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

Guillaume Chauvet (ENSAI)

En collaboration avec H. Juillard (INED)

Colloque Francophone sur les Sondages
Gatineau, le 13/10/16

ENSAI
École nationale de la statistique et de l'analyse de l'information

Résumé

Nous nous intéressons aux enquêtes par cohorte, avec suivi d'un panel d'individus liés par un évènement commun. La cohorte est généralement sujette à une non-réponse initiale, puis à un problème d'attrition à tous les temps d'enquête.

Nous considérons le cas d'une non-réponse totale monotone, traitée par repondération à chaque temps. En étendant les résultats de Kim et Kim (2007), nous donnons des estimateurs de variance applicables à chaque temps d'enquête.

Nous comparons les résultats obtenus avec un estimateur simplifié de la variance de non-réponse, traitant les probabilités de réponse comme connues.

ENSAI
École nationale de la statistique et de l'analyse de l'information

Notation

Traitement de la non-réponse initiale

Traitement de l'attrition

Etude par simulations

Notation

Notation

Soit une population d'individus U. Un échantillon s_0 est sélectionné selon un plan de sondage $p(\cdot)$, avec $\pi_i > 0$ pour tout $i \in U$. Dans le cadre de ELFE, sélection d'un échantillon de nouveaux-nés selon un plan de sondage produit (Juillard et al., 2015).

En l'absence de non-réponse, le total t_y peut être estimé sans biais par l'estimateur de Horvitz-Thompson

$$\hat{Y}_0 = \sum_{i \in s_0} \frac{y_i}{\pi_i} \quad \text{avec} \quad V(\hat{Y}_0) = \sum_{i,j \in U} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

Le terme $V(\hat{Y}_0)$ représente la variance d'échantillonnage (incompressible). Nous noterons $\hat{V}_t^p(\cdot)$ pour un estimateur de cette variance calculé au temps t.

ENSAI École nationale de la statistique et de l'analyse de l'information

Traitement de la non-réponse initiale

Traitement de la non-réponse initiale

En raison de la non-réponse initiale, seul un sous-échantillon de répondants $s_1 \subset s_0$ est observé. Le mécanisme de réponse est modélisé selon un plan de Poisson, avec p_i^1 la probabilité de réponse de l'individu i au temps $t = 1$.

On postule un modèle paramétrique pour la probabilité de réponse, logistique pour simplifier :

$$\text{logit}(p_i^1) \quad = \quad (z_i^1)^\top \alpha^1,$$

avec z_i^1 un vecteur de variables auxiliaires connues au temps $t = 1$.

L'estimation de α^1 conduit aux probabilités estimées $\hat{p}_i^1$ et à l'estimateur corrigé de la non-réponse initiale

$$\hat{Y}_1 \quad = \quad \sum_{i \in s_1} \frac{y_i}{\pi_i \ \hat{p}_i^1}.$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Traitement de la non-réponse initiale

Estimateur de variance

La variance de cet estimateur peut s'écrire

$$V(\hat{Y}_1) \quad = \quad V^p(\hat{Y}_1) + V^{nr}(\hat{Y}_1).$$

Variance d'échantillonnage : $V^p(\hat{Y}_1) = V(\hat{Y}_0)$.

Estim. de la variance de non-réponse $V^{nr}(\hat{Y}_1)$ (Kim and Kim, 2007) :

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

$$\text{avec } \hat{\gamma}_1^1 \quad = \quad \left\{ \sum_{i \in s_1} k_i^\delta (1 - \hat{p}_i^1)(z_i^1)^\top z_i^1 \right\}^{-1} \sum_{i \in s_1} \frac{1 - \hat{p}_i^1}{\hat{p}_i^1} (z_i^1) \frac{y_i}{\pi_i}.$$

Terme de centrage $k_i^1 (z_i^1)^\top \hat{\gamma}_1^1$: prédiction de $\frac{y_i}{\pi_i}$ par z_i^1.

Traitement de la non-réponse initiale

Estimateur simplifié de variance

Estimateur de la variance de non-réponse $V^{nr}(\hat{Y}_1)$:

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

Estimateur simplifié de la variance de non-réponse $V^{nr}(\hat{Y}_1)$:

$$\hat{V}_{1,simp}^{nr}(\hat{Y}_1) = \sum_{i \in s_1} (1 - \hat{p}_i^1) \left(\frac{y_i}{\pi_i \hat{p}_i^1} - 0 \right)^2 .$$

Terme de centrage $k_i^1 (z_i^1)^\top \hat{\gamma}_1^1$ ignoré.
Estime la variance que l'on aurait en repondérant par les vraies probabilités de réponse.
Tend à surestimer la variance de non-réponse.

