ESTIMATION DE PARAMETRES NON LINEAIRES PAR DES METHODES NON-PARAMETRIQUES EN POPULATION FINIE

Camelia Goga (*), Anne Ruiz-Gazen (**)
(*) IMB, Université de Bourgogne
(**)TSE, Université de Toulouse

Introduction

Dans les enquêtes par sondages certains paramètres d'intérêt tels que les ratios ou les indices économiques de type Gini sont des fonctions non-linéaires de totaux. L'estimation de la variance des estimateurs de tels paramètres nécessite une méthodologie adaptée et on trouve dans la littérature deux grandes familles de méthodes que sont la linéarisation et le rééchantillonnage (Wolter, 2007). Par ailleurs, en présence d'information auxiliaire, il est bien connu qu'il existe des méthodes d'estimation susceptibles d'améliorer la précision. Ainsi, dans un cadre paramétrique, l'estimateur « GREG » basé sur la régression linéaire a été largement étudié (voir Särndal et al., 1992). Dans le cas de l'estimation de paramètres fonctions linéaires de totaux, plusieurs approches non-paramétriques ont aussi été envisagées. Dans un cadre « model-assisted », Breidt et Opsomer (2000, respectivement 2005) ont proposé d'utiliser les polynômes locaux (respectivement des P-splines) tandis que Goga (2005) a proposé une approche B-splines. Des approches « calage » ont aussi été envisagées notamment par Montanari et Ranalli (2005). Nous proposons dans ce papier une nouvelle classe d'estimateurs de paramètres non-linéaires prenant en compte l'information auxiliaire à l'aide d'un modèle non-paramétrique. Nous nous intéressons à l'estimation de paramètres fonctions de totaux et nous donnons des résultats asymptotiques basés sur l'approche par la fonction d'influence (Deville, 1999). La première section rappelle, dans un cadre général, les propriétés des méthodes d'estimation d'un total par régression non-paramétrique obtenues dans Breidt et Opsomer (2000, 2005) et Goga (2005). La deuxième section donne les propriétés asymptotiques des estimateurs que l'on propose pour un paramètre quelconque fonction de totaux, par des méthodes de régression non-paramétriques de type polynômes locaux, P ou B-splines. La section 3 donne une application des résultats précédents au cas où l'on estime un ratio par régression B-splines. L'approche B-splines est rappelée dans cette troisième section ; on insiste en particulier sur les nombreux points communs qu'elle entretient avec l'approche par régression linéaire. Toujours dans le cadre de l'estimation d'un ratio par B-splines, la section 4 présente une étude empirique qui permet de valider l'approche proposée. Enfin la dernière section donne des perspectives.

1. Estimation d'un total par régression non-paramétrique

On considère une population finie $U = \{1, \ldots, k, \ldots, N\}$ et on suppose connues les valeurs d'une variable auxiliaire unidimensionnelle $\mathcal{Z}$ pour toutes les unités k dans U. On note ces valeurs z_k pour $k \in U$. Un échantillon s de taille fixe n est sélectionné dans U selon un plan de sondage non-informatif quelconque $p(s)$ et la valeur de la variable d'intérêt $\mathcal{Y}$ est observée pour chaque unité dans l'échantillon; on obtient y_k pour $k \in s$. Pour chaque individu dans la population, les probabilités d'inclusion du premier ordre (respectivement du second ordre) dans s, $\pi_k = Pr(k \in s)$ pour tout $k \in U$ (respectivement $\pi_{kl} = Pr(k, l \in s) > 0$ pour tous $k, l \in U$) sont supposées strictement positives. Nous voulons estimer le total t_y d'une variable $\mathcal{Y}$ sur U :

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

En l'absence d'information auxiliaire, ce total est estimé par l'estimateur de Horvitz-Thompson :

$$\hat{t}_{y,HT} = \sum_s \frac{y_k}{\pi_k}.$$

En présence d'information auxiliaire, on peut introduire un modèle de superpopulation :

$$\xi : \quad y_k = f(z_k) + \varepsilon_k \tag{1.1}$$

et considérer la classe d'estimateurs assistés par un modèle introduite par Cassel et al. (1976) :

$$\hat{t}_y = \sum_{k \in s} \frac{y_k - f_k}{\pi_k} + \sum_{k \in U} f_k \tag{1.2}$$

avec $f_k = f(z_k)$. Dans le cas d'un modèle linéaire, on obtient la classe d'estimateurs GREG présentée par Särndal et al. (1992). Néanmoins, si la vraie relation n'est pas linéaire, l'efficacité en terme de variance, de l'estimateur par la régression généralisée $\hat{t}_{GREG}$ sous un modèle linéaire peut se révéler mauvaise, même comparée à l'estimateur de Horvitz-Thompson qui pourtant ne prend pas en compte l'information auxiliaire.

L'utilisation de modèles non-paramétriques permet de couvrir une classe beaucoup plus large de relations entre information auxiliaire et variable d'intérêt, en imposant uniquement des conditions de régularité (dérivabilité) sur la fonction de régression. Il faut toutefois remarquer que, contrairement aux estimateurs dérivés du modèle linéaire, les estimateurs basés sur des modèles non-paramétriques nécessitent que l'information auxiliaire soit connue pour tout les individus de la population (ce qui permet d'utiliser aussi la variable auxiliaire pour établir le plan de sondage). Récemment, Breidt & Opsomer (2000) ont proposé des estimateurs assistés par un modèle non-paramétrique en utilisant une approche par polynômes locaux pour des plans de sondage à une ou deux phases. Ultérieurement, Breidt and Opsomer (2005) ont utilisé une décomposition de la fonction de régression dans une base de fonctions splines des polynômes tronqués (P-splines) et Goga (2005) a utilisé une décomposition de f_k dans une base de B-splines. Dans la suite, nous proposons d'étudier les propriétés asymptotiques d'une classe générale d'estimateurs de paramètres complexes qui contient l'approche par polynômes locaux mais aussi l'approche par P et par B-splines.

