilde{\bar{\rho}}_U},$$

où $\tilde{\bar{\rho}}_U = \sum_r d_i / N$.

Bien qu'il s'agisse d'estimateurs simples, les indicateurs R basés sur la population qui s'appuient sur (2.3) et (2.7) posent problème. Leurs erreurs-types et leurs biais augmentent quand les taux de réponse sont élevés. Nous donnerons la preuve de cette tendance dans l'étude d'évaluation, à la section 4.2. Manifestement, un plus grand nombre de répondants devrait fournir de plus petites erreurs-types et réduire le biais, puisque les variables auxiliaires ne varieront pas autant parmi les non-répondants restants. Les estimateurs (2.3) et (2.7) possèdent ces propriétés parce qu'il s'agit d'estimateurs naturels, mais naïfs, qui ne tiennent pas compte de l'échantillonnage, qui fait varier les covariances dans le dénominateur des propensions à répondre estimées en même temps que le numérateur. En « substituant » (*plugging in*) une covariance de population fixe dans le dénominateur, on évite la variation due à l'échantillonnage.

Un moyen de modérer cet effet consisterait à utiliser un estimateur composite, c'est-à-dire à employer une combinaison linéaire de la propension estimée et du taux de réponse,

$$\tilde{\rho}_{i,T1}^C = (1-\lambda)\,\tilde{\rho}_{i,T1} + \lambda \tilde{\bar{\rho}}_U, \tag{3.7}$$

avec $\tilde{\bar{\rho}}_U = \sum_r d_i / N$, et de façon similaire pour le type 2. L'estimation composite en (3.7) est semblable à un estimateur « à rétrécissement », par exemple, Copas (1983 et 1993), pour la variance des propensions à répondre $\tilde{S}_\rho^2$ donnée par (3.6). Dans ce cas, le λ optimal est habituellement choisi de manière à minimiser l'EQM en résolvant la dérivée de l'EQM par rapport à λ. Nous examinons le choix de λ à la section 3.4 et notons ici que, sachant le biais observé et les propriétés de variance, λ devrait être une fonction croissante du taux de réponse et devrait converger vers 1 quand les taux de réponse augmentent. Les propensions à répondre estimées plus grandes que 1 seront rapprochées de 1 par un tel λ, en raison de l'utilisation de la fonction de lien linéaire sous taux de réponse élevé.

Nous avons exploré plusieurs autres options pour remplacer l'estimateur composite en (3.7), par exemple, un estimateur composite de la matrice de covariance de population et de la matrice de covariance des réponses des $\mathbf{x}_i$, ainsi que la troncation des propensions à répondre pour qu'elles soient contenues dans l'intervalle [0, 1] pour les taux de réponse élevés, mais ces solutions ont donné de moins bons résultats que l'estimateur composite en (3.7). En outre, nous avons étudié une estimation de type Hájek, mais celle-ci a donné des résultats similaires à ceux produits par l'estimateur proposé en (3.6). Un autre avantage de l'utilisation de l'estimateur composite en (3.7) est que nous pouvons facilement construire des ajustements du biais des indicateurs R similaires aux ajustements du biais construits en se basant sur les propensions à répondre en (2.3) ou (2.7).

Une option prometteuse pourrait consister à adopter une approche d'algorithme EM dans laquelle les variables auxiliaires manquantes pour les non-répondants sont imputées. Une telle approche est, cependant, de nature très différente et nous réservons son étude à de futurs travaux de recherche.

3.4 Biais et erreur-type des indicateurs R basés sur la population

Shlomo et coll. (2012) obtiennent des approximations analytiques pour le biais et l'erreurs-type de l'estimation de l'indicateur R basée sur l'échantillon (3.4). Le biais de cet estimateur est dû principalement à la « substitution » des propensions à répondre estimées dans les variances d'échantillons. Cette source de biais est appelée biais de petit échantillon. Une contribution nettement plus faible, et habituellement négligeable au biais découle de l'utilisation des moyennes d'échantillon au lieu des moyennes de population. Même si la réponse est représentative, c'est-à-dire que les propensions à répondre sont égales, on observe une certaine variation des propensions à répondre estimées. Le biais est inversement proportionnel à la taille d'échantillon, c'est-à-dire qu'il est d'autant plus petit que l'échantillon est grand. Schouten et coll. (2009) étudient le biais pour différentes tailles d'échantillon. Il découle de leurs analyses que le biais est relativement petit pour les tailles d'échantillon habituelles dans les enquêtes de grande portée, par comparaison à l'erreur-type des indicateurs R. En outre, l'ajustement pour tenir compte du biais permet d'éliminer ce dernier.

Pour les indicateurs R estimés en se basant sur la population, nous nous attendons à ce que les propriétés statistiques soient assez différentes de celles des indicateurs R basés sur l'échantillon. Comme les estimateurs utilisent moins d'information, les erreurs-types seront plus grandes. Le biais des estimateurs basés sur la population peut aussi être plus grand, puisqu'en plus du biais observé pour les petites tailles d'échantillon dans les estimateurs basés sur l'échantillon, les estimateurs basés sur la population présenteront vraisemblablement un biais dû à l'estimation des moyennes et des covariances d'échantillon et au fait de se restreindre aux moyennes des réponses (pondérées par les propensions à répondre).

Afin de réduire le biais des estimateurs basés sur la population, nous proposons d'ajuster $\tilde{S}^2_{\tilde{\rho}_{T1}}$ et $\tilde{S}^2_{\tilde{\rho}_{T2}}$ pour le biais. Cela mène à la version ajustée qui suit de l'estimateur de l'indicateur R sous l'information de type 1 :

$$\tilde{R}^{\text{AJUS}}_{\tilde{\rho}_{T1}} = 1 - 2\left[\tilde{S}^2_{\tilde{\rho}_{T1}} - \tilde{B}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)\right]^{1/2}. \tag{3.8}$$

Nous donnons à l'annexe A la dérivation de l'expression générale pour $\tilde{B}_{\bar{\rho}T1}\left(\tilde{S}^2_{\bar{\rho}T1}\right)$ sous échantillonnage aléatoire simple, ainsi qu'une expression plus générale sous échantillonnage complexe. Provenant de l'annexe A, l'estimateur du biais basé sur l'ensemble de réponses sous échantillonnage aléatoire simple est :

$$\tilde{B}^{\text{EAS}}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1 - \frac{n-1}{N-1}\tilde{\rho}_{i,T1}\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i \right.$$

$$\left. + \frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T1} - \left(1 - \frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T1}}}{n} - \frac{n_r}{n^2}\right], \tag{3.9}$$

où n_r désigne la taille de l'ensemble de réponses r.

Dans le cas de l'information de type 2, la version ajustée de l'estimateur de l'indicateur R est la même qu'en (3.8) avec les termes de type 2 remplaçant l'information de type 1.

À l'annexe B, nous dérivons l'expression générale pour le biais de $\tilde{S}^2_{\tilde{\rho}_{T2}}$, $\tilde{B}_{\bar{\rho}T2}\left(\tilde{S}^2_{\bar{\rho}T2}\right)$, sous échantillonnage aléatoire simple et sous le cas plus général de l'échantillonnage complexe. Provenant de l'annexe B, l'estimateur du biais basé sur l'ensemble de réponses sous échantillonnage aléatoire simple est :

$$\tilde{B}^{\text{EAS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\left\{\frac{1}{n_r^2}\sum_{i\in r}\left\{1 - \frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}\right.$$

$$- \frac{N}{nn_r}\sum_{i\in r}\left\{1 - \frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{z}_i\mathbf{z}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}$$

$$+ \frac{N}{n^2}\sum_{i\in r}\left\{1 - \frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{x}_i$$

$$\left. + \frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T2} - \left(1 - \frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T2}}}{n} - \frac{n_r}{n^2}\right\},$$

où $\hat{\mathbf{F}} = Nn^{-1}\sum_r \mathbf{z}_k\mathbf{z}_k^T$, $\hat{\mathbf{t}} = Nn^{-1}\sum_r \mathbf{x}_k$, et $\mathbf{z}_i = (\mathbf{x}_i - \bar{\mathbf{x}}_U)$.

En se tournant vers l'estimateur composite, il est simple de montrer que (3.7) peut se réécrire sous la forme

$$\tilde{S}^2_{\tilde{\rho}^C_{T1}} = (1-\lambda)\tilde{S}^2_{\tilde{\rho}_{T1}}, \tag{3.10}$$

et que son biais égale

$$B\left(\tilde{S}^2_{\tilde{\rho}^C_{T1}}\right) = (1-\lambda)B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) - \lambda S^2_{\rho}. \tag{3.11}$$

Un estimateur de $B\left(\tilde{S}^2_{\tilde{\rho}^C_{T1}}\right)$ basé sur l'ensemble de réponses s'obtient en utilisant l'estimateur basé sur l'ensemble de réponses établi pour $B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)$. Pour l'estimateur de type 1 et sous échantillonnage aléatoire simple :

$$\begin{aligned}\tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}^C_{T1}}\left(\tilde{S}^2_{\tilde{\rho}^C_{T1}}\right) &= (1-\lambda)\,\tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}^C_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) - \lambda\tilde{S}^2_{\tilde{\rho}^C_{T1}} \\ &= (1-\lambda)\frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}^C_{i,T1}\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i\right. \\ &\qquad \left. + \frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}^C_{i,T1} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}^C_{T1}}}{n} - \frac{n_r}{n^2}\right] - \lambda\tilde{S}^2_{\tilde{\rho}^C_{T1}}. \qquad (3.12)\end{aligned}$$

La même approche s'applique à l'estimateur de type 2.

La variance de (3.10) est égale à

$$V\left(\tilde{S}^2_{\tilde{\rho}^C_{T_1}}\right) = (1-\lambda)^2\,V\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right). \qquad (3.13)$$

Pour estimer la variance de $\tilde{R}^{\mathrm{AJUS}}_{\tilde{\rho}T1}$ dans (3.8), ainsi que la variance de l'estimateur composite dans (3.13), nous devons estimer la variance de $\tilde{S}^2_{\tilde{\rho}_{T_1}}$ définie en (3.6) et notée $V\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$. Pour estimer cette variance, nous utilisons des méthodes de rééchantillonnage. Plus précisément, nous appliquons des méthodes bootstrap (voir Efron et Tibshirani, 1993; Booth, Butler et Hall, 1994 et Wolter, 2007 pour l'utilisation de méthodes bootstrap en population finie) et étudions leur performance dans l'étude d'évaluation présentée à la section 4.

Revenons maintenant au choix de λ pour l'estimateur composite en (3.7). Le λ optimal peut être calculé en combinant (3.11) et (3.13), puis en prenant les dérivées. Si nous désignons par B et V les termes $B\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$ et $V\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$, respectivement, il s'ensuit que le λ optimal est

$$\lambda_{\mathrm{opt}} = \frac{B\left(B+S^2_\rho\right)+V}{\left(B+S^2_\rho\right)^2+V}. \qquad (3.14)$$

Nous notons qu'à mesure qu'augmente la taille d'échantillon, les termes B et V tendent tous deux vers zéro et qu'il se peut que λ_{opt} soit négatif. Cependant, selon l'étude d'évaluation, pour les types d'enquêtes nationales à grande échelle pris en considération, ce problème ne se pose pas en pratique.

