Estimation de paramètres non-linéaires par des méthodes non-paramétriques en population finie

Camelia Goga[1] et Anne Ruiz-Gazen[2]

(1) IMB, Université de Bourgogne-Dijon,
camelia.goga@u-bourgogne.fr

(2) TSE, Université des Sciences Sociales-Toulouse,
ruiz@cict.fr

JMS Paris-2009

Paramètre non-linéaire : estimation sans information auxiliaire

- une population finie $U = \{1, \ldots, k, \ldots, N\}$
- un échantillon $s \in \mathcal{S}$ et $s \subset U$ selon un plan $p(s)$; les probabilités d'inclusion π_k et π_{kl}.

Objectif : estimer un paramètre non-linéaire avec information auxiliaire à l'aide d'un modèle non-paramètrique.

Exemples :

- $R = t_y/t_x$ et $\text{Cov} = \sum_U x_k y_k/N - \sum_U x_k \sum_U y_k/N^2$
- fonctions de la fonction de répartition ou des quantiles : l'indice de Gini $G = \sum_U y_k \left(2F(y_k) - 1\right)/t_y$, de Theil ...
- les éléments propres d'une ACP fonctionnelle : $\Gamma v_j(t) = \lambda_j v_j(t)$, $t \in [0, 1]$ (Cardot *et al.*, 2009),
- les quantiles multidimensionnels (Chaouch & Goga, 2009)

Survol des approches

Estimateur non-linéaire dans deux cas :

1. Paramètre non-linéaire sans information auxiliaire :
   - méthodes de rééchantillonnage (Berger & Skinner, 2005, Gross, 1980, Chauvet, 2007)
   - méthodes de linéarisation : équations estimantes (Kovačević & Binder, 1997), Taylor linéarisation (Särndal et al., 1992, Demnati & Rao, 2004) et fonction d'influence (Deville, 1999)
2. Paramètre linéaire avec information auxiliaire :
   - calage (Deville & Särndal 1992), model-calibration (Wu & Sitter, 2001), hyper-calage (Deville, 2007),
   - "model-assisted" (Särndal *et al.*, 1992), "model-based" (Chambers *et al.*, 1996)

Särndal *et al.* (1992) : ratio et estimateurs GREG ;
Deville (1999) : paramètre non-linéaire et calage ;
Harms & Duchesne (2006) : calage sur quantiles.

Linéarisation par la fonction d'influence

- Paramètre non-linéaire Φ
- On considère la mesure $M = \sum_{k \in \mathbf{U}} \delta_{x_k}$ sur $\mathbf{R}^p$ et

$$\phi = T(M) \quad \text{pour une fonctionnelle homogène} \quad T \quad \text{de degré } \alpha$$

(Ex : un total est de degré 1 et un ratio de degré 0).

- Estimer M par $\hat{M} = \sum_{k \in \mathbf{s}} w_k \delta_{x_k}$, $w_k = \frac{1}{\pi_k}$ (ou poids de calage)

$$\hat{\phi} = T(\hat{M}) \quad \text{l'estimateur par } \textbf{substitution de} \quad \phi$$

Objectif : donner un développement asymptotique de $T(\widehat{M}/N)$ autour de M/N

Résultat asymptotique

Définition

La fonction d'influence de $T(M)$ est définie comme la dérivée au sens de Gateaux 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

La variable linéarisée u_k pour $k \in U$ est $u_k = IT(M, x_k), \quad k \in U.$

Résultat

Sous des conditions générales,

$$\sqrt{n}N^{-\alpha}(\hat{\phi} - \phi) = \sqrt{n}N^{-\alpha} \sum_U u_k(w_k - 1) + o(1).$$

$$w_k = 1/\pi_k \quad Var(\hat{\phi}) \simeq Var\left(\frac{\sum_s u_k}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$$

Le total : estimation avec information auxiliaire et régression non-paramètrique

Objectif : utiliser l'information auxiliaire $\mathcal{Z}$ pour estimer

$$t_y = \sum_U y_k$$

Approche "model-assisted" :

modèle de superpopulation ξ : y_k, $k \in U$ sont des variables *iid*,

$$\xi : \left\{ \begin{array}{rcl} E_\xi(y_k) & = & f(z_k) \\ V_\xi(y_k) & = & v(z_k) \end{array} \right.$$

L'estimateur par la différence généralisée (Cassel *et al.*, 1976)

$$\hat{t}_{y,diff} = \sum_{k \in s} \frac{y_k - f(z_k)}{\pi_k} + \sum_{k \in U} f(z_k)$$

Si $f = Z'\beta$ alors estimateur GREG.