Soit f_k estimé par $\hat{f}_{y,k}$ qui dépend de la variable $\mathcal{Y}$ lorsqu'on utilise une méthode non-paramétrique telle que les polynômes locaux, les P ou les B-splines. Si on remplace dans (1.2), les f_k par $\hat{f}_{y,k}$, on obtient un estimateur pour le total t_y qui est toujours p-sans biais (sans biais par rapport au plan) mais approximativement ξ-sans biais (approximativement sans biais par rapport au modèle) :

$$\hat{t}_{y,\mathrm{diff}} = \sum_{k \in s} \frac{y_k - \hat{f}_{y,k}}{\pi_k} + \sum_{k \in U} \hat{f}_{y,k}. \tag{1.3}$$

Les $\hat{f}_{y,k}$ étant inconnus pour $k \in U$, ils sont estimés par $\tilde{f}_{y,k}$ obtenus à l'aide du plan d'échantillonnage p qui fournit s et on obtient :

$$\hat{t}_{y,np} = \sum_{k \in s} \frac{y_k - \tilde{f}_{y,k}}{\pi_k} + \sum_{k \in U} \tilde{f}_{y,k}. \tag{1.4}$$

Les estimateurs assistés par un modèle non-paramétrique, qu'ils soient obtenus par polynômes locaux, P ou B-splines, peuvent s'écrire comme une somme pondérée des valeurs de $\mathcal{Y}$ avec des **poids indépendants de la variable d'intérêt** et contenant l'information auxiliaire :

$$\hat{t}_{y,np} = \sum_s w_{ks} y_k \tag{1.5}$$

où l'expression de w_{ks} dépend de la méthode non-paramétrique utilisée.

Résultat 1 : Sous certaines hypothèses (voir Breidt & Opsomer, 2000, 2005 et Goga, 2005), l'estimateur $\hat{t}_{y,np}$ satisfait :

$$\frac{1}{N}(\hat{t}_{y,np} - t_y) = O_p(n^{-1/2})$$

et

$$n^{1/2}N^{-1}(\hat{t}_{y,np} - t_y) = n^{1/2}N^{-1}(\hat{t}_{y,\text{diff}} - t_y) + o_p(1) \tag{1.6}$$

Par conséquent, $\hat{t}_{y,np}$ est asymptotiquement sans biais et convergent pour t_y.

De plus, si la distribution asymptotique de $n^{1/2}N^{-1}(\hat{t}_{y,\text{diff}} - t_y)$ est normale, alors la variance asymptotique de $n^{1/2}N^{-1}(\hat{t}_{y,np} - t_y)$ est donnée par la variance de $n^{1/2}N^{-1}(\hat{t}_{y,\text{diff}} - t_y)$ qui a pour expression :

$$\frac{n}{N^2}\sum_U\sum_U \Delta_{kl}\frac{y_k - \hat{f}_{y,k}}{\pi_k}\frac{y_k - \hat{f}_{y,l}}{\pi_l}.$$

avec $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$.

La variance asymptotique est d'autant plus petite que les résidus $y_k - \hat{f}_k$ sont petits en valeur absolue. Ce résultat justifie l'usage des techniques non-paramétriques qui permettent de fournir de bons estimateurs d'une large classe de fonctions f qui ne sont pas toujours paramétrables simplement.

Remarquons que l'estimateur $\hat{t}_{y,np}$ dans le cas d'une régression par des P ou B-splines possède la plupart des propriétés des estimateurs GREG sous un modèle linéaire ; notamment le fait que l'estimateur de Horvitz-Thompson pour les résidus $y_k - \tilde{f}_{y,k}$ est nul. Par conséquent, pour des P ou B-splines :

$$\hat{t}_{y,np} = \sum_{k\in U}\tilde{f}_{y,k}.$$

2. Estimation d'un paramètre non-linéaire par régression non-paramétrique

Nous souhaitons prendre en compte l'information auxiliaire $\mathcal{Z}$ pour l'estimation d'un paramètre Φ fonction **non-linéaire** de totaux. Pour simplifier, nous supposons que Φ est une fonction de seulement deux totaux :

$$\Phi = \Phi(t_x, t_y).$$

Les poids w_{ks} donnés dans la relation (1.5) ont été obtenus en utilisant $\mathcal{Z}$ pour estimer le total t_y. Ces poids ne dépendent pas de la variable d'intérêt et par conséquent, ils peuvent être utilisés pour estimer d'autres totaux.

Dans la suite, nous estimons les totaux t_x et t_y par des estimateurs pondérés avec les poids w_{ks},

$$\hat{t}_{x,np} = \sum_{k \in s} w_{ks} x_k, \quad \hat{t}_{y,np} = \sum_{k \in s} w_{ks} y_k.$$

L'estimateur par substitution de Φ est donné par :

$$\widehat{\Phi}_{np} = \Phi(\hat{t}_{x,np}, \hat{t}_{y,np}). \tag{1.7}$$

Pour calculer la variance asymptotique de $\widehat{\Phi}_{np}$, on utilise l'approche par linéarisation basée sur la fonction d'influence (Deville, 1999). On considère la mesure discrète :

$$M = \sum_U \delta_{(x_k, y_k)}$$

et on suppose que Φ peut s'écrire comme une fonctionnelle T de M :

$$\Phi = T(M).$$

La mesure M est estimée par :

$$\hat{M}_{np} = \sum_s w_{ks} \delta_{(x_k, y_k)}$$

avec des poids w_{ks} donné par (1.5). L'estimateur par substitution non-paramétrique $\widehat{\Phi}_{np}$, donné par la relation (1.7) est obtenu en remplaçant M par $\hat{M}_{np}$:

$$\widehat{\Phi}_{np} = T(\hat{M}_{np}).$$

Pour obtenir la variance asymptotique de $\widehat{\Phi}_{np}$, nous faisons un développement au premier ordre de la fonctionnelle T. La différentielle de T s'appelle fonction d'influence et elle est définie ci-dessous.