Afin d'estimer λ_{opt}, il faut estimer les quantités B, V et S^2_ρ. Sous information de type 1 et échantillonnage aléatoire simple, nous proposons d'estimer B par $\tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)$ comme en (3.9), S^2_ρ par $\tilde{S}^2_{\tilde{\rho}_{T1}}$, et V par l'estimateur bootstrap de la variance de $\tilde{S}^2_{\tilde{\rho}_{T1}}$. Cela mène à l'estimateur basé sur la population de type 1 pour λ_{opt}, que nous notons $\tilde{\lambda}_{\mathrm{opt},\,T1}$, et aux propensions composites basées sur la population

$$\tilde{\rho}^{\mathrm{PC}}_{i,T1} = \left(1-\tilde{\lambda}_{\mathrm{opt},T1}\right)\tilde{\rho}_{i,T1} + \tilde{\lambda}_{\mathrm{opt},T1}\tilde{\bar{\rho}}_U.$$

L'indicateur R basé sur la population correspondant est alors calculé comme en (3.5) et sa version ajustée pour le biais, comme en (3.8), où l'ajustement pour le biais est donné par (3.12).

Nous proposons d'estimer la variance de l'estimateur composite basé sur la population par linéarisation

$$\frac{\tilde{V}^{\mathrm{BT}}\left(\tilde{S}^{2}_{\tilde{\rho}_{T1}}\right)\left(1-\tilde{\lambda}_{\mathrm{opt},T1}\right)^{2}}{\tilde{S}^{2}_{\tilde{\rho}_{T1}}},$$

où $\tilde{V}^{\mathrm{BT}}\left(\tilde{S}^{2}_{\tilde{\rho}_{T1}}\right)$ est l'estimateur bootstrap de la variance pour $V\left(\tilde{S}^{2}_{\tilde{\rho}_{T1}}\right)$.

La même approche s'applique pour l'information de type 2.

4 Étude d'évaluation

À la présente section, nous réalisons une étude d'évaluation sur des données réelles de recensement provenant de l'échantillon du recensement d'Israël de 1995 pour évaluer les propriétés d'échantillonnage des procédures d'estimation présentées à la section 3.

L'objectif de l'évaluation est double, à savoir a) étudier les propriétés d'échantillonnage des indicateurs R basés sur la population ajustés et non ajustés pour le biais, en les comparant aux propriétés des indicateurs R correspondants basés sur l'échantillon, et évaluer l'effet de la taille d'échantillon, du nombre de variables auxiliaires dans le modèle et du taux de réponse; b) étudier la performance de l'estimateur bootstrap pour estimer la variance de l'indicateur R basé sur la population.

4.1 Données et plan de l'étude d'évaluation

L'échantillon à 20 % du recensement d'Israël de 1995 contient 753 711 personnes de 15 ans et plus réparties dans 322 411 ménages. L'échantillon du recensement est tiré selon un plan d'échantillonnage aléatoire systématique où un ménage sur cinq reçoit un questionnaire détaillé portant sur une gamme de questions socioéconomiques. Les unités d'échantillonnage sont les ménages et toutes les personnes de plus de 15 ans dans les ménages échantillonnés sont interviewées. Habituellement, un questionnaire à remplir par procuration est utilisé, si bien qu'il n'y a pas de non-réponse individuelle dans le ménage. Dans la présente étude, nous supposons que chaque ménage a une même probabilité d'être inclus dans l'échantillon. L'étude porte sur les données au niveau du ménage $(N = 322\ 411)$.

Nous avons mis en œuvre un plan en deux étapes pour définir les propensions à répondre dans les données de population (recensement). Cette procédure permet de s'assurer d'avoir un modèle connu pour générer les propensions à répondre. En outre, afin d'étudier l'effet de la variation des taux de réponse et du nombre de variables auxiliaires dans le modèle sur la performance des estimateurs, nous avons considéré six scénarios définis par le niveau des taux de réponse (3 catégories) et le type de modèle (2 catégories).

A. Pour commencer, les probabilités de réponse ont été définies en fonction des variables suivantes : type de localité (4 catégories définies par rurale/urbaine et type de population), nombre de personnes dans le ménage (3 catégories : 1 ou 2, 3 à 5, 6 et plus), indicateur de présence d'enfants dans le ménage (oui, non), région (7 catégories divisant le pays du nord au sud), et densité (3 catégories : moins de 1,5, 1,5 à 3,0, supérieure à 3,0). Ces variables définissent des groupes pour lesquels on sait que les taux de réponse aux enquêtes sociales diffèrent en pratique. Pour étudier l'effet des taux de réponse sur la performance des estimateurs, les probabilités de réponse p ont été définies comme étant $p = p_1 p_2 p_3 p_4 p_5 + \alpha$ avec trois choix, $\alpha = 0{,}15$ (TR1), $\alpha = 0{,}55$ (TR2) et $\alpha = 0{,}75$ (TR3), où les probabilités p_1, p_2, p_3, p_4, p_5 sont données dans le tableau 4.1. Nous avons généré trois variables indicatrices de réponse en utilisant la loi de Bernoulli pour chacun des scénarios de réponse définis sous TR1, TR2, et TR3.

B. Pour chacun des scénarios de réponse de l'étape (A), nous utilisons l'indicateur de réponse comme variable dépendante et ajustons un modèle de régression linéaire ainsi qu'un modèle de régression logistique à la population, afin de prédire les « vraies » propensions à répondre pour notre étude d'évaluation sous les deux fonctions de lien. Nous avons considéré deux modèles différents pour prédire les « vraies » propensions à répondre. Dans le modèle 1, les variables indépendantes sont exactement les variables explicatives utilisées à l'étape A pour la définition des probabilités de réponse (indicateur de la présence d'enfants, nombre de personnes dans le ménage, région, type de localité, densité). Dans le modèle 2, les variables indépendantes sont le type de localité, le nombre de personnes dans le ménage, et l'indicateur de présence d'enfants. Soulignons que nous utilisons les mêmes variables indicatrices de réponse pour ajuster les deux modèles. Cela nous permet d'isoler l'effet du modèle, à l'exclusion des différences dues à la variabilité aléatoire de l'indicateur de réponse.

Les taux de réponse pour les variables définissant les probabilités, ainsi que les taux de réponse globaux et les vraies valeurs de population de l'indicateur R sous les deux modèles sont présentés au tableau 4.1. Aux fins de comparaison, nous donnons les valeurs de population de l'indicateur R basé sur les modèles de régression linéaire ainsi que logistique, où les taux de réponse varient entre 25,1 % et 35,1 % sous TR1, entre 64,7 % et 75,4 % sous TR2 et entre 84,7 % et 94,6 % sous TR3. TR2 représente le type de taux de réponse observé dans les enquêtes sociales nationales à grande échelle. Comme le montre le tableau 4.1, la différence est faible entre les valeurs de population des indicateurs R basés sur les fonctions de lien linéaire et logistique pour TR1 et TR2, et la différence est légère pour TR3 sous les deux modèles où les taux de réponse sont situés dans la queue supérieure de la distribution. Nous constatons aussi que les valeurs de population de l'indicateur R sont généralement élevées sur l'ensemble des taux de réponse globaux très différents.

Tableau 4.1
Probabilités de réponse et pourcentage de réponses générées dans l'ensemble de données de population utilisé pour l'évaluation en fonction des variables auxiliaires

			Pourcentage de réponses		
Variable	**Catégorie**	**Probabilité de réponse**	**TR1**	**TR2**	**TR3**
Enfants dans le ménage	Aucun	0,6	25,7	65,6	85,7
	1 et plus	0,8	35,1	75,4	94,6
Nombre de personnes dans le ménage	1 ou 2	0,5	24,6	64,5	84,7
	3 à 5	0,8	32,9	72,8	92,5
	6 et plus	0,7	29,9	70,3	90,0
Type de localité	Type 1	0,6	25,1	64,9	85,0
	Type 2	0,7	28,3	68,5	88,4
	Type 3	0,8	31,5	71,7	91,2
	Type 4	0,75	28,9	69,2	88,9
Région	1	0,6	25,1	65,1	84,7
	2	0,8	31,2	71,5	91,0
	3	0,7	28,1	67,6	87,8
	4	0,6	26,7	66,5	86,4
	5	0,6	24,8	64,7	84,9
	6	0,7	27,6	67,8	88,0
	7	0,8	30,3	70,4	90,9
Densité	≤ 1,5	0,6	26,1	66,0	86,2
	1,5 à 3,0	0,8	28,9	68,9	88,8
	> 3	0,7	24,7	64,7	84,7
Taux de réponse global			27,1	67,0	87,0
« Vrai » indicateur R de population (logistique)	Modèle 1		0,9031	0,9005	0,9063
	Modèle 2		0,9103	0,9074	0,9137
« Vrai » indicateur R de population (linéaire)	Modèle 1		0,9033	0,9006	0,9076
	Modèle 2		0,9104	0,9074	0,9145

Lorsqu'on utilise le modèle 2, le vrai indicateur R est systématiquement environ 0,007 point plus grand que la valeur correspondante sous le modèle 1. Cela tient au fait que le modèle 2 pour l'estimation des propensions à répondre est mal spécifié. Il contient un moins grand nombre de variables auxiliaires et donc, la variation des propensions à répondre estimées est plus petite, ce qui donne un indicateur R plus élevé. En conséquence, nous obtenons un indicateur R un peu plus élevé pour le modèle 2, car une partie de la variation n'est pas saisie par ce dernier. C'est pourquoi il est toujours important de communiquer les indicateurs R accompagnés de l'information auxiliaire utilisée pour les calculer, puisque leurs valeurs dépendent du modèle de non-réponse. En outre, nous devrions utiliser des covariables qui sont corrélées aux variables d'enquête (Schouten et coll., 2012).

Pour chaque scénario de réponse, nous avons tiré 500 échantillons de la population sous échantillonnage aléatoire simple (EAS) pour trois taux d'échantillonnage différents, soit 1 % $(n = 3\,224)$, 2 % $(n = 6\,448)$ et 4 % $(n = 12\,896)$. Pour chaque échantillon tiré, nous avons généré un indicateur de

réponse dans l'échantillon à partir des « vraies » probabilités de réponse de population basées sur la fonction de lien logistique. Cela détermine l'ensemble de réponses r. Nous avons ensuite estimé les propensions à répondre et les indicateurs R à partir de chaque échantillon pour les variables auxiliaires basées sur l'échantillon ainsi que pour les variables auxiliaires basées sur la population. Nous avons estimé les propensions à répondre dans l'échantillon en utilisant le « vrai » modèle (modèle 1 ou modèle 2, selon le scénario).

Afin d'estimer la variance des estimateurs basés sur la population, nous employons un algorithme bootstrap non paramétrique. De chaque ensemble de réponses, nous avons tiré $B = 500$ échantillons bootstrap par échantillonnage aléatoire simple (EAS) avec remise. Ensuite, nous avons généré la non-réponse dans l'échantillon bootstrap en copiant les valeurs 0 ou 1 de l'indicateur de réponse dans l'échantillon. Une répétition de l'estimateur a été calculée sur chaque échantillon bootstrap.

