Consistance sous un modèle de réponse de la fonction de répartition estimée en présence de données manquantes

Hélène Boistard (Université Toulouse I)
Guillaume Chauvet (Ensai, Crest)
David Haziza (Université de Montréal)

Colloque francophone sur les sondages
Rennes, 07/11/2012

Plan de l'exposé

Notations

Notation

On considère une population finie d'individus $U = \{1, \ldots, k, \ldots, N\}$, où chaque individu est supposé identifiable par son label. Pour chaque individu k, soit :

- y_k la valeur prise par une variable d'intérêt y,
- π_k sa probabilité de sélection (> 0),
- $d_k = 1/\pi_k$ le poids de sondage.

Notation

On considère une population finie d'individus $U = \{1, \ldots, k, \ldots, N\}$, où chaque individu est supposé identifiable par son label. Pour chaque individu k, soit :

- y_k la valeur prise par une variable d'intérêt y,
- π_k sa probabilité de sélection (> 0),
- $d_k = 1/\pi_k$ le poids de sondage.

En situation de réponse complète, on estime sans biais

$$t_y = \sum_{k \in U} y_k \quad \text{par} \quad \hat{t}_{y\pi} = \sum_{k \in S} d_k y_k,$$

$$F_N(t) = N^{-1} \sum_{k \in U} 1(y_k \leq t) \quad \text{par} \quad \hat{F}_N(t) = \hat{N}^{-1} \sum_{k \in S} d_k 1(y_k \leq t).$$

Estimateur imputé

En cas de non-réponse pour y, une valeur manquante y_k est généralement remplacée par une valeur imputée y_k^* (e.g. Haziza, 2009) :

$$\begin{aligned}\hat{t}_{yI} &= \sum_{k\in S_r} d_k y_k + \sum_{k\in S_m} d_k y_k^*, \\ \hat{F}_I(t) &= \hat{N}^{-1}[\sum_{k\in S_r} d_k 1(y_k \leq t) + \sum_{k\in S_m} d_k 1(y_k^* \leq t)].\end{aligned}$$

Estimateur imputé

En cas de non-réponse pour y, une valeur manquante y_k est généralement remplacée par une valeur imputée y_k^* (e.g. Haziza, 2009) :

$$\begin{aligned} \hat{t}_{yI} &= \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k y_k^*, \\ \hat{F}_I(t) &= \hat{N}^{-1} [\sum_{k \in S_r} d_k 1(y_k \leq t) + \sum_{k \in S_m} d_k 1(y_k^* \leq t)]. \end{aligned}$$

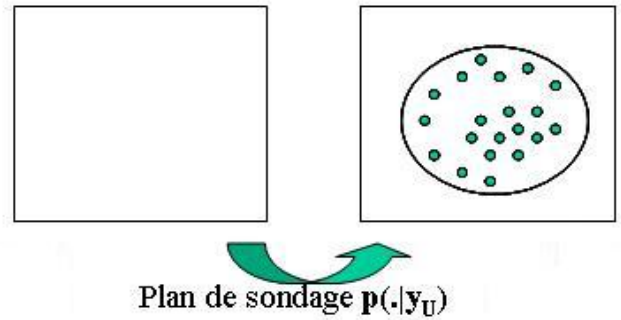
Deux mécanismes aléatoires supplémentaires interviennent : le *méca. de non-réponse* ($S \Rightarrow S_r$), et le *méca. d'imputation* ($y_k \Rightarrow y_k^*$).

Estimateur imputé

En cas de non-réponse pour y, une valeur manquante y_k est généralement remplacée par une valeur imputée y_k^* (e.g. Haziza, 2009) :

$$\begin{aligned} \hat{t}_{yI} &= \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k y_k^*, \\ \hat{F}_I(t) &= \hat{N}^{-1} [\sum_{k \in S_r} d_k 1(y_k \leq t) + \sum_{k \in S_m} d_k 1(y_k^* \leq t)]. \end{aligned}$$

Deux mécanismes aléatoires supplémentaires interviennent : le *méca. de non-réponse* ($S \Rightarrow S_r$), et le *méca. d'imputation* ($y_k \Rightarrow y_k^*$).

