L'imputation non-paramétrique par des fonctions B-spline pour le traitement de la non-réponse partielle dans les enquêtes par sondages

Camelia Goga
LMB - Univ. de Bourgogne Franche-Comté
en collaboration avec
David Haziza - Univ. de Montréal

Journée Besançon-Dijon
25 juin 2018

Plan de l'exposé

- Motivation : estimation de totaux en présence de la non-réponse partielle ;
- Un nouvel estimateur imputé basé sur la régression non-paramétrique par des fonctions B-splines ;
- Propriétés asymptotiques et une étude par simulations ;

Population, échantillon et l'estimateur d'Horvitz-Thompson

- Soit $U = \{1, \ldots, k, \ldots, N\}$ une population finie et soit $s \subset U$ un échantillon sélectionné selon un plan de sondage sample $p(\cdot)$;
- Soient $\pi_k = Pr(k \in s) = \sum_{k \in s} p(s)$ et $\pi_{kl} = Pr(k, l \in s) = \sum_{k,l \in s} p(s)$ les probabilités d'inclusion de premier et deuxième degré ;
- Soit $\mathcal{Y}$ une variable d'intérêt et l'objectif est l'estimation du total :

$$t_y = \sum_{k \in U} y_k$$

- Si tous les individus échantillonnés **répondent** et si $\pi_k > 0$ pour tous les $k \in U$, alors le total t_y est estimé sans biais par l'estimateur d'Horvitz-Thompson :

$$\hat{t}_{HT} = \sum_{k \in s} \frac{1}{\pi_k} y_k = \sum_{k \in U} \frac{1}{\pi_k} y_k 1_{\{k \in s\}}$$

$$\text{Var}(\hat{t}_{HT}) = \sum_{k \in U} \sum_{k \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Et en présence de la non-réponse ...

Malheureusement, il y a des personnes échantillonnées

- qui n'ont répondu à aucune question : non-réponse totale
- qui ont répondu qu'à certaines questions : non-response partielle ;

indiv.	Y_1	Y_2	Y_3
1	y_{11}	y_{12}	y_{13}
2	?	?	?
3	?	y_{32}	y_{33}
4	y_{41}	y_{42}	y_{43}
5	?	?	?
6	?	y_{62}	?

- La non-réponse arrive dans quasiment toutes les enquêtes d'individus, la classe d'âge la plus "touchée" est 18-25 ans (jusqu'à 80%-90% de non-réponse) ;
- Dans d'autres domaines comme l'agriculture, "wildlife surveys", ... le terme "données manquantes" est plus utilisé.

Pourquoi "traiter" les non-répondants ?

- La pratique la plus courante est d'enlever les non-répondants du fichier.
- Dans ce cas les estimateurs construits à partir de répondants seulement peuvent être très biaisés :

	taille	total	moyenne	variance
répondants	N_r	t_r	$\overline{y}_{rU}$	S_r^2
non-répondants	N_m	t_m	$\overline{y}_{mU}$	S_m^2
population	N	t	$\overline{y}_U$	S_U^2

Soit $\overline{y}_r$ un estimateur (approx.) sans biais de $\overline{y}_{rU}$, on a

$$\overline{y}_U = \frac{N_r}{N}\overline{y}_{rU} + \frac{N_m}{N}\overline{y}_{mU}$$

et

$$\mathbb{E}(\overline{y}_r) - \overline{y}_U \simeq \frac{N_m}{N}(\overline{y}_{rU} - \overline{y}_{mU}).$$

Ce biais est petit si $N_m \simeq 0$ ou $\overline{y}_{rU} - \overline{y}_{mU} \simeq 0$ ce qui est très rarement satisfait !

Mécanisme de non-réponse

- Soit la variable réponse

$$r_k = \begin{cases} 1 & \text{si l'individu k répond} \\ 0 & \text{sinon} \end{cases}$$

- La probabilité de réponse est $\phi_k = P(r_k = 1)$ qui est inconnue.