4.2 Résultats

Le tableau 4.2 donne les résultats de l'étude d'évaluation pour chaque scénario de taux de réponse, type de modèle et taux d'échantillonnage. Nous contrastons les indicateurs R basés sur l'échantillon (sous les deux fonctions de lien pour mettre en relief toute différence) avec les indicateurs R basés sur la population. Dans l'évaluation, nous examinons aussi la performance de l'estimateur composite basé sur la population (PC) donnée en (3.7).

Pour chaque estimateur, le tableau 4.2 montre : a) le biais relatif en pourcentage (BR %) calculé comme étant $100\left\{\left[\sum_{j=1}^{500}\left(\hat{R}_{\hat{\rho}j} - R_\rho\right)\Big/R_\rho\right]\Big/500\right\}$, où $\hat{R}_{\hat{\rho}j}$ est la valeur de l'estimateur calculé pour le j^{e} échantillon et R_ρ est le vrai indicateur R basé sur le modèle linéaire de régression (tiré du tableau 4.1), et similairement pour $\tilde{R}_{\tilde{\rho}_{T1}}$, $\tilde{R}_{\tilde{\rho}_{T2}}$, et l'estimateur composite; b) la racine carrée de l'erreur quadratique moyenne relative (REQMR) calculée comme étant

$$100\left\{R_\rho^{-1}\sqrt{\sum_{j=1}^{500}\left(\hat{R}_{\hat{\rho}j} - R_\rho\right)^2\Big/500}\right\}.$$

Le tableau 4.2 montre que les différences entre les estimateurs basés sur l'échantillon calculés en utilisant les fonctions de lien linéaire et logistique sont très petites en général, sauf quand le taux de réponse est très proche de 1 (TR3).

Tant pour les estimateurs basés sur l'échantillon que pour les estimateurs basés sur la population de type 1 et de type 2, nous observons un biais à la baisse général dans le cas des indicateurs R non ajustés, et voyons que ce biais tend à diminuer à mesure que la taille d'échantillon augmente, aussi bien pour le modèle 1 que pour le modèle 2. Cette constatation n'est pas étonnante. L'erreur d'échantillonnage a tendance à donner lieu à une surestimation de la variabilité des propensions à répondre estimées, ce qui mène à une sous-estimation de l'indicateur R. Le degré de sous-estimation est généralement plus important pour les estimateurs basés sur la population que pour ceux basés sur l'échantillon, surtout pour les taux de réponse élevés. La variation des propensions à répondre est plus grande dans ce cas que la variation sous

les variables auxiliaires basées sur l'échantillon. En outre, la REQMR des estimateurs diminue à mesure qu'augmente la taille de l'échantillon et elle est généralement plus grande pour les estimateurs basés sur la population. Donc, les indicateurs R basés sur la population sont en général moins précis que leurs analogues basés sur l'échantillon, et donnent lieu à des conclusions plus faibles concernant la nature de la réponse.

Tableau 4.2
Propriétés des indicateurs R estimés pour les variables auxiliaires basées sur l'échantillon et basées sur la population pour 500 échantillons dans l'étude d'évaluation

Taux de réponse	Taux d'échantillonnage	Estimateur	Modèle 1				Modèle 2			
			Non ajusté		Ajusté		Non ajusté		Ajusté	
			BR %	REQMR %	BR %	REQMR %	BR %	REQMR %	BR %	REQMR %
TR1	1 %	Basé sur l'éch. (log)	-1,73	2,39	0,32	2,01	-0,77	1,88	0,34	1,96
		Basé sur l'éch. (lin)	-1,71	2,37	0,33	2,01	-0,75	1,87	0,35	1,95
		Type 1	-2,32	3,08	0,32	2,54	-1,08	2,32	0,30	2,39
		Type 1 - PC	0,04	2,28	0,22	2,42	0,59	2,44	0,38	2,41
		Type 2	-1,47	2,29	1,06	2,50	-0,20	1,74	1,01	2,27
		Type 2 - PC	0,71	2,11	0,94	2,34	1,19	2,32	1,05	2,28
	2 %	Basé sur l'éch. (log)	-0,90	1,53	0,14	1,36	-0,41	1,30	0,14	1,31
		Basé sur l'éch. (lin)	-0,89	1,51	0,16	1,36	-0,40	1,29	0.15	1,31
		Type 1	-1,24	1,89	0,12	1,61	-0,51	1,57	0,17	1,59
		Type 1 - PC	0,04	1,56	0,10	1,59	0,38	1,68	0,21	1,62
		Type 2	-0,45	1,30	0,84	1,64	0,26	1,31	0,86	1,63
		Type 2 - PC	0,72	1,53	0,82	1,61	1,02	1,75	0,89	1,66
	4 %	Basé sur l'éch. (log)	-0,48	1,00	0,05	0,93	-0,27	0,90	0,00	0,88
		Basé sur l'éch. (lin)	-0,46	0,99	0,06	0,92	-0,26	0,89	0,01	0,88
		Type 1	-0,63	1,23	0,05	1,12	-0,34	1,13	-0,01	1,11
		Type 1 - PC	0,15	1,14	0,07	1,12	0,18	1,18	0,01	1,13
		Type 2	0,12	0,92	0,78	1,25	0,40	1,01	0,69	1,19
		Type 2 - PC	0,83	1,29	0,79	1,26	0,83	1,30	0,70	1,20
TR2	1 %	Basé sur l'éch. (log)	-1,81	2,44	0,33	2,01	-0,76	1,83	0,34	1,94
		Basé sur l'éch. (lin)	-1,79	2,42	0,34	2,01	-0,75	1,82	0,35	1,94
		Type 1	-5,17	5,95	-0,01	3,95	-2,45	3,77	0,25	3,43
		Type 1 - PC	-1,50	3,58	-0,47	3,69	0,69	3,37	0,49	3,46
		Type 2	-4,76	5,50	0,27	3,75	-1,95	3,29	0,58	3,23
		Type 2 - PC	-1,13	3,28	-0,12	3,51	0,74	3,13	0,71	3,25
	2 %	Basé sur l'éch.(log)	-1,00	1,59	0,08	1,37	-0,40	1,29	0,14	1,30
		Basé sur l'éch. (lin)	-0,98	1,57	0,09	1,36	-0,40	1,28	0,14	1,30
		Type 1	-2,89	3,55	0,07	2,59	-1,19	2,58	0,37	2,72
		Type 1 - PC	-0,57	2,37	-0,12	2,49	0,53	2,67	0,41	2,69
		Type 2	-2,52	3,19	0,39	2,50	-0,79	2,28	0,69	2,63
		Type 2 - PC	-0,26	2,19	0,19	2,37	0,81	2,58	0,71	2,60
	4 %	Basé sur l'éch. (log)	-0,48	0,98	0,07	0,90	-0,16	0,81	0,12	0,83
		Basé sur l'éch. (lin)	-0,46	0,97	0,08	0,90	-0,15	0,81	0,12	0,82
		Type 1	-1,42	2,12	0,13	1,81	-0,60	1,66	0,16	1,67
		Type 1 - PC	0,16	1,77	0,14	1,80	0,37	1,76	0,20	1,69
		Type 2	-1,07	1,82	0,46	1,78	-0,25	1,47	0,47	1,63
		Type 2 - PC	0,45	1,72	0,46	1,75	0,65	1,73	0,50	1,66
TR3	1 %	Basé sur l'éch. (log)	-1,07	1,59	0,10	1,30	-0,52	1,21	0,02	1,16
		Basé sur l'éch. (lin)	-0,85	1,40	0,24	1,26	-0,41	1,13	0,10	1,13
		Type 1	-6,60	7,32	-0,76	4,24	-3,20	4,61	0,06	4,12
		Type 1 - PC	-2,22	4,15	-0,88	4,16	-0,28	3,70	0,09	3,92
		Type 2	-6,29	6,99	-0,53	4,08	-2,85	4,25	0,27	3,95
		Type 2 - PC	-2,12	3,97	-0,67	4,02	-0,04	3,52	0,33	3,78
	2 %	Basé sur l'éch. (log)	-0,73	1,13	-0,14	0,92	-0,30	0,88	-0,03	0,85
		Basé sur l'éch. (lin)	-0,54	0,98	0,01	0,87	-0,20	0,82	0,06	0,82
		Type 1	-3,70	4,31	0,12	2,93	-1,74	2,98	0,20	2,86
		Type 1 - PC	-0,78	2,60	-0,15	2,78	0,42	2,81	0,36	2,94
		Type 2	-3,46	4,07	0,30	2,87	-1,46	2,73	0,41	2,77
		Type 2 - PC	-0,61	2,47	0,02	2,70	0,64	2,74	0,57	2,87
	4 %	Basé sur l'éch. (log)	-0,46	0,77	-0,16	0,66	-0,18	0,57	-0,05	0,55
		Basé sur l'éch. (lin)	-0,29	0,66	-0,01	0,61	-0,09	0,53	0,04	0,53
		Type 1	-1,96	2,62	0,12	2,12	-0,89	1,81	0,13	1,76
		Type 1 - PC	-0,03	1,97	0,07	2,06	0,38	1,84	0,19	1,79
		Type 2	-1,74	2,42	0,31	2,07	-0,66	1,65	0,31	1,71
		Type 2 - PC	0,11	1,89	0,25	2,00	0,56	1,81	0,38	1,75

En général, les estimateurs composites basés sur la population (PC) non ajustés ont de meilleures propriétés que les estimateurs basés sur la population non ajustés correspondants, tant en ce qui concerne le BR que la REQMR en pourcentage, surtout pour les taux de réponse élevés. Ils présentent encore un certain degré de surestimation sous le modèle 1 correct quand les taux de réponse sont faibles, et de sous-estimation quand les taux de réponse sont élevés. Toutefois, pour le modèle 2, nous voyons une surestimation.

Passons maintenant aux indicateurs R estimés et ajustés pour le biais du tableau 4.2. Pour le type 1, l'ajustement pour le biais permet d'éliminer ce dernier. La correction analytique du biais pour l'estimateur basé sur la population de type 1 fonctionne bien et donne généralement de meilleurs résultats que la correction analytique du biais pour l'estimateur basé sur la population de type 2. Il semble capter la plupart du biais et fournit des estimations ajustées qui sont plus proches des indicateurs R basés sur l'échantillon. La REQMR de l'estimateur ajusté pour le biais est généralement similaire à la REQMR correspondante de l'estimateur non ajusté, ce qui signifie que l'accroissement de la variabilité est compensé par la réduction du biais. Pour les taux de réponse élevés, l'estimation composite ajustée basée sur la population réduit le biais et la REQMR des indicateurs R basés sur la population correspondants.

Non ajusté, l'indicateur R de type 2 a un meilleur comportement que l'indicateur R de type 1. Ce résultat est plutôt surprenant, car il semble que nous arrivons à obtenir une estimation plus précise du vrai indicateur R lorsque nous utilisons moins d'information. Cette situation tient au fait que, pour l'estimateur de type 1, nous n'incluons aucune part de la variation d'échantillonnage quand nous « substituons » la matrice de covariance de population, tandis que pour l'estimateur de type 2, nous utilisons uniquement l'information de marge et « substituons » la matrice de covariance des réponses qui tient compte d'une plus grande part de la variation d'échantillonnage. Après ajustement pour le biais, les estimateurs de type 2 ont un BR % plus élevé (surtout pour les faibles taux de réponse), mais une REQMR similaire. L'ajustement du biais pour le type 2 donne de moins bons résultats que l'ajustement du biais pour le type 1 et surcompense le biais. Nous nous attendions à ce résultat, car la correction du biais pour l'estimateur de type 2 est fondée sur une approximation linéaire, tandis qu'elle est calculée exactement pour l'estimateur de type 1.