Définition 1 : La fonction d'influence $IT(M, x)$ de $T(M)$ est définie comme la dérivée au sens de Gâteaux de T par rapport à M dans la direction de la masse de Dirac en x :

$$IT(M, x) = \lim_{\varepsilon \longrightarrow 0} \frac{T(M + \varepsilon \delta_x) - T(M)}{\varepsilon}$$

lorsque cette limite existe.

Définition 2 : La variable linéarisée u_k pour $k \in U$ est la valeur de IT dans (x_k, y_k) :

$$u_k = IT(M, (x_k, y_k)), \quad k \in U.$$

Soient les estimateurs $\hat{t}_{u,\text{diff}}$ et $\hat{t}_{u,np}$ obtenus en remplaçant dans les expressions (1.3) et (1.4) les valeurs y_k de la variable $\mathcal{Y}$, par celles des variables linéarisées, u_k. Le résultat suivant donne une linéarisation de la statistique complexe $\widehat{\Phi}_{np}$ par un estimateur par la différence généralisée du total des variables linéarisées, $t_u = \sum_U u_k$.

Résultat 2 : Supposons que la fonctionnelle T est dérivable au sens de Fréchet et de degré α, c'est à dire $T(rM) = r^\alpha T(M)$ et de plus $\lim_{N \to \infty} T(M/N) < \infty$. Supposons que les variables $N^{1-\alpha} u_k$ satisfont la relation (1.6), c'est-à-dire

$$N^{-\alpha}(\hat{t}_{u,np} - \hat{t}_{u,\text{diff}}) = o_p(n^{-1/2}).$$

Alors,

$$N^{-\alpha}\left(\widehat{\Phi}_{np} - \Phi\right) = N^{-\alpha}(\hat{t}_{u,np} - t_u) + o_p(n^{-1/2}) = N^{-\alpha}(\hat{t}_{u,\text{diff}} - t_u) + o_p(n^{-1/2}).$$

Supposons que la distribution de $N^{-\alpha}(\hat{t}_{u,\text{diff}} - t_u)$ est normale, alors la variance asymptotique de $\sqrt{n}N^{-\alpha}\left(\widehat{\Phi}_{np} - \Phi\right)$ est donnée par

$$\frac{\sqrt{n}}{N^{2\alpha}} \sum_{k \in U} \sum_{i \in U} \Delta_{kl} \frac{u_k - \hat{f}_{u,k}}{\pi_k} \frac{u_l - \hat{f}_{u,l}}{\pi_l}$$

avec $\hat{f}_{u,k}$ donné par la relation (1.3) pour la variable u.

Preuve : On considère un développement de Von-Mises de premier ordre de la fonctionnelle $T(\widehat{M}_{np}/N) = N^{-\alpha}T(\widehat{M}_{np})$ au point $T(M/N) = N^{-\alpha}T(M)$:

$$N^{-\alpha}\left(T(\widehat{M}_{np}) - T(M)\right) = \int IT\left(\frac{M}{N}, y\right) d\left(\frac{\widehat{M}_{np}}{N} - \frac{M}{N}\right) + o(d(N^{-1}\widehat{M}_{np}, N^{-1}M))$$

$$= N^{-\alpha}\left(\sum_s w_{ks}u_k - \sum_U u_k\right) + o_p(n^{-1/2})$$

car $IT\left(\frac{M}{N}, y\right) = N^{-\alpha+1}IT(M, y)$ par linéarité de la différentielle de Fréchet IT. Nous avons que $d(N^{-1}\widehat{M}_{np}, N^{-1}M)$ est égale à la distance entre $\hat{t}_{np,y}/N$ et t_y/N pour toute variable $\mathcal{Y}$ qui est d'ordre $O_p(n^{-1/2})$ (voir le résultat 1).

Les conditions dans lesquelles l'approximation $N^{-\alpha}(\hat{t}_{u,np} - \hat{t}_{u,\text{diff}}) = o_p(n^{-1/2})2$ est valable, dépendent du type d'estimateur non-paramétrique utilisé (par polynômes locaux, par P ou B-splines).

La variance asymptotique donnée par le résultat 2 sera d'autant plus petite que les résidus $u_k - \hat{f}_{u,k}$ seront petits c'est-à-dire que le modèle expliquera bien la variable linéarisée. Ce résultat déjà mis en avant dans Deville (1999) montre que dans le cas d'un paramètre complexe, la variable à modéliser est la variable linéarisée et sera illustré dans la section 5.

3. Estimation d'un ratio par régression non-paramétrique basée sur des B-splines

Considérons maintenant l'estimation d'un ratio $R = t_y/t_x$ quand une variable auxiliaire $\mathcal{Z}$ est disponible pour chaque individu dans la population. Cette situation a déjà été traitée par Särndal et al. (1992) en utilisant un modèle de régression multiple pour améliorer l'estimation des totaux t_x et t_y. Une approche par calage a aussi été proposée pour un paramètre quelconque, et donc notamment un ratio, par Deville (1999). Nous proposons ici une alternative en introduisant un modèle non paramétrique et une estimation par B-splines. Rappelons d'abord la construction de l'estimateur non-paramétrique par B-splines dans le cadre des sondages introduit par Goga (2005).

