Estimation de variance sous une non-réponse monotone pour une enquête de type cohorte

Hélène Juillard [1] & Guillaume Chauvet [2]

[1] *INED, 133 boulevard Davout, 75020 Paris, France, helene.juillard@ined.fr*
[2] *ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France, chauvet@ensai.fr*

Résumé. Dans les enquêtes répétées dans le temps, plusieurs types de plans de sondage peuvent être utilisés selon les objectifs, tels que les enquêtes par panel ou par échantillons rotatifs. Dans ce travail, nous nous intéressons au cas particulier des enquêtes par cohorte, où le panel est constitué d'un échantillon d'individus liés par un évènement commun (par exemple, une naissance) et qui sont suivis au cours du temps. La cohorte est généralement sujette à un problème de non-réponse initiale, et à un problème d'attrition à tous les temps d'enquête, ce qui complique la production d'estimations et de mesures de précision associées.
Dans ce travail, nous considérons le cas d'une non-réponse totale monotone. La non-réponse est traitée à chaque temps d'enquête par une modélisation du mécanisme de réponse, puis par repondération par l'inverse des probabilités de réponse estimées. En nous basant sur les résultats de Kim et Kim (2007), nous donnons des estimateurs de variance approximativement sans biais et applicables à chaque temps d'enquête. Les résultats obtenus seront illustrés dans le cas particulier des groupes homogènes de réponse, souvent utilisés en pratique. Nous considérons également le cas d'estimateurs complexes et/ou d'estimateurs calés. Nous comparerons les résultats obtenus avec un estimateur simplifié de la variance due à la non-réponse, traitant les probabilités de réponse comme connues. Nous proposons également une application sur les données de l'Etude Longitudinale Française depuis l'enfance (ELFE).

Mots-clés. Enquête par cohorte, estimation de variance, groupe homogène de réponse, non-réponse totale.

1 Contexte

Dans les enquêtes répétées dans le temps, plusieurs types de plans de sondage peuvent être utilisés selon les objectifs, tels que les enquêtes par panel ou par échantillons rotatifs. Kalton (2009) définit les panels comme des enquêtes dans lesquelles les mêmes éléments sont suivis dans le temps. Par exemple, les enquêtes sur des panels de ménage consistent à sélectionner initialement un échantillon de ménages, puis à suivre dans le temps les

individus constituant ces ménages. Dans ce travail, nous nous intéressons au cas particulier des enquêtes par cohorte pour lesquelles le panel est constitué d'individus liés par un évènement commun (par exemple, une naisance) et qui sont suivis au cours du temps.

Nous nous intéressons à une population finie U. Un échantillon s_0 est tout d'abord sélectionné selon un plan de sondage $p(\cdot)$, et nous supposons que les probabilités d'inclusion π_i sont strictement positives pour tout $i \in U$. La première phase d'échantillonnage consiste en la sélection initiale des unités de l'échantillon. Par exemple, dans le cadre de l'Etude Longitudinale Française depuis l'Enfance (ELFE), un échantillon de nouveaux-nés a été sélectionné selon un plan de sondage produit (cross-classified sampling), où un échantillon de maternités et un échantillon de jours ont été sélectionnés indépendamment. L'enquête a été conduite dans les maternités sélectionnées lors des jours sélectionnés (Juillard et al., 2015). Nous notons également π_{ij} la probabilité que les unités i et j soient sélectionnées conjointement dans l'échantillon, et $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$.

Nous considérons le cas d'une enquête par cohorte, où seules les unités de l'échantillon d'origine sont suivies dans le temps, sans réintroduction de nouveaux individus à des dates ultérieures pour représenter de possibles naissances. Nous nous intéressons donc à un paramètre défini sur la population U, pour une variable d'intérêt y prenant la valeur y_i pour l'individu i. Les unités de l'échantillon s_0 sont ensuite suivies aux temps $\delta = 1, \ldots, t$, et l'échantillon peut être sujet à de la non-réponse totale à chaque temps. Nous notons r_i^δ l'indicatrice de réponse pour l'individu i au temps δ, et s_δ le sous-échantillon de répondants observé au temps δ. Nous supposons que la non-réponse est monotone, ce qui se traduit par la suite imbriquée

$$s_0 \supset s_1 \supset ... \supset s_t. \tag{1}$$

Pour $\delta = 1, \ldots, t$, nous notons

$$p_i^\delta \quad = \quad Pr\left(i \in s_\delta | s_{\delta-1}\right) \tag{2}$$

la probabilité pour l'individu i d'être répondant au temps δ. Nous supposons qu'à chaque temps δ, les unités répondent indépendamment les unes des autres, et nous notons

$$p_{ij}^\delta \quad = \quad Pr\left(i, j \in s_\delta | s_{\delta-1}\right) = p_i^\delta p_j^\delta \tag{3}$$

la probabilité que deux unités i et j répondent conjointement au temps δ.