En ce qui concerne l'augmentation des taux de réponse, étonnamment, dans le cas des estimateurs basés sur la population non ajustés, nous observons de meilleurs résultats pour les faibles taux de réponse, tant en ce qui concerne le biais relatif en pourcentage (BR %) que la REQMR. Celle-ci est deux à trois fois plus grande pour TR3 que pour TR1. Les ajustements analytiques du biais donnent de très bons résultats sous tous les taux de réponse, quoique avec des REQMR plus élevées pour les taux de réponse élevés. L'utilisation de l'estimateur composites réduit ces REQMR.

Pour ce qui est de l'effet du nombre de variables dans le modèle, le BR % et la REQMR observés sont plus faibles sous le modèle 2 que sous le modèle 1 pour les estimateurs basés sur la population non ajustés. Les estimateurs composites présentent en général une tendance opposée. La performance des versions ajustées pour le biais est similaire sous les deux modèles.

Le tableau 4.3 donne la moyenne du λ_{opt} estimé pour les estimateurs composites basés sur la population de type 1 et de type 2 comparativement à la vraie valeur obtenue pour la population sous les deux scénarios

de taux de réponse extrêmes, TR1 et TR3. L'examen du tableau révèle que le λ_{opt} estimé moyen s'écarte peu des vraies valeurs dans l'étude d'évaluation.

Tableau 4.3
λ_{opt} moyen pour les variables auxiliaires basées sur la population pour 500 échantillons dans l'étude d'évaluation

Taux de réponse	Taux d'échantillonnage	Modèle 1				Modèle 2			
		Type 1		Type 2		Type 1		Type 2	
		Vrai	Basé sur la pop.	Vrai	Basé sur la pop.	Vrai	Basé sur la pop.	Vrai	Basé sur la pop.
TR1	1 %	0,40	0,33	0,36	0,33	0,31	0,29	0,26	0,28
	2 %	0,25	0,21	0,22	0,21	0,19	0,22	0,15	0,19
	4 %	0,14	0,13	0,13	0,13	0,10	0,10	0,08	0,09
TR3	1 %	0,68	0,44	0,67	0,44	0,57	0,51	0,55	0,48
	2 %	0,51	0,39	0,50	0,38	0,41	0,43	0,39	0,41
	4 %	0,35	0,27	0,34	0,27	0,25	0,23	0,24	0,22

Au tableau 4.4, nous analysons la performance des estimateurs bootstrap pour l'estimation de la variance des indicateurs R basés sur la population sous les deux scénarios de taux de réponse extrêmes, TR1 et TR3. Des expressions analytiques ont été établies pour la variance des indicateurs R basés sur l'échantillon et utilisées dans l'étude d'évaluation (voir Shlomo et coll., 2012). Les moyennes de simulation des estimateurs de variance sont comparées au tableau 4.4 avec les variances de simulation (calculées sur l'ensemble des échantillons répétés), en utilisant le biais relatif en pourcentage. Le tableau inclut aussi le taux de couverture défini comme étant le pourcentage de fois que le vrai R_ρ est contenu dans l'intervalle de confiance $100\left\{\left[\sum_{j=1}^{500} I\left(R_\rho \in \hat{R}_{\hat{\rho}j} \pm 1{,}96\sqrt{\hat{V}_j\left(\hat{R}_{\hat{\rho}j}\right)}\right)\right] \Big/ 500\right\}$, où $\hat{V}_j\left(\hat{R}_{\hat{\rho}j}\right)$ est la variance estimée pour le j^{e} échantillon (estimateur de variance par linéarisation pour l'estimateur basé sur l'échantillon et estimateur de variance par le bootstrap pour les estimateurs basés sur la population) et I est la fonction indicatrice. Les estimateurs de variance bootstrap pour les estimateurs basés sur la population donnent de bons résultats. L'estimateur basé sur l'échantillon présente une meilleure couverture que les versions basées sur la population correspondantes. Les estimateurs basés sur la population de type 1 et de type 2 ont des couvertures similaires. La couverture s'améliore toujours quand la taille d'échantillon augmente.

Pour ce qui est du comportement sous les différents taux de réponse, les résultats sont partagés. Il semble exister une interaction entre la taille de l'échantillon et le taux de réponse. Le nombre de variables dans le modèle n'a pas grand effet sur la couverture. Cependant, nous constatons des problèmes concernant la couverture pour les estimateurs basés sur la population sous le taux de réponse le plus élevé (TR3), en particulier pour le taux d'échantillonnage de 1 %.

Les figures 4.1, 4.2 et 4.3 présentent les boîtes à moustaches comparant les estimateurs et leurs versions corrigées du biais sous le modèle 1, pour les différents scénarios de taux de réponse TR1, TR2 et TR3, respectivement. Les gains découlant des ajustements pour le biais sont évidents pour les indicateurs R de type 1 et de type 2. Les erreurs-types sont beaucoup plus grandes pour TR3 que pour TR1 sous les mêmes taux d'échantillonnage. La variabilité de l'estimateur ajusté pour le biais augmente et est plus grande pour les petites tailles d'échantillon.

Tableau 4.4
Propriétés des estimateurs de variance pour les indicateurs R sous les variables auxiliaires basées sur l'échantillon et basées sur la population pour 500 échantillons

Taux de réponse	Taux d'échantillonnage	Estimateur	Modèle 1		Modèle 2	
			BR %	Couverture	BR %	Couverture
TR1	1 %	Basé sur l'échantillon	1,84	0,95	-5,74	0,95
		Type 1	4,35	0,95	11,12	0,96
		Type 2	4,99	0,94	7,72	0,95
	2 %	Basé sur l'échantillon	1,43	0,96	1,15	0,95
		Type 1	8,62	0,96	5,31	0,95
		Type 2	7,03	0,93	2,10	0,92
	4 %	Basé sur l'échantillon	7,93	0,97	-4,58	0,95
		Type 1	13,23	0,96	3,42	0,95
		Type 2	13,38	0,89	2,53	0,90
TR3	1 %	Basé sur l'échantillon	-1,05	0,95	-9,48	0,92
		Type 1	2,87	0,78	11,47	0,86
		Type 2	4,97	0,78	10,26	0,85
	2 %	Basé sur l'échantillon	-4,34	0,94	-7,96	0,94
		Type 1	-7,61	0,92	2,37	0,91
		Type 2	-8,07	0,92	1,02	0,90
	4 %	Basé sur l'échantillon	3,31	0,94	-3,54	0,95
		Type 1	-8,33	0,93	12,32	0,96
		Type 2	-8,13	0,93	10,89	0,96

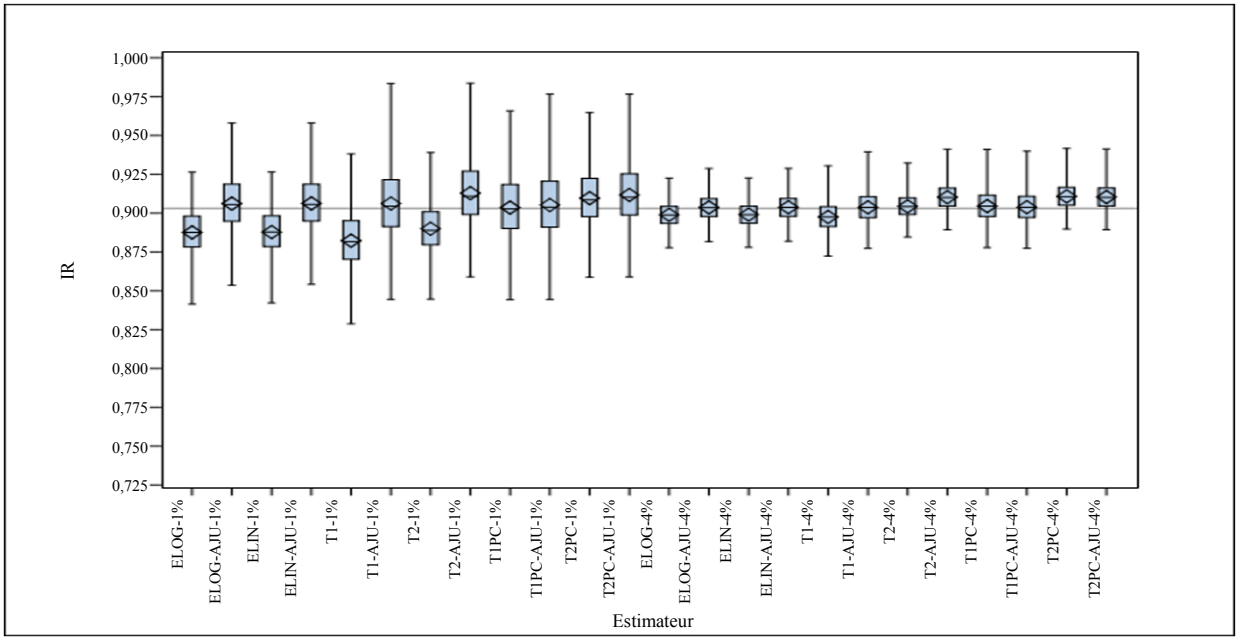

Figure 4.1 Boîtes à moustaches pour 500 indicateurs R estimés pour des échantillons à 1 % et 4 % pour le modèle 1 et TR1. (ELOG) désigne l'indicateur R logistique basé sur l'échantillon, (ELIN) l'indicateur R linéaire basé sur l'échantillon, (T1) l'indicateur R basé sur la population de type 1, (T2) l'indicateur R basé sur la population de type 2, et (T1PC) et (T2PC) les estimateurs composites basés sur la population de type 1 et de type 2. AJU désigne les estimateurs ajustés pour le biais correspondants.