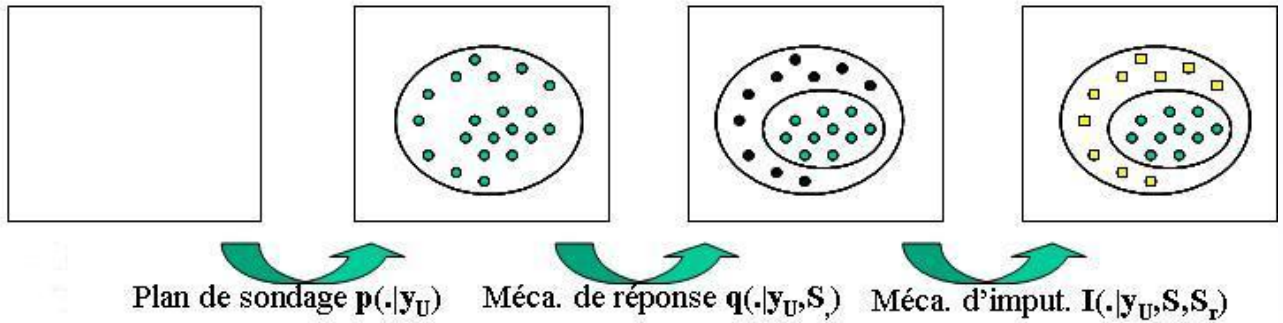


Estimateur imputé

En cas de non-réponse pour y, une valeur manquante y_k est généralement remplacée par une valeur imputée y_k^* (e.g. Haziza, 2009) :

$$\begin{aligned}\hat{t}_{yI} &= \sum_{k\in S_r} d_k y_k + \sum_{k\in S_m} d_k y_k^*, \\ \hat{F}_I(t) &= \hat{N}^{-1}[\sum_{k\in S_r} d_k 1(y_k \leq t) + \sum_{k\in S_m} d_k 1(y_k^* \leq t)].\end{aligned}$$

Deux mécanismes aléatoires supplémentaires interviennent : le *méca. de non-réponse* ($S \Rightarrow S_r$), et le *méca. d'imputation* ($y_k \Rightarrow y_k^*$).



Estimateur imputé du total

Modèle d'imputation

Le mécanisme d'imputation est motivé par un *modèle d'imputation*
$\Rightarrow$ prédiction de y_k à l'aide d'une information auxiliaire $\mathbf{x}_k$:

$$m \; : \; y_k \quad = \quad \mathbf{x}_k^\top \, \boldsymbol{\beta} + \sigma \sqrt{v_k} \epsilon_k.$$

Dans ce modèle :

- $\boldsymbol{\beta}$ et σ^2 sont des paramètres inconnus,
- v_k est une constante connue,
- les résidus ϵ_k sont des variables aléatoires **iid**, centrées réduites.

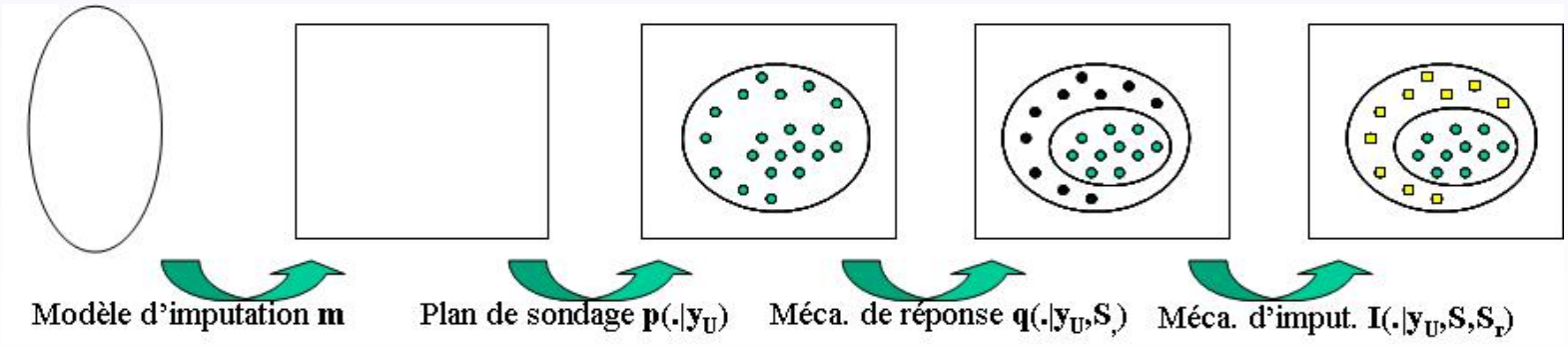
Modèle d'imputation

Le mécanisme d'imputation est motivé par un *modèle d'imputation* $\Rightarrow$ prédiction de y_k à l'aide d'une information auxiliaire $\mathbf{x}_k$:

$$m \; : \; y_k \quad = \quad \mathbf{x}_k^\top \, \boldsymbol{\beta} + \sigma \sqrt{v_k} \epsilon_k .$$

Dans ce modèle :

- $\boldsymbol{\beta}$ et σ^2 sont des paramètres inconnus,
- v_k est une constante connue,
- les résidus ϵ_k sont des variables aléatoires **iid**, centrées réduites.



