Survey sampling and nonparametric models for taking into account the auxiliary information : a B-splines approach

Camelia Goga

IMB, Université de Bourgogne
e-mail : camelia.goga@u-bourgogne.fr

Workshop on Survey Sampling Theory and Methodology
Kuressaare - August 25-29 2008

Population, sample.

- a finite population $U = \{1, \dots, k, \dots, N\}$
- a sample $s \in \mathcal{S}$ and $s \subset U$
- the sampling design $p(s)$: a probability distribution on the set $\mathcal{S}$;
  $p(s)$ is controlled by the statistician.
- the inclusion probabilities
  - of first degree : $\pi_k = Pr(k \in s)$
  - of second degree : $\pi_{kl} = Pr(k, l \in s)$ for $k \neq l$ and $\pi_{kk} = \pi_k$

Examples

Example 1 : Simple random sampling without replacement of size n :

- the sampling design $p(s) = 1/C_N^n$,
- $\pi_k = \frac{n}{N}$ and $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ for $k \neq l$.

Example 2 : Simple random sampling with replacement and proportional to the size :

- the sampling design $p(s) = p_{k_1} p_{k_2} \dots p_{k_m}$ and p_{k_i} the probability of selecting the individual k_i at the ith selection ;
- $p_k = x_k / \sum_U x_k$

Estimator of Finite Population Total : the Horvitz-Thompson Estimator

Let us consider :

- the variable of interest $\mathcal{Y}$,
  $y_k =$ the value of $\mathcal{Y}$ for the k-th individual,
- we know y_k for $k \in s$
- we want to estimate the population total of $\mathcal{Y}$, namely

$$t_y = \sum_U y_k$$

- $\hat{t} = \sum_U w_k(s) y_k \quad \text{with} \quad w_k(s) = 0 \quad \text{if} \quad k \in U - s,$

- The Horvitz-Thompson (HT) estimator :

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

$$I_k = \mathbf{1}_{\{k \in s\}} \quad \text{the sample membership}$$

Properties

The estimator HT for a total is

- the only homogeneous and linear in y_k estimator being unbiased and with weights not depending on the variable of interest $\mathcal{Y}$ and on the sample s ;
- the HT variance is

$$V(\hat{t}_\pi) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

- the HT variance estimator is

$$\hat{V}(\hat{t}_\pi) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Drawbacks :

1. The HT estimator contains little auxiliary information (the π_k) !
2. The variance as well as the variance estimator contain double sums.

Variance reduction - Auxiliary information

An auxiliary information $\mathcal{Z}$, (uni ou multidimensional)
z_k the value for the k-th individual from U.
We know

- the total $\sum_U z_k$ or
- z_k for all $k \in U$

Approaches for improving the HT estimator

- **Calibration** : we improve the HT estimator without considering any super-population model (Deville & Särndal 1992)
- **The super-population model** ξ : y_k are independent and identically distributed random variables with

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

  - **"model assisted"** : we construct the estimator based on the sampling design and assisted by the super-population model (Särndal, Swensson & Wretman 1992)
  - **"model bassed"** : we predict the population total by using the super-population model without taking into account the sampling design (Royall & Cumberland 1978)

The super-population model with "model-assisted" estimator :

$$\underbrace{Y_1, Y_2, \ldots \Longrightarrow Y_1, Y_2,}_{Y_i \quad \text{random variables} \quad (\xi)} \; \underbrace{\ldots, Y_N \Longrightarrow y_k, k \in s}_{I_k \quad \text{random} \quad (p)}$$

Goal : We search for an estimator $\hat{t}$ for the total t_Y which takes into account $\mathcal{Z}$ such that

$$E_\xi E_p(\hat{t} - t_y) = 0$$

and minimizing the "anticipated variance"

$$Var_{\xi,p}(\hat{t} - t_y) = E_\xi E_p(\hat{t} - t_y)^2 - \left[E_\xi E_p(\hat{t} - t_y)\right]^2.$$

Remark : $E_{(\xi,p)} = E_\xi E_p = E_p E_\xi$ because the sampling design p does not depend on the variables $\mathcal{Y}$ and $\mathcal{Z}$ (non-informatif sampling design) ;

We consider the HT estimator for the total

$$\hat{t}_\pi = \sum_{k\in s} \frac{y_k}{\pi_k}$$

which is p-unbiased but ξ-biased :

$$E_p(\hat{t}_\pi) = t_y \quad \text{et} \quad E_\xi(\hat{t}_\pi - t_y) = \sum_s \frac{f(z_k)}{\pi_k} - \sum_U f(x_k).$$