Trois types de mécanisme de non-réponse :

1. **Missing completely at random (MCAR)** : ϕ ne dépend ni de Y ni des variables auxiliaires $\mathbf{Z}$ (age, sexe, ...) ;
2. **Missing at random (MAR)** ou mécanisme de réponse ignorable : ϕ dépend de $\mathbf{Z}$ (ex. age) mais pas de Y ; la non-réponse dépend que des valeurs observées de Y ;
3. **Not Missing at random (NMAR)** ou mécanisme de réponse non-ignorable : ϕ dépend de Y (ex. revenu) donc des valeurs observées et non-observées de Y ;

Traitement de la non-réponse partielle

- Soit s_m l'ensemble des *non-répondants*= l'ensemble des individus pour lesquels $\mathcal{Y}$ est manquant et s_r l'ensemble des *répondants* ;
- La valeur manquante y_k, pour $k \in s_m$, est remplacée par une valeur imputée y_k^* ;
- Le total t_y est estimé par l'estimateur imputé :

$$\begin{aligned} \hat{t}_I &= \sum_{k \in s_r} \frac{y_k}{\pi_k} + \sum_{k \in s_m} \frac{y_k^*}{\pi_k} \\ &= \sum_{k \in s} \frac{1}{\pi_k} (r_k y_k + (1 - r_k) y_k^*) \end{aligned}$$

où r_k est la variable de réponse.

- Le biais et la variance de $\hat{t}_I$ plus compliqués à calculer : le plan d'échantillonnage, le mécanisme de non-réponse et d'imputation ;

Méthodes d'imputation dans le cas MAR

- Des variables auxiliaires connues pour les répondants et les non-répondants sont utilisées pour construire un modèle pour prédire les valeurs manquantes (mais elles ne seront jamais aussi bien que les vraies valeurs) ;

indiv.	Y_1	Y_2	Y_3	Z_1 (âge)	Z_2 (sexe)
1	y_{11}	y_{12}	y_{13}	x_{11}	x_{12}
2	?	?	?	x_{21}	x_{22}
3	?	y_{32}	y_{33}	x_{31}	x_{32}
4	y_{41}	y_{42}	y_{43}	x_{41}	x_{42}
5	?	?	?	x_{51}	x_{52}
6	?	y_{62}	?	x_{61}	x_{62}

- Méthodes d'imputation aléatoires ou pas, paramétriques telles que l'imputation par la moyenne ou par le ratio, ou l'imputation hot-deck (survol dans Haziza, 2009) ;
- Méthode d'imputation non-paramétriques par le plus proche-voisin proposée par Chen and Shao (2000) ; Beaumont and Bocci (2009) ;

Estimateur non-paramétrique basé sur les B-splines

- On considère que la population U est divisée dans J classes disjointes $C_j, j = 1, \ldots, J$ (Chen & Shao, 2000 ; Cardot, De Moliner, Goga, 2018) ;

- On considère qu'une variable auxiliaire $\mathcal{Z}$ (uni-variée) est disponible et z_k est connu pour tous les $k \in U$;

- Soit un modèle non-paramétrique à l'intérieur de chaque classe :

$$y_k = f_j(z_k) + \varepsilon_k, \quad k \in C_j$$

où f_j est une fonction inconnue et lisse ; les résidus ε_k sont indépendants de moyenne 0 et variance σ_k^2.

