Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes

par Phillip S. Kott

Juin 2012

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2012
Vol. 38, N° 1, pp. 103-107
Statistique Canada, N° 12-001-X au catalogue

Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes

Phillip S. Kott [1]

Résumé

En cas de non-réponse totale d'une unité dans un échantillon tiré suivant les principes de l'échantillonnage probabiliste, une pratique courante consiste à diviser l'échantillon en groupes mutuellement exclusifs de manière qu'il soit raisonnable de supposer que toutes les unités échantillonnées dans un groupe ont la même probabilité de ne pas répondre. De cette façon, la réponse d'une unité peut être traitée comme une phase supplémentaire de l'échantillonnage probabiliste en se servant de l'inverse de la probabilité de réponse estimée d'une unité dans un groupe comme facteur de correction pour calculer les poids finaux pour les répondants du groupe. Si l'objectif est d'estimer la moyenne de population d'une variable d'enquête qui se comporte plus ou moins comme une variable aléatoire dont la moyenne est constante dans chaque groupe indépendamment des poids de sondage originaux, il est habituellement plus efficace d'intégrer les poids de sondage dans les facteurs de correction que de ne pas le faire. En fait, si la variable d'enquête se comportait exactement comme une telle variable aléatoire, l'estimation de la moyenne de population calculée en se servant des facteurs de correction pondérés selon le plan de sondage serait presque sans biais dans un certain sens (c'est-à-dire sous la combinaison du mécanisme d'échantillonnage probabiliste original et d'un modèle de prédiction), même si les unités échantillonnées dans un groupe n'ont pas toutes la même probabilité de répondre.

Mots clés : Double protection ; modèle de prédiction ; échantillonnage probabiliste ; modèle de réponse ; phase d'échantillonnage ; échantillonnage bernoullien stratifié.

1. Introduction

En l'absence de non-réponse, il est possible d'estimer la moyenne d'une population finie d'après un échantillon sans avoir à recourir à un modèle statistique qui, aussi raisonnable qu'il soit, pourrait ne pas être vérifié. Pour cela, on attribue à chaque unité de la population une probabilité positive de sélection dans l'échantillon et l'on crée des estimateurs en s'appuyant sur ce mécanisme de sélection aléatoire. Malheureusement, dans des conditions réelles, les enquêtes souffrent souvent de non-réponse.

Deux types distincts de modèle peuvent être utilisés pour faire face à la non-réponse totale d'une unité. L'un est un modèle de prédiction, ou de résultat, dans lequel on suppose que la variable d'enquête se comporte comme une variable aléatoire dont on connaît les caractéristiques, mais non les paramètres. L'autre est un modèle de réponse, ou de sélection, dans lequel le simple fait qu'une unité réponde à une enquête est traité comme une phase supplémentaire de la sélection aléatoire de l'échantillon.

Habituellement, les statisticiens d'enquête préfèrent les modèles de réponse pour deux raisons. Outre le fait que la modélisation de la réponse est commode parce qu'elle permet de traiter la réponse d'une unité comme une phase supplémentaire de l'échantillonnage aléatoire, une enquête est habituellement conçue afin de recueillir des renseignements sur plusieurs variables auprès des unités échantillonnées. La modélisation de la prédiction requiert que l'on formule un modèle hypothétique différent pour chaque variable d'enquête, chacun de ces modèles pouvant ne pas être vérifié. Par contre, la modélisation de la réponse ne nécessite l'hypothèse que d'un seul modèle. Il n'en est toutefois plus ainsi en cas de non-réponse partielle (pour une variable particulière de l'enquête). Par conséquent, les modèles de prédiction sont souvent préférés pour traiter la non-réponse partielle par imputation. Cela étant dit, la non-réponse partielle dépasse le cadre du présent article.

Sous un modèle de réponse hypothétique, les probabilités de réponse des unités sont traitées comme étant inconnues, ce qui signifie qu'elles doivent être estimées d'après l'échantillon. Habituellement, on suppose que le mécanisme de réponse des diverses unités est indépendant et qu'il ne dépend pas de la sélection de l'unité dans l'échantillon (chaque unité possède une probabilité a priori de réponse qui devient opérationnelle si elle est sélectionnée dans l'échantillon). Le modèle de réponse le plus simple et le plus fréquemment utilisé consiste à diviser l'échantillon et, implicitement la population complète, en groupes mutuellement exclusifs, appelés « groupes de réponse homogènes » par Särndal, Swensson et Wretman (1992) (le terme « classes de pondération » est plus courant ; voir, par exemple, Lohr [2009, pages 340-341]), et à supposer que chaque unité d'un groupe possède la même probabilité de ne pas répondre, quelle que soit sa probabilité de sélection dans l'échantillon original, π_k. Donc, le mécanisme de réponse produit un sous-échantillon bernoullien stratifié dans lequel les groupes constituent les strates.