Imputation déterministe

L'*imputation par la régression déterministe* est obtenue en prenant $y_k^* = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_r$, avec

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{x}_k y_k$$

un estimateur du paramètre $\boldsymbol{\beta}$ inconnu, et ω_k un **poids d'imputation** (> 0) associé à l'unité k (Haziza, 2009).

Imputation déterministe

L'*imputation par la régression déterministe* est obtenue en prenant $y_k^* = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_r$, avec

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{x}_k y_k$$

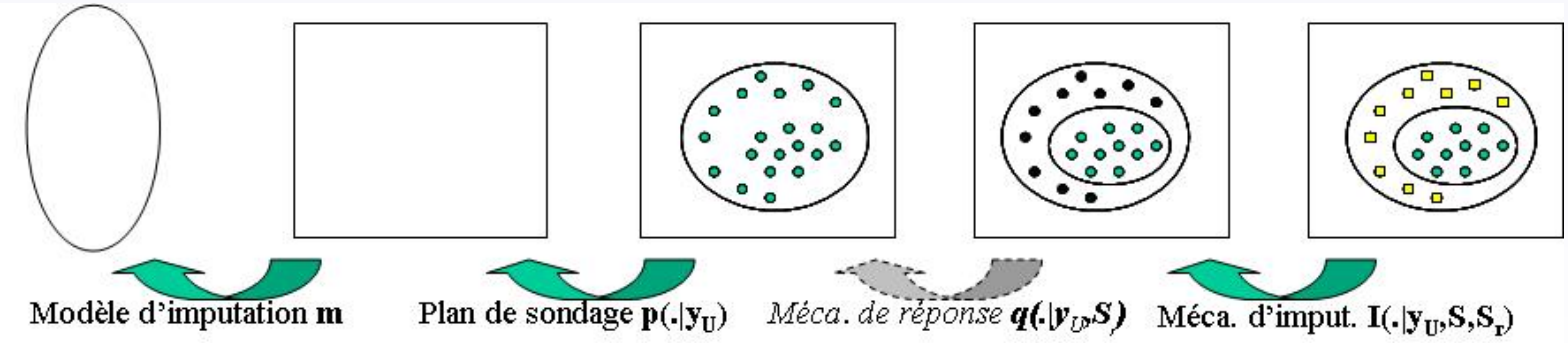
un estimateur du paramètre $\boldsymbol{\beta}$ inconnu, et ω_k un **poids d'imputation** (> 0) associé à l'unité k (Haziza, 2009).

Dans ce cas, l'estimateur imputé du total est égal à

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k \left[\mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_r \right].$$

Approche sous le Modèle d'Imputation (IM)

L'inférence se fait sous le modèle (postulé) d'imputation. Le mécanisme de non-réponse n'est pas explicitement modélisé.

