Techniques d'enquête, décembre 1986
Vol. 12, n° 2, pp. 215–224
Statistique Canada

Estimation par la méthode de régression en situation de non-réponse

CARL ERIK SÄRNDAL[1]

RÉSUMÉ

Dans les cas de non-réponse totale, deux genres de variables se dégagent parfois des unités de l'échantillon initial *s;* ce sont les variables servant à estimer le mécanisme de réponse (les probabilités de réponse) et les variables (en l'occurrence, les covariables) qui expliquent la variable d'intérêt dans le sens normal de la théorie de la régression. Dans le présent article, fondé sur l'ouvrage de Sarndal et Swensson (1985 a, b), nous analysons les estimateurs corrigés en fonction de la non-réponse en faisant tantôt intervenir des covariables et tantôt non. Cette étude nous amène à conclure qu'un estimateur qui renferme de fortes covariables tend à présenter plusieurs caractéristiques favorables. Par exemple, les estimateurs qui utilisent des covariables sont beaucoup moins exposés au biais dû à la non-réponse. Nous examinons aussi le calcul d'erreurs types et d'intervalles de confiance valables pour les estimateurs comprenant des covariables. Enfin, nous analysons la structure de l'erreur type.

MOTS CLÉS: Mécanisme de réponse; méthode des groupes de correction; covariable; robustesse.

1. INTRODUCTION

Nous considérons une population finie $U = \{1, \ldots, k, \ldots, N\}$, de laquelle un échantillon s de taille n est prélevé à l'aide d'un plan de sondage selon lequel la probabilité de sélection (strictement positive) de la k-ième unité est π_k. Le poids d'échantillonnage de la k-ième unité est donc π_k^{-1}. Nous pouvons supposer un plan de sondage complexe, qui n'est pas nécessairement auto-pondéré, par exemple, un plan de sondage à 3 degrés avec sélection stratifiée des unités primaires. Selon ce plan, la probabilité que les unités k et l soient toutes deux incluses dans l'échantillon est désignée π_{kl} ($\pi_{kl} > 0$ pour tout $k \neq l$ et π_{kk} est considérée égale à π_k).

Étant donné l'échantillon s, nous supposons qu'il y aura un certain taux de non-réponse totale. Nous désignons r le sous-ensemble de s que constituent les répondants et m la taille de ce sous-ensemble. La variable d'intérêt, y, est observée uniquement pour $k \in r$. Afin de compenser le biais dû à la non-réponse, nous supposons, pour les besoins de la présente étude, que la méthode très connue des groupes de correction est utilisée: l'échantillon s est subdivisé en H groupes $s_1, \ldots, s_h, \ldots, s_H$ de tailles respectives $n_1, \ldots, n_h, \ldots, n_H$. De même, nous subdivisons l'ensemble de réponses r en sous-ensembles $r_1, \ldots, r_h, \ldots, r_H$ de tailles respectives $m_1, \ldots, m_h, \ldots, m_H$. Le taux de réponse pour le groupe h est défini comme $f_h = m_h/n_h$. La méthode précitée comporte une opération qui consiste à affecter les observations du groupe h d'un poids de correction f_h^{-1}, en plus du poids de sondage. (Dans le présent modèle, nous supposons que la taille et la composition des groupes de correction au niveau de la population sont inconnues.) Nous avons:

$$n = \sum_{h=1}^{H} n_h;\ m = \sum_{h=1}^{H} m_h.$$

[1] Carl Erik Särndal, département de mathématique et de statistique, Université de Montréal, Montréal (Québec), Canada, H3C 3J7.