1. Phillip S. Kott, RTI International, Suite 902, 6100 Executive Blvd., Rockville, MD 20852, États-Unis. Courriel : pkott@rti.org.

Conditionnellement aux tailles des échantillons de répondants dans les groupes, un sous-échantillon bernoullien stratifié dont les probabilités de sélection (réponse) sont inconnues est converti en un sous-échantillon aléatoire simple stratifié dont les probabilités de sélection sont connues : r_g / n_g pour les unités d'un groupe g quand ce groupe contient n_g unités échantillonnées, dont r_g fournissent une réponse.

Bien que la probabilité conditionnelle de réponse dans le groupe g sous le modèle de réponse bernoullien stratifié soit r_g / n_g, nous verrons qu'il est souvent préférable de multiplier le poids de sondage, $d_k = 1/\pi_k$, d'une unité répondante dans le groupe non pas par n_g / r_g, mais par

$$f_g = \frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k}, \qquad (1)$$

où S_g est l'échantillon original et R_g, le sous-échantillon de répondants dans le groupe g. Ce *facteur de correction* peut différer de n_g / r_g quand les d_k dans le groupe g varient.

Little et Vartivarian (2003) affirment que f_g est habituellement utilisé en pratique. Toutefois, ils soutiennent qu'intégrer de cette façon les poids de sondage dans le facteur de correction peut « accroître la variance ».

À la section 2, nous établissons la notation pour estimer la moyenne de population d'une variable d'enquête. L'utilisation du rapport n_g / r_g produit un estimateur à facteur d'extension double tandis que l'utilisation de f_g produit un estimateur à facteur d'expansion repondéré. Nous pouvons exprimer les deux estimateurs en nous servant d'une formulation donnée dans Kim, Navarro et Fuller (2006). Cette expression permet de voir que, si la variable d'enquête se comporte approximativement comme une variable aléatoire de moyenne constante à l'intérieur de chaque groupe, quels que soient les poids de sondage, l'utilisation de f_g est souvent plus efficace que l'utilisation de n_g / r_g. En fait, si la variable d'enquête se comporte exactement comme une telle variable aléatoire, l'estimation de la moyenne de population calculée en se servant de f_g sera presque sans biais sous la combinaison du plan de sondage original et de ce modèle de prédiction, même si le modèle de réponse ne tient pas.

À la section 3, nous montrons que les résultats empiriques présentés dans Little et Vartivarian (2003) concordent avec ces arguments et nous offrons certaines conclusions.

2. Les deux estimateurs

Supposons que nous voulions estimer la moyenne de population d'une variable d'enquête y_k :

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{\sum_{g=1}^{G} \sum_{k \in U_g} y_k}{\sum_{g=1}^{G} N_g} = \frac{\sum_{g=1}^{G} N_g \bar{y}_{U_g}}{\sum_{g=1}^{G} N_g},$$

où la population U est divisée en G groupes, $U_1, ..., U_G$, chaque U_g contenant N_g unités, et $N = N_1 + ... + N_G$. En l'absence de non-réponse, chaque N_g est estimé sans biais sous la théorie de l'échantillonnage probabiliste par $\hat{N}_g = \sum_{k \in S_g} d_k$, et chaque $\bar{y}_{U_g}$ est estimé de façon presque sans biais (c'est-à-dire asymptotiquement)

$$\bar{y}_{S_g} = \frac{\sum_{k \in S_g} d_k y_k}{\sum_{k \in S_g} d_k}, \qquad (2)$$

sous des contraintes faibles quand n_g est suffisamment grand. Nous émettons ces deux hypothèses ici.