- On propose d'imputer la valeur manquante $y_k, k \in s_m$ par :

$$\hat{y}_k = \hat{f}_j(z_k), \quad k \in s_m^{(j)} = s_m \cap C_j$$

où $\hat{f}_j(z_k)$ est l'estimateur de f_j obtenu par une régression non-paramétrique par des B-splines et considéré seulement sur les répondants ;

Nouveaux challenges dans ce cadre non-paramétrique

- On peut écrire :

$$\hat{t}_I - t_y = \underbrace{\hat{t}_{HT} - t_y}_{\text{erreur d'échantillonnage}} + \underbrace{\hat{t}_I - \hat{t}_{HT}}_{\text{erreur d'imputation}}$$

- L'inférence statistique est par au modèle de superpopulation, m, au plan d'échantillonnage p et le mécanisme de non-réponse q ;
- On a $\mathbb{E}_p(\hat{t}_{HT} - t_y) = 0$, alors le biais total de l'estimateur imputé se réduit à (dans le cas d'une seule classe) :

$$\mathbb{E}_{mpq}(\hat{t}_I - \hat{t}_{HT}) = \mathbb{E}_{pq}\mathbb{E}_m(\hat{t}_I - \hat{t}_{HT}|s, s_r) = \mathbb{E}_{pq}\sum_{s_m}\frac{1}{\pi_k}\mathbb{E}_m(\hat{y}_k - y_k|s, s_r)$$

- Dans le cas des modèles non-paramétriques, les estimateurs de f sous le modèle sont toujours biaisé (Sarda & Vieu, 2000), contrairement au cas paramétrique.