L'idée qui est à la base de l'estimation non–paramétrique par B-splines de f est en quelque sorte une généralisation de la régression linéaire. Nous avons le nuage des points (z_k, y_k) pour $k \in U$ et nous voulons trouver la « vraie » relation f qui gouverne ce nuage des points. Faire passer une courbe exactement par les points (z_k, y_k) peut s'avérer trop coûteux ou simplement indésirable pour diverses raisons. Dans ce cas, on remplace un problème d'interpolation par un problème d'approximation : on se donne un espace de dimension beaucoup plus petite que le nombre de points z_k et l'on cherche une fonction de cet espace qui approche aussi bien que possible les valeurs y_k. La régression linéaire (simple ou multiple) est obtenue pour l'approximation par une fonction affine de $\mathcal{Z}$ et en utilisant le critère des moindres carrés. Une situation plus générale consiste à prendre l'approximation polynomiale par morceaux qui permet de mieux contrôler les problèmes liés au manque de régularité totale de la fonction f et la stabilité dans les calculs. Une telle fonction est appelée spline et le critère pour trouver l'approximation de f dans cet espace est le critère des moindres carrés. La méthode d'estimation par B-splines a ainsi beaucoup des propriétés en commun avec la régression linéaire qui est juste un cas particulier.

Définissons l'espace des fonctions spline $S_{K,m}$ d'ordre m ($m \geq 2$) avec K nœuds intérieurs équidistants $0 = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = 1$ par :

$$S_{K,m} = \{u \in C^{m-2}[0,1] : u(x) \text{ est un polynôme de degré } m-1 \text{ sur } (\xi_j, \xi_{j+1})\}. \quad (1.8)$$

Pour $m = 1$, $S_{K,1}$ est l'ensemble des fonctions en escalier sur les sous intervalles de $[0,1]$ définis par les nœuds et pour $m = 2$, $S_{K,2}$ est l'ensemble de fonctions continues sur $[0,1]$ et linéaires par morceaux. L'espace $S_{K,m}$ est un espace linéaire de dimension $q = K + m$ dont une base est constituée des fonctions B–splines $(B_j(\cdot))_{j=1}^{q}$ (Schumaker, 1981 and Dieckx, 1993). Chaque fonction B_j pour $j = 1, \ldots, q$ a comme support un nombre petit et fixé d'intervalles entre les nœuds. De plus, ces fonctions sont positives de somme un :

$$\sum_{j=1}^{q} B_j(x) = 1 \ , \qquad x \in [0,1]. \qquad (1.9)$$

Un estimateur non-paramétrique de la fonction de régression f du modèle ξ est la meilleure approximation du vecteur $(y_1, \ldots, y_N)'$ au sens des moindres carrés dans l'espace engendré par les fonctions $B_j(z_k)$ pour $j = 1, \ldots, q$ et $k \in U$. Cela revient à prendre comme estimation de f la projection du vecteur $(y_1, \ldots, y_N)'$ dans cet espace :

$$\hat{f}_{y,k} = \textstyle\sum_{j=1}^{q} \hat{\theta}_{jy} B_j(x)$$

où

$$\hat{\boldsymbol{\theta}}_y = (\hat{\theta}_{1y}, \ldots, \hat{\theta}_{qy})' = \operatorname{Arg\,min}_{\theta_j} \sum_{k=1}^{N} \left(y_k - \sum_{j=1}^{q} \theta_j B_j(z_k) \right)^2 . \qquad (1.10)$$

Pour $\boldsymbol{b}'(z_k) = (B_1(z_k), \ldots, B_q(z_k))$, l'estimateur par B–splines de f_k s'écrit pour chaque $z_k, k \in U$:

$$\begin{cases} \hat{f}_{y,k} = \boldsymbol{b'}(z_k)\hat{\boldsymbol{\theta}}_y \ , \qquad k \in U \\ \hat{\boldsymbol{\theta}}_y = \left(\sum_U \boldsymbol{b}(z_k)\boldsymbol{b'}(z_k)\right)^{-1} \left(\sum_U \boldsymbol{b}(z_k) y_k\right) \end{cases} \qquad (1.11)$$

en supposant que la matrice $\sum_U \boldsymbol{b}(x_k)\boldsymbol{b}'(z_k)$ est inversible.

Les estimateurs de type *design–based* de $\hat{f}_{y,k}$ sont donnés par :

$$\begin{cases} \tilde{f}_{y,k} = \boldsymbol{b}'(z_k)\tilde{\boldsymbol{\theta}}_y, \qquad k \in U \\ \tilde{\boldsymbol{\theta}}_y = \left(\sum_s \dfrac{\boldsymbol{b}(x_k)\boldsymbol{b}'(z_k)}{\pi_k}\right)^{-1} \left(\sum_s \dfrac{\boldsymbol{b}(z_k)y_k}{\pi_k}\right). \end{cases} \quad (1.12)$$

Remarquons que $\tilde{\boldsymbol{\theta}}_y$ ne dépend pas du point k où on calcule $\tilde{f}_{y,k}$ et connaître $\tilde{\boldsymbol{\theta}}_y$ nous permet de calculer rapidement la valeur de $\tilde{f}_{y,k}(z)$ en n'importe quel point $z \in [0,1]$. Finalement, l'estimateur B-splines dans l'échantillon s du total t_y est obtenu en remplaçant dans (1.4) les $\tilde{f}_{y,k}$ données par (1.12) :

$$\hat{t}_{BS} = \sum_{k\in s} \frac{y_k - \tilde{f}_{y,k}}{\pi_k} + \sum_{k\in U} \tilde{f}_{y,k}. \quad (1.13)$$

La relation (1.12) nous permet d'écrire $\hat{t}_{BS}$ comme un estimateur par la régression généralisée (GREG) pour un vecteur d'information auxiliaire $\boldsymbol{b}(z_k)$ dont la dimension $q = K + m$ dépend du nombre de nœuds :

$$\hat{t}_{BS} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\boldsymbol{b}'(z_k)}{\pi_k} - \sum_U \boldsymbol{b}'(z_k)\right)\tilde{\boldsymbol{\theta}}_y = \sum_s w_{ks}y_k$$

avec les poids

$$w_{ks} = \frac{1}{\pi_k}\left(\sum_U \boldsymbol{b}'(z_i)\right)\left(\sum_{i\in s} \boldsymbol{b}(z_i)\boldsymbol{b}'(z_i)/\pi_i\right)^{-1} \boldsymbol{b}(z_k). \quad (1.14)$$