Posons $t = \Sigma_U y_k$ comme le total de population à estimer. (Si A est un ensemble d'unités arbitraire, nous écrirons systématiquement $\Sigma_A y_k$ au lieu de $\Sigma_{k\epsilon A} y_k$.). Selon la méthode des groupes de correction, l'estimateur régulier de t devient alors:

$$\hat{t} = \sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}. \tag{1.1}$$

Théoriquement, cette méthode est fondée sur l'hypothèse selon laquelle la probabilité de réponse (inconnue) est la même pour chaque unité du même groupe. (Cette hypothèse est une des conditions du modèle A défini à la section 3 plus loin.) La méthode des groupes de correction exige formellement que chaque unité $k \epsilon s$ soit classée. Les variables (qualitatives) servant à cette classification peuvent donc être considérées comme un moyen d'estimer un mécanisme de réponse fondamental.

Un autre genre de variable peut être observée pour chaque $k \epsilon s$; il s'agit des variables qui expliquent y dans le sens habituel de la théorie de la régression. Ces variables seront appelées covariables. Lorsque celles-ci sont incluses dans un estimateur, non seulement elles réduisent sa variance, mais elles le permettent de mieux contrebalancer le biais dû à la non-réponse. (Les covariables ne sont pas des variables auxiliaires dans le sens normal puisqu'elles ne se rapportent pas à la population globale U mais uniquement à l'échantillon visé s.)

Nous faisons donc une nette distinction dans cette étude entre les deux genres de variables observées pour $k \epsilon s$, soit celles qui servent à estimer le mécanisme de réponse et celles qui expliquent la variable d'intérêt y. Dans la description d'un modèle général pour des données en situation de non-réponse, Little (1983) définit plusieurs genres de variables. Afin de situer notre étude en fonction du modèle de Little, disons que ce que celui-ci appelle l'ensemble des variables de réponse complètes est, dans notre étude, divisé en deux sous-ensembles: un sous-ensemble de variables servant à définir le modèle du mécanisme de non-réponse et un sous-ensemble de variables (les covariables) servant à expliquer la variable de réponse incomplète y. La méthode d'inférence que nous utilisons est la méthode de "quasi-randomisation" (Oh et Scheuren 1983), où "quasi" signifie que l'étape de la sélection du mécanisme de non-réponse doit être définie par un modèle tandis que l'étape de l'échantillonnage est régie par les échantillons.

2. ESTIMATEURS SIMPLES DU TOTAL DE POPULATION CORRIGÉS EN FONCTION DE LA NON-RÉPONSE

En développant un peu l'équation courante (1.1), nous obtenons une autre équation, généralement plus efficace, où les poids de sondage π_k^{-1} ont, semble-t-il, une application plus complète:

$$\hat{t}_{\text{EXP}} = \left(\sum_s \frac{1}{\pi_k}\right) \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{1}{\pi_k}}.$$

Cette équation, qui se ramène à (1.1) lorsque le plan de sondage est auto-pondéré, peut être exprimée comme la formule développée de la moyenne de l'ensemble de réponses:

$$\hat{t}_{\text{EXP}} = \hat{N}\tilde{y}_r,$$

pourvu que l'on définisse $\hat{N} = \Sigma_s \, 1/\pi_k$, et

$$\tilde{y}_r = \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{1}{\pi_k}}. \tag{2.1}$$

Le tilde ($\sim$) sert à indiquer une statistique moyenne qui est pondérée convenablement. On détermine la statistique $\tilde{y}_r$, qui est la moyenne d'un ensemble de réponses, en affectant la k-ième unité du poids multiplicatif:

$$\text{poids de sondage} \times \text{poids de correction pour la non-réponse} = \pi_k^{-1} f_h^{-1}$$

pour chaque unité k incluse dans le h-ième groupe de correction.