Pour un énoncé formel des conditions dans lesquelles chaque $\bar{y}_{S_g}$ est convergeant sous la théorie de l'échantillonnage probabiliste et par conséquent presque sans biais, voir Fuller (2009, page 115). Le lecteur que cela intéresse est invité à consulter Fuller chaque fois qu'un résultat du présent exposé dépend d'hypothèses au sujet du plan de sondage et de la population à mesure que la taille d'échantillon devient arbitrairement grande. Un traitement plus rigoureux de la plupart de la matière discutée ici sous le modèle de réponse peut être consulté dans Kim, Navarro et Fuller (2006).

Désignons l'estimateur de $\bar{y}_U$ sous échantillon complet dont nous avons discuté par $\bar{y}_S = \sum^G \hat{N}_g \bar{y}_{S_g}$. Il existe des moyens plus directs de rendre $\bar{y}_S$, mais la version susmentionnée servira mieux nos objectifs.

Si nous faisons une correction pour tenir compte de la non-réponse en utilisant le facteur f_g dans l'équation (1), nous obtenons l'estimateur à facteur d'extension repondéré :

$$\hat{\bar{y}}_{rw} = \frac{\sum_{g=1}^{G} \left(f_g \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \left(f_g \sum_{k \in R_g} d_k \right)}$$

$$= \frac{\sum_{g=1}^{G} \left(\frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \sum_{k \in S_g} d_k} = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k} \right)}{\sum_{g=1}^{G} \hat{N}_g}.$$

Techniquement, $\hat{\bar{y}}_{rw}$ est le ratio de deux estimateurs à facteur d'extension repondéré, mais nous utilisons la terminologie plus simple ici.

L'emploi de n_g / r_g donne l'estimateur à facteur d'extension double :

$$\hat{\bar{y}}_{de} = \frac{\sum_{g=1}^{G}\left(\frac{n_g}{r_g}\sum_{k\in R_g} d_k y_k\right)}{\sum_{g=1}^{G}\left(\frac{n_g}{r_g}\sum_{k\in R_g} d_k\right)}.$$

Pour les besoins de notre étude, cet estimateur peut également être exprimé sous la forme

$$\hat{\bar{y}}_{de} = \frac{\sum_{g=1}^{G}\left(\frac{\sum_{k\in S_g} d_k p_k}{\sum_{k\in R_g} d_k p_k}\sum_{k\in R_g} d_k y_k\right)}{\sum_{g=1}^{G}\left(\frac{\sum_{k\in S_g} d_k p_k}{\sum_{k\in R_g} d_k p_k}\sum_{k\in R_g} d_k\right)} = \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k y_k}{\sum_{k\in R_g} d_k p_k}\right)}{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k p_k}\right)},$$

où

$$p_k = \frac{1}{d_k}\frac{\sum_{j\in S_g} d_j}{n_g} \quad \text{pour } k \in S_g \tag{3}$$

(de sorte que $\sum_{S_g} d_k p_k = \sum_{S_g} d_k = \hat{N}_g$).

Les estimateurs $\hat{\bar{y}}_{rw}$ et $\hat{\bar{y}}_{de}$ peuvent s'écrire tous deux sous la forme :

$$\hat{\bar{y}}_{S,\mathbf{q}} = \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k y_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k q_k}\right)}. \tag{4}$$

Pour l'estimateur à facteur d'extension repondéré, tous les $q_k = 1$, tandis que pour l'estimateur à facteur d'extension double, $q_k = p_k$ tel qu'il est défini par l'équation (3).

Nous nous servirons bientôt de l'expression qui suit pour nos deux estimateurs :

$$\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_S = \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k e_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k q_k}\right)} \approx \frac{\sum_{g=1}^{G}\left(\hat{N}_g \frac{\sum_{k\in R_g} d_k e_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G}\hat{N}_g}, \tag{5}$$

où $e_k = y_k - \bar{y}_S$. L'équation (5) est vérifiée exactement quand tout $q_k = 1$. Si $q_k = p_k$, la presque égalité dépendra du fait que r_g est suffisamment grand, ainsi que d'autres contraintes faibles.

Supposons maintenant que le modèle de réponse qui suit est vérifié : Chaque unité k d'un groupe a une probabilité positive de réponse qui ne varie pas en fonction de π_k ni de y_k. Autrement dit, l'indicateur de réponse ρ_k, qui vaut 1 quand l'unité k répond si elle est échantillonnée et 0 autrement, est une variable aléatoire de Bernoulli dont la moyenne est commune dans U_g indépendamment des valeurs de π_k et y_k.