We modify $\hat{t}_\pi$ in order to obtain a ξ-unbiased estimator : the generalized difference estimator (Cassel, Särndal & Wretman 1976)

$$\begin{aligned} \hat{t}_{diff} &= \sum_{k\in s} \frac{y_k - f(z_k)}{\pi_k} + \sum_{k\in U} f(z_k) \\ &= \underbrace{\sum_{k\in s} \frac{y_k}{\pi_k}}_{\hat{t}_{HT}} - \underbrace{\left(\sum_s \frac{f(z_k)}{\pi_k} - \sum_{k\in U} f(z_k) \right)}_{E_\xi(\hat{t}_{HT} - t_y)} \end{aligned}$$

Properties of $\hat{t}_{diff}$

- $\hat{t}_{diff}$ is p and ξ-unbiased ;
- The variance under the sampling design is

$$V_p(\hat{t}_{diff}) = \sum_{k \in U} \sum_{i \in U} \Delta_{ki} \frac{y_k - f(z_k)}{\pi_k} \frac{y_i - f(z_i)}{\pi_i};$$

- The variance under the model and the sampling design is

$$E_\xi E_p(\hat{t}_{diff} - t_y)^2 = \sum_U \frac{1 - \pi_k}{\pi_k} v(z_k)$$

  the Godambe-Joshi lower bound (1965).

Drawback : in practice, we do not know $f(z_k)$.

Estimation of the regression function

We estimate the $f(z_k)$ in two steps :

•First step - at the population level :
we estimate $f(z_k)$ by $\hat{f}(z_k)$ using parametric or nonparametric methods ;

The estimators $\hat{f}(z_k)$ depend on the whole population U, so they are unknown.
At this level, the sampling design does not appear.

•Second step at the sample level : we estimate $\hat{f}(z_k)$ by $\hat{\hat{f}}(z_k)$ using the sampling design p.
The difference estimator becomes

$$\sum_{k \in s} \frac{y_k - \hat{\hat{f}}(z_k)}{\pi_k} + \sum_{k \in U} \hat{\hat{f}}(z_k).$$

The GREG Estimator

- If $f(z_k) = \mathbf{z}_k'\boldsymbol{\beta} \rightarrow$ the generalized regression estimator (GREG) (Särndal, Swensson & Wretman 1992)

$$\text{Population level}: \quad \hat{\boldsymbol{\beta}} = \left(\sum_U \frac{\mathbf{z}_k\mathbf{z}_k'}{v_k}\right)^{-1} \sum_U \frac{\mathbf{z}_k y_k}{v_k}$$

$$\text{Sample level}: \quad \hat{\boldsymbol{\beta}}_s = \left(\sum_s \frac{\mathbf{z}_k\mathbf{z}_k'}{\pi_k v_k}\right)^{-1} \sum_s \frac{\mathbf{z}_k y_k}{\pi_k v_k}$$

Then, $\hat{t}_{GREG} = \hat{t}_{y\pi} - (t_z - t_{z\pi})'\hat{\boldsymbol{\beta}}_s$.

We need only $\sum_{k\in U} \mathbf{z}_k$.

Nonparametric estimation by regression B-spline

- Suppose we know z_k for all $k \in U$.
  How can we use this supplementary information ?
- We can only suppose that f is a smooth function (differentiable) without a specific parametric expression.
- Breidt & Opsomer (2000, 2005) propose a class of estimators based on local polynomial regression and respectively, on penalised splines.
- I propose a B-spline approach. (The Canadian Journal of Statistics, 2005)

B-spline regression estimator

- The set of spline functions of degree m ($m \geq 2$) with K interiors equidistant knots

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

$$S_{K,m} = \{s \in C^{m-2}[0,1] : s(x) \quad \text{is a polynomial of degree} \quad m-1 \quad \text{sur} \quad (\xi_j, \xi_{j+1})\}.$$

- $\{B_1(\cdot), \ldots, B_q(\cdot)\}$ form a basis for $S_{K,m}$ of dimension $q = K + m$ (Schumaker 1981, Dieckx 1993)
  $B_1(\cdot), \ldots, B_q(\cdot)$ B-splines ;