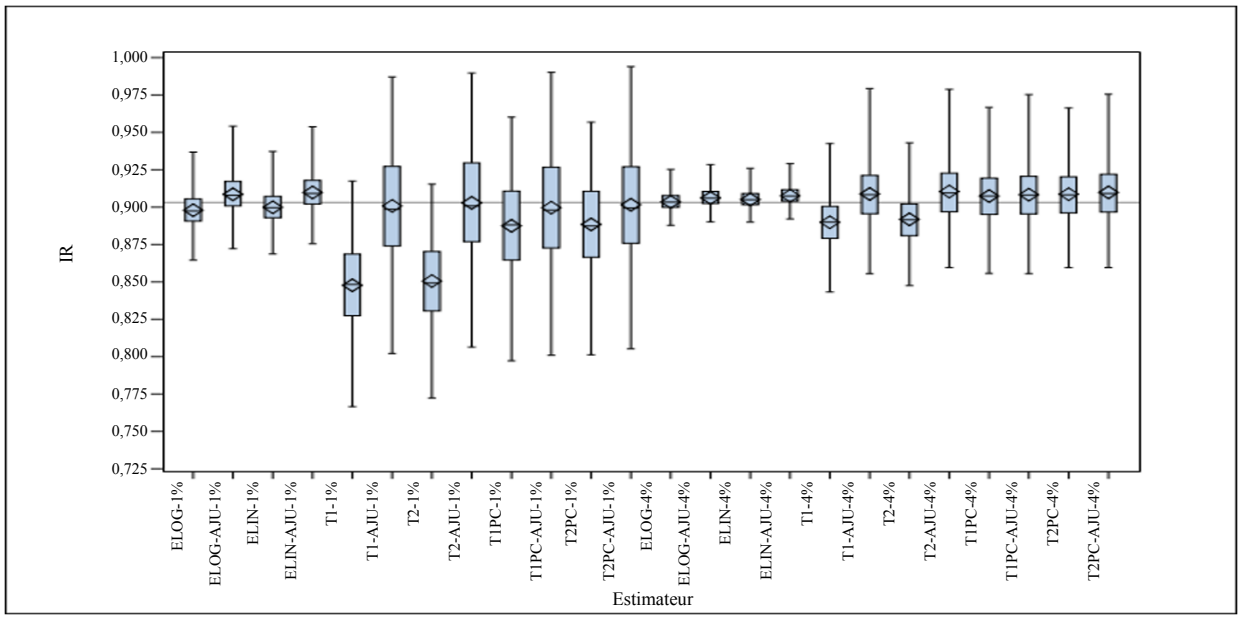

Figure 4.2 Boîtes à moustaches pour 500 indicateurs R estimés pour des échantillons à 1 % et 4 % pour le modèle 1 et TR2. (ELOG) désigne l'indicateur R logistique basé sur l'échantillon, (ELIN) l'indicateur R linéaire basé sur l'échantillon, (T1) l'indicateur R basé sur la population de type 1, (T2) l'indicateur R basé sur la population de type 2, et (T1PC) et (T2PC) les estimateurs composites basés sur la population de type 1 et de type 2. AJU désigne les estimateurs ajustés pour le biais correspondants.

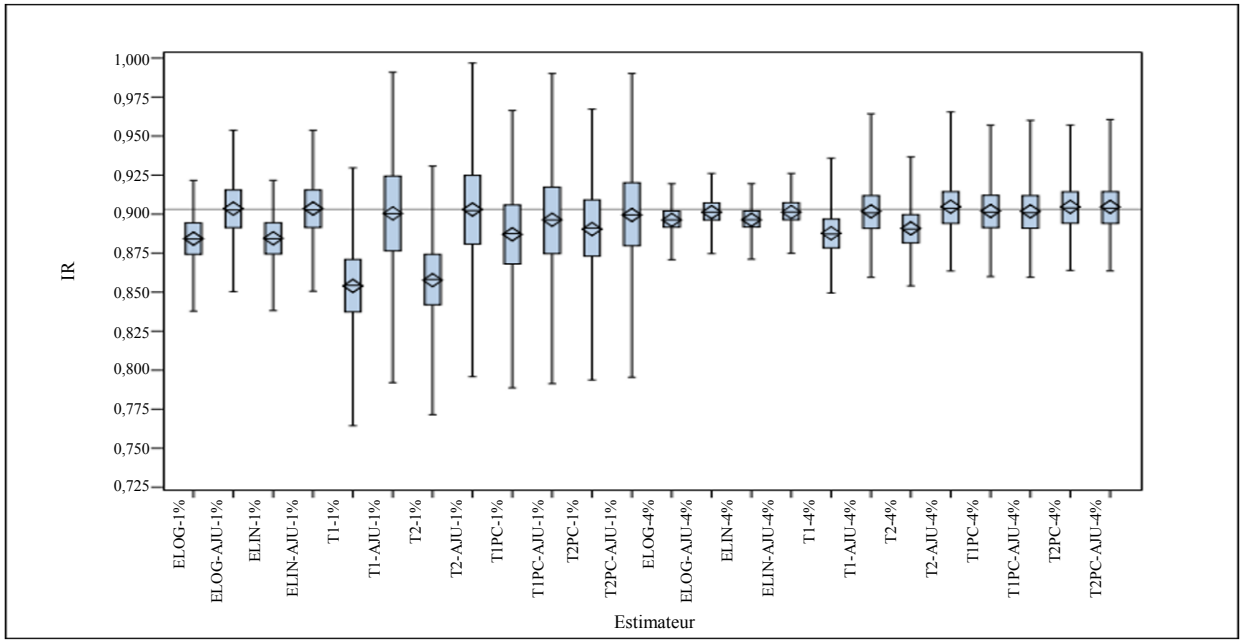

Figure 4.3 Boîtes à moustaches pour 500 indicateurs R estimés pour des échantillons à 1 % et 4 % pour le modèle 1 et TR3. (ELOG) désigne l'indicateur R logistique basé sur l'échantillon, (ELIN) l'indicateur R linéaire basé sur l'échantillon, (T1) l'indicateur R basé sur la population de type 1, (T2) l'indicateur R basé sur la population de type 2, et (T1PC) et (T2PC) les estimateurs composites basés sur la population de type 1 et de type 2. AJU désigne les estimateurs ajustés pour le biais correspondants.

5 Application à la *Dutch Health Survey*

À la présente section, nous appliquons les estimateurs basés sur la population de type 1 et de type 2 à la *Dutch Health Survey* réalisée par *Statistics Netherlands*. Nous utilisons trois variables auxiliaires qui font partie de la norme de référence adoptée par les entreprises néerlandaises d'études de marché et comparons la performance des estimateurs basés sur la population à celle des estimateurs basés sur l'échantillon.

La *Dutch Health Survey* (HS) a été commanditée en 1998 sous forme d'une enquête transversale répétée auprès de l'entièreté de la population inscrire dans le Registre de la population des Pays-Bas, à l'exclusion de la population vivant en établissement. L'enquête est réalisée selon un plan de sondage autopondéré à deux degrés dans lequel le premier degré correspond aux municipalités et le deuxième, aux personnes vivant dans les municipalités sélectionnées. Jusqu'à 2012, la HS était une enquête avec interviews sur place. En 2012, elle est passée à un plan de collecte à mode mixte comprenant des interviews en ligne et des interviews sur place. Au fil des ans, la taille de l'échantillon a été réduite considérablement, pour passer d'environ 35 000 à environ 18 000. Nous utilisons les données de la HS de 2002, l'une des dernières années où l'échantillon avait sa taille originale. La taille d'échantillon nette est de 33 584 personnes.

Pour les échantillons nationaux et régionaux, les entreprises néerlandaises d'études de marché utilisent les statistiques de population qu'il est convenu d'appeler norme de référence produite par *Statistics Netherlands* (MOA, 2015). La norme de référence est un ensemble explicitement défini de variables auxiliaires que les entreprises affiliées incluent dans leurs questionnaires d'enquête. Trois de ces variables sont l'âge, le sexe et l'état matrimonial. Nous nous concentrons sur ces trois variables dans l'application.

Le tableau 5.1 contient l'échantillon de la HS et les distributions des réponses, ainsi que les distributions des trois variables dans la population fournies par *Statistics Netherlands*. Les distributions de populations conjointes, nécessaires pour estimer les matrices de covariance basées sur la population de type 1, sont également disponibles, mais ne sont pas présentées ici. En pratique, la distribution dans l'échantillon est, évidemment, inconnue. Le tableau diffère pour les trois variables : pour l'âge et l'état matrimonial, la distribution des réponses est plus proche de la distribution dans l'échantillon que de la distribution dans la population, et les propensions à répondre basées sur la population donnent une plus grande variation. Pour le sexe, la distribution dans la population est plus proche de la distribution des réponses et nous observons moins de variation.

Tableau 5.1
Distributions de l'âge, du sexe et de l'état matrimonial pour l'échantillon, les répondants et la population

Variables	Catégories	Répondants	Échantillon	Population
Âge	20 à 24	7,5	7,9	8,1
	25 à 29	7,3	8,2	8,9
	30 à 34	9,9	10,2	10,9
	35 à 39	10,9	10,8	11
	40 à 44	10,3	10,3	10,4
	45 à 49	9,7	9,4	9,6
	50 à 54	9,4	9,6	9,5
	55 à 59	8,8	8,9	8
	60 à 64	7,1	6,7	6,3
	65 à 69	5,9	5,6	5,4
	70 à 74	5,4	4,7	4,6
	75 et +	7,7	7,8	7,2

Tableau 5.1 (suite)
Distributions de l'âge, du sexe et de l'état matrimonial pour l'échantillon, les répondants et la population

Variables	Catégories	Répondants	Échantillon	Population
Sexe	Hommes	48,9	49,8	49,2
	Femmes	51,1	50,2	50,8
État matrimonial	Non marié(e)	23,7	26,8	26,9
	Marié(e)	63,3	59,3	58,8
	Veuf(ve)	6,5	6,7	6,7
	Divorcé(e)	6,4	7,2	7,6

Nous donnons les estimations des indicateurs R basés sur la population de type 1 et de type 2 au tableau 5.3. Pour l'estimateur composite, nous avons utilisé le paramètre de lissage estimé $\tilde{\lambda}_{\text{opt}}$ fondé sur les propensions à répondre basées sur la population. Nous incluons aussi une estimation de λ_{opt} calculée en utilisant les propensions à répondre basées sur l'échantillon. Cette dernière estimation, qui ne peut ordinairement pas être calculée, est incluse aux fins de comparaison. Le tableau 5.2 donne le paramètre de lissage estimé $\tilde{\lambda}_{\text{opt}}$ fondé sur les propensions à répondre basées sur la population, ainsi que basées sur l'échantillon. Les $\tilde{\lambda}_{\text{opt}}$ basés sur l'échantillon sont plus grands et ont tendance à avoir un effet de lissage plus prononcé. Cependant, tous les $\tilde{\lambda}_{\text{opt}}$ sont relativement petits.

Le tableau 5.3 présente les divers indicateurs R basés sur la population. Aux fins de comparaison, nous donnons aussi l'indicateur R basé sur l'échantillon en utilisant la fonction de lien logistique. La fonction de lien linéaire a produit les mêmes résultats. Nous pouvons conclure que les indicateurs R basés sur la population, en utilisant uniquement les distributions des réponses et de la population, diffèrent des indicateurs R basés sur l'échantillon, en utilisant les distributions des réponses et de l'échantillon. Cette différence augmente, comme prévu, quand les indicateurs de type 2 sont utilisés. Les estimateurs composites donnent d'un peu meilleurs résultats que les estimateurs non composites, mais une différence importante persiste encore. Cela n'est pas attribuable à un paramètre de lissage biaisé, car la différence n'est qu'un peu plus faible quand les propensions à répondre basées sur l'échantillon sont utilisées pour estimer le paramètre de lissage. En outre, après ajustement pour tenir compte du biais, la différence entre les estimateurs composites disparaît pour les propensions à répondre basées sur l'échantillon et basées sur la population.