Estimation d'un total par une régression non-paramètrique (NP)

1. "Population level" : Construire $\hat{f}_y$ en utilisant $(y_k)_{k\in U}$ et une régression non-paramétrique :
   - méthodes à noyau : les polynômes locaux (Breidt & Opsomer, 2000),
   - splines en utilisant les bases des polynômes tronqués et une pénalisation (Breidt & Opsomer, 2005) ou B-splines (Goga, 2005).
2. "Sample level" : Construire des estimateurs $\tilde{f}_y$ basés sur le plan s pour $\hat{f}_y$ et estimer le total t_y par

$$\begin{aligned}\hat{t}_{y,np} &= \sum_{k\in s}\frac{y_k-\tilde{f}_y(z_k)}{\pi_k}+\sum_{k\in U}\tilde{f}_y(z_k)\\ &= \sum_{k\in s} w_{ks}y_k\end{aligned}$$

Important : pour les trois méthodes NP les poids w_{ks} ne dépendent pas de $\mathcal{Y}$ et contient $\mathcal{Z}$

Paramètre non-linéaire : estimation avec information auxiliaire et régression non-paramètrique

- Objectif : estimer $\Phi = \Phi(t_x, t_y)$ en prenant en compte $\mathcal{Z}$.
- Méthode :
  - écrire $\Phi = T(M)$
  - estimer M en utilisant des poids w_{ks} obtenus en faisant une régression non-paramétrique :

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

  - et construire l'estimateur non-paramétrique par substitution

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

Résultat asymptotique

Sous des conditions générales,

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

$$\hat{t}_{u,\text{diff}}=\sum_{k\in s}\frac{u_k-\hat{f}_u(z_k)}{\pi_k}+\sum_{k\in U}\hat{f}_u(z_k)$$

où $\hat{f}_u$ est l'estimateur (inconnu) non-paramétrique de f obtenu en régressant le vecteur de variables linéarisées $(u_1,\ldots,u_N)$ sur $\mathcal{Z}$.

$$Var\left(\widehat{\Phi}_{np}-\Phi\right)\simeq\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{u_k-\hat{f}_u(z_k)}{\pi_k}\frac{u_l-\hat{f}_u(z_l)}{\pi_l}$$

Important : l'usage des modèles NP est encore plus justifié car la relation entre u_k et z_k est inconnue et peut être plus compliquée que la relation entre y_k et z_k.

Estimation d'un total par une régression non-paramètrique par des B-splines (Goga, 2005)

Le modèle ξ pour $k = 1, \ldots, N$:

$$\xi : \begin{cases} E_\xi(y_k) &= f(z_k) \\ V_\xi(y_k) &= v(z_k) \end{cases}$$

Objectif : "trouver" la vraie relation f qui gouverne le nuage de points (z_k, y_k).
Méthode : approcher f par $\hat{f}$ appartenant à un sous-espace de dimension beaucoup plus petite que N :

- si l'espace des fonctions affines de $\mathcal{Z}$, régression simple ou multiple
- plus générale : l'espace des fonctions polynômes par morceaux : fonctions splines

Critère : moindres carrés

Base de B-splines

- L'ensemble $S_{K,m}$ de fonctions splines de degré m ($m \geq 2$) avec K noeuds intérieurs équidistants