L'estimateur $\hat{t}_{\text{EXP}}$ convient bien en situation de non-réponse puisqu'il tient compte du plan de sondage et qu'il tend à corriger l'effet de la non-réponse. Il est toutefois possible d'améliorer $\hat{t}_{\text{EXP}}$, si l'on dispose de plus d'informations. Supposons qu'une covariable simple (et toujours positive) x soit également observée pour $k \in s$. Sur le modèle de l'estimateur par quotient classique, nous pouvons donc définir

$$\hat{t}_{\text{RA}} = \left(\sum_s \frac{x_k}{\pi_k}\right) \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{x_k}{\pi_k}} = \hat{N} \tilde{x}_s \frac{\tilde{y}_r}{\tilde{x}_r},$$

où la statistique moyenne $\tilde{x}_r$ est définie par l'équation (2.1) en substituant x_k à y_k, et

$$\tilde{x}_s = \frac{\sum_s \frac{x_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}}.$$

Comme elle est définie au niveau de l'échantillon visé s, la statistique moyenne $\tilde{x}_s$ n'utilise que des poids de sondage. (Ce genre de moyenne peut être calculée pour la variable x, qui est observée pour tout $k \in s$, mais ne peut évidemment pas l'être pour la variable y qui est observée uniquement pour $k \in r$.)

Dans le cadre de notre analyse, l'équation classique de l'estimateur par régression s'exprime:

$$\hat{t}_{\text{REG}} = \hat{N}\{\tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r)\}$$

où

$$b = \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} (y_k - \tilde{y}_r)(x_k - \tilde{x}_r)/\pi_k}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} (x_k - \tilde{x}_r)^2/\pi_k}.$$

(Note: Les poids de sondage ainsi que les poids de correction pour la non-réponse sont utilisés pour déterminer b.)

En résumé, nous avons trois estimateurs

$$\hat{t}_{\text{EXP}} = \hat{N}\tilde{y}_r, \tag{2.2a}$$

$$\hat{t}_{\text{RA}} = \hat{N}\tilde{x}_s \frac{\tilde{y}_r}{\tilde{x}_r}, \tag{2.2b}$$

$$\hat{t}_{\text{REG}} = \hat{N}\{\tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r)\}. \tag{2.2c}$$

Tous les trois sont pondérés convenablement avec des poids de sondage et des poids de correction pour la non-réponse. Le principal élément de distinction entre eux est la covariable: $\hat{t}_{\text{EXP}}$ n'en renferme aucune alors que $\hat{t}_{\text{RA}}$ et $\hat{t}_{\text{REG}}$ en contiennent. De plus, $\hat{t}_{\text{RA}}$ dénote clairement une relation sous-jacente entre y et la covariable x sous la forme d'une droite passant par l'origine, la pente de cette droite étant estimée par $\tilde{y}_r/\tilde{x}_r$. En ce qui concerne l'estimateur $\hat{t}_{\text{REG}}$, la relation sous-jacente est définie par une droite de régression ayant une ordonnée à l'origine non nulle. Nous approfondirons plus loin le rôle de la covariable.

Si l'on connaît la taille de la population N, il est préférable, en règle générale, de substituer N à $\hat{N}$ dans les trois équations précédentes, ce qui donne

$$\hat{t}^*_{\text{EXP}} = N\tilde{y}_r, \tag{2.3a}$$

$$\hat{t}^*_{\text{RA}} = N\tilde{x}_s \frac{\tilde{y}_r}{\tilde{x}_r}, \tag{2.3b}$$

$$\hat{t}^*_{\text{REG}} = N\{\tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r)\}. \tag{2.3c}$$

Pour estimer le total de la population, il faut connaître la valeur de N dans chacun des trois estimateurs ci-dessus, ce qui n'est pas toujours le cas. En revanche, pour estimer la moyenne de la population $\bar{Y}$, on divise chacune de ces équations par N, ce qui donne les équations simplifiées suivantes

$$\hat{\bar{Y}}_{\text{EXP}} = \tilde{y}_r, \tag{2.4a}$$

$$\hat{\bar{Y}}_{\text{RA}} = \tilde{x}_s \frac{\tilde{y}_r}{\tilde{x}_r}, \tag{2.4b}$$

$$\hat{\bar{Y}}_{\text{REG}} = \tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r). \tag{2.4c}$$