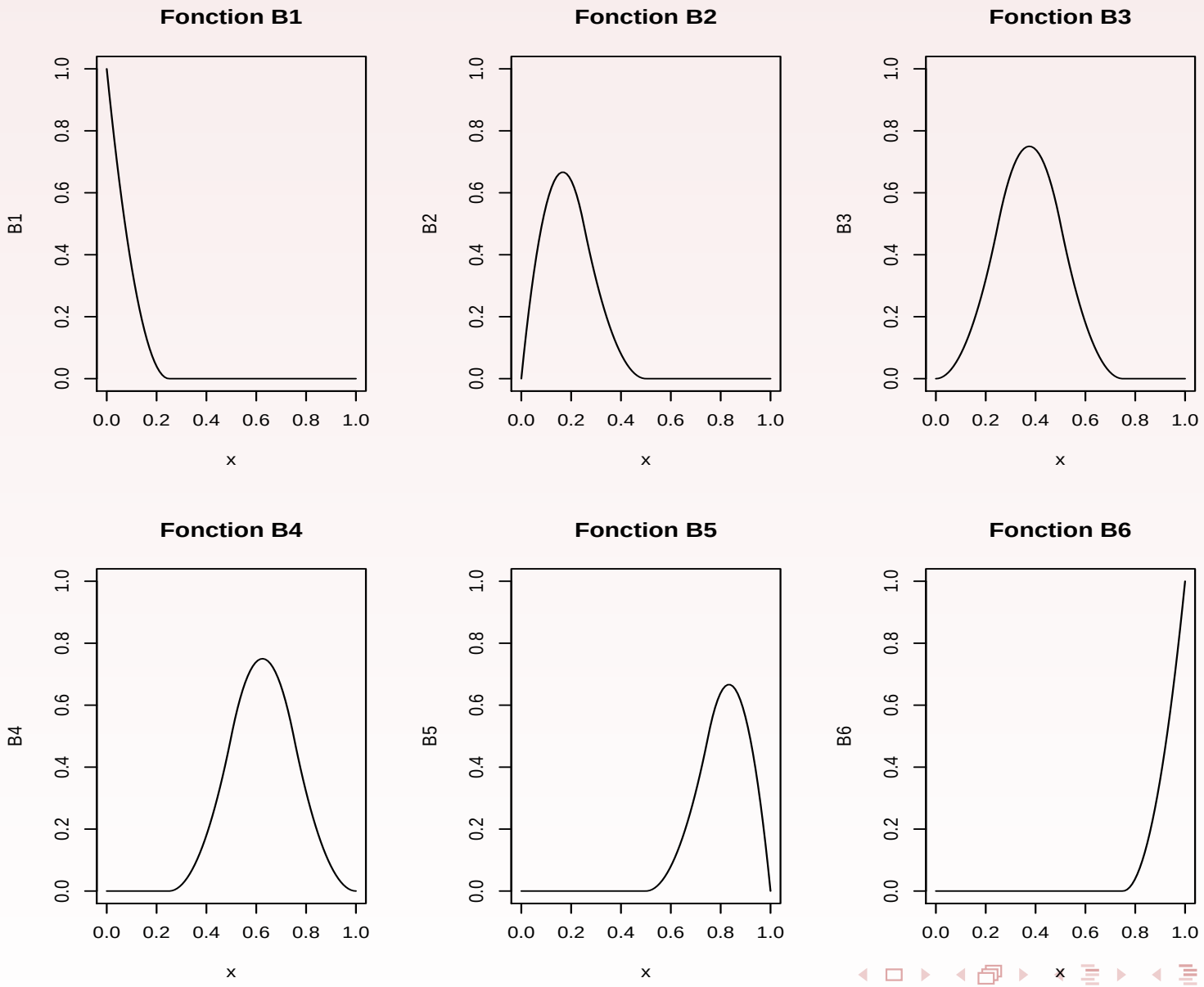
Les fonctions B-splines

- L'ensemble de fonctions spline d'ordre m ($m \geq 2$) avec K noeuds intérieurs : polynômes de degré $m-1$ sur les intervalles entre les noeuds. Les splines sont des fonctions très flexibles pour s'adapter aux données statistiques ;
- Soient $B_1, \ldots, B_q$ la base des fonctions B-spline de dimension $q = K + m$:

$$\sum_{\ell=1}^{q} B_\ell(z) = 1, \quad z \in [0,1].$$

- En pratique, on prend $m = 3$ ou 4.
- Le nombre de noeuds et leur position sont deux questions délicates. Les noeuds sont le plus souvent placés aux quantiles de $\mathcal{Z}$ et pour palier le choix du nombre de noeuds, on peut prendre beaucoup de noeuds et ajouter une pénalité (Ruppert *et al.*, 2003).

Les fonctions B-spline pour $m = 3$ et $K = 3$



Estimation de la fonction f_j

- Sous le modèle, f_j est estimée par (Zhou *et al.*, 1998) :

$$\tilde{f}_j(z) = \sum_{\ell=1}^{q} \tilde{\beta}_\ell^{(j)} B_\ell(z) = \boldsymbol{b'}(z)\tilde{\boldsymbol{\beta}}^{(j)}, \quad \boldsymbol{b'}(z) = (B_\ell(z))_{\ell=1}^{q}$$

où $\tilde{\beta}_\ell^{(j)}$ est déterminé par un critère des moindres carrés :

$$\tilde{\boldsymbol{\beta}}^{(j)} = \text{Arg} \min_{\boldsymbol{\beta}\in\mathbb{R}^q} \sum_{i\in C_j} \left(y_i - \sum_{\ell=1}^{q} \beta_\ell^{(j)} B_\ell(z_i) \right)^2 = \left(\sum_{i\in C_j} \boldsymbol{b}(z_i)\boldsymbol{b'}(z_i) \right)^{-1} \sum_{i\in C_j} \boldsymbol{b}(z_i) y_i$$

- Avec des données d'enquête, $\hat{f}_j(z) = \boldsymbol{b'}(z)\hat{\boldsymbol{\beta}}^{(j)}$ (Goga, 2005 ; Goga & Ruiz-Gazen, 2014) où

$$\hat{\boldsymbol{\beta}}^{(j)} = \left(\sum_{i\in s^{(j)}} \frac{\boldsymbol{b}(z_i)\boldsymbol{b'}(z_i)}{\pi_i} \right)^{-1} \sum_{i\in s^{(j)}} \frac{\boldsymbol{b}(z_i) y_i}{\pi_i}$$

où $s^{(j)} = s \cap C_j, \quad j = 1, \ldots, J.$

Estimation de la fonction f_j

- Sous le modèle, f_j est estimée par (Zhou *et al.*, 1998) :