$$0 = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = 1$$

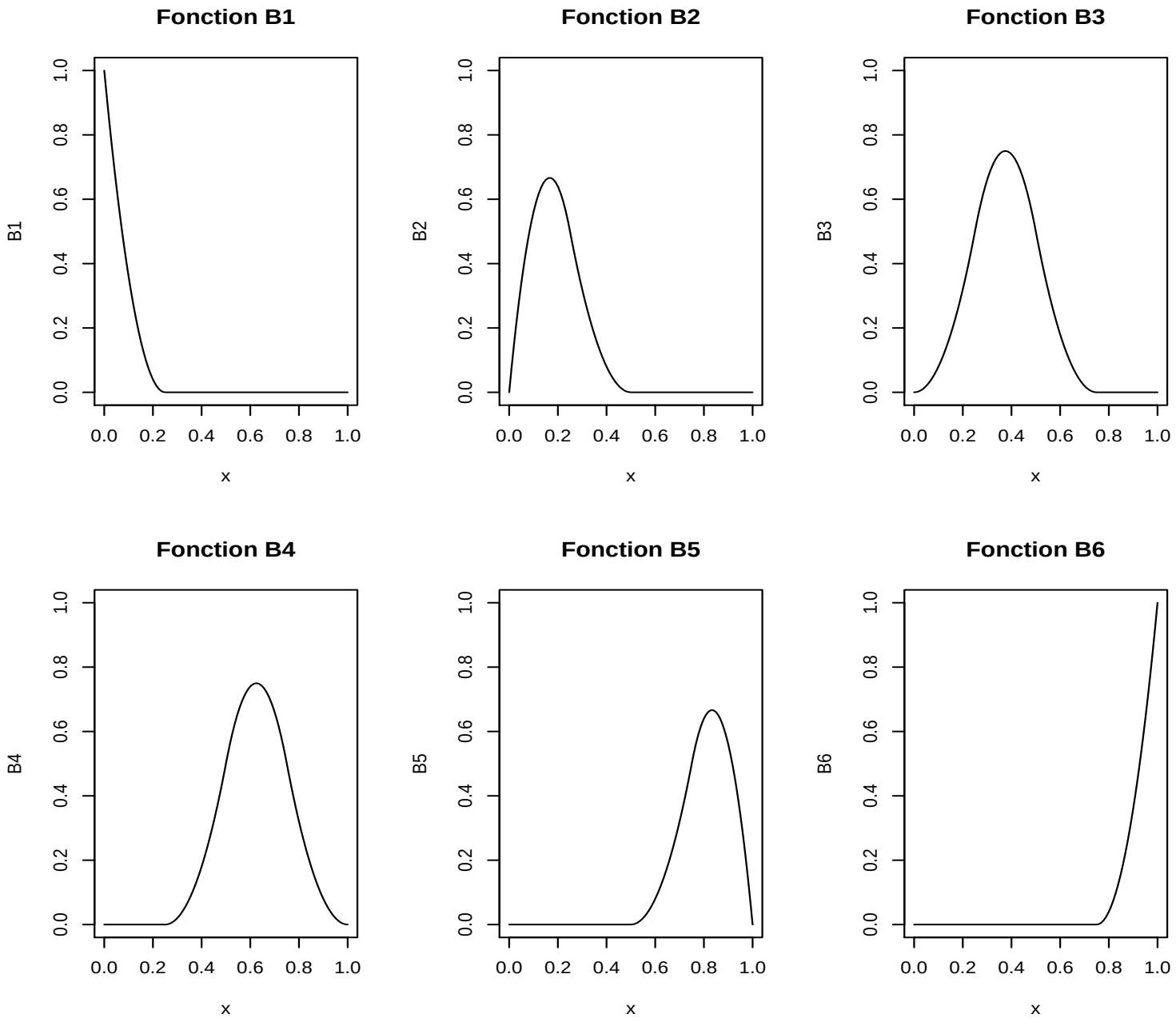
$$S_{K,m} = \{s \in C^{m-2}[0,1] : s(x) \quad \text{est un polynôme de degré } m-1 \quad \text{sur} \quad (\xi_j, \xi_{j+1})\}.$$

- $S_{K,m}$ est de dimension $q = K + m$ (Schumaker 1981, Dieckx 1993) et une base est donnée par les fonctions B-splines

$$B_1, \ldots, B_q \quad \text{et} \quad \sum_{j=1}^{q} B_j = 1$$

  - pour $m = 2$, $S_{K,2}$: les fonctions continues et linéaires par morceaux ;
  - pour $m = 2$ et $K = 0$, $S_{0,2}$: les droites $\rightarrow$ la régression simple.