En traitant de cette façon la réponse d'une unité comme une deuxième phase de l'échantillonnage probabiliste, nous pouvons exprimer la variance/l'erreur quadratique moyenne supplémentaire due à la non-réponse, sachant l'échantillon original et les r_g pour les deux estimateurs, sous la forme

$$\begin{aligned} A_{\mathbf{q}} &= E_\rho[(\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_s)^2 \mid S, \{r_g\}] \\ &\approx \frac{\sum_{g=1}^{G} \hat{N}_g{}^2 \operatorname{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} \mid S_g, r_g)}{\left(\sum_{g=1}^{G}\hat{N}_g\right)^2}, \end{aligned} \tag{6}$$

où $\hat{\bar{e}}_{S_g,\mathbf{q}} = \hat{\bar{y}}_{S_g,\mathbf{q}} - \bar{y}_S$, $\bar{e}_{S_g} = \bar{y}_{S_g} - \bar{y}_S$, et

$$\begin{aligned} &\operatorname{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} \mid S_g, r_g) \\ &\quad\approx \left(\frac{n_g}{r_g} - 1\right)\frac{\sum_{k\in S_g} d_k{}^2 (e_k - q_k \bar{e}_{S_g})^2}{\left(\sum_{k\in S_g} d_k q_k\right)^2} \\ &\quad= \left(\frac{n_g}{r_g} - 1\right)\frac{\sum_{k\in S_g} d_k{}^2 ([y_k - \bar{y}_S] - q_k[\bar{y}_{S_g} - \bar{y}_S])^2}{\left(\sum_{k\in S_g} d_k q_k\right)^2}, \end{aligned} \tag{7}$$

en appliquant à la population et au plan de sondage original des contraintes faibles que nous supposons être vérifiées, y compris (de nouveau) le fait que les tailles r_g sont suffisamment grandes. Ces conditions rendent les deux estimateurs presque sans biais sous la théorie de l'échantillonnage quasi probabiliste (théorie probabiliste augmentée d'un modèle de réponse) et rend discutable la distinction entre la variance et l'erreur quadratique moyenne en grand échantillon. La théorie de l'échantillonnage quasi probabiliste est également appelée théorie de l'échantillonnage « quasi fondée sur un plan » ou « quasi aléatoire ».

Si nous examinons les équations (6) et (7), nous constatons qu'à l'un des extrêmes, $\hat{\bar{y}}_{rw}$ possède une variance supplémentaire due à la non-réponse (approximativement) nulle quand tous les y_k échantillonnés originalement dans un groupe sont égaux, tandis qu'à l'autre extrême, $\hat{\bar{y}}_{de}$ présente une variance supplémentaire nulle quand tous les

$d_k e_k$ (ou, exprimés d'une autre façon, les $d_k[y_k - \bar{y}_S]$) échantillonnés au départ dans un groupe sont égaux.

Sur le plan heuristique, l'estimateur à facteur d'extension repondéré est plus efficace que l'estimateur à facteur d'extension double quand $\bar{e}_{S_g}$ est un meilleur prédicteur de e_k que $p_k \bar{e}_{S_g}$ pour $k \in S_g$. Donc, quand les groupes sont construits comme l'ont recommandé Little et Vartivarian (2003), et auparavant Little (1986), de façon à ce que la réponse y_k dans un groupe soit homogène (par opposition au fait que $d_k[y_k - \bar{y}_S]$ soit homogène), l'estimateur à facteur d'extension repondéré calculé en se servant de f_g sera généralement plus efficace que l'estimateur à facteur d'extension double calculé en se servant de n_g / r_g.

L'observation heuristique peut être exprimée formellement en utilisant une autre justification de l'utilisation de l'estimateur à facteur d'extension repondéré. Supposons que le modèle de prédiction qui suit est vérifié : Chaque y_k dans U_g est une variable aléatoire de moyenne commune, μ_g, indépendamment de π_k et ρ_k. Alors, $\hat{\bar{y}}_{rw}$ est presque sans biais sous des contraintes faibles en ce qui concerne la combinaison du mécanisme d'échantillonnage original (qui traite les d_k comme étant aléatoires, où $d_k = 0$ pour $k \neq S$) et du modèle de prédiction (qui traite les y_k comme étant aléatoires). C'est-à-dire que $\mathrm{E}_d[\mathrm{E}_y(\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] \approx 0$, puisque l'espérance double de $\hat{\bar{y}}_{rw}$ ainsi que $\bar{y}_U$ est presque $\sum^G N_g \mu_g / \sum^G N_g$. Cette absence combinée de biais est exacte quand le plan est tel que $\sum_S d_k \equiv N$. L'échantillonnage aléatoire simple stratifié est un exemple de ce type de plan. L'échantillonnage non stratifié avec probabilités inégales et de nombreux plans à plusieurs degrés ne le sont pas.