$$\tilde{f}_j(z) = \sum_{\ell=1}^{q} \tilde{\beta}_\ell^{(j)} B_\ell(z) = \boldsymbol{b'}(z)\tilde{\boldsymbol{\beta}}^{(j)}, \quad \boldsymbol{b'}(z) = (B_\ell(z))_{\ell=1}^{q}$$

où $\tilde{\beta}_\ell^{(j)}$ est déterminé par un critère des moindres carrés :

$$\tilde{\boldsymbol{\beta}}^{(j)} = \text{Arg} \min_{\boldsymbol{\beta}\in\mathbb{R}^q} \sum_{i\in C_j} \left(y_i - \sum_{\ell=1}^{q} \beta_\ell^{(j)} B_\ell(z_i) \right)^2 = \left(\sum_{i\in C_j} \boldsymbol{b}(z_i)\boldsymbol{b'}(z_i) \right)^{-1} \sum_{i\in C_j} \boldsymbol{b}(z_i) y_i$$

- Avec des données d'enquêtes et données manquantes, on a $\hat{f}_j(z) = \boldsymbol{b'}(z)\hat{\boldsymbol{\beta}}^{(j)}$ où :

$$\hat{\boldsymbol{\beta}}^{(j)} = \left(\sum_{i\in s_r^{(j)}} \frac{\boldsymbol{b}(z_i)\boldsymbol{b'}(z_i)}{\pi_i \phi_i^{(j)}} \right)^{-1} \sum_{i\in s_r^{(j)}} \frac{\boldsymbol{b}(z_i) y_i}{\pi_i \phi_i^{(j)}}$$

où $s_r^{(j)} = s_r \cap C_j$ et $\phi_i^{(j)}$ est la probabilité de réponse de l'individu $i \in C_j, \quad j = 1, \ldots, J.$

L'estimateur imputé par des fonctions B-splines

- Nous suggérons imputer les valeurs manquantes par

$$\hat{y}_k^{(j)} = \mathbf{b}'(z_k)\hat{\boldsymbol{\beta}}^{(j)} = \sum_{i \in s_r^{(j)}} \left(\mathbf{b}'(z_k)\hat{\mathbf{T}}_{\phi j}^{-1} \frac{\boldsymbol{b}(z_i)}{\pi_i \phi_i^{(j)}} \right) y_i,$$

où $\hat{\mathbf{T}}_{\phi j} = \sum_{i \in s_r^{(j)}} \boldsymbol{b}(z_i)\boldsymbol{b'}(z_i)/\pi_i\phi_i^{(j)}$; cette méthode d'imputation est "linéaire" dans $\mathcal{Y}$ (Beaumont & Bissonnette, 2011).

- L'estimateur imputé par des B-spline est donné par :

$$\begin{aligned} \hat{t}_I &= \sum_{k \in s_r} \frac{y_k}{\pi_k} + \sum_{j=1}^{J} \sum_{k \in s_m^{(j)}} \frac{\hat{y}_k^{(j)}}{\pi_k} \\ &= \sum_{j=1}^{J} \left(\sum_{k \in s_r^{(j)}} \frac{y_k - \hat{y}_k^{(j)}}{\pi_k} + \sum_{k \in s^{(j)}} \frac{\hat{y}_k^{(j)}}{\pi_k} \right) \end{aligned}$$

Résultat

Si les probabilités de réponse sont constantes à l'intérieur des classes alors :

$$\sum_{k \in s_r^{(j)}} \frac{y_k - \hat{y}_k^{(j)}}{\pi_k} = 0 \quad \textit{pour tout} \quad j.$$

- Alors, l'estimateur imputé peut s'écrire sour une forme de *projection* :

$$\hat{t}_I = \sum_{j=1}^{J} \sum_{k \in s^{(j)}} \frac{\hat{y}_k^{(j)}}{\pi_k},$$