Ces poids ne dépendent pas de la variable d'intérêt $\mathcal{Y}$ et satisfont les équations du calage pour les variables auxiliaires $\boldsymbol{b}(z_k)$ (Goga, 2005). Par conséquent, un estimateur du total de $\mathcal{Y}$ calé sur $\boldsymbol{b}(z_k)$ peut être obtenu. Goga (2005) démontre que l'estimateur $\hat{t}_{BS}$ partage d'autres propriétés avec les estimateurs GREG comme le fait que le total des résidus $E_k = y_k - \hat{f}_{y,k}$ ainsi que l'estimateur de Horvitz-Thompson des résidus $y_k - \tilde{f}_{y,k}$ soient nuls. Ces propriétés découlent de la relation (1.9). Nous avons également que $\sum_U w_{ks} = N$.

Construisons un estimateur du ratio

$$R = \sum_U y_k / \sum_U x_k$$

par régression non–paramétrique basée sur des B-splines. Les totaux t_x (respectivement t_y) sont estimés par $\sum_s w_{ks}x_k$ (respectivement $\sum_s w_{ks}y_k$) où les poids w_{ks} sont donnés par (1.14).

On obtient l'estimateur par des B-splines du ratio R comme suit :

$$\hat{R}_{BS} = \frac{\sum_s w_{ks} y_k}{\sum_s w_{ks} x_k} = \frac{\sum_{k\in s}(y_k - \boldsymbol{b}'(z_k)\tilde{\boldsymbol{\theta}}_y)/\pi_k + \sum_U \boldsymbol{b}'(z_k)\tilde{\boldsymbol{\theta}}_y}{\sum_{k\in s}(x_k - \boldsymbol{b}'(z_k)\tilde{\boldsymbol{\theta}}_x)/\pi_k + \sum_U \boldsymbol{b}'(z_k)\tilde{\boldsymbol{\theta}}_x}$$

pour $\tilde{\boldsymbol{\theta}}_y$ donné par (1.12) et $\tilde{\boldsymbol{\theta}}_x$ obtenu de (1.12) pour x_k. La variable linéarisée associée à R est :

$$u_k = \frac{1}{t_x}(y_k - Rx_k)$$

et la variance asymptotique de $\hat{R}_{BS}$ est égale à :

$$\sum_{k\in U}\sum_{i\in U} \Delta_{ki} \frac{u_k - \hat{f}_{u,k}}{\pi_k} \frac{u_l - \hat{f}_{u,l}}{\pi_l}$$

avec :

$$\hat{f}_{u,k} = \boldsymbol{b}'(z_k)\left(\sum_U \boldsymbol{b}(z_k)\boldsymbol{b}'(z_k)\right)^{-1} \sum_U \boldsymbol{b}(z_k) u_k.$$

Le résultat 1 est obtenu sous des hypothèses concernant le plan de sondage (sur les π_k et π_{kl}), l'inversion de la matrice $\sum_U \boldsymbol{b}(x_k)\boldsymbol{b}'(z_k)$ et pour des variables d'intérêt $\mathcal{Y}$ ayant un moment d'ordre deux fini. Il est également supposé que le nombre de nœuds K tend vers l'infini et $K = o(n^{1/3})$, $K = o(N)$. Pour obtenir la variance asymptotique de $\hat{R}_{BS}$, on utilise le résultat 2 et le cadre asymptotique de Goga (2005).

4. Comparaison empirique

Dans cette étude comparative, nous considérons des données simulées selon le même principe que dans Breidt et Opsomer (2000). Nous générons des variables Z, X et Y pour une population de taille $N = 1000$ telles que :

z_k suit une loi uniforme sur [0 ; 1], ε_k suit une loi normale centrée d'écart-type $\sigma = 0.1$ ou $\sigma = 0.4$,

$$y_k = f(z_k) + \varepsilon_k,$$

$$x_k = g(z_k) + \varepsilon_k,$$

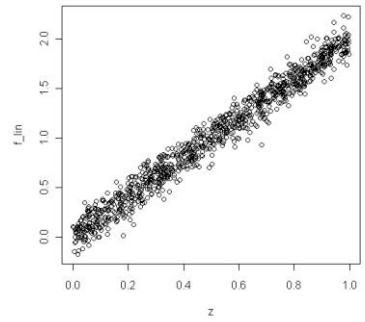
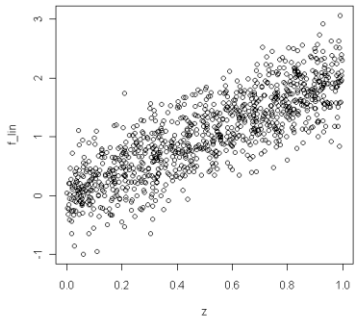
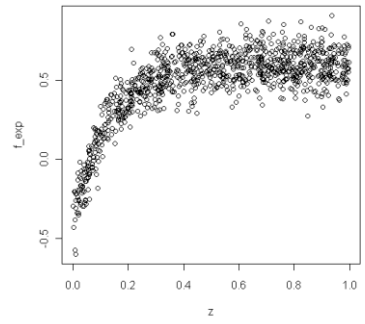
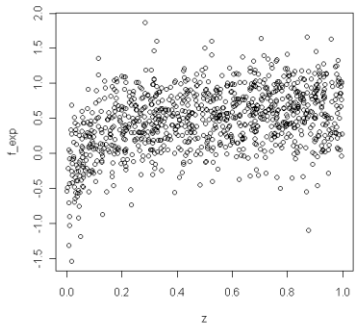
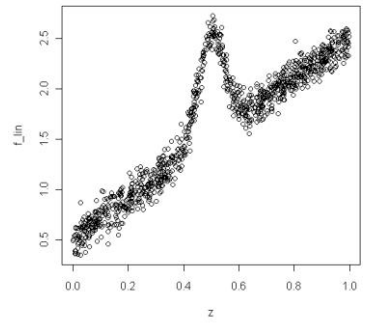
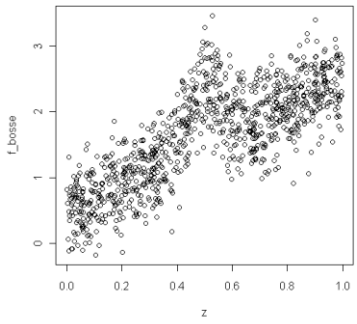
où les fonctions f et g sont choisies parmi les fonctions suivantes :