1. $0 \leq B_j(\cdot) \leq 1$, $\sum_{j=1}^{q} B_j(\cdot) = 1$.
2. $\mathbf{B}_U = (B_j(z_k))_{k \in U, j=1,\ldots,q} = (\mathbf{b}'(z_k))_{k \in U}$

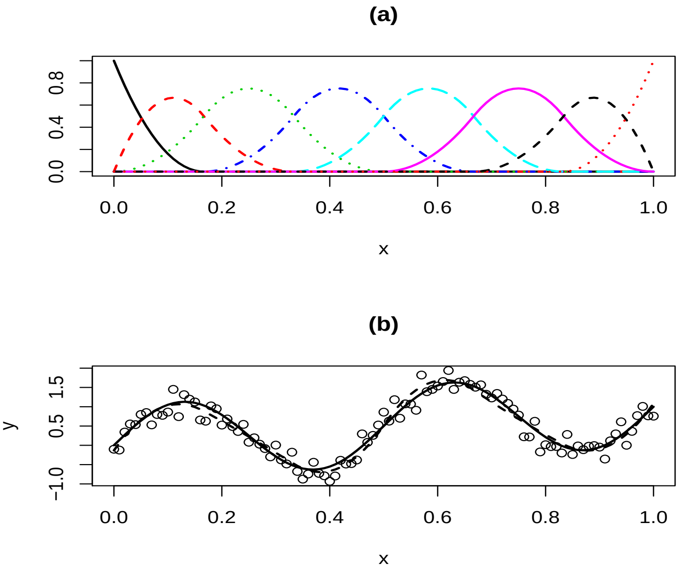

FIG.: Growth curves.

- **under the model** ξ,

  $\hat{f}(t) = \sum_{j=1}^{q} \hat{\theta}_j B_j(t)$ with

  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)$ is obtained by least squares

$$\hat{\boldsymbol{\theta}} = \operatorname{Arg}\min_{\boldsymbol{\theta} \in \mathbb{R}^q} \sum_{k=1}^{N} \left(y_k - \sum_{j=1}^{q} \theta_j B_j(z_k) \right)^2 .$$

$$\begin{cases} \hat{f}_k & = & \mathbf{b}'(z_k)\hat{\boldsymbol{\theta}} \quad k \in U \\ \hat{\boldsymbol{\theta}} & = & (\mathbf{B}'_U \mathbf{B}_U)^{-1} \mathbf{B}'_U \mathbf{y}_U \\ & = & \left(\sum_{i \in U} \mathbf{b}(z_i)\mathbf{b}'(z_i) \right)^{-1} \sum_{i \in U} \mathbf{b}(z_i) y_i \end{cases}$$

- **under the sampling design,** $p(\cdot)$:
  $\hat{f}(z_k)$ is estimated by substitution of each total with the HT estimator :

$$\begin{aligned}
\hat{\hat{f}}(z_k) &= \mathbf{b}'(z_k)\hat{\hat{\theta}} \\
&= \mathbf{b}'(z_k)(\mathbf{B}'_s\mathbf{\Pi}_s^{-1}\mathbf{B}_s)^{-1}\mathbf{B}'_s\mathbf{\Pi}_s^{-1}\mathbf{y}_s \quad k \in U \\
&= \mathbf{b}'(z_k)\left(\sum_{i\in s}\frac{\mathbf{b}(z_i)\mathbf{b}'(z_i)}{\pi_i}\right)^{-1}\left(\sum_{i\in s}\frac{\mathbf{b}(z_i)y_i}{\pi_i}\right)
\end{aligned}$$

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

The B-spline estimator for the total

We replace the $\hat{\hat{f}}(z_k)$ in $\hat{t}_{diff}$ for obtaining the estimator for t_y :

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

- **Population fit residuals** $E_k = y_k - \hat{f}(x_k)$ for all $k \in U$ satisfy

$$\sum_U E_k = 0.$$

- **Sample fit residuals** $e_k = y_k - \hat{\hat{f}}(x_k)$ for all $k \in U$ satisfy

$$\sum_s \frac{e_k}{\pi_k} = 0.$$

Properties of $\hat{t}_{BS}$

- $\hat{t}_{BS}$ is the population total of $\hat{\hat{f}}(x_k)$,

$$\hat{t}_{BS} = \sum_{k \in U} \hat{\hat{f}}(z_k) = \sum_{k \in s} w_{ks} y_k$$

  with

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

- the weights w_{ks} contain the auxiliary information and they not depend on the variable of interest ; as a consequence, they may used for estimating the population total of another variable of interest.