Fonctions B-splines de degré 3 et 3 noeuds équidistants



Estimation de la fonction de régression f

On note $\mathbf{B}_U = (B_j(z_k))_{k\in U, j=1,\dots q}$ et $\mathbf{b}'(z_k)$ les vecteurs lignes ;

$$\hat{f}_y(z_k) = \mathbb{P}_{S_{K,m}}\mathbf{y}_U = \sum_{j=1}^{q} \hat{\theta}_j B_j(z_k) = \mathbf{b}'(z_k)\hat{\boldsymbol{\theta}}_y$$

$\mathbb{P}_{S_{K,m}}$ est le projecteur sur $S_{K,m}$ qui est fixe une fois la base choisie

$$\begin{cases} \hat{\boldsymbol{\theta}}_y & = & \text{Arg}\min_{\boldsymbol{\theta}\in\mathbb{R}^q} \sum_{k=1}^{N} \left(y_k - \sum_{j=1}^{q} \theta_j B_j(z_k)\right)^2 \\ \hat{\boldsymbol{\theta}}_y & = & (\mathbf{B}'_U\mathbf{B}_U)^{-1}\mathbf{B}'_U\mathbf{y}_U = \left(\sum_{i\in U} \mathbf{b}(z_i)\mathbf{b}'(z_i)\right)^{-1} \sum_{i\in U} \mathbf{b}(z_i)y_i \end{cases}$$

$$\tilde{f}_y(z_k) \quad = \quad \mathbf{b}'(z_k)\tilde{\boldsymbol{\theta}}_{\mathbf{y}} = \mathbf{b}'(z_k)(\mathbf{B}'_s\boldsymbol{\Pi}_s^{-1}\mathbf{B}_s)^{-1}\mathbf{B}'_s\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s$$

$$\boldsymbol{\Pi}_s = \text{diag}(\pi_k)_{k\in s} \text{ et } \mathbf{B}'_s = (\mathbf{b}'(z_k))_{k\in s}.$$

Les poids non-paramètriques w_{ks}

Le total $t_y = \sum_U y_k$ est estimé par :

$$\begin{aligned}\hat{t}_{y,BS} &= \underbrace{\sum_{k\in s} \frac{y_k - \tilde{f}_y(z_k)}{\pi_k}}_{=0} + \sum_{k\in U} \tilde{f}_y(z_k) = \left(\sum_U \mathbf{b}'(z_k)\right)\tilde{\theta}_y \\ &= \sum_s w_{ks} y_k\end{aligned}$$

Propriétés

- un GREG sur $\mathbf{b}' = (B_1, \ldots, B_q)$ et $q \to \infty$,
- calage : $\sum_s w_{ks} B_j(z_k) = \sum_U B_j(z_k)$ pour $j = 1, \ldots, K$
- $\sum_U w_{ks} = N$
- asymptotiquement sans biais et convergent.

Estimation par des B-splines d'un paramètre non-linéaire

- Le paramètre non-linéaire $\Phi = T(M)$;
- L'estimateur non-paramètrique $\hat{\Phi}_{np} = T(\hat{M}_{np})$ qui utilise les poids

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

- $Var(\hat{\Phi}_{np}) \simeq Var \sum_s \frac{u_k - \hat{f}_u(z_k)}{\pi_k}$ avec
  - u_k la variable linéarisée de Φ
  - $\hat{f}_u(z_k) = \mathbb{P}_{S_{K,m}} \mathbf{u}_U = \mathbf{b}'(z_k)(\mathbf{B}'_U \mathbf{B}_U)^{-1} \mathbf{B}'_U \mathbf{u}_U$

Étude sur des données simulées : le cas d'un ratio

- $R = \frac{\sum_U y_k}{\sum_U x_k}$
- $y_k = f(z_k) + \epsilon_k, \quad \epsilon \sim N(0, 0.1)$
- $x_k = g(z_k) + \epsilon_k, \quad \epsilon \sim N(0, 0.1)$
  $z \in [0, 1]$, distribution uniforme.
- 3 fonctions différentes
  $f_{\text{lin}}(x) = 1 + 2(x - 0.5)$,
  $f_{\text{exp}}(x) = 0.6 + \exp(-8x)$,
  $f_{\text{saut}}(x) = 1.5 + [0.35 + 2(x - 0.5)^2] I_{\{x \leq 0.65\}}$
- Une population $\mathcal{U}$, $N = 1000$.
- Sondage aléatoire simple sans remise de taille $n = 100$,
  $\pi_k = n/N$.
- Splines avec 5 noeuds intérieurs positionnés aux quantiles de Z dans la population et $m = 3$.