- et égal à une somme pondérée de valeurs de Y pour les répondants :

$$\hat{t}_I = \sum_{j=1}^{J} \sum_{k \in s_r^{(j)}} \frac{1}{\pi_k} g_k^{(j)} y_k,$$

où $g_k^{(j)} = \left(\sum_{i \in s^{(j)}} \pi_i^{-1} \mathbf{b}_i'\right) \hat{\mathbf{T}}_j^{-1} \mathbf{b}_k', \quad k \in s_r^{(j)}.$

Hypothèses

Sur le plan d'échantillonnage (Robinson and Särndal, 1983, Breidt and Opsomer, 2000) et **le mécanisme de non-réponse** :

- $\lim_{N\to\infty}(n/N) \in (0,1)$; $N_j = O(N), j = 1, \ldots, J$;
- $\min_{k\in U} \pi_k \geq \tilde{\lambda} > 0$, $\min_{k\leq l\in U} \pi_{kl} \geq \lambda^* > 0$, $\overline{\lim}_{N\to\infty} n \max_{k\neq l\in U} |\Delta_{kl}| < \infty$
- $\phi_i^{(j)} = \phi^{(j)} \geq c > 0$ for all $i \in C_j, j = 1, \ldots, J$.

Sur les modèles d'imputation : f_j est supposée a fois diférentiable et $\overline{\lim}_{N\to\infty} \sum_{k\in U} \sigma_k^2/N < \infty$;

Sur les fonctions B-spline

- Il existe une distribution $Q(z)$ avec une densité strictement positive sur $[0,1]$ telle que $\sup_{z\in[0,1]} |Q_N(z) - Q(z)| = o(1/K)$, avec $Q_N(z)$ la distribution empirique de $(z_i)_{i\in U}$ (Zhou *et al.*, 1998).
- $K = o(N)$ et $K = O(n^a)$ avec $a < 1/3$;

Propriétés asymptotiques de l'estimateur imputé

On suppose que le mécanisme de non-réponse est ignorable.

Résultat

On montre que :

$$\frac{1}{N}\mathbb{E}|\hat{t}_I - t_y| = O\left(\frac{1}{\sqrt{n}}\right),$$

donc l'estimateur imputé par des B-spline est asymptotiquement sans biais et convergent ; $\mathbb{E}(\cdot)$ est considérée par rapport au modèle d'imputation, le plan de sondage et le mécanisme de non-réponse.

Résultat

Le biais total est :

$$\frac{1}{N}\mathbb{E}(\hat{t}_I - t_y) = o\left(\frac{1}{\sqrt{n}}\right).$$

Le calcul de la variance

On a :

$$\hat{t}_I - t_y = \underbrace{\hat{t}_{HT} - t_y}_{\text{erreur d'échantillonnage}} + \underbrace{\hat{t}_I - \hat{t}_{HT}}_{\text{erreur d'imputation}}$$

et la décomposition de la variance (Sarndal, 1992) :

$$\begin{aligned}\mathbb{V}_{tot}(\hat{t}_I - t_y) &\simeq \mathbb{E}_{mpq}\left(\hat{t}_I - t_y\right)^2 \\ &= \mathbb{V}_{SAM} + \mathbb{V}_{IMP} + \mathbb{V}_{MIX}\end{aligned}$$

où $\mathbb{V}_{SAM}$ est la variance due à l'échantillonnage, $\mathbb{V}_{IMP}$ la variance due à l'imputation et $\mathbb{V}_{MIX}$ le terme de covariance mixte.

Etude par simulation

- On considère une population de taille $N = 5000$;
- On génère une variable auxiliaire $\mathcal{Z}$ dans une distribution uniforme sur $[0, 1]$ et la variable d'intérêt $\mathcal{Y}$ selon quatre modèles (Breidt and Opsomer (2000)) :

  (Linear) : $y_i = 1 + 2(z_i - 0.5) + \epsilon_i$;
  (Quadratic) : $y_i = 1 + 2(z_i - 0.5)^2 + \epsilon_i$;
  (Bump) : $y_i = 1 + 2(z_i - 0.5) + \exp(-200\{(z_i - 0.5)\}^2) + \epsilon_i$;
  (Exponential) : $y_i = \exp(-8z_i) + \epsilon_i$,

  et les résidus ϵ_i sont générés selon une loi normale.