2 Estimateur redressé de la non-réponse totale

En pratique, les probabilités de réponse à chaque temps sont inconnues et doivent être estimées. Nous supposons qu'à chaque temps δ, la probabilité de réponse est modélisée

paramétriquement sous la forme

$$p_i^\delta \;=\; p^\delta(z_i^\delta, \alpha^\delta) \tag{4}$$

pour une fonction connue $p^\delta(\cdot,\cdot)$, où z_i^δ est un vecteur de variables auxiliaires observé pour toutes les unités du sous-échantillon $s_{\delta-1}$, et α^δ désigne un paramètre inconnu. En suivant l'approche de Kim and Kim (2007), nous supposons que le vrai paramètre est estimé par $\hat{\alpha}^\delta$, qui est la solution de l'équation estimante

$$\frac{\partial}{\partial\alpha}\sum_{i\in s_{\delta-1}} k_i^\delta \left\{ r_i^\delta \ln(p_i^\delta) + (1-r_i^\delta)\ln(1-p_i^\delta)\right\} = 0, \tag{5}$$

avec k_i^δ un poids pour l'individu i dans l'équation estimante. Parmi les choix habituels, on trouve $k_i^\delta = 1$ et $k_i^\delta = \pi_i^{-1}$, voir Fuller et An (1998), Beaumont (2005) et Kim et Kim (2007).

La probabilité de réponse estimée au temps δ est

$$\hat{p}_i^\delta \;=\; p^\delta(z_i^\delta, \hat{\alpha}^\delta). \tag{6}$$

L'estimateur corrigé de la non-réponse au temps t est

$$\hat{Y}_t = \sum_{i\in s_\delta} \frac{y_i}{\pi_i \hat{p}_i^{1\to t}} \quad \text{avec} \quad \hat{p}_i^{1\to t} = \prod_{\delta=1}^{t} \hat{p}_i^\delta. \tag{7}$$

Sous de faibles hypothèses sur les mécanismes de réponse, et sous des conditions de régularité standard pour les fonctions $p^\delta(\cdot,\cdot)$ (voir les conditions R.1 et R.2 dans Kim et Kim, 2007), l'estimateur corrigé de la non-réponse totale est approximativement non biaisé pour le total Y.

3 Estimation de variance

La variance de $\hat{Y}_t$ est approximativement donnée par

$$V(\hat{Y}_t) \;\simeq\; V\left(\sum_{i\in s_0}\frac{y_i}{\pi_i}\right) + E\left\{\sum_{\delta=1}^{t} V(\hat{Y}_\delta | s_{\delta-1})\right\}. \tag{8}$$

Le premier terme du membre de droite de (8) représente la variance due au plan de sondage, que nous notons $V^p(\hat{Y}_t)$. Le second terme du membre de droite de (8) représente la variance due à la non-réponse, que nous notons $V^{nr}(\hat{Y}_t)$.

Au temps t, un estimateur approximativement sans biais pour la variance sous le plan de sondage est donné par

$$\hat{V}_t^p(\hat{Y}_t) = \sum_{i,j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \tag{9}$$

où $\hat{p}_{ij}^{1\to t} \equiv \prod_{\delta=1}^{t} \hat{p}_{ij}^{\delta}$, et où

$$\hat{p}_{ij}^{\delta} = \begin{cases} \hat{p}_i^{\delta} & \text{si } i = j, \\ \hat{p}_i^{\delta}\hat{p}_j^{\delta} & \text{sinon.} \end{cases} \tag{10}$$

En adaptant l'équation (25) de Kim et Kim (2007), $V^{nr}(\hat{Y}_t)$ peut être estimée approximativement sans biais au temps t par

$$\hat{V}_t^{nr}(\hat{Y}_t) = \sum_{\delta=1}^{t} \hat{V}_t^{nr\delta}(\hat{Y}_t) \tag{11}$$

avec

$$\hat{V}_t^{nr\delta}(\hat{Y}_t) = \sum_{i\in s_t} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} \left(\frac{y_i}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta}(\hat{h}_i^{\delta})^{\top}\hat{\gamma}_t^{\delta} \right)^2, \tag{12}$$

$$\hat{h}_i^{\delta} = h(z_i, \hat{\alpha}^{\delta}), \tag{13}$$

$$\hat{\gamma}_t^{\delta} = \left\{ \sum_{i\in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} \hat{h}_i^{\delta}(\hat{h}_i^{\delta})^{\top} \right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}} \hat{h}_i^{\delta} \frac{y_i}{\pi_i}. \tag{14}$$

Cela conduit à l'estimateur de variance global au temps t

$$\hat{V}_t(\hat{Y}_t) = \hat{V}_t^p(\hat{Y}_t) + \hat{V}_t^{nr}(\hat{Y}_t). \tag{15}$$