Traitement de l'attrition

Traitement de l'attrition

Au temps t, seul un sous-échantillon $s_t \subset s_{t-1} \cdots \subset s_0$ (non-réponse monotone) est observé. Chaque mécanisme de réponse est modélisé selon un plan de Poisson, avec p_i^δ la probabilité de réponse de l'individu i au temps δ.

On postule un modèle paramétrique (logistique) pour la probabilité de réponse :

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

L'estimation des α^δ conduit à l'estimateur corrigé de la non-réponse initiale+attrition

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

ENSAI École nationale de la statistique et de l'analyse de l'information

Traitement de l'attrition

Estimateur de variance

La variance de cet estimateur peut s'écrire

$$V(\hat{Y}_t) \quad = \quad V^p(\hat{Y}_t) + \sum_{\delta=1}^{t} V^{nr\delta}(\hat{Y}_t).$$

Variance d'échantillonnage : $V^p(\hat{Y}_t) = V(\hat{Y}_0)$.

Estimateur de la variance de non-réponse $V^{nr}(\hat{Y}_t)$:

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

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

Traitement de l'attrition

Estimateur simplifié de variance

Estimateur de la variance de non-réponse $V^{nr}(\hat{Y}_t)$:

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

Estimateur simplifié de la variance de non-réponse $V^{nr}(\hat{Y}_t)$:

$$\begin{aligned} \hat{V}_t^{nr}(\hat{Y}_1) &= \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^{\delta}(1-\hat{p}_i^{\delta})}{\hat{p}_i^{\delta \to t}} \left(\frac{y_i}{\pi_i \hat{p}_i^{1 \to \delta}} - 0 \right)^2 \\ &= \sum_{i \in s_t} \frac{1-\hat{p}_i^{1 \to t}}{\left(\hat{p}_i^{1 \to t}\right)^2} \left(\frac{y_i}{\pi_i} \right)^2 . \end{aligned}$$

Estime la variance que l'on aurait en repondérant par les vraies probabilités de réponse à chaque temps : variance surestimée.
Tout se passe comme si s_t était obtenu par tirage poissonien de probabilités $p_i^{1 \to t} = \prod_{\delta=1}^{t} p_i^{\delta}$ dans s_0.

Discussion

Est-ce que la surestimation est importante ? Dépend du paramètre et de l'estimateur utilisé :

- Pour un total estimé par Horvitz-Thompson :
  $\rightarrow$ dépend du pouvoir explicatif des z_i^δ sur y_i,
  $\rightarrow$ potentiellement fort, mais HT peu utilisé en pratique.
- Pour un total estimé par calage :
  $\rightarrow$ dépend du pouvoir explicatif des z_i^δ sur les résidus $E_i = y_i - x_i^\top B_y$,
  $\rightarrow$ peut être fort si les variables de calage sont peu explicatives de la variable d'intérêt.
- Pour un paramètre complexe θ estimé par calage :
  $\rightarrow$ dépend du pouvoir explicatif des z_i^δ sur les résidus $E_{\theta i} = u_i - x_i^\top B_u$, avec u_i la linéarisée de θ
  $\rightarrow$ normalement faible

ENSAI École nationale de la statistique et de l'analyse de l'information

Etude par simulations

Cadre des simulations

Population de taille $10,000$ avec 3 variables d'intérêt :

$$\begin{aligned} y_{1i} &= 10 + 5x_{ai} + 5x_{bi} + 10u_{1i}, \quad (R^2 = 0.50) \\ y_{2i} &= 0.8y_{1i} + 10u_{2i}, \\ y_{3i} &= 0.8y_{2i} + 10u_{3i}. \end{aligned}$$

L'échantillon initial est sélectionné par sondage aléatoire simple de taille $n = 1,000$.

A chacun des temps $t = 1, 2, 3$, on considère un mécanisme de réponse de Poisson avec

$$\text{logit}(p_i^{\delta}) = -1 + \beta^{\delta a} x_{ai} + \beta^{\delta b} x_{bi}.$$

Probabilités de réponse moyennes de 0.75, 0.81 et 0.81.

ENSAI École nationale de la statistique et de l'analyse de l'information

Cadre des simulations

On réalise $5,000$ simulations. On s'intéresse à l'estimation

- des totaux Y_t,
- des ratios $R_t = Y_t / Y_1$,
- des évolutions $\Delta(1 \to t) = Y_t - Y_1$.

Les ratios et évolutions sont estimés par substitution.