$$f_{\text{lin}} = 1 + 2(x - 0.5)$$

$$f_{\text{exp}} = 0.6 + \exp\{-8x\}$$

$$
\begin{aligned}
f_{\text{bosse}} &= -1.5 + 2(x - 0.5) + \exp\{-200(x - 0.5)^2\} \\
f_{\text{saut}} &= 1.5 + \left[0.35 + 2(x - 0.5)^2\right] I_{\{x \leq 0.65\}} \\
f_{\text{sinus}} &= 2 + \sin(2\pi x)
\end{aligned}
$$

La figure 1 donne les diagrammes de dispersion des variables $f(z_k) + \varepsilon_k$ pour les cinq différentes fonctions f en fonction des z_k pour les 1000 observations de la population générée et pour les deux valeurs de σ (0.1 et 0.4).

$\sigma = 0.1$	$\sigma = 0.4$
	
	
	

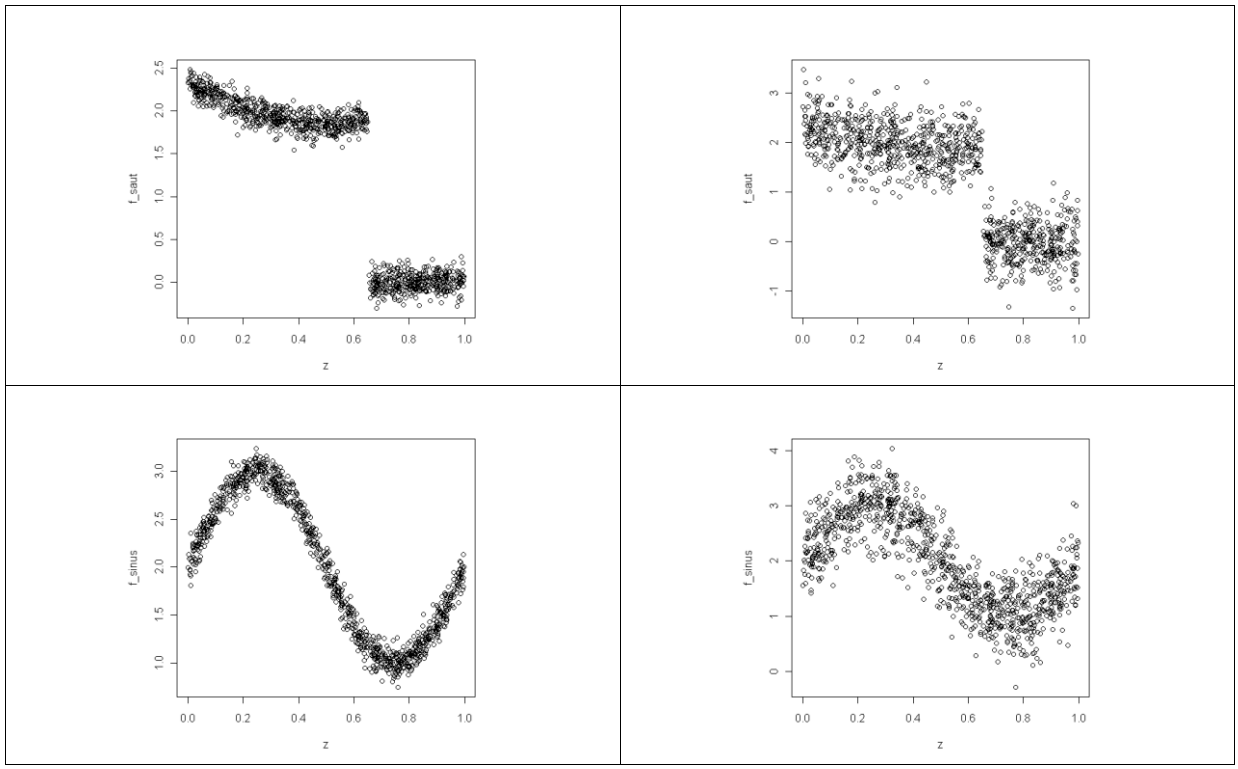


Figure 1 : Diagrammes de dispersion des 1000 observations de la population selon les modèles « linéaire », « exponentiel », « bosse » , « saut » ou « sinus » avec $\sigma = 0.1$ à gauche et $\sigma = 0.4$ à droite

Nous considérons 10000 tirages d'échantillons de taille $n = 100$ selon un plan simple sans remise et nous nous intéressons à l'estimation du ratio $R = t_y / t_x$. Nous considérons trois estimateurs du ratio selon que les totaux t_y et t_x sont des estimateurs de Horvitz-Thompson (HT) qui ne prennent pas en compte l'information auxiliaire Z, des estimateurs GREG basés sur la régression linéaire ou des estimateurs BS basés sur les B-splines d'ordre 4. Les estimateurs GREG et BS prennent en compte Z. Pour l'approche par splines de régression, nous considérons la base des B-splines d'ordre 4 avec 5 nœuds intérieurs positionnés aux quantiles de Z dans la population. D'autres choix pour les nœuds, en particulier des choix adaptatifs en fonction des données, pourraient être envisagés.