Il n'est pas difficile de voir que $\hat{\bar{y}}_{rw}$ est également sans biais par rapport à cette espérance double (c'est-à-dire $\mathrm{E}_d[\mathrm{E}_y(\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] = 0$) quand tous les μ_g sont égaux. En fait, l'espérance de $\hat{\bar{y}}_{rw}$ ainsi que de $\hat{\bar{y}}_{de}$ sous le modèle de prédiction est égale à cette moyenne commune, de même que l'espérance sous le modèle de prédiction d'un estimateur sans aucune correction pour tenir compte de la non-réponse totale, c'est-à-dire avec le remplacement de f_g dans $\hat{\bar{y}}_{rw}$ par 1. L'avantage de $\hat{\bar{y}}_{rw}$ par rapport à $\hat{\bar{y}}_{de}$ sous le modèle de prédiction s'obtient uniquement quand les μ_g varient, c'est-à-dire quand la moyenne de prédiction de la variable d'enquête varie d'un groupe à l'autre.

Il convient de souligner que si le modèle de réponse *ou* le modèle de prédiction tient, l'estimateur à facteur d'extension repondéré est presque sans biais dans un certain sens (c'est-à-dire sous la combinaison du plan de sondage original et du modèle de réponse ou sous le plan de sondage original et le modèle de prédiction). Cette propriété a été appelée « double protection » contre le biais de non-réponse. Voir, par exemple, Bang et Robins (2005).

3. Conclusion

Le présent exposé porte sur deux types de modèles distincts. Nous avons décrit un modèle de réponse dans lequel les indicateurs de réponse, ρ_k, sont traités comme une variable aléatoire de Bernoulli dans chaque groupe, mais dont les paramètres sont inconnus. Nous avons également décrit un modèle de prédiction dans lequel les valeurs observées, y_k, sont traitées comme des variables aléatoires de moyenne inconnue pouvant varier entre les groupes mais non à l'intérieur de ceux-ci.

Dans le modèle de réponse, nous supposons qu'à l'intérieur d'un groupe, les ρ_k ne dépendent pas des y_k. Par analogie, dans le modèle de prédiction, nous supposons que, dans un groupe, les y_k ne dépendent pas des ρ_k. Quand ρ_k ainsi que y_k sont traités comme des variables aléatoires, la première hypothèse, à savoir que les non-répondants *manquent au hasard*, équivaut à la seconde hypothèse, à savoir que le mécanisme de réponse est *ignorable* (voir, par exemple, Little et Rubin 1987). Il convient toutefois de bien saisir que les y_k n'ont pas à être traités comme des variables aléatoires sous le modèle de réponse et que les ρ_k n'ont pas à être traités comme des variables aléatoires sous le modèle de prédiction. Les deux concepts (réponse manquant au hasard et non-réponse ignorable) sont peut-être équivalents dans un certain sens, mais ils ne sont pas identiques.

L'exposé de Little et Vartivarian (2003) a pour élément central une série de simulations comportant une variable d'enquête binaire, deux groupes de réponses possibles et deux probabilités de sélection originales. La variable d'enquête ainsi que les indicateurs de réponse sont générés sous cinq modèles. La valeur prévue de chacune de ces variables 1) dépend du groupe de réponse seulement, 2) dépend de la probabilité de sélection seulement, 3) ne dépend ni de l'un ni de l'autre, ou 4) et 5) dépend de l'une de deux combinaisons égales de groupe de réponse et de probabilité de sélection. Cela produit 25 scénarios dont dix nous intéressent tout spécialement. Ces dix scénarios sont ceux dans lesquels la variable d'enquête est une fonction du groupe de réponse seulement ou n'est une fonction ni du groupe de réponse ni de la probabilité de sélection.