On considère pour un total l'estimateur :

- de Horvitz-Thompson : $\hat{Y}_t$,
- calé sur les variables $1, x_{ai}, x_{bi}$ du modèle : $\hat{Y}_{wt}$,
- calé sur d'autres variables $1, x_{ci}, x_{di}$: $\hat{Y}_{\tilde{w}t}$.

Résultats

	t			t			t		
	1	2	3	1	2	3	1	2	3
	$\hat{Y}_t$			$\hat{Y}_{wt}$			$\hat{Y}_{\tilde{w}t}$		
$\mathsf{RB}_{\mathsf{mc}}(\hat{V})$	-0	-1	-2	-1	-1	-2	-1	-1	-3
$\mathsf{RB}_{\mathsf{mc}}(\hat{V}^{nr}_{simp})$	559	188	80	0	-1	-2	83	34	15
	$\hat{R}_t$			$\hat{R}_{wt}$			$\hat{R}_{\tilde{w}t}$		
$\mathsf{RB}_{\mathsf{mc}}(\hat{V})$	-	-0	-2	-	-1	-2	-	-1	-2
$\mathsf{RB}_{\mathsf{mc}}(\hat{V}^{nr}_{simp})$	-	0	0	-	-1	-2	-	-1	-1
	$\hat{\Delta}_{tt}(1 \to t)$			$\hat{\Delta}_{tt,w}(1 \to t)$			$\hat{\Delta}_{tt,\tilde{w}}(1 \to t)$		
$\mathsf{RB}_{\mathsf{mc}}(\hat{V})$	-	-0	-2	-	-0	-2	-	-1	-3
$\mathsf{RB}_{\mathsf{mc}}(\hat{V}^{nr}_{simp})$	-	19	30	-	-1	-2	-	3	5

ENSAI
École nationale de la statistique et de l'analyse de l'information

Références

Juillard, H., and Chauvet, G. (2016). Variance estimation under monotone non-response for a panel survey.

Juillard, H., Chauvet, G. and Ruiz-Gazen, A. (2016). Estimation under cross-classified sampling with application to a childhood survey. To appear in Journal of the American Statistical Association.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. Canadian Journal of Statistics 35, 501-514.

Lynn, P. (2009). Methods for Longitudinal Surveys. Methodology of Longitudinal Surveys, 1-19.

Skinner, C. (2015). Cross-classified sampling : some estimation theory. Statistics and Probability Letters 104, 163-168.

Annonce

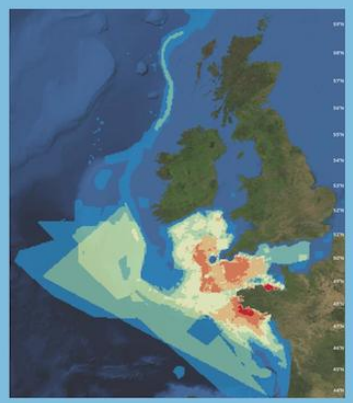
8èmes Journées des **METHODES AVANCÉES POUR L'ANALYSE DE SONDAGES COMPLEXES**

ECHANTILLONNAGE SPATIAL ET ESTIMATION SPATIALISÉE

Nantes le 08 et 09 novembre 2016, MSH Ange-Guépin
1ère journée : présentations orales théoriques et appliquées
2ème journée : tutoriel sur les méthodes de sondages avec *R*

CONFÉRENCIERS ET FORMATEURS :

Roberto BENEDETTI (Univ of Chieti-Pescara, dpt of Economics)
Guillaume CHAUVET (ENSAI)
Sébastien DEMANECHE (IFREMER, Brest)
Eric LESAGE (INSEE, Division Etudes Territoriales)
Alina MATEI (Univ Neuchâtel, Institut de Statistique)
Thomas MERLY-ALPA (INSEE, DMCSI)
Elodie PLISSONNEAU (GIS VALPENA, Univ Nantes)
Audrey-Anne VALLÉE (Univ Neuchâtel, Institut de Statistique)
Aurélie VANHEUVERZWYN (Médiamétrie)



CRPMEM Bretagne

CDPMEM 29

COMITÉ D'ORGANISATION :

Lise BELLANGER (Univ de Nantes)
Guillaume CHAUVET (ENSAI)
Elodie PLISSONNEAU (GIS VALPENA, Univ de Nantes)
Brice TROUILLET (Univ de Nantes)

Participation gratuite. Nombre de places limité.
Informations et inscriptions : http://maasc2016.sciencesconf.org

ENSAI École nationale de la statistique et de l'analyse de l'information