- Nous sommes intéressés à estimer le total

$$t_y = \sum_{i \in U} y_i$$

pour les quatre modèles ;

- Nous considérons $R = 10,000$ échantillons aléatoire simple sans remise de taille $n = 250$
- Les indicateurs de réponse r_i sont générés selon Bernoulli(p_i), où p_i est généré selon

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 z_i,$$

et les paramètres γ_0 et γ_1 ont été fixés de façon à que le taux de réponse général est d'approximativement 70%.

On considère les procédures d'imputation suivantes :

(i) Imputation par la régression linéaire (LRI), pour laquelle les valeur imputées sont données par :

$$y_i^* = \mathbf{z}_i' \left(\sum_{i \in s} \pi_i^{-1} r_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in s} \pi_i^{-1} r_i \mathbf{z}_i y_i$$

avec $\mathbf{z}_i = (1, z_i)'$.

(ii) Imputation par le plus-proche voisin basée sur la variable z, pour laquelle les valeurs imputées sont $y_i^* = z_j, j \in s_r$ telles que $(z_j - z_i)^2$ est minimisé.

(iii) L'imputation par B-spline (SI). Le nombre de noeuds est de $K = 2; 5$ et 10, $m = 3$.

Mesures de comparaison

- Biais relatif

$$RB_{MC}(\widehat{\theta}) = \frac{100}{R}\sum_{r=1}^{R}\frac{(\widehat{t}_{I(r)} - t_y)}{t_y},$$

où $\widehat{t}_{I(r)}$ est l'estimateur imputé $\widehat{t}_I$ dans le r-ème échantillon.

- L'efficacité relative (RE) de $\widehat{t}_I$ par rapport à l'estimateur basé sur les données complètes $\widehat{t}_{comp}$:

$$RE(\cdot) = 100 \times \frac{MSE_{MC}\left(\widehat{t}_I(\cdot)\right)}{MSE_{MC}\left(\widehat{t}_{comp}\right)}.$$

où $\widehat{t}_I(\cdot)$ est l'estimateur imputé $\widehat{t}_I$ et l'erreur quadratique moyenne Monte Carlo est donnée par :

$$MSE_{MC}(\widehat{t}_I) = \frac{1}{R}\sum_{r=1}^{R}(\widehat{t}_{I(r)} - t_y)^2.$$

TABLE : Biais relatif et efficacité relative de type Monte Carlo pour les totaux et $\mathbf{n = 250}$

	LRI	NNI	SI $K = 2$	SI $K = 5$	SI $K = 10$
(Linear)	0.0 (129)	0.1 (161)	0.1 (140)	0.1 (141)	0.1 (143)
(Quadratic)	-3.2 (305)	-0.1 (230)	-0.1 (182)	-0.1 (185)	-0.1 (189)
(Bump)	3.4 (195)	0.1 (150)	0.1 (137)	0.1 (135)	0.1 (135)
(Exponential)	-26.1 (262)	-1.7 (217)	-1.8 (177)	-1.7 (179)	-1.7 (182)

Estimation de la fonction de répartition et des quantiles

- La fonction de répartition $F_N(t) = \sum_{i \in U} \mathbf{1}_{\{y_i \leq t\}}/N$ est estimée par l'estimateur imputé :

$$\widehat{F}_I(t) = \left(\sum_{i \in s} \pi_i^{-1} \right)^{-1} \left(\sum_{i \in s_r} \frac{\mathbf{1}_{\{y_i \leq t\}}}{\pi_i} + \sum_{i \in s_m} \frac{\mathbf{1}_{\{y_i^* \leq t\}}}{\pi_i} \right).$$