Tableau 5.2
Valeurs du paramètre de lissage λ_{opt} fondé sur les propensions à répondre basées sur la population et sur les propensions à répondre basées sur l'échantillon pour les estimateurs composites de type 1 et de type 2

	Paramètre de lissage $\tilde{\lambda}_{\text{opt}}$	
	Type 1	Type 2
Propensions à répondre basées sur la population	0,043	0,038
Propensions à répondre basées sur l'échantillon	0,076	0,095

Tableau 5.3
Indicateurs R basés sur l'échantillon et basés sur la population de type 1 et de type 2, non ajustés et ajustés pour le biais, pour les données de la HS de 2002. Les indicateurs R composites basés sur la population sont fondés sur le paramètre de lissage λ_{opt} obtenu en utilisant les propensions à répondre basées sur la population et basées sur l'échantillon. Les intervalles de confiance (IC) à 95 % obtenus par approximation normale sont présentés

	Non ajusté			Ajusté pour le biais		
Estimateur	**Indicateur R**	**IC à 95 %**		**Indicateur R**	**IC à 95 %**	
Basé sur l'échantillon	0,899	0,888	0,909	0,901	0,890	0,912
Type 1 – original	0,876	0,860	0,891	0,879	0,864	0,895
Type 1 – composite basé sur la population	0,880	0,865	0,896	0,880	0,864	0,895
Type 1 – composite basé sur l'échantillon	0,883	0,868	0,898	0,880	0,865	0,895
Type 2 – original	0,873	0,858	0,889	0,877	0,861	0,894
Type 2 – composite basé sur la population	0,878	0,863	0,894	0,878	0,862	0,893
Type 2 – composite basé sur l'échantillon	0,881	0,866	0,897	0,878	0,863	0,893

L'une des conclusions de l'application est que les valeurs plus faibles des indicateurs R basés sur la population résultent des grands écarts entre les distributions des variables auxiliaires dans l'échantillon et dans la population. Pour une taille d'échantillon de 33 584 personnes, un test des différences entre les distributions dans l'échantillon et dans la population est significatif pour chacune des trois variables au seuil de signification de 5 %. L'échantillon net disponible de la *Dutch Health Survey* ne contient pas les unités d'échantillonnage présentant des erreurs de base de sondage ou d'autres erreurs administratives, ni les populations hors du champ d'observation de l'enquête, comme les personnes vivant en établissement. Cette modification, ainsi que certains petits ajustements des charges de travail des intervieweurs, ont très vraisemblablement fait s'écarter les distributions d'échantillon des chiffres de population originaux. Cela met en relief le « tendon d'Achille » des indicateurs R basés sur la population, à savoir qu'il est impératif qu'il n'y ait pas de disparité entre les définitions et les populations.

6 Discussion

L'extension des estimateurs des indicateurs R basés sur l'échantillon à ceux basés sur la population comprend deux étapes, à savoir 1) l'estimation des propensions à répondre, et 2) l'estimation des indicateurs R basée sur ces propensions. L'estimation des propensions à répondre basées sur la population est simple quand on émet l'hypothèse de modèles linéaires pour les propensions à répondre et les influences des réponses. La fonction de lien linéaire est raisonnable lorsqu'on estime les propensions à répondre sous des taux de réponse habituellement observés pour les grandes enquêtes sociales nationales, comme le montre l'étude d'évaluation présentée à la section 4. Les estimateurs basés sur l'échantillon contiennent des matrices de covariance d'échantillon et des fréquences d'échantillon qui peuvent être remplacées par des matrices de covariance de population ou des fréquences de population. Nous avons cerné deux types de conditions, d'une part, les cas où les produits croisés de population sont disponibles et d'autre part, les cas où l'information auxiliaire est limitée uniquement aux chiffres de marge de la population. Nous avons dénommé les estimateurs correspondants estimateurs de type 1 et de type 2, respectivement. Les conditions de type 2 sont plus contraignantes que celles de type 1.

Après avoir estimé les propensions à répondre basées sur la population, nous avons construit des estimateurs basés sur la population pour l'indicateur R et examiné leurs propriétés théoriquement ainsi qu'empiriquement. Les estimateurs sont appliqués à des échantillons tirés de données réelles provenant du Recensement d'Israël de 1995, où les « vraies » propensions à répondre ont été calculées en s'appuyant sur des hypothèses réalistes pour les enquêtes sociales nationales auprès des ménages. Nous avons donc abordé les deux premières questions de recherche du début de l'article : comment étendre les propensions à répondre et les indicateurs R basés sur l'échantillon aux propensions à répondre et aux indicateurs R basés sur la population ? Quelles sont les propriétés statistiques des indicateurs R basés sur la population ?

De nombreuses options pour l'estimation des indicateurs R existent, selon la réponse à l'enquête. Nous avons utilisé les moyennes des réponses pondérées par la propension à répondre, car les propensions sont disponibles. Cependant, n'importe quelle méthode de calage peut être utilisée, dont la méthode de pondération linéaire ou la méthode des classes d'ajustement. En fait, l'ensemble de variables auxiliaires utilisé pour l'estimation des indicateurs R peut être un sous-ensemble des variables auxiliaires employées pour l'estimation des propensions et des influences. Des modèles parcimonieux peuvent s'avérer plus efficaces, car il est connu que la pondération par les propensions à répondre peut influencer fortement la précision des estimateurs. Il s'agit d'un sujet pour de futurs travaux de recherche.

Les deux propriétés que nous avons examinées sont le biais et l'erreur-type des indicateurs R basés sur la population proposés. Comme prévu, le biais et l'erreur-type dépendent de la taille de l'échantillon et du type d'information auxiliaire disponible, le biais et l'erreur-type étant d'autant plus grands que l'échantillon est petit. Quand les échantillons sont petits, il devient plus difficile de faire la distinction entre la variation d'échantillonnage et la variation de la réponse. Clairement, les intervalles de confiance deviennent plus grands, car les petits échantillons fournissent moins d'information.

Les estimateurs de type 1 ajustés pour le biais (produits croisés de population) donnent de meilleurs résultats que les estimateurs de type 2 ajustés pour le biais (chiffres de marge de la population). Ce résultat n'est pas étonnant étant donné que les premiers utilisent plus d'information. Cependant, les estimateurs de type 2 non ajustés ont de meilleures propriétés de REQMR que les estimateurs de type 1 non ajustés. Ce résultat est surprenant et souligne un usage sous-optimal des produits croisés de population quand ils sont utilisés comme des « valeurs de substitution » et ne tiennent compte d'aucune variation d'échantillonnage. Les erreurs-types des estimateurs basés sur la population sont plus grandes que celles de leurs analogues basés sur l'échantillon.

L'étude d'évaluation montre dans le scénario TR3 que, pour les taux de réponse très élevés, les indicateurs R basés sur la population donnent des erreurs-types plus élevées et de plus grands biais, principalement attribuables au fait que les propensions à répondre sont estimées en dehors de l'intervalle [0, 1]. Pour cette raison, nous proposons un estimateur composite utilisant divers paramètres de lissage en fonction du taux de réponse. Les erreurs-types sont réduites, mais au prix d'un accroissement du biais.

Les analyses révèlent que le biais des estimateurs de type 1 et de type 2 dépend du nombre de variables auxiliaires; toutefois, cette dépendance était modeste dans nos évaluations. Le biais peut augmenter si l'on utilise des modèles détaillés contenant de nombreuses variables pour l'estimation des propensions à

répondre. Cela s'explique par le fait que les modèles détaillés permettent de saisir une plus grande part de la variation d'échantillonnage sous forme de biais.

Un certain nombre de mises en garde s'appliquent aux indicateurs R basés sur la population.

Premièrement, le choix d'information auxiliaire disponible au niveau national peut être plus limité que l'information auxiliaire basée sur l'échantillon, selon l'existence de registres et de données administratives. La sélection des variables auxiliaires doit se faire en fonction de leur corrélation avec les variables cibles de l'enquête. En outre, il est vivement recommandé d'utiliser les statistiques de population fondées sur des registres ou des données administratives plutôt que celles basées sur des chiffres pondérés provenant d'autres enquêtes, puisque ces statistiques pourraient ne pas refléter précisément la vraie distribution dans la population. On tirerait des conclusions erronées au sujet de la représentativité des réponses si les estimations de population sont biaisées.

Deuxièmement, nous formulons l'hypothèse que l'enquête mesure les mêmes quantités que celles dans l'information sur la population et nous n'examinons pas l'effet d'écarts éventuels par rapport à cette hypothèse. Cependant, nous soulignons qu'il existe un risque d'erreurs de mesure lorsqu'on compare la représentativité des questions d'enquête aux statistiques de population. Il convient de confirmer que les questions d'enquête utilisées ont les mêmes définitions et classifications que les tableaux démographiques. Donc, il est préférable d'éviter les questions propices aux erreurs de mesure, comme celles qui requièrent un important effort cognitif ou qui peuvent entraîner des réponses socialement désirables.

Troisièmement, dans les situations où l'information n'est disponible que pour la population, les options en vue d'améliorer la représentativité durant la collecte des données sont nettement plus limitées, puisqu'on ne dispose d'aucune information auxiliaire individuelle pour les non-répondants. Néanmoins, dans ces conditions, les évaluations de la représentativité peuvent encore être utiles pour la conception des lettres préalables et de rappel, la formation des intervieweurs et la collecte de paradonnées.

Enfin, nous ne considérons pas les conditions hybrides où l'indicateur R est basé sur des données appariées et des tableaux de données de population. En outre, nous ne traitons pas le cas où nous pourrions utiliser des estimations d'enquête pondérées s'il n'existe pas d'information agrégée sur la population. Cela aura une incidence tant sur le biais que sur les estimations de la variance pour les indicateurs R basés sur la population. Ces extensions sont relativement simples, mais seront abordées dans de futurs articles.

La recherche sur les indicateurs R basés sur la population en est encore à ses débuts et il est trop tôt pour fournir une réponse catégorique à la dernière question de recherche présentée dans l'introduction concernant la faisabilité et la praticabilité des indicateurs R basés sur de l'information auxiliaire de population agrégée. Comme il est mentionné dans l'introduction, d'autres usages de ces indicateurs R sont étudiés dans le contexte de l'évaluation et de la surveillance des données administratives recueillies en continu et de l'évaluation de la représentativité des enregistrements appariés. De plus, Schouten et coll. (2011) ont introduit des indicateurs R partiels sous information auxiliaire basée sur l'échantillon pour évaluer le manque de représentativité dû à une variable ou à une catégorie auxiliaire particulière. Ces indicateurs partiels ont été utilisés pour surveiller et évaluer la collecte des données. Schouten et Shlomo (2017)

illustrent l'utilisation d'indicateurs R partiels pour les plans de collecte de données adaptatifs. Similairement, il est facile de définir des indicateurs R partiels basés sur la population et cela sera le sujet de futurs travaux.

Quant à l'étude d'évaluation présentée à la section 4 portant sur la représentativité des enquêtes, elle est fondée sur des données réelles sous des hypothèses réalistes au sujet des probabilités de réponse habituellement retrouvées dans les enquêtes sociales réalisées par les instituts nationaux de statistique. De futurs travaux de recherche seront nécessaires afin de voir s'il est possible de construire des estimateurs de rechange plus précis et, par conséquent, permettant de tirer des conclusions plus fermes concernant la nature de la réponse. Une piste naturelle à explorer serait une approche itérative s'appuyant sur une modification de l'algorithme EM, dans laquelle le score des non-répondants sur les variables auxiliaires est estimé et utilisé pour mettre à jour les estimations des propensions à répondre.