- **Calibration** (Deville & Särndal 1992) : Suppose that $\sum_{k\in U} B_j(z_k)$ are known. The weights w_{ks} satisfy the calibrating equation on the B-splines :

$$\sum_s w_{ks} B_j(z_l) = \sum_U B_j(z_k) \quad j = 1, \ldots, q$$

- $\hat{t}_{BS} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{b}'(x_k)}{\pi_k} - \sum_U \mathbf{b}'(x_k) \right) \hat{\hat{\theta}}$ is a kind of GREG-estimator for the auxiliary information vector $\mathbf{z}_k = \mathbf{b}'(x_i)$ of dimension $K + m$.
- **Poststratification** : $U = \bigcup_{h=1}^{H} U_h$ and $s \subset U$ a SRSwr sample "stratified", $s = \bigcup_{h=1}^{H} s_h$
for $m = 1$ and knots at the strata bounds, we obtain the poststratified estimator $\hat{t}_{BS} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{s_h} y_k$;

Model - Robustness of $\hat{t}_{BS}$

We verify that $\hat{t}_{BS}$ is (Särndal 1980, Särndal & Robinson 1982) :

- asymptotically design unbiased (ADU) and consistent (ADC) for t_y;

$$\lim_{N\to\infty} \frac{1}{N} E_p(\hat{t}_{BS} - t_y) = 0$$

$$\varepsilon > 0, \quad \lim_{N\to\infty} Pr(\frac{1}{N} \mid \hat{t}_{BS} - t_y \mid > \varepsilon) = 0$$

- robust (Godambe 1982) : the estimator reaches asymptotically the Godambe-Joshi lower bound (1965) :

$$E_\xi E_p(\frac{1}{N}(\hat{t}_{BS} - t_y))^2 \simeq \frac{1}{N^2} \sum_U v(x_k) \frac{1-\pi_k}{\pi_k}$$

Population :

- $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0,1)$.
- $\limsup_{N\to\infty} \frac{1}{N} \sum_{k\in U} y_k^2 < \infty$ with ξ probability 1.
- $\sup_{z\in[0,1]} |Q_N(z) - Q(z)| = o(K^{-1})$.

Sampling design : $\min_{k\in U} \pi_k \geq \lambda > 0, \min_{i,k\in U} \pi_{ik} \geq \lambda* > 0,$

$$\overline{\lim_{N\to\infty}}\, n \max_{i\neq k\in U} |\pi_{ik} - \pi_i\pi_k| < \infty.$$

Result

Under the above assumptions and for $K = o(N)$, $K = o(\sqrt{n})$:

$$N^{-1}E_p|\hat{t}_{BS} - \hat{t}_{HT}| = O(n^{-1/2})$$

for $\hat{t}_{HT} = \sum_s \frac{y_k}{\pi_k}$. *It results then*

- $\hat{t}_{BS}$ *is ADU and ADC,*
- $N^{-1}(\hat{t}_{BS} - \hat{t}_{HT}) = O_p(n^{-1/2})$.

We have also, $N^{-1}(\hat{t}_{BS} - t_y) = O_p(n^{-1/2})$.

Asymptotic properties of $\hat{t}_{BS}$ under the design p : the variance

Result

Under the above assumptions and for $K = o(N)$, $K = o(\sqrt{n})$:

$$n^{1/2}N^{-1}(\hat{t}_{BS} - t_y) = n^{1/2}N^{-1}(\hat{t}_y - t_y) + o_p(1)$$

for

$$\hat{t}_y = \sum_s \frac{y_k - \hat{f}_k}{\pi_k} + \sum_U \hat{f}_k$$

Consequence :

$$Var_p(\frac{1}{N}(\hat{t}_{BS} - t_y)) \simeq \frac{1}{N^2} \sum_{k \in U} \sum_{i \in U} \Delta_{ki} \frac{y_k - \hat{f}(z_k)}{\pi_k} \frac{y_i - \hat{f}(z_i)}{\pi_i}.$$

Asymptotic framework : assumptions 2

Population

- $\limsup_{N\to\infty} \frac{1}{N} \sum_{k\in U} \varepsilon_k^2 < \infty$ with ξ probability 1.
- the noise variance $v(\cdot)$ is bounded : $\sup_{k\in U} v(z_k) < \infty$.

Regularity of f : f is m-times continuously differentiable in $[0,1]$.