Intuitivement, les trois séries d'estimateurs (2.2), (2.3) et (2.4) sont facilement concevables puisqu'elles reposent sur des règles de pondération élémentaires et les notions fondamentales du quotient ou de la régression. Il est toutefois plus difficile de déterminer l'effet qu'auront ces séries d'estimateurs sur l'estimation de la variance et la construction d'intervalles de confiance valables. Ces questions sont traitées à la section 4. (Contrairement à ce que pourrait laisser croire la présentation plutôt informelle des séries d'estimateurs (2.2) à (2.4), il ne s'agit pas d'équations "ad hoc" mais plutôt d'équations qui découlent d'une méthode d'estimation générale et formalisée (avec une régression à plusieurs variables) pour deux phases de sélection; à cet effet, voir Särndal et Swensson (1985a). Ce qui est encore plus important, c'est que les estimateurs de la variance et les intervalles de confiance découlent directement de cette théorie.)

3. MODÈLES DE RÉPONSE

Les poids de correction pour la non-réponse servant au calcul des estimateurs décrits dans la section précédente peuvent être justifiés au moyen d'un modèle de réponse qui prévoit une probabilité de réponse constante pour chaque unité d'un groupe donné. En termes plus formels, examinons le mécanisme de réponse suivant.

MODÈLE A:

1) probabilité de réponse est constante (et égale à une valeur inconnue Θ_h) pour toutes les unités $k \in s_h$; $h = 1, \ldots, H$;
2) les unités répondent indépendamment l'une de l'autre.

Les probabilités de réponse théoriques Θ_h peuvent varier considérablement d'un groupe à un autre. (Le fait qu'il peut exister de fortes différences de probabilité de réponse entre divers sous-ensembles porte évidemment à créer des groupes de correction et à faire les pondérations nécessaires.)

Considérons un échantillon permanent s. Les fréquences de groupe $n_1, \ldots, n_h, \ldots, n_H$ sont alors fixes. Considérons également une valeur constante pour le vecteur des fréquences de réponse des groupes $\underset{\sim}{m} = (m_1, \ldots, m_h, \ldots, m_H)$. Étant donné que s et $\underset{\sim}{n}$ sont fixes, il est possible de montrer que le prélèvement d'un ensemble de réponses r_h suivant le modèle A équivaut à l'échantillonnage aléatoire simple où l'on choisit m_h unités de n_h. La probabilité de réponse conditionnelle d'une unité k appartenant au h-ième groupe est donc définie:

$$\pi_{k|s,\underset{\sim}{m}} = \frac{m_h}{n_h} = f_h, \text{ pour tout } k \in s_h. \tag{3.1}$$

(Ce raisonnement est à la base du poids f_h^{-1} utilisé dans le calcul des estimateurs.) De même, étant donné s et $\underset{\sim}{m}$, il est possible de montrer que la probabilité que les unités k et l répondent toutes deux à l'enquête suivant le Modèle A est

$$\pi_{kl|s,\underset{\sim}{m}} = \begin{cases} f_h & \text{si } k = l \\ \dfrac{f_h(m_h - 1)}{n_h - 1} & \text{si } k \neq l \in s_h \\ f_h f_{h'} & \text{si } k \in s_h \,;\, l \in s_{h'} \;\; (h \neq h') \end{cases} \tag{3.2}$$

(par définition, $\pi_{kk|s,\underset{\sim}{m}}$ est égale à $\pi_{k|s,\underset{\sim}{m}}$). Ces valeurs (qui nous rappellent un échantillonnage aléatoire stratifié de m_h unités parmi n_h unités dans la h-ième strate) sont importantes pour le calcul des estimations de la variance et des erreurs types (voir ci-dessous).

En pratique, l'analyste décide de la méthode de formation des groupes s_h. Cette décision est cruciale puisqu'elle déterminera les poids de correction f_h^{-1} et, par voie de conséquence, la valeur numérique de l'estimation de t, l'estimation de la variance et l'intervalle de confiance. Deux modes de classement différents peuvent produire des estimations ponctuelles et des intervalles de confiance entièrement différents.