Dépendence de la variable linéarisée u de R en fonction de la variable auxiliaire z

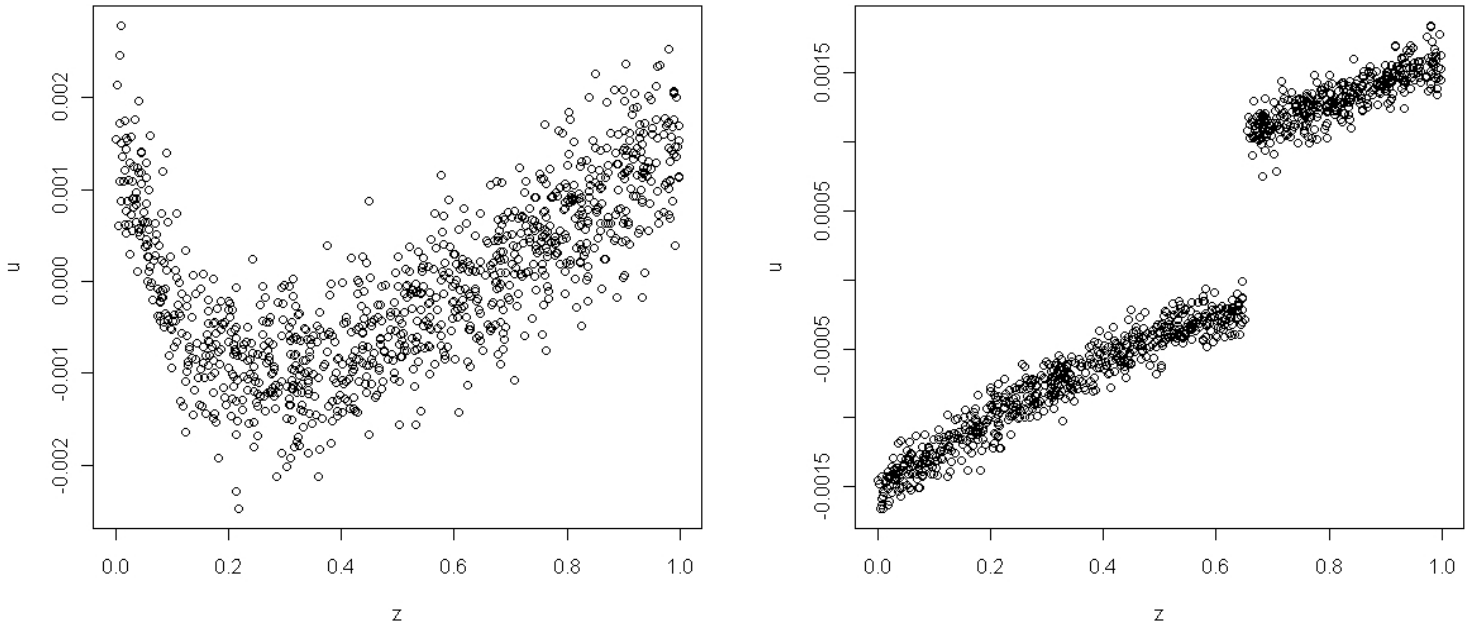


FIG.: Gauche : $R = \frac{y_{lin}}{x_{exp}}$ et droite : $R = \frac{y_{lin}}{x_{saut}}$

Comparaison de trois estimateurs de R

Le ratio $R = \frac{t_y}{t_x}$ estimé par R_{HT}, R_{GREG} et R_{BS}

Modèles	f : lin, g : exp	f : lin, g : saut
HT	100	100
GREG	77	85
BS	32	39

Ratio EQM estimateur sur EQM HT

$\text{EQM}(\hat{\theta}) = \frac{1}{b}\sum_{r=1}^{b}(\hat{\theta}_r - \theta)^2$ pour $b = 10000$ simulations.

Étude sur des données réelles : le cas de l'indice de Gini

On considère une base de sondage constituée de 22741 salariés pour lesquels on dispose des salaires en 1999 et en 2000 (extrait enquêtes emploi INSEE).

On s'intéresse à l'estimation de l'indice de Gini en 2000 et on suppose que le salaire en 1999 constitue l'information auxiliaire.

