Nonparametrics for complex parameters estimation in finite population. Application to the estimation of the Gini index with survey data.

Goga, Camelia
Université de Bourgogne, IMB UMR CNRS 5584
9 Avenue Alain Savary
Dijon (21000), France
camelia.goga@u-bourgogne.fr

Ruiz-Gazen, Anne
Université Toulouse 1 Capitole, TSE
21 Allée de Brienne
Toulouse (31000), France
ruiz@cict.fr

1 Introduction

Estimating complex parameters such as Gini indices or other measures of inequality with good precision is particularly important nowadays. In the European Statistics on Income and Living Conditions (EU-SILC) survey, several indicators for studying social inequalities and poverty are considered including the Gini index, the at-risk-of-poverty rate, the quintile share ratio and the low-income proportion. Estimating totals or means by taking into account auxiliary information has been extensively studied in a finite population context. In the model-assisted approach, both linear and non-parametric models have been explored and non-parametric models have shown their superiority in terms of precision if the linear model fitting the study variable using the auxiliary variable is misspecified. However, there is much less literature on complex parameters (see Harms and Duchesne, 2006, for quantiles and Berger and Skinner, 2003, for the low-income proportion).

In the present paper, we consider complete univariate quantitative auxiliary information and we introduce a general class of complex parameter estimators with weights derived using a nonparametric model-assisted approach. These weights do not depend on the study variable and can be used for any other survey variable but also for any survey parameter to estimate. Having a unique system of weights is very important in multipurpose surveys such as the EU-SILC. Using these weights, we define a class of substitution estimators for complex parameters and we derive their asymptotic variance in a general context using the influence function approach by Deville (1999). Interestingly, the asymptotic variance and consequently the precision of the proposed estimators depend on the residuals from the fitted values of the linearized variable of the parameter of interest.

Because linearized variables may be quite complex, linear models are unlikely to perform well and are outperformed by non-parametric models even if the study variable is linearly related with the auxiliary one. The theory developed in Goga and Ruiz-Gazen (2011) is presented using a general non-parametric framework. Many details are given for B-spline estimators of the Gini index. Furthermore, the asymptotic and finite-sample performances of the proposed estimators are illustrated using two real data sets. Specifically, point and confidence intervals estimation of the Gini index are derived for measurements of television audience.

In section 2, we briefly recall the nonparametric model-assisted estimator for finite population totals and in section 3, we generalize the method for complex parameters. Section 4 gives the B