L'analyste n'ira pas s'imaginer que les groupes qu'il vient de créer ont tous la même probabilité de réponse. Il est toutefois convaincu (le plus souvent à juste titre) que ces groupes (et, par conséquent, les poids f_h^{-1}) lui permettront d'obtenir des estimations ponctuelles et des intervalles de confiance plus valables. La méthode des groupes de correction est une méthode fiable et solidement établie.

Après une analyse plus minutieuse, on peut découvrir plusieurs irrégularités dans un modèle de réponse comme le modèle A: la probabilité de réponse n'est peut-être pas constante à

l'intérieur des groupes. Et même si elle l'était, les groupes sont peut-être mal définis, c'est-à-dire que le modèle aurait dû prévoir un plus grand nombre de groupes, etc. Pour la suite de notre analyse, il est donc nécessaire de distinguer deux situations:

a) Le mécanisme de réponse hypothétique (MRH, en l'occurrence le modèle A) est exact (en pratique, il est peu probable qu'il en soit réellement ainsi).
b) Le MRH est plus ou moins fondé. C'est ce qu'on observe malheureusement dans la plupart des cas et il en découle un biais dû à la non-réponse. En ce qui concerne le modèle A, les groupes peuvent être formés plus ou moins correctement.

Comme cela se fait couramment en statistique, l'analyste définira son modèle en fonction de ce qu'il croit être les meilleurs critères; par conséquent, il fera certaines inférences (jugements de confiance par exemple). Il s'interrogera ensuite sur la robustesse de ses conclusions, c'est-à-dire qu'il cherchera à tester leur validité dans le cas où le modèle ne serait pas fondé. Dans le même ordre d'idées, appliquons ces questions au cas qui nous occupe.

4. ESTIMATEURS DE LA VARIANCE FONDÉS SUR UN MÉCANISME DE RÉPONSE HYPOTHÉTIQUE DONNÉ

Étant donné un ensemble de groupes précis, nous supposons que le modèle A est fondé. Les taux de réponses, $f_h = m_h/n_h$ $(h = 1, \ldots, H)$, ont été déterminés. À l'aide de ces renseignements préliminaires, nous allons étudier les estimateurs de la variance qui doivent servir à construire un intervalle de confiance à un niveau de $100(1 - \alpha)\%$. Si $\hat{t}$ est un des estimateurs définis à la section 2 et que le modèle A est vraiment exact, nous constatons ce qui suit:

a) $\hat{t}$ est non biaisé (sauf pour un biais technique habituellement négligeable)
b) un intervalle de confiance pour t à un seuil d'environ $100(1 - \alpha)\%$ est

$$\hat{t} \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{t})},$$

où la probabilité que la variable unitaire normale soit supérieure à la constante $z_{1-\alpha/2}$ est de $\alpha/2$.

Par des prélèvements successifs d'échantillons s et la formation répétée d'ensembles de réponses r pour chaque échantillon permanent s (conformément au modèle hypothétique A) l'intervalle contiendra le total réel de la population dans $100(1 - \alpha)\%$ des cas.

La variance et sa valeur estimée seront déterminées par deux séries de probabilités de sélection:

1. π_k et π_{kl}, les probabilités d'inclusion (du premier et du second ordre) qui se rattachent à l'étape de l'échantillonnage;
2. $\pi_{k|s,\underline{m}}$, $\pi_{kl|s,\underline{m}}$ les probabilités de réponse conditionnelles (du premier et du second ordre) qui se rattachent au modèle A (étape de la non-réponse).

Dans le cas qui nous occupe et en raison du modèle A, les probabilités $\pi_{k|s,\underline{m}}$, et $\pi_{kl|s,\underline{m}}$, sont définies respectivement par (3.1) et (3.2). En ce qui concerne π_k et π_{kl}, nous supposons qu'elles ont un caractère tout à fait général; nous pouvons utiliser n'importe quel plan de sondage.