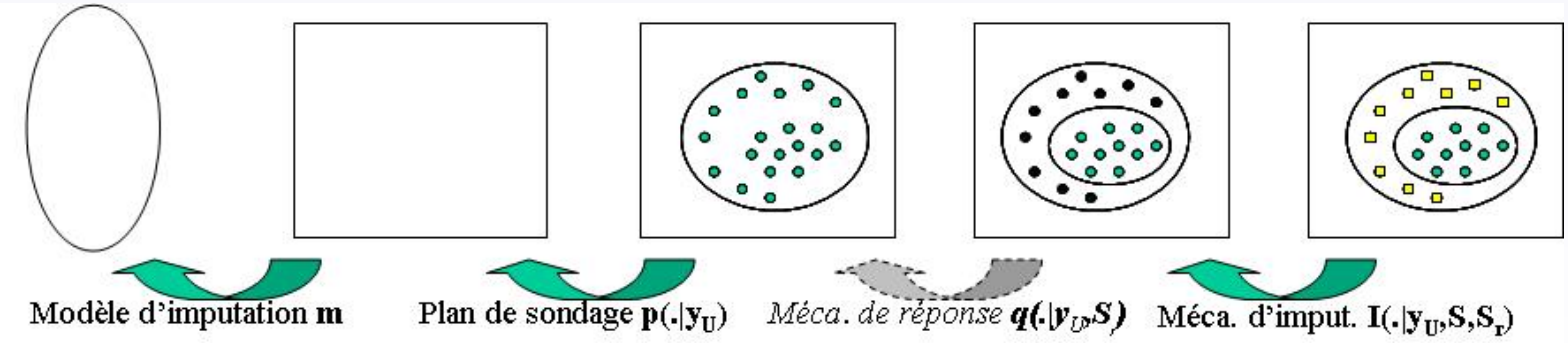


Approche sous le Modèle d'Imputation (IM)

L'inférence se fait sous le modèle (postulé) d'imputation. Le mécanisme de non-réponse n'est pas explicitement modélisé.

Un choix quelconque des poids d'imputation ω_k conduit à un estimateur du total approximativement mpq non biaisé :

$$E_m E_p E_q(\hat{t}_{yI} - t_y) \quad \simeq \quad 0.$$



Approche sous le Modèle de Non-Réponse (NM)

Pour se prémunir contre une mauvaise spécification du modèle m, il est intéressant de disposer d'un méca. d'imputation donnant une estimation non biaisée sous un modèle de non-réponse.

Approche sous le Modèle de Non-Réponse (NM)

Pour se prémunir contre une mauvaise spécification du modèle m, il est intéressant de disposer d'un méca. d'imputation donnant une estimation non biaisée sous un modèle de non-réponse.

Modèle de non-réponse : jeu d'hypothèses sur le mécanisme (inconnu) de non-réponse. On suppose ici que la probabilité de réponse à la variable y_k suit le modèle logistique

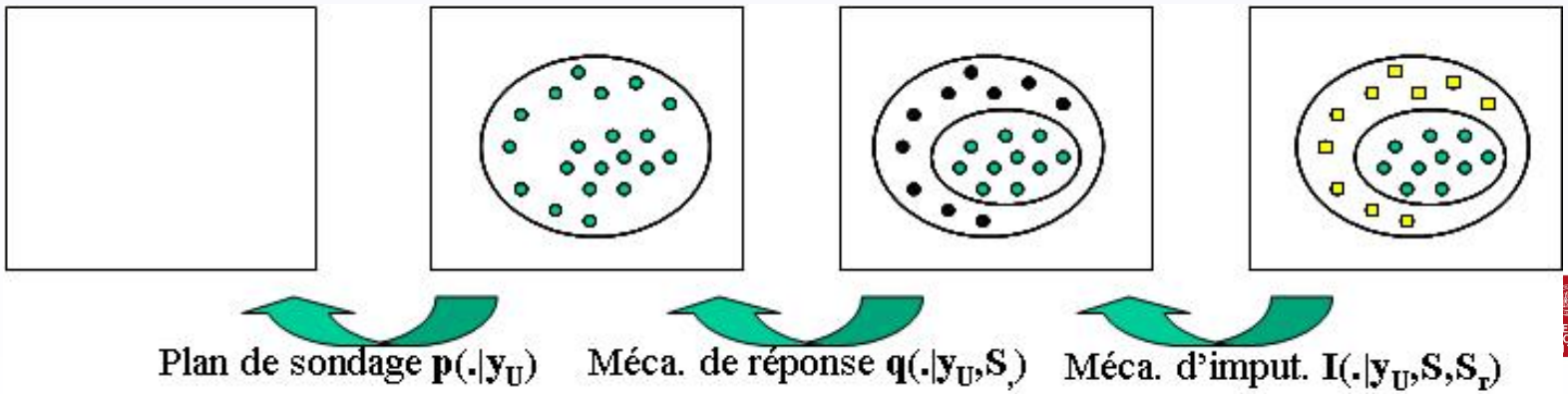
$$p_k \equiv \Pr\left(r_k = 1\right) \quad = \quad \frac{\exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_k\right)}{1 + \exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_k\right)}.$$

Approche sous le Modèle de Non-Réponse (NM)

Pour se prémunir contre une mauvaise spécification du modèle m, il est intéressant de disposer d'un méca. d'imputation donnant une estimation non biaisée sous un modèle de non-réponse.