Nous n'avons pas envisagé l'estimation basée sur la population pour d'autres types de modèles, comme la régression logistique ou probit. Comme l'indique l'évaluation numérique de la section 4, les différences entre les estimateurs basés sur l'échantillon avec fonction de lien linéaire, d'une part, et fonction de lien logistique, d'autre part, sont en général faibles, mais deviennent plus importantes quand les taux de réponse sont très proches de 1. Pour ces cas, l'élaboration d'autres fonctions de lien pour l'estimation basée sur la population est un sujet de futures études. Il s'agirait d'une extension utile et naturelle de la théorie des indicateurs R, car ces modèles sont souvent utilisés en pratique et permettent d'éviter les propensions à répondre en dehors de l'intervalle [0, 1].

Remerciements

L'étude présentée ici a été élaborée en partie dans le contexte du projet RISQ (Representativity Indicators for Survey Quality, www.risq-project.eu), financé par le 7e programme-cadre européen. Nous remercions les membres du projet RISQ : Katja Rutar du Statistični Urad Republike Slovenije, Geert Loosveldt et Koen Beullens de la Katholieke Universiteit, Leuven, Øyvin Kleven, Johan Fosen et Li-Chun Zhang du Statistisk Sentralbyrå, Norvège, Ana Marujo de la University of Southampton, Royaume-Uni et Paul Knottnerus, du Centraal Bureau voor de Statistiek, de leur contribution précieuse.

Les travaux du premier auteur ont été financés par une subvention STSM du programme COST Action IS1004 et par la subvention ex 60% University of Bergamo, Biffignandi.

Annexe A

Approximation analytique du biais des estimateurs de type 1 $\tilde{R}_{\tilde{\rho}_{T1}}$

Pour commencer, nous calculons le biais de $\tilde{S}^2_{\tilde{\rho}_{T1}}$ sous un plan d'échantillonnage général. En posant que $\hat{m}_1 = N^{-1}\sum_r d_i$ et $\hat{m}_2 = N^{-1}\sum_r d_i \tilde{\rho}_{i,T1}$, nous pouvons écrire

$$B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = E\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) - S^2_{\rho} = \frac{N}{N-1}\left\{E\left(\hat{m}_2\right) - V\left(\hat{m}_1\right) - \left[E\left(\hat{m}_1\right)\right]^2\right\} - \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in U}\rho_i^2 - \bar{\rho}_U^2\right\}. \quad (A.1)$$

Notons que

$$E\left(\hat{m}_2\right) = E\left(\frac{1}{N}\sum_{i\in U} d_i s_i r_i \tilde{\rho}_{i,T1}\right) = \frac{1}{N}\sum_{i\in U}\mathbf{x}_i^T\mathbf{T}_1^{-1}E_s\left\{E_m\left(d_i^2 s_i r_i \mathbf{x}_i + \sum_{\substack{k\in U\\ k\neq i}} d_i d_k s_i s_k r_i r_k \mathbf{x}_k \mid s\right)\right\}$$

$$= \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T\mathbf{T}_1^{-1}\sum_{\substack{k\in U\\ k\neq i}} d_k \pi_{ik}\rho_k \mathbf{x}_k,$$

$$E\left(\hat{m}_1\right) = E\left(\frac{1}{N}\sum_{i\in U} d_i s_i r_i\right) = E_s\left(\frac{1}{N}\sum_{i\in U} d_i s_i \rho_i\right) = \bar{\rho}_U,$$

et

$$V\left(\hat{m}_1\right) = V_s\left\{E_m\left(\hat{m}_1 \mid s\right)\right\} + E_s\left\{V_m\left(\hat{m}_1 \mid s\right)\right\}$$

$$= V_s\left\{\frac{1}{N}\sum_{i\in U} d_i s_i \rho_i\right\} + E_s\left\{\frac{1}{N^2}\sum_{i\in U} d_i^2 s_i \rho_i\left(1-\rho_i\right)\right\}$$

$$= \frac{1}{N^2}\sum_{i\in U}\sum_{k\in U} d_i d_k \Delta_{ik}\rho_i\rho_k + \frac{1}{N^2}\sum_{i\in U} d_i \rho_i\left(1-\rho_i\right),$$

où $\Delta_{ik} = \pi_{ik} - \pi_i\pi_k$ et π_{ik} sont les probabilités d'inclusion dans l'échantillon d'ordre deux. Donc, le biais de $\tilde{S}^2_{\tilde{\rho}_{T1}}$ par rapport à la distribution conjointe du plan d'échantillonnage et du mécanisme de réponse est donné par

$$B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T\mathbf{T}_1^{-1}\sum_{\substack{k\in U\\ k\neq i}} d_k \pi_{ik}\rho_k \mathbf{x}_k\right.$$

$$\left. - \frac{1}{N^2}\sum_{i\in U}\sum_{k\in U} d_i d_k \Delta_{ik}\rho_i\rho_k - \frac{1}{N^2}\sum_{i\in U} d_i \rho_i\left(1-\rho_i\right) - \frac{1}{N}\sum_{i\in U}\rho_i^2\right]. \quad (A.2)$$

Sous échantillonnage aléatoire simple sans remise, (A.2) peut se simplifier en

$$B^{\mathrm{EAS}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{1}{n}\sum_{i\in U}\rho_i\left\{1 - \frac{n-1}{N-1}\rho_i\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i + \frac{n-1}{n(N-1)N}\sum_{i\in U}\rho_i^2 - \frac{\bar{\rho}_U}{n} - \left(1-\frac{n}{N}\right)\frac{S^2_{\rho}}{n}\right].$$

Un estimateur de $B^{\mathrm{EAS}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)$ basé sur l'ensemble de réponses s'écrit

$$\tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1 - \frac{n-1}{N-1}\tilde{\rho}_{i,T1}\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i + \frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T1} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T1}}}{n} - \frac{n_r}{n^2}\right].$$

Plus généralement, l'estimateur de Horvitz-Thompson basé sur l'ensemble de réponses pour (A.2) sous échantillonnage complexe est donné par

$$\tilde{B}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in r} d_i\left(d_i - \tilde{\rho}_{i,T1}\right)\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i - \frac{1}{N^2}\sum_{i\in r} d_i^3\Delta_{ii}\tilde{\rho}_{i,T1}\right.$$
$$- \frac{1}{N^2}\sum_{i\in r}\sum_{\substack{k\in r\\ k\neq i}} d_i d_k \frac{\Delta_{ik}}{\pi_{ik}} - \frac{1}{N^2}\sum_{i\in r} d_i^2\left(1-\tilde{\rho}_{i,T1}\right)$$
$$\left. + \frac{1}{N}\sum_{i\in r} x_i^T T_1^{-1}\sum_{\substack{k\in r\\ k\neq i}} x_k\left(d_i d_k - \frac{1}{\pi_{ik}}\right)\right\}.$$

Annexe B

Approximation analytique du biais des estimateurs de type 2 $\tilde{R}_{\tilde{\rho}_{T2}}$

La stratégie de calcul d'un ajustement analytique du biais pour $\tilde{S}^2_{\tilde{\rho}_{T2}}$ consiste d'abord à approximer $\tilde{\rho}_{i,T2}$ par un estimateur linéaire en utilisant les techniques de linéarisation de Taylor. Puis, nous calculons un ajustement approximatif du biais pour $\tilde{S}^2_{\tilde{\rho}_{T2}}$, en insérant l'approximation linéaire pour $\tilde{\rho}_{i,T2}$ dans $\hat{m}_2$.

Ensuite, nous définissons, pour $j = 1, \ldots, p$ et $j' = 1, \ldots, p$, les totaux estimés

$$\hat{t}_0 = \sum_s d_k r_k, \qquad \hat{t}_{jj'} = \sum_s d_k r_k z_{jk} z_{j'k}, \qquad \text{et} \qquad \hat{t}_j = \sum_s d_k r_k x_{jk},$$

où $\mathbf{z}_k = \left(\mathbf{x}_k - \overline{\mathbf{x}}_U\right)$ et $z_{jk} = \left(x_{jk} - \overline{x}_{jU}\right)$. Soit $\hat{t}$ le vecteur p des composantes $\hat{t}_j$, et $\hat{\mathbf{F}}$ la matrice $(p \times p)$ symétrique contenant les éléments $\hat{t}_{jj'}$. Nous pouvons écrire

$$\tilde{\rho}_{i,T2} = \mathbf{x}_i^T\left[N\hat{t}_0^{-1}\hat{\mathbf{F}} + N\overline{\mathbf{x}}_U\overline{\mathbf{x}}_U^T\right]^{-1}\hat{\mathbf{t}} = \mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}.$$

Définissons maintenant les totaux de population

$$t_0 = \sum_U \rho_k, \qquad \mathbf{F} = \sum_U \rho_k \mathbf{z}_k\mathbf{z}_k^T, \qquad \text{et} \qquad \mathbf{t} = \sum_U \rho_k \mathbf{x}_k.$$

Notons que $\hat{t}_0$ est sans biais pour t_0, $\hat{\mathbf{F}}$ est sans biais pour $\mathbf{F}$, et $\hat{\mathbf{t}}$ est sans biais pour $\mathbf{t}$. Soit $\mathbf{T}_2 = Nt_0^{-1}\mathbf{F} + N\overline{\mathbf{x}}_U\overline{\mathbf{x}}_U^T$.

Proposition 1. L'estimateur $\tilde{\rho}_{i,T2}$ défini en (2.7) peut être approximé par

$$\tilde{\rho}_{i,T2} \cong \mathbf{x}_i^T\mathbf{T}_2^{-1}\left(Nt_0^{-2}\mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t}\left(\hat{t}_0 - t_0\right) - \mathbf{x}_i^T\mathbf{T}_2^{-1}Nt_0^{-1}\left(\hat{\mathbf{F}} - \mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t} + \mathbf{x}_i^T\mathbf{T}_2^{-1}\hat{\mathbf{t}}.$$

Preuve. En suivant la linéarisation de Taylor classique (voir Särndal, Swensson et Wretman, 1992 et Bethlehem, 1988), l'estimateur $\tilde{\rho}_{i,T2}$ peut être approximé par

$$\tilde{\rho}_{i,T2} \cong \rho^*_{i,T2} + a_0\left(\hat{t}_0 - t_0\right) + \sum_{j=1}^{p}\sum_{j'\leq j} a_{jj'}\left(\hat{t}_{jj'} - t_{jj'}\right) + \sum_{j=1}^{p} a_j\left(\hat{t}_j - t_j\right), \qquad \text{(B.1)}$$

où $\rho^*_{i,T2} = \mathbf{x}_i^T \mathbf{T}_2^{-1}\mathbf{t}$, et

$$a_0 = \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_0}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = \mathbf{x}_i^T\left[-\hat{\mathbf{T}}_2^{-1}\left(-N\hat{t}_0^{-2}\hat{\mathbf{F}}\right)\hat{\mathbf{T}}_2^{-1}\right]\hat{\mathbf{t}}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = \mathbf{x}_i^T\mathbf{T}_2^{-1}\left(Nt_0^{-2}\mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t},$$

$$a_{jj'} = \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_{jj'}}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = -\mathbf{x}_i^T\mathbf{T}_2^{-1}\left(Nt_0^{-1}\mathbf{\Lambda}_{jj'}\right)\mathbf{T}_2^{-1}\mathbf{t},$$

$$a_j = \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_j}\Big|_{\substack{\hat{t}_0=t_0 \\ \hat{\mathbf{F}}=\mathbf{F} \\ \hat{\mathbf{t}}=\mathbf{t}}} = \mathbf{x}_i^T\mathbf{T}_2^{-1}\boldsymbol{\lambda}_j,$$

où $\mathbf{\Lambda}_{jj'}$ est une matrice $(p \times p)$ contenant des valeurs 1 aux positions (j, j') et (j', j) et des zéros ailleurs, et $\boldsymbol{\lambda}_j$ est un vecteur p dont la j^e composante est égale à 1 et les autres sont nulles. L'insertion des dérivées partielles dans (B.1) donne le résultat.