Par une analyse détaillée, nous constaterons que la variance totale de n'importe quel estimateur $\hat{t}$ définit à la section 2 se divise en deux éléments:

$$V(\hat{t}) = V_1(\hat{t}) + V_2(\hat{t})$$

où $V_1(\hat{t})$ désigne la variance d'échantillonnage et $V_2(\hat{t})$ la variance due à la non-réponse. Bien que les formules exactes, définies dans Särndal et Swensson (1985a), ne soient pas reproduites ici, on constate que ces deux éléments ont des propriétés valables:

1. $V_1(\hat{t}) = 0$ si l'ensemble de la population U est observée (c'est-à-dire, s'il s'agit d'un recensement plutôt que d'une enquête par sondage);

2. $V_2(\hat{t}) = 0$ si le taux de non-réponse est nul ($r = s$);
3. $V_2(\hat{t})$ est réduite sensiblement si nous disposons d'une covariable forte; celle-ci n'a toutefois aucun effet sur $V_1(\hat{t})$ et avec raison puisqu'elle n'est observée que pour $k \in s$.

Regardons de plus près les estimateurs de la variance. Si $\hat{V}_i(\hat{t})$ désigne l'estimateur de $V_i(\hat{t})$, $i = 1, 2$, la variance totale $V(\hat{t})$ peut être estimée au moyen d'une expression du genre

$$\hat{V}(\hat{t}) = \hat{V}_1(\hat{t}) + \hat{V}_2(\hat{t}).$$

Dans l'équation ci-dessus, la variance d'échantillonnage estimée est

$$\hat{V}_1(\hat{t}) = \sum_{k \in r} \sum_{l \in r} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) \frac{1}{\pi_{kl|s,\underline{m}}} u_k u_l,$$

où $\pi_{kl|s,\underline{m}}$ est définie par (3.2), et π_k, π_{kl} sont les probabilités d'inclusion du plan de sondage. La variance de non-réponse estimée est

$$\hat{V}_2(\hat{t}) = \sum_{h=1}^{H} n_h^2 \left(\frac{1}{m_h} - \frac{1}{n_h} \right) S_{wr_h}^2$$

où

$$S_{wr_h}^2 = \frac{1}{m_h - 1} \sum\nolimits_{r_h} (w_k - \bar{w}_{r_h})^2.$$

Les quantités u_k et w_k varient d'un estimateur $\hat{t}$ à un autre. Examinons tout d'abord la variance de non-réponse estimée $\hat{V}_2(\hat{t})$. La définition de cette composante rappelle un échantillonnage stratifié. En effet, le facteur $n_h^2(1/m_h - 1/n_h)$ est une expression qui caractérise un échantillonnage aléatoire simple stratifié suivant lequel m_h unités sont prélevées parmi n_h unités dans la h-ième strate. Cette structure s'explique par les probabilités de réponse conditionnelles $\pi_{kl|s,\underline{m}}$ définies en (3.2).

Les quantités w_h sont définies de la façon suivante:

$$\text{Pour } \hat{t}_{\text{EXP}} \text{ et } \hat{t}^*_{\text{EXP}}\text{: } w_k = \frac{y_k - \tilde{y}_r}{\pi_k},$$

$$\text{pour } \hat{t}_{\text{RA}} \text{ et } \hat{t}^*_{\text{RA}}\text{: } \quad w_k = \frac{y_k - (\tilde{y}_r/\tilde{x}_r)x_k}{\pi_k},$$

$$\text{pour } \hat{t}_{\text{REG}} \text{ et } \hat{t}^*_{\text{REG}}\text{: } w_k = \frac{y_k - \tilde{y}_r - b(x_k - \tilde{x}_r)}{\pi_k}.$$

Les expressions ci-dessous sont des valeurs résiduelles de régression affectées d'un poids de sondage. Par conséquent, si x_k est une variable explicative puissante pour y_k, la variance de w_k (et, par voie de conséquence, $\hat{V}_2(\hat{t})$) sera moins élevée pour les estimateurs de type RA et REG que pour l'estimateur de type EXP, où la quantité w_k n'est que l'écart entre y_k et la moyenne de l'ensemble de réponse $\tilde{y}_r$. Ainsi, dans des conditions favorables, la portion de l'erreur type qui est attribuable à la non-réponse tendra vers zéro, notamment lorsqu'il aura une corrélation quasi parfaite entre x et y.