Pour comparer les trois estimateurs (HT, GREG, BS), nous considérons différents couples de fonctions f et g :

- (1) : $f = f_{\text{lin}}$ et $g = f_{\text{exp}}$,
- (2) : $f = f_{\text{lin}}$ et $g = f_{\text{saut}}$,
- (3) : $f = f_{\text{bosse}}$ et $g = f_{\text{saut}}$,
- (4) : $f = f_{\text{saut}}$ et $g = f_{\text{sinus}}$,
- (5) : $f = f_{\text{bosse}}$ et $g = f_{\text{sinus}}$.

Dans le tableau suivant, nous donnons pour chaque type d'estimateur et pour les différents couples de fonctions f et g, les biais absolus relatifs en % ainsi que les rapports entre l'écart quadratique

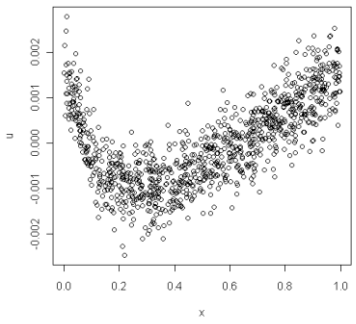
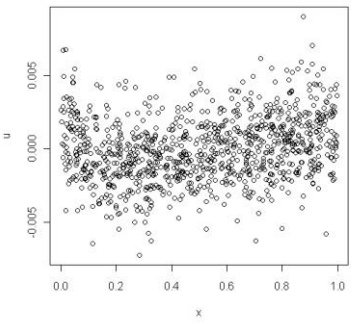
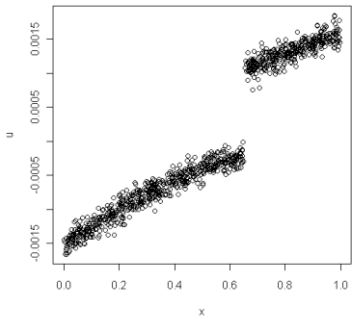
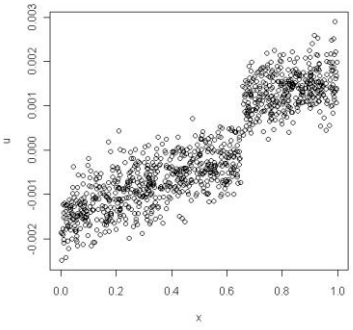
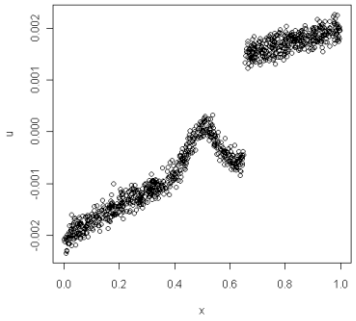
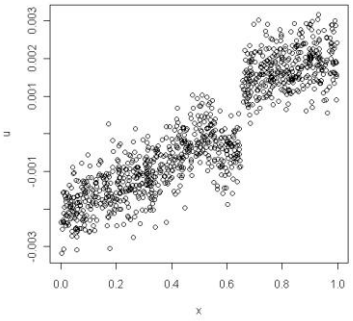
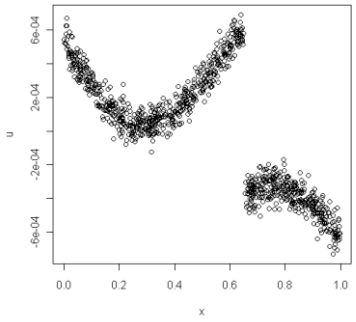
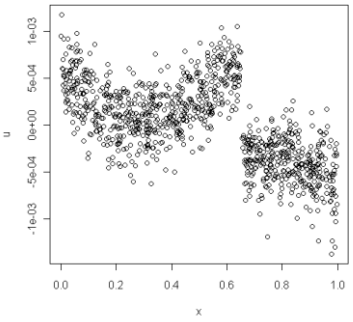
moyen (EQM) de l'estimateur considéré et l'EQM de l'estimateur HT en % calculés à partir des 10000 simulations.

Comme on pouvait s'y attendre, les biais relatifs sont faibles pour les trois estimateurs quels que soient les modèles considérés. Par contre, en terme d'erreur quadratique moyenne, si on omet le cas (1) avec $\sigma = 0.4$, les résultats illustrent clairement l'intérêt d'utiliser l'information auxiliaire puisque les estimateurs GREG et BS donnent de meilleurs voire des bien meilleurs résultats que HT. Concernant la comparaison entre GREG et BS, on peut remarquer que BS conduit à des résultats meilleurs que GREG pour (4) et (5) et pour (2) et (3) lorsque $\sigma = 0.1$, mais que les résultats sont équivalents pour (2) et (3) lorsque $\sigma = 0.4$.

Modèles	***(1)*** ***f : linéaire*** ***g : exp***		***(2)*** ***f : linéaire*** ***g : saut***		***(3)*** ***f : bosse*** ***g : saut***		***(4)*** ***f : saut*** ***g : sinus***		***(5)*** ***f : bosse*** ***g : sinus***	
	Biais relatif	**Ratio EQM**	**Biais relatif**	**Ratio EQM**	**Biais relatif**	**Ratio EQM**	**Biais relatif**	**Ratio EQM**	**Biais relatif**	**Ratio EQM**
	Sigma=0.1									
HT	0.1	100	0.8	100	0.8	100	0.1	100	0.1	100
GREG	0.1	77	0.1	85	0.1	88	0.2	49	0.1	51
BS	0.1	**32**	0.1	**39**	0.1	**65**	0.1	**19**	0.03	**19**
	Sigma=0.4									
HT	0.5	100	1	100	0.8	100	0.04	100	0.2	100
GREG	0.5	**97**	0.1	18	0.1	**17**	0.2	65	0.1	29
BS	0.8	101	0.2	**16**	0.2	**17**	0.1	**48**	0.04	**18**

Tableau 1 : Biais relatif et ratio entre EQM de l'estimateur et EQM de HT en %
(les meilleurs résultats sont en gras)

Les diagrammes de dispersion de la variable linéarisée en fonction de la variable auxiliaire sont représentés sur la figure 2 et permettent d'illustrer l'adéquation d'un modèle linéaire. Ainsi, pour le cas (1) avec $\sigma = 0.4$, on voit qu'il n'existe pas de relation claire (linéaire ou pas) entre la variable linéarisée et la variable auxiliaire : il n'y a donc pas d'amélioration lorsque on utilise le GREG ou BS comparativement à HT. Pour les cas (2) et (3) avec $\sigma = 0.4$, on remarque que le modèle linéaire ne s'ajuste pas trop mal, ce qui explique que l'on n'ait pas d'amélioration en utilisant BS au lieu de GREG. Cette étude empirique est toutefois limitée et il serait intéressant d'étudier l'estimation d'autres paramètres que le ratio comme par exemple l'indice de Gini.