Asymptotic properties of $\hat{t}_{BS}$ under the design p and the model ξ

Result

Under the above assumptions and for $K = o(N)$, $K = o(n^{1/2})$

- $\hat{t}_{BS}$ *est asymptotically pξ-unbiased and*
- $\hat{t}_{BS}$ *is robust :*

$$E_\xi E_p \left(\frac{1}{N} (\hat{t}_{BS} - t_y) \right)^2 = \frac{1}{N^2} \sum_{k \in U} v(z_k) \frac{1 - \pi_k}{\pi_k} + o(1)$$

- *the anticipated variance is minimum for* $\pi_k \propto v(z_k)^{1/2}$

$$E_\xi E_p \left(\frac{1}{N} (\hat{t}_{BS} - t_y) \right)^2 \simeq \frac{1}{nN^2} \left[\left(\sum_U \sqrt{v(z_k)} \right)^2 - n \sum_U v(z_k) \right]$$

A simulation study

- Population $\mathcal{U}$, $N = 1000$.
- $y_k = f(x_k) + \epsilon_k, \quad \epsilon \sim N(0, \sigma)$
  $x \in [0, 1]$, uniform distribution .
- 3 different functions f
  $f_{lin}(x) = 1 + 2(x - 0.5)$,
  $f_{\exp}(x) = \exp(-8x)$,
  $f_{\sin}(x) = 2 + \sin(2\pi x)$
- Simple random sampling without replacement of size $n = 100$, $\pi_k = n/N$.
- Splines with 5 interiors knots at the population quantile and $m = 3$.

MSE	f	$\hat{t}_{HT}$	$\hat{t}_{GREG}$	$\hat{t}_{BS}$
$\sigma = 0.1$	f_{lin}	2980	**94**	99
	$f_{\exp}$	513	281	**100**
	$f_{\sin}$	4706	1835	**102**
$\sigma = 0.4$	f_{lin}	4504	**1515**	1633
	$f_{\exp}$	1788	1638	**1552**
	$f_{\sin}$	5476	3103	**1565**

- for f linear, $\hat{t}_{BS}$ is almost apresque aussi bon que $\hat{t}_{GREG}$;
- for f nonlinear, $\hat{t}_{BS}$ has a better behaviour ;
- $\hat{t}_{HT}$ conducts not very well ;

$MSE(\hat{\theta}) = \frac{1}{b}\sum_{r=1}^{b}(\hat{\theta}_r - \theta)^2$ for b simulations.

The estimation of the empirical distribution function and of quantiles in presence of auxiliary information

The empirical distribution function (edf)

$$F_Y(t) = \frac{1}{N} \sum_U I_{\{y_k \le t\}}$$

and the α-th quantile :

$$q_\alpha = \inf\{t : F_Y(t) \ge \alpha\}$$

Method : we derive $\hat{F}_Y(t)$ and then $\hat{q}_\alpha = \inf\{t : \hat{F}_Y(t) \ge \alpha\}$
Without auxiliary information and N known : the Horvitz-Thompson estimator

$$\hat{F}_{HT,Y}(t) = (1/N) \sum_s \frac{I_{\{y_k \le t\}}}{\pi_k}$$

The estimation of the empirical distribution function using the B-splines approach

- Supposing that N is known, we propose (Aragon, Goga & Ruiz, 2005) the following estimator

$$\hat{F}_{BS}(t) = \frac{1}{N}\sum_{s} w_{ks} I_{\{y_k \leq t\}}$$

with w_{ks} independent of $I_{\{y_k \leq t\}}$ and given by

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

- If N is unknown, then $F_Y(t)$ is a nonlinear parameter.

A simulation study

- The population $\mathcal{U}$, $N = 1000$.
- $y_k = f(x_k) + v^{1/2}(z_k)\epsilon_k, \quad \epsilon \sim N(0, \sigma)$ and $\sigma = 0.2$
  $x \in [0, 1]$, the uniform distribution.
- 4 different functions f (Breidt & Opsomer, 2000)
  $f_{lin}(x) = 1 + 2(x - 0.5)$,
  $f_{\exp}(x) = 0.6 + \exp(-8x)$,
  $f_{bump}(x) = 1.5 + 2(x - 0.5) + \exp(-200(x - 0.5)^2)$
  $f_{jump}(x) = 1.5 + \left(O.35 + 2(x - 0.5)^2\right) I_{\{x \leq 0.65\}}$
- Simple random sampling without replacement of size $n = 100$, $\pi_k = n/N$.
- Splines with 5 interiors knots at the population quantile and $m = 3$.