La variance d'échantillonnage $\hat{V}(\hat{t})$ présente moins d'intérêt dans cette analyse puisqu'elle n'est pas directement influencée par la covariable. Soulignons toutefois que les valeurs de u_k sont déterminées comme suit: pour les estimateurs $\hat{t}_{\text{EXP}}$, $\hat{t}_{\text{RA}}$, et $\hat{t}_{\text{REG}}$, $u_k = y_k$, tandis que pour la série d'estimateurs $\hat{t}^*_{\text{EXP}}$, $\hat{t}^*_{\text{RA}}$, et $\hat{t}^*_{\text{REG}}$, $u_k = y_k - \tilde{y}_s$, où $\tilde{y}_s = (\Sigma_s\, y_k/\pi_k)/(\Sigma_s\, 1/\pi_k)$ est la moyenne des valeurs estimées par la fonction de régression, de sorte que pour $\hat{t}^*_{\text{EXP}}$, $\hat{y}_k = \tilde{y}_r$ pour tout k; pour $\hat{t}^*_{\text{RA}}$, $\hat{y}_k = (\tilde{y}_r/\tilde{x}_r)x_k$; et pour $\hat{t}^*_{\text{REG}}$, $\hat{y}_k = \tilde{y}_r - b(x_k - \tilde{x}_r)$.

Un cas particulier se présente lorsque $m_h = n_h$ pour toutes les valeurs de h (c'est-à-dire lorsqu'il n'y a pas de non-réponse). Dans ce cas, $\hat{V}_2(\hat{t}) = 0$ (comme on peut s'y attendre), et $\pi_{kl|s,\underline{m}} = 1$ pour tout k et l, ce qui laisse la composante non nulle

$$\hat{V}_1(\hat{t}) = \sum_{k \in r} \sum_{l \in r} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) u_k u_l$$

qui est l'estimateur de la variance bien connu pour les cas où le taux de non-réponse est nul.

5. ROBUSTESSE DES ESTIMATEURS LORSQUE LE MÉCANISME DE RÉPONSE HYPOTHÉTIQUE N'EST PAS FONDÉ

Les estimateurs précités peuvent produire des estimations sans biais et des intervalles de confiance valables à la condition que le MRH (défini par le modèle A) soit exact. La présence d'une covariable forte a pour effet de réduire la variance due à la non-réponse.

Il est plus intéressant d'analyser ce qui se produit dans la réalité lorsque le MRH ne tient plus. Il est nécessaire d'envisager cette situation car il est impossible que les critères de jugement ayant servi à la formation de groupes de correction soient parfaits. L'efficacité des divers estimateurs dépendra alors de l'écart entre le mécanisme de réponse observé et le MRH. En d'autres mots, les propriétés statistiques (biais, étendue des intervalles de confiance, etc.) dépendent du degré de disparité entre le mécanisme de réponse réel et le mécanisme hypothétique.

Särndal et Swensson (1985a) contient une simulation de Monte-Carlo abrégée qui avait pour but d'analyser l'incidence de certaines formes de dérogation au modèle A. Pour les besoins de la présente étude, nous rapportons ici quelques résultats de cette simulation.

Dans le cas d'un MRH exact, il y avait 4 groupes de correction ($H = 4$) ayant chacun une probabilité de réponse différente (quoique les unités d'un même groupe avaient toutes la même probabilité de réponse). Mille échantillons aléatoires simples ont été prélevés et chacun a été soumis à une simulation de non-réponse suivant le MRH exact, (ce qui est tenu pour acquis étant donné qu'il s'agit d'une simulation contrôlée).