Modèle de non-réponse : jeu d'hypothèses sur le mécanisme (inconnu) de non-réponse. On suppose ici que la probabilité de réponse à la variable y_k suit le modèle logistique

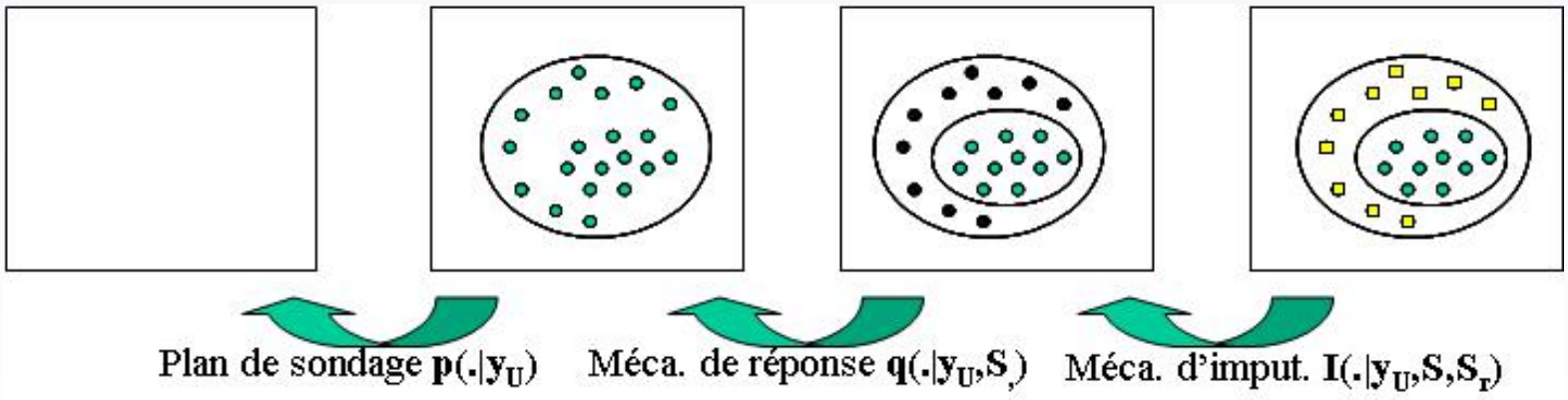
$$p_k \equiv \Pr(r_k = 1) = \frac{\exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_k\right)}{1 + \exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_k\right)}.$$



Approche sous le Modèle de Non-Réponse (NM)

L'utilisation du mécanisme d'imputation par la régression déterministe, avec les **poids d'imputation** $\omega_k = d_k \frac{1-p_k}{p_k}$, conduit à un estimateur imputé du total pq non biaisé (Haziza et Rao, 2006) :