Proposition 2. Sous échantillonnage aléatoire simple, une approximation du biais pour $\tilde{S}^2_{\tilde{\rho}_{T2}}$ par rapport à la distribution conjointe du plan d'échantillonnage et du mécanisme de réponse est donnée par

$$\begin{aligned} B^{\mathrm{EAS}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\Bigg\{ & t_0^{-2}\frac{N}{n}\sum_U c_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} \\ & - t_0^{-1}\frac{N}{n}\sum_U \mathbf{b}_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\}\mathbf{z}_i\mathbf{z}_i^T\mathbf{T}_2^{-1}\mathbf{t} + \frac{1}{n}\sum_U \rho_i\mathbf{x}_i^T\mathbf{T}_2^{-1}\mathbf{x}_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} \\ & + \frac{n-1}{n(N-1)}\sum_U \rho_i\rho^*_{i,T2} - \left(1-\frac{n}{N}\right)\frac{S^2_\rho}{n} - \frac{\overline{\rho}_U}{n} + \frac{1}{nN}\sum_U \rho_i^2 - \frac{1}{N}\sum_U \rho_i^2\Bigg\}, \end{aligned}$$

où $c_i = \mathbf{x}_i^T\mathbf{T}_2^{-1}\mathbf{F}\mathbf{T}_2^{-1}\mathbf{t}$, $\mathbf{b}_i = \mathbf{x}_i^T\mathbf{T}_2^{-1}$ et $\rho^*_{i,T2} = \mathbf{x}_i^T\mathbf{T}_2^{-1}\mathbf{t}$.

Un estimateur de $B^{\mathrm{EAS}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ basé sur l'ensemble de réponses est donné par

$$\begin{aligned} \tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\Bigg\{ & \frac{1}{n_r^2}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} \\ & - \frac{N}{nn_r}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{z}_i\mathbf{z}_i^T\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} \\ & + \frac{N}{n^2}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\mathbf{x}_i^T\hat{\mathbf{T}}_2^{-1}\mathbf{x}_i \\ & + \frac{n-1}{n^2(N-1)}\sum_r\tilde{\rho}_{i,T2} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T2}}}{n} - \frac{n_r}{n^2}\Bigg\}. \end{aligned}$$

Preuve. Grâce à la proposition 1, $\hat{m}_2$ défini à l'annexe A peut être approximé comme il suit

$$\begin{aligned}\hat{m}_2 &= \frac{1}{N}\sum_U d_i s_i r_i \tilde{\rho}_{i,T2} \\ &\cong \frac{1}{N}\sum_U d_i s_i r_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\left(Nt_0^{-2}\mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t}\left(\hat{t}_0 - t_0\right) \\ &\quad -\frac{1}{N}\sum_U d_i s_i r_i \mathbf{x}_i^T \mathbf{T}_2^{-1} Nt_0^{-1}\left(\hat{\mathbf{F}} - \mathbf{F}\right)\mathbf{T}_2^{-1}\mathbf{t} + \frac{1}{N}\sum_U d_i s_i r_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\hat{\mathbf{t}} \\ &=: A + B + C.\end{aligned}$$

Les espérances des termes A, B et C sont

$$\begin{aligned}E(A) &= t_0^{-2}\sum_{i\in U} c_i d_i \rho_i + t_0^{-2}\sum_{i\in U} c_i d_i \sum_{k\neq i} d_k \rho_i \rho_k \pi_{ik} - t_0^{-1}\sum_{i\in U} c_i \rho_i, \\ E(B) &= -t_0^{-1}\sum_{i\in U} d_i b_i \rho_i \mathbf{z}_i \mathbf{z}_i^T \mathbf{T}_2^{-1}\mathbf{t} - t_0^{-1}\sum_{i\in U} d_i \mathbf{b}_i \sum_{k\neq i} d_k \rho_i \rho_k \pi_{ik} \mathbf{z}_k \mathbf{z}_k^T \mathbf{T}_2^{-1}\mathbf{t} + t_0^{-1}\sum_{i\in U} \rho_i \mathbf{b}_i \mathbf{F}\mathbf{T}_2^{-1}\mathbf{t},\end{aligned}$$

et

$$E(C) = \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\sum_{k\neq i} d_k \rho_k \pi_{ik}\mathbf{x}_k.$$

Il s'ensuit que, sous échantillonnage aléatoire simple, $E(\hat{m}_2)$ devient

$$\begin{aligned}E^{\mathrm{EAS}}(\hat{m}_2) &= t_0^{-2}\frac{N}{n}\sum_U c_i \rho_i \left\{1 - \frac{n-1}{N-1}\rho_i\right\} - t_0^{-1}\frac{N}{n}\sum_U \mathbf{b}_i \rho_i \left\{1 - \frac{n-1}{N-1}\rho_i\right\}\mathbf{z}_i \mathbf{z}_i^T \mathbf{T}_2^{-1}\mathbf{t} \\ &\quad + \frac{1}{n}\sum_U \rho_i \mathbf{x}_i^T \mathbf{T}_2^{-1}\mathbf{x}_i \left\{1 - \frac{n-1}{N-1}\rho_i\right\} + \frac{n-1}{n(N-1)}\sum_U \rho_i \rho_{i,T2}^*.\end{aligned}$$

Donc, le biais total sous échantillonnage aléatoire simple est obtenu en insérant $E^{\mathrm{EAS}}(\hat{m}_2)$ calculé ci-dessus dans (A.1) et en suivant la preuve donnée à l'annexe A pour les autres termes.

L'estimateur $\tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ de $B^{\mathrm{EAS}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ basé sur l'ensemble de réponses s'obtient en remplaçant t_0 par $\hat{t}_0 = Nn_r/n$, F par $\hat{\mathbf{F}} = Nn^{-1}\sum_r \mathbf{z}_k \mathbf{z}_k^T$, $\mathbf{T}_2$ par $\hat{\mathbf{T}}_2 = N\hat{t}_0^{-1}\hat{\mathbf{F}} + N\overline{\mathbf{x}}_U \overline{\mathbf{x}}_U^T$, et t par $\hat{\mathbf{t}} = Nn^{-1}\sum_r \mathbf{x}_k$.

Notons que l'ajustement pour le biais $\tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)$ correspond à la « substitution » des quantités de type 2 ($\tilde{\rho}_{i,T2}$ au lieu de $\tilde{\rho}_{i,T1}$, matrice $\hat{\mathbf{T}}_2$ au lieu de $\mathbf{T}_1$, et $\tilde{S}^2_{\tilde{\rho}_{T2}}$ au lieu de $\tilde{S}^2_{\tilde{\rho}_{T1}}$) dans l'ajustement analytique du biais $\tilde{B}^{\mathrm{EAS}}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)$ développé pour $\tilde{S}^2_{\tilde{\rho}_{T1}}$ avec deux termes supplémentaires dus à la linéarisation de $\hat{\mathbf{T}}_2$.

Plus généralement, l'estimateur de Horvitz-Thompson pour l'ensemble de réponses sous échantillonnage complexe pour l'ajustement du biais de l'indicateur R basé sur la population de type 2 est donné par

$$\begin{aligned}\tilde{B}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) &= \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in r} d_i\left(d_i - \tilde{\rho}_{i,T2}\right)\mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\mathbf{x}_i - \frac{1}{N^2}\sum_{i\in r} d_i^3 \Delta_{ii}\tilde{\rho}_{i,T2} - \frac{1}{N^2}\sum_{i\in r}\sum_{\substack{k\in r\\k\neq i}} d_i d_k \frac{\Delta_{ik}}{\pi_{ik}}\right. \\ &\quad - \frac{1}{N^2}\sum_{i\in r} d_i^2\left(1 - \tilde{\rho}_{i,T2}\right) + \frac{1}{N}\sum_{i\in r} x_i^T \hat{T}_2^{-1}\sum_{\substack{k\in r\\k\neq i}} x_k\left(d_i d_k - \frac{1}{\pi_{ik}}\right) \\ &\quad + \left(\sum_{k\in r} d_k\right)^{-2}\sum_{i\in r} d_i^2 \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} + \left(\sum_{k\in r} d_k\right)^{-2}\sum_{i\in r} d_i \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{F}}\hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}\sum_{k\neq i} d_k \\ &\quad \left. - \left(\sum_{k\in r} d_k\right)^{-1}\sum_{i\in r} d_i^2 \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\mathbf{z}_i \mathbf{z}_i^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}} - \left(\sum_{k\in r} d_k\right)^{-1}\sum_{i\in r} d_i \mathbf{x}_i^T \hat{\mathbf{T}}_2^{-1}\sum_{k\neq i} d_k \mathbf{z}_k \mathbf{z}_k^T \hat{\mathbf{T}}_2^{-1}\hat{\mathbf{t}}\right\}.\end{aligned}$$

Bibliographie

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

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Booth, J.G., Butler, R.W. et Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89 (428), 1282-1289.

Brick, J.M., et Jones, M.E. (2008). Propensity to respond and nonresponse bias. *METRON – International Journal of Statistics*, LXVI (1), 51-73.

Copas, J.B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B*, 45, 311-354.

Copas, J.B. (1993). The shrinkage of point scoring methods. *Journal of the Royal Statistical Society, Series C*, 42, 315-331.

De Heij, V., Schouten, B. et Shlomo, N. (2015). RISQ manual 2.1. Tools in SAS and R for the computation of R-indicators and partial R-indicators, accessible à l'adresse www.risq-project.eu.

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

Efron, B., et Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Kreuter, F. (2013). *Improving Surveys with Process and Paradata*, Edited monograph, New Jersey: John Wiley & Sons, Inc.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Lundquist, P., et Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29 (4), 557-582.

MOA (2015). User Instruction Gold Standard, Dutch Market Research Association, accessible à l'adresse www.moaweb.nl/sevrices/services/gouden-standaard.html.

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

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

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 (4), 361-387.

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

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer.

Schouten, B., et Shlomo, N. (2017). Selecting adaptive survey design strata with partial R-indicators. *Revue Internationale de Statistique*, 85 (1), 143-163.

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, 29-58. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-fra.pdf.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 101-113. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10887-fra.pdf.

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

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 (3), 727-748.

Schouten, B., Bethlehem, J., Beulens, K., Kleven, Ø., Loosveldt, G., Rutar, K., Shlomo, N. et Skinner, C. (2012). Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators. *Revue Internationale de Statistique*, 80 (3), 382-399.

Shlomo, N., Skinner, C. et Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference,* 142, 201-211.

Van der Laan, D., et Bakker, B. (2015). Indicators for the representativeness of linked sources, NTTS 2015 Proceedings, accessible à l'adresse https://ec.europa.eu/eurostat/cros/system/files/NTTS2015%20 proceedings.pdf.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, 76 (3), 555-575.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7 (1), 45-55.

Wagner, J., et Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology*, 2, 323-342.

Wolter, K.M. (2007). *Introduction to Variance Estimation,* 2e Éd. New York: Springer.