- La médiane est estimée par :

$$\hat{\theta}_{1/2} = \inf_t \{ \widehat{F}_I(t) \geq \frac{1}{2} \}$$

L'imputation aléatoire

L'imputation aléatoire peut être vue comme une imputation déterministe plus un résidu aléatoire :

$$y_i^* = \widehat{f}(z_i) + \epsilon_i^*,$$

où ϵ_i^* est sélectionné aléatoirement parmi les résidus standardisés des répondants $\{\tilde{e}_j; j \in s_r\}$, avec la probabilité

$$P(\epsilon_i^* = \tilde{e}_j) = \frac{\pi_j^{-1}}{\sum_{l \in s_r} \pi_l^{-1}},$$

où $\tilde{e}_j = e_j - \bar{e}_r$ et $e_j = y_j - \widehat{f}(z_j)$ avec $\bar{e}_r = \frac{\sum_{j \in s_r} \pi_j^{-1} e_j}{\sum_{j \in s_r} \pi_j^{-1}}$.

TABLE : Biais relatif et efficacité relative de type Monte Carlo pour les quantiles et $\mathbf{n = 500}$

Quantile	Modèle	LRI	SI $K = 5$	NNI	RLRI	RSI $K = 5$
$\widehat{\theta}_{1/2}$	Linear	-5.6	-5.6	-0.2	0.3	0.4
		(318)	(336)	(160)	(132)	(**129**)
	Quadratic	-7.5	-0.4	0.0	-3.3	0.0
		(1181)	(252)	(246)	(377)	(**186**)
	Bump	-4.6	0.1	-0.1	2.0	0.7
		(247)	(138)	(135)	(149)	(**120**)
	Exponential	29.3	-0.7	-1.7	-23.3	-3.0
		(283)	(185)	(197)	(250)	(**157**)

Conclusion et travail en cours

- Une nouvelle méthode d'imputation basée sur des fonctions B-spline new non-parametric imputation method based on B-spline functions ;
- Cette méthode est simple à mettre en oeuvre (très similaire à l'imputation par la regression), très bonnes propriétés asymptotiques et performances pratiques ;

Travail en cours :

- résultats asymptotic sur l'imputation aléatoire et équilibrée par des B-spline ;
- plus d'études de simulation sur l'estimation de la variance ;
- extension au cas multi-varié en considérant des modèles additifs.

Courte bibliographie

- Beaumont, J-F. and Bissonnette, J. (2011). Variance estimation under composite imputation : The methodology behind SEVANI. *Survey Methodology*, **37**, 171-179.
- Beaumont, J-F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *The Canadian Journal of Statistics*, **37**, 400-416.
- Cardot, H., De Moliner, A. and Goga, C. (2018). Estimation of total electricity consumption curves by sampling in a finite population when some trajectories are partially unobserved (in revision for Canadian Journal of Statistics, special edition for CANSSI project),
- Chen, J. and Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. *Journal of Official Statistics*, **16**, 113-131.
- Goga, C. (2005), Réduction de la variance dans les sondages en présence d'information auxiliaire : une approche non paramétrique par splines de régression, The Canadian Journal of Statistics, **33**, 1-18.
- Goga, C. and Ruiz-Gazen (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. (*JRSSB*, 76, 113-140).
- Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of statist., 29A, Sample surveys : design methods and applications*. Elsevier/North Holland, Amsterdam 215-246.
- Ruppert, D., Wand, M.P. and Caroll, R. (2003). *Semiparametric regression*, Cambridge : Cambridge University Press.
- Sarda, P. & Vieu, P. (2000). Kernel regression, in *Smoothing and regression : approaches, computation and application (ed. M.G. Schimek)*, New-York Wiley.
- Sarndal, C.-R. (1992). Methods for estimating the precision of survey estimates when imputation is has been used, *Survey Methodology*, 18, 2, 241-252.