$\sigma = 0.1$	$\sigma = 0.4$
	
	
	
	

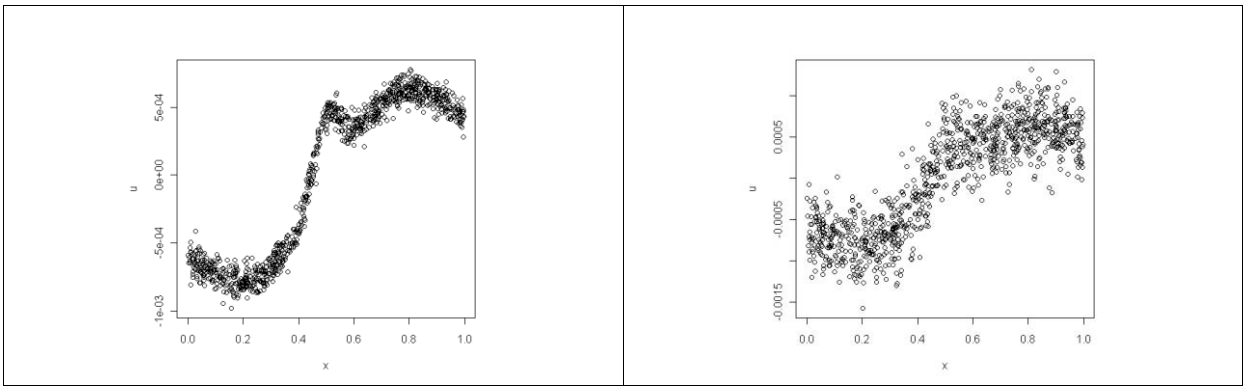


Figure 2 : Diagrammes de dispersion des variables linéarisée en fonction de la variable auxiliaire pour les cas (1), (2), (3), (4) et (5) de haut en bas avec $\sigma = 0.1$ à gauche et $\sigma = 0.4$ à droite

5. Perspectives

Nous avons proposé dans ce travail une méthode générale pour prendre en compte l'information auxiliaire par des méthodes non-paramétriques quand l'estimation d'un paramètre non-linéaire est désirée. Le cadre de l'estimation non-paramétrique par des B-splines est décrit plus en détail et appliqué à l'estimation d'un ratio. La méthode proposée est validée lors d'une étude par simulation.
Nous envisageons deux extensions de ce travail. Tout d'abord, nous souhaitons étudier l'approche par calage non-paramétrique notamment dans le cas des B-splines et les liens qu'entretient cette approche avec l'approche « model-assisted ». Dans le cas de l'estimation d'une moyenne, Wu and Sitter (2001) introduisent l'approche « model-calibration » tandis que Montanari and Ranalli (2005) effectuent un calage non-paramétrique par des polynômes locaux. Notre travail est une généralisation de ces travaux puisqu'il pourra s'appliquer à des paramètres non-linéaires plus généraux que la moyenne comme l'indice de Gini par exemple. L'autre extension consiste à construire des estimateurs basés sur plusieurs variables auxiliaires. Plus précisément, pour $\Phi = \Phi(t_x, t_y)$ nous pourrons envisager d'estimer t_x en utilisant une variable auxiliaire $\mathcal{Z}_1$, et t_y en utilisant une autre variable $\mathcal{Z}_2$. L'estimateur obtenu pourra être linéarisé en utilisant les fonctions d'influence partielles introduites par Goga, Deville and Ruiz-Gazen (2009).

Bibliographie

[1] Cassel, C.M., Särndal, C. E. and Wretman, J.H. (1976), Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika*, 63, 615 ‑ 620.
[2] Breidt, F.J. and Opsomer, J. (2000), Local Polynomial Regression Estimators in Survey Sampling, *The Annals of Statistics*, 28, 1026 ‑ 1053.
[3] Breidt, F.J. and Opsomer, J. (2005), Model-assisted estimation for complex surveys using penalised splines, *Biometrika*, 92, 831 ‑ 846.
[4] Deville, J.C. (1999), Variance estimation for complex statistics and estimators: linearization and residual techniques, *Survey Methodology*, 25, 193 ‑ 203.
[5] Dierckx, P. (1993). Curves and surface fitting with splines. Oxford, Clarendon Press.
[6] 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.
[7] Goga, C., Deville, J.C. and Ruiz-Gazen, A. (2009), Use of functionals in linearization and composite estimation with application to two-sample survey data, a paraître dans *Biometrika*.
[8] Montanari, G.E. and Ranalli, M.G. (2005), Nonparametric model calibration estimation in survey sampling, Journal of the American Statistical Association, 100, 1429-1442.
[9] Särndal C.E., Swensson B. and Wretman J. (1992), Model Assisted Survey Sampling, Springer, Berlin.

[10] Schumaker, L. L. (1981). Spline Functions: Basic Theory. Wiley, New-York.
[11] Wu, C. and Sitter, R.R. (2001), A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, 96, 185-193.
[12] Wolter, K. M. (2007). Introduction to variance estimation. Springer, second edition.