MSE/MSE_{HT}	f	RKM_{ratio}	RA_{diff}	$BS(3)$	$Postrat$
q_{25}	f_{lin}	.51	.39	.38	.40
	f_{exp}	17.7	1	.97	.97
	f_{bump}	2.38	.38	**.34**	.38
	f_{jump}	21.4	.66	**.56**	.57
q_{50}	f_{lin}	.41	.37	.24	.25
	f_{exp}	8.02	.93	**.83**	.83
	f_{bump}	.66	.40	**.20**	.23
	f_{jump}	9.54	.87	**.58**	.60
q_{75}	f_{lin}	.38	.40	.31	.35
	f_{exp}	3.56	.96	**.50**	.52
	f_{bump}	2.72	.76	**.59**	.65
	f_{jump}	5.47	.98	**.61**	.62

The estimation of a nonlinear parameter of population totals with auxiliary information (in work with Anne Ruiz-Gazen)

Let us consider θ a nonlinear parameter.
Linearization by the influence function approach

• Write $\theta = T(M)$ with T a homogeneous functional of degree α and $M = \sum_U \delta_{y_k}$;

• Use the estimator $\hat{\theta} = T(\hat{M})$ with

$$\hat{M} = \sum_s w_{ks} \delta_{y_k}$$

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

Variance asymptotique

Result

Let us consider the influence function defined as follows :

$$IT(M,y) = \lim_{\varepsilon \to \infty} \frac{T(M+\varepsilon\delta_y) - T(M)}{\varepsilon}.$$

Under broad assumptions we have

$$\begin{aligned} \sqrt{n}N^{-\alpha}(T(\hat{M}) - T(M)) &= \sqrt{n}N^{-\alpha}\int IT(M,y)d(\hat{M}-M) + o_p(1) \\ &= \sqrt{n}N^{-\alpha}(\sum_s w_{ks}u_k - \sum_U u_k) + o_p(1) \end{aligned}$$

with $u_k = IT(M, y_k)$ the linearized variables.

Example : $F_Y(t) = \frac{1}{N}\sum_U I_{\{y_k \le t\}}$ is estimated by

$$\hat{F}_Y(t) = \frac{\sum_s w_{ks} I_{\{y_k \le t\}}}{\sum_s w_{ks}}.$$

Consequence

$$\frac{\sqrt{n}}{N^{2\alpha}} Var_p(T(\hat{M}) - T(M)) \simeq \frac{\sqrt{n}}{N^{2\alpha}} Var_p(\sum_s w_{ks} u_k - \sum_U u_k)$$

$$\simeq \frac{\sqrt{n}}{N^{2\alpha}} \sum_{k \in U} \sum_{i \in U} \Delta_{ki} \frac{u_k - \hat{f}(z_k)}{\pi_k} \frac{u_i - \hat{f}(z_i)}{\pi_i}$$

$$\hat{f}_u(z_k) = \mathbf{b}'(z_k)(\mathbf{B}'_U \mathbf{B}_U)^{-1} \mathbf{B}'_U \mathbf{u}_U$$

RB	f	$\hat{t}_{HT}$	$\hat{t}_{BS}$
$\sigma = 0.2$	f_{exp}	-0.04	-0.0005
	f_{lin}	-0.0016	-0.0009
	$f_{\sin}$	0.0005	0.0001
$\sigma = 0.4$	f_{exp}	0.002	0.0005
	f_{lin}	0.0023	0.0002
	$f_{\sin}$	0.0008	0.0001

pour $RB(\hat{\theta}) = \frac{E(\hat{\theta})-\theta}{\theta}$ le biais relatif ;

RMSE	f	$\hat{t}_{HT}$	$\hat{t}_{BS}$
$\sigma = 0.2$	f_{exp}	0.0016	0.0008
	f_{lin}	0.0033	0.0003
	$f_{\sin}$	0.0023	0.0002
$\sigma = 0.4$	f_{exp}	0.0039	0.0034
	f_{lin}	0.0045	0.0014
	$f_{\sin}$	0.003	0.0008

pour $RMSE = \frac{MSE(\hat{\theta})}{\theta}$ l'erreur quadratique moyenne relative.

Conclusion

A simple method for taking into account the auxiliary information.

Further questions :

1. Choice of smoothing parameter K,
2. Extension to multivariate variable $\mathcal{Z}$.