Étude sur des données réelles : le cas de l'indice de Gini

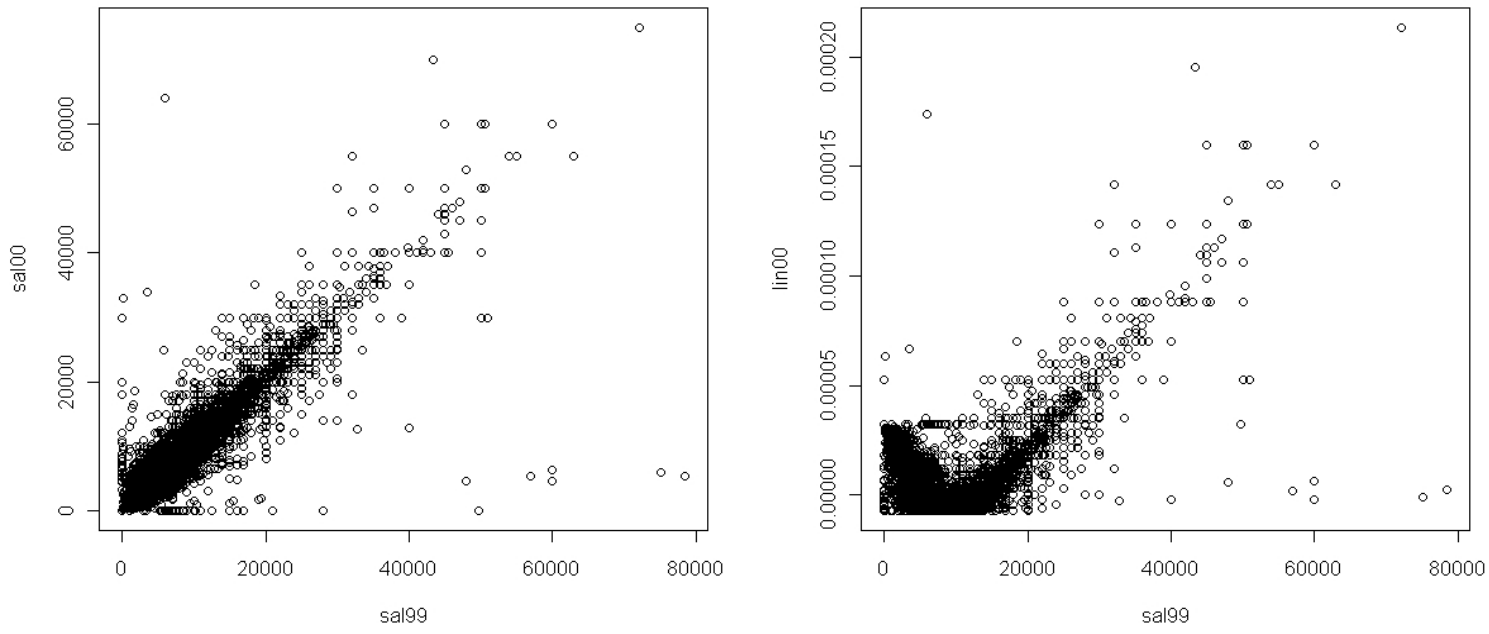


FIG.: Diagrammes de dispersion (salaire 2000 à gauche (linéarisée de Gini à droite) en fonction de salaire 1999.

Étude sur des données réelles : le cas de l'indice de Gini

Lorsque l'on modélise la variable salaire en 2000 en fonction du salaire en 1999, le gain en terme de rapport des écarts-type des résidus entre régression linéaire et B-splines (mêmes paramétrage que précédemment) est de 7% (Exemple de l'estimation du total ou de la moyenne).

Lorsque l'on modélise la variable linéarisée pour le coefficient de Gini du salaire en 2000 en fonction du salaire en 1999, le gain en terme de rapport des écarts-type des résidus entre régression linéaire et B-splines (mêmes paramétrage que précédemment) est de 50%.

Conclusion et perspectives

- Estimation NP par des B-splines : calage non-paramétrique et choix des knoeuds ;
- Extension : utilisation des informations auxiliaires différentes par la techniques de linéarisation avec des fonction d'influence partielles (Goga, Deville, Ruiz-Gazen, 2009) ;
- Estimation de la variance