Conformément à la théorie, lorsque le MRH sur lequel reposent les estimateurs $\hat{t}_{\text{EXP}}$ et $\hat{t}_{\text{RA}}$ est exact, le biais est pratiquement inexistant dans les deux cas et le taux de couverture empirique des intervalles de confiance est fondamentalement conforme au seuil nominal de 95%. La chose qui distingue les deux estimateurs est que la variance due à la non-réponse est plus faible dans le cas de $\hat{t}_{\text{RA}}$. (Voir "MRH exact" dans le tableau 1.)

Pour créer un MRH non fondé, nous avons combiné des groupes du vrai MRH. L'estimateur et l'intervalle de confiance (qui reposent sur le MRH non fondé) seront donc calculés en fonction d'un nombre de groupes inférieur à la normale. La section inférieure du tableau 1 (MRH non fondé) correspond à la situation extrême où les quatre groupes du premier MRH ont été réunis en un seul, ce qui suppose que, pour les besoins d'estimation, toutes les unités de la population ont la même probabilité de réponse (inconnue mais estimée). D'après le tableau 1, nous pouvons voir que l'estimateur avec covariable présente certains avantages (peu surprenants) par rapport à l'estimateur sans covariable $\hat{t}_{\text{RA}}$: a) meilleure résistance au biais dû à la non-réponse (4.85 comparativement à 1.26); b) seuil de confiance beaucoup plus près du niveau nominal de 95% (taux de couverture empirique de 92.6% contre 46.3% pour $\hat{t}_{\text{EXP}}$).

Tableau 1

Comparaison de $\hat{t}_{EXP}$ et de $\hat{t}_{RA}$

	Estimateur	Biais absolu	Moyenne de la variance due à la la non-réponse $\hat{V}_2$	Taux de couverture empirique (seuil nominal de 95%)
MRH exact	$\hat{t}_{EXP}$	0.00	1.99	95.2%
	$\hat{t}_{RA}$	−0.01	0.78	95.5%
MRH non fondée	$\hat{t}_{EXP}$	4.85	2.55	46.3%
	$\hat{t}_{RA}$	1.26	0.78	92.6%

De plus, la variance due à la non-réponse est plus faible dans le cas de $\hat{t}_{RA}$, ce qui signifie, en moyenne, des intervalles de confiance plus étroits.

6. CONCLUSION

En résumé, nous avons vu dans cet article qu'il existe deux catégories de variables importantes (observées pour k dans l'échantillon visé s):

a) les variables servant à estimer le mécanisme de réponse (en ce qui concerne le modèle A, ces variables servent à former les groupes de correction);
b) les variables (en l'occurrence, les covariables) qui sont de puissants estimateurs de la variable y; lorsqu'elles sont utilisées dans la formule d'estimation, elles permettent de réduire la variance et d'accroître la robustesse.

Aussi, dans la mesure du possible, il serait souhaitable de trouver les covariables appropriées. Il convient aussi de souligner que, lorsqu'on doit estimer plusieurs totaux y, les covariables correspondantes peuvent différer d'une variable y à une autre, tandis que les groupes de pondération sont vraisemblablement formés de telle sorte qu'ils s'appliquent uniformément à toutes les variables à l'étude.

BIBLIOGRAPHIE

LITTLE, R.J.A. (1983). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

SÄRNDAL, C.E., et SWENSSON, B. (1985a). A general view of estimation for two phases of selection. Part I: Randomized subsample selection (Two-phase sampling). Part II: Nonrandomized subsample selection (Nonresponse). Promemorior fran P/STM no. 20, Statistics Sweden.

SÄRNDAL, C.E., et SWENSSON, B. (1985b). Incorporating nonresponse modelling in a general randomization theory approach. *Bulletin of the International Statistical Institute* (45th session), 51:3, 15.2.1-16.

OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit non-response. Dans *Incomplete Data in Sample Surveys*, Vol. 2, (éd. W.G. Madow, I. Olkin, et D.B. Rubin), New York: Academic Press, 143-183.