Comme le prédit notre théorie quand la variable d'enquête n'est fonction ni du groupe de réponse ni de la probabilité de sélection, l'estimateur à facteur d'extension repondéré et l'estimateur à facteur d'extension double ont tous deux un biais empirique presque nul (tableau 5 dans Little et Vartivarian) parce qu'ils sont tous deux presque sans biais sous la combinaison du plan d'échantillonnage original et d'un modèle de prédiction valide : toutes les unités de population ont la même moyenne. Quand la variable d'enquête est une fonction du groupe de réponse *et* que l'indicateur de réponse est entièrement ou partiellement une fonction de la probabilité de sélection, seul l'estimateur à

facteur d'extension repondéré est presque sans biais empiriquement, puisqu'il est le seul à être sans biais sous la combinaison du plan d'échantillonnage original et d'un modèle de prédiction valide. Par conséquent, $\hat{\bar{y}}_{rw}$ donne aussi une racine carrée de l'erreur quadratique moyenne empirique plus faible et une erreur absolue moyenne significativement plus faible qu'un estimateur de $\bar{y}_S$ (tableaux 4 et 6 dans Little et Vartivarian, respectivement ; le test de signification traite la valeur moyenne sur l'ensemble des simulations de $|\hat{\bar{y}}_{rw} - \bar{y}_S| - |\hat{\bar{y}}_{de} - \bar{y}_S|$ comme étant asymptotiquement normale).

Quand la variable d'enquête ainsi que les indicateurs de réponse sont des fonctions du groupe de réponse seulement, l'estimateur à facteur d'extension repondéré produit une racine carrée de l'erreur quadratique moyenne empirique et une erreur absolue moyenne légèrement plus faibles que l'estimateur à facteur d'extension double, mais la différence n'est pas significative pour le second paramètre.

Il n'est guère surprenant que la réduction de la racine carrée de l'erreur quadratique moyenne empirique soit modeste. La contribution de la non-réponse à la variance sous le modèle de réponse exprimé par les équations (6) et (7) est conditionnelle à l'échantillon original (techniquement, la contribution de la non-réponse à la variance totale sous échantillonnage quasi probabiliste de $\hat{\bar{y}}_{S,\mathbf{q}}$ est l'espérance de $A_{\mathbf{q}}$ dans l'équation (6) sous le mécanisme d'échantillonnage original). Dans les applications où les taux de réponse sont assez élevés (dans les simulations ils valent en moyenne 0,5), cette contribution peut être dominée par la variance/l'erreur quadratique moyenne sous échantillonnage probabiliste de l'estimateur sous échantillon complet, $\hat{\bar{y}}_U$.

Deux mises en garde sont de rigueur. La taille de l'échantillon de répondants dans chaque groupe doit être suffisamment grande pour que l'estimateur à facteur d'extension répondéré soit presque sans biais sous la théorie de l'échantillonnage quasi probabiliste. Pour l'estimateur à facteur d'extension double, il suffit que chaque taille r_g soit positive. En outre, le fait que l'estimateur à facteur d'extension repondéré est doublement protégé contre le biais de non-réponse n'est utile que si le modèle de réponse ou le modèle de prédiction hypothétique est correct. Si les probabilités de réponse *ainsi que* les valeurs des variables observées varient avec les poids de sondage, l'estimateur à facteur d'extension repondéré peut présenter un biais significatif. Malgré la perspective adoptée dans le présent exposé, il s'agit du message que Little et Vartivarian (2003) souhaitaient communiquer et il ne peut pas être contesté.

Remerciements

Je remercie le rédacteur associé et deux examinateurs de leur lecture attentive de versions antérieures du manuscrit qui m'a permis d'améliorer considérablement la qualité et l'exactitude des travaux résultants. Toute erreur qui subsiste dans le texte est entièrement imputable à l'auteur.

Bibliographie

Bang, H., et Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.

Fuller, W. (2009). *Sampling Statistics*, Hoboken, New Jersey : Wiley.

Kim, J.K., Navarro, A. et Fuller, W. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Little, R. (1986). Survey nonresponse adjustments. *Revue Internationale de Statistique*, 54, 139-157.

Little, R., et Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York : John Wiley & Sons, Inc.

Little, R., et Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Lohr, S (2009). *Sampling: Design and Analysis*, *Second Edition*, Boston : Brooks/Cole.

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