3.1 Application au modèle de régression logistique

Dans le cas particulier où un modèle de régression logistique est utilisé à chaque temps δ, le modèle (4) peut se réécrire

$$\text{logit}(p_i^{\delta}) = (z_i^{\delta})^{\top}\alpha^{\delta}. \tag{16}$$

Nous obtenons $\hat{h}_i^{\delta} = z_i^{\delta}$, et l'estimateur de la variance due à la non-réponse est donné par (11), avec

$$\hat{V}_t^{nr\delta}(\hat{Y}_t) = \sum_{i\in s_t} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} \left(\frac{y_i}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta}(z_i^{\delta})^{\top}\hat{\gamma}_t^{\delta} \right)^2, \tag{17}$$

$$\hat{\gamma}_t^{\delta} = \left\{ \sum_{i\in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta\to t}} z_i^{\delta}(z_i^{\delta})^{\top} \right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}} z_i^{\delta} \frac{y_i}{\pi_i}. \tag{18}$$

Si l'estimateur corrigé de la non-réponse est calculé au temps $t = 1$, l'estimateur en (11) de la variance due à la non-réponse peut se réécrire sous la forme

$$\begin{aligned}\hat{V}_1^{nr}(\hat{Y}_1) &= \hat{V}_1^{nr,1}(\hat{Y}_1) \\ &= \sum_{i\in s_1}(1-\hat{p}_i^1)\left(\frac{y_i}{\pi_i\hat{p}_i^1} - k_i^1(z_i^1)^\top\hat{\gamma}_1^1\right)^2. \end{aligned} \tag{19}$$

Si l'estimateur corrigé de la non-réponse est calculé au temps $t = 2$, l'estimateur en (11) de la variance due à la non-réponse peut se réécrire sous la forme

$$\begin{aligned}\hat{V}_2^{nr}(\hat{Y}_2) &= \hat{V}_2^{nr,1}(\hat{Y}_2) + \hat{V}_2^{nr,2}(\hat{Y}_2) \\ &= \sum_{i\in s_2}\frac{(1-\hat{p}_i^1)}{\hat{p}_i^2}\left(\frac{y_i}{\pi_i\hat{p}_i^1} - k_i^1(z_i^1)^\top\hat{\gamma}_2^1\right)^2 \\ &+ \sum_{i\in s_2}(1-\hat{p}_i^2)\left(\frac{y_i}{\pi_i\hat{p}_i^1\hat{p}_i^2} - k_i^2(z_i^2)^\top\hat{\gamma}_2^2\right)^2. \end{aligned} \tag{20}$$

4 Travail en cours

Le modèle des groupes homogènes de réponse est couramment utilisé en pratique pour corriger de la non-réponse totale. Ce modèle suppose que l'échantillon peut être découpé en sous-échantillons au sein desquels la probabilité de réponse est constante. Nous proposerons une application de l'estimateur de variance donné en formule (15) au cas des groupes homogènes de réponse.

Une dernière étape est généralement appliquée à l'estimateur $\hat{Y}_t$ du total obtenu au temps t. Si un vecteur x_i de variables auxiliaires est disponible pour chaque individu $i \in s_t$, et si le vecteur de leurs totaux X est connu, l'étape de calage (Deville et Särndal, 1992) permet de modifier les poids $\pi_i^{-1}(\hat{p}_i^{1\to t})^{-1}$ pour obtenir des poids calés w_{ti} permettant de reproduire les vrais totaux X. Nous proposerons une extension de l'estimateur de variance donné en formule (15) au cas d'un estimateur calé.

On peut également s'intéresser à des paramètres plus complexes qu'un total. Supposons que la variable d'intérêt y possède q composantes, et que le paramètre d'intérêt soit $\theta = f(Y)$ avec $f(\cdot)$ une fonction connue. Au temps t, en substituant $\hat{Y}_t$ dans θ, on obtient l'estimateur par substitution $\hat{\theta}_t = f(\hat{Y}_t)$ qui est approximativement non biaisé pour θ (see Deville, 1999; Goga, Deville and Ruiz-Gazen, 2011). Nous proposerons une extension de l'estimateur de variance donné en formule (15) pour le cas d'un estimateur par substitution. Nous considérerons également le cas où un estimateur calé est utilisé dans l'estimateur par substitution.

Nous comparerons également les performances de l'estimateur de variance proposé à l'estimateur naïf de la variance due à la non-réponse, obtenu en considérant les probabilités de réponse comme connues. Nous comparerons ces deux estimateurs au travers d'une étude par simulations, et proposerons une application sur des données issues de l'enquête ELFE.

Bibliographie

[1] Beaumont, J-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach, Journal of the Royal Statistical Society B, 67, pp. 445–458.
[2] Deville, J-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques, Survey Methodology, 25, pp. 193–203.
[3] Deville, J-C., et Särndal, C-E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), pp. 376–382. [4] Fuller, W.A., and An, A.B. (1998). Regression adjustment for non-response, Journal of the Indian Society of Agricultural Statistics, 51, pp. 331–342.
[5] Goga, C., Deville, J-C. and Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data, Biometrika, 96, pp. 691–709.
[6] Juillard, H., Chauvet, G., et Ruiz-Gazen (2016). Estimation under cross-classified sampling with application to a childhood survey. En révision pour Journal of the American Statistical Association.
[7] Kalton, G. (2009). Designs for surveys over time. Handbook of Statistics, 29, pp. 89–108.
[8] Kim, J. K., et Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. Canadian Journal of Statistics, 35(4), pp. 501–514.