$$E_p E_q(\hat{t}_{yI} - t_y) \simeq 0.$$

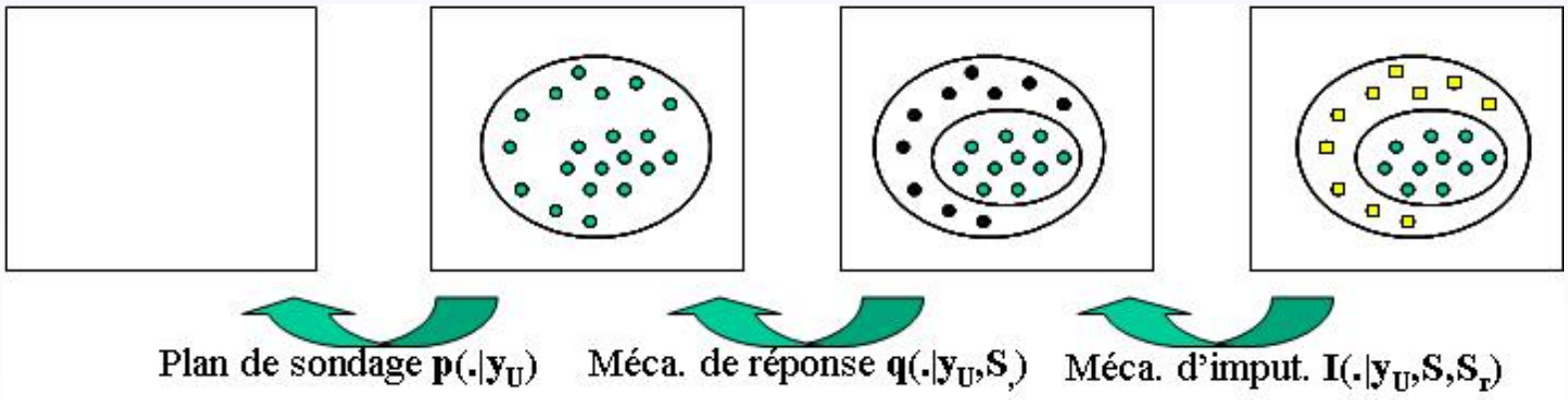


Approche sous le Modèle de Non-Réponse (NM)

L'utilisation du mécanisme d'imputation par la régression déterministe, avec les **poids d'imputation** $\omega_k = d_k \frac{1-p_k}{p_k}$, conduit à un estimateur imputé du total pq non biaisé (Haziza et Rao, 2006) :

$$E_p E_q(\hat{t}_{yI} - t_y) \simeq 0.$$

On parle d'**estimation doublement robuste** du total, voir également Kott (1994), Kim et Park (2006).



Estimateur imputé de la fonction de répartition

Imputation déterministe

L'utilisation d'une imputation par la régression déterministe conduit à l'estimateur

$$\hat{F}_I(t) = \frac{1}{\hat{N}} \left[\sum_{k \in S_r} d_k 1(y_k \leq t) + \sum_{k \in S_m} d_k 1(\mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_r \leq t) \right].$$

Cependant, cet estimateur est mpq biaisé. Pour résoudre ce problème, Chambers et Dunstan (1986) ont proposé un estimateur corrigé du biais a posteriori.

Autre solution : méthode d'imputation aléatoire.

Imputation aléatoire : approche IM

L'imputation par la régression aléatoire est obtenue en prenant

$$y_k^* \quad = \quad \mathbf{x}_k^\top \, \hat{\beta}_r + \hat{\sigma}\sqrt{v_k}\epsilon_k^*,$$

i.e. en rajoutant à la prédiction de y_k un terme aléatoire.

Les résidus ϵ_k^* sont tirés au hasard et avec remise, **avec des probabilités proportionnelles aux poids d'imputation** ω_k, parmi les résidus observés sur les répondants.

L'estimateur imputé $\hat{t}_{yI}$ est approximativement $mpqI$ non biaisé :

$$E_m E_p E_q E_I(\hat{t}_{yI} - t_y) \quad \simeq \quad 0.$$

Imputation aléatoire : approche IM

Sous des hypothèses standard, Chauvet, Deville et Haziza (2011) montrent que la fonction de répartition imputée par la régression aléatoire est consistante sous l'approche IM :

$$\hat{F}_I(t) - F_N(t) \quad \rightarrow_{\mathbb{P}} \quad 0.$$

Ce résultat s'étend au cas où les résidus aléatoires sont sélectionnés de façon équilibrée :

$$\sum_{k \in S} (1 - r_k) \sqrt{v_k} \epsilon_k^* = 0 \quad \Rightarrow \quad \hat{t}_{yI} = E_I(\hat{t}_{yI} | y_U, S, S_r)$$

$$\Rightarrow \quad \text{variance d'imputation annulée.}$$

Imputation aléatoire : approche NM

Pour se prémunir contre une mauvaise spécification du modèle d'imputation, on cherche un mécanisme d'imputation donnant une estimation non biaisée sous l'approche NM.

Nous nous restreignons ici au cas du hot-deck aléatoire : une valeur manquante y_k est remplacée en sélectionnant au hasard et avec remise un donneur $y_j \in S_r$, avec des probabilités proportionnelles aux poids d'imputation ω_j.

Théorème (BCH, 2012)

Le choix $\omega_k = d_k \frac{1-p_k}{p_k}$ *dans le hot-deck aléatoire donne une estimation consistante de* $F_N(\cdot)$ *sous l'approche NM :*

$$\hat{F}_I(t) - F_N(t) \quad \to_{\mathbb{P}} \quad 0.$$

Etude par simulations

Cadre

Population de taille $N = 10\ 000$, générée selon le modèle

$$y_k \quad = \quad 10 + x_{1i} + x_{2i} + \eta_i,$$

où les x_{1i}, x_{2i} sont générés selon une loi gamma et les η_i selon une loi normale centrée. On utilise $R^2 = 0.70$.

Echantillon S de taille $n = 500$ sélectionné par sondage aléatoire simple. La non-réponse est générée selon un mécanisme poissonien, avec

$$Pr(r_i = 1|x_{1i}, x_{2i}) \quad = \quad \frac{\exp\left(-1 + 1.6\ x_{1i} + 1.6\ x_{2i}\right)}{1 + \exp\left(-1 + 1.6\ x_{1i} + 1.6\ x_{2i}\right)}.$$

Probabilité de réponse moyenne de 0.60.

Imputation par la régression aléatoire

On réalise $B = 1000$ simulations. On s'intéresse à l'estimation de $F_N(t_\alpha)$, avec $\alpha = 0.05,\ 0.25,\ 0.50,\ 0.75,\ 0.95$.

Pour illustrer les risques d'une mauvaise spécification du modèle d'imputation, on examine les performances de l'**imputation par la régression aléatoire non pondérée** :

- avec le modèle correct : $\mathbf{x} = (1, x_1, x_2)$,
- avec un modèle incomplet : $\mathbf{x} = (1, x_1)$.

On calcule le biais relatif (RB), et le MSE relatif

$$\text{RMSE}\{\hat{F}_I(t)\} = \frac{\sqrt{\text{MSE}\{\hat{F}_I(t)\}}}{F_N(t)} \times 100.$$

Résultats obtenus

			α				
			0.05	0.25	0.50	0.75	0.95
$\mathbf{x} = (1, x1, x2)$	REGI	RB	0.9	0.8	0.1	-0.1	0.0
		RMSE	23.9	9.3	5.0	2.7	1.1
$\mathbf{x} = (1, x1)$	REGI	RB	-18.8	-12.9	-8.5	-4.6	-1.0
		RMSE	29.4	16.0	10.2	5.6	1.7

Tab.: Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour la fonction de répartition imputée par la régression aléatoire

Résultats obtenus

			α				
			0.05	0.25	0.50	0.75	0.95
$\mathbf{x} = (1, x1, x2)$	REGI	RB	0.9	0.8	0.1	-0.1	0.0
		RMSE	23.9	9.3	5.0	2.7	1.1
$\mathbf{x} = (1, x1)$	REGI	RB	-18.8	-12.9	-8.5	-4.6	-1.0
		RMSE	29.4	16.0	10.2	5.6	1.7

Tab.: Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour la fonction de répartition imputée par la régression aléatoire

Imputation par hot-deck

On s'intéresse également à l'estimation de $F_N(t_\alpha)$, avec $\alpha = 0.05,\ 0.25,\ 0.50,\ 0.75,\ 0.95$. On modélise le mécanisme de non-réponse de façon :

- correcte : $\mathbf{x} = (1, x_1, x_2) \Rightarrow \hat{p}_1$,
- incomplète : $\mathbf{x} = (1, x_1) \Rightarrow \hat{p}_2$.

On examine les performances de l'**imputation par hot-deck** :

- non-pondéré : $\omega_k = 1$,
- pondéré avec les bonnes probas de réponse : $\omega_k = d_k \ \frac{1-\hat{p}_1}{\hat{p}_1}$,
- pondéré avec les mauvaises probas de réponse : $\omega_k = d_k \ \frac{1-\hat{p}_2}{\hat{p}_2}$.

Résultats obtenus

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Tab.: Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour trois méthodes de hot-deck aléatoire

Résultats obtenus

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Tab.: Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour trois méthodes de hot-deck aléatoire

Résultats obtenus

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Tab.: Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour trois méthodes de hot-deck aléatoire

Bibliographie

Chambers, R.L., Dunstan, R. (1986). *Estimating distribution functions from survey data*, Biometrika, vol 73, pp. 597–604.

Chauvet, G., Deville, J.-C., and Haziza, D. (2011). *On balanced random imputation in surveys*, Biometrika, vol 98, 459-471.

Deville, J.-C., and Tillé, Y. (2004). *Efficient balanced sampling : the cube method*, Biometrika, 91, pages 893-912.

Haziza, D. (2009). *Imputation and inference in the presence of missing data*, Handbook of Statistics, vol.29, chap. 10.

Kim, J.K. and Park, H.A. (2006). *Imputation using response probability*. Canadian Journal Statistics 34, 171-182.

Kott, P.S. (1994). *A note on handling nonresponse in sample surveys*. Journal of the American Statistical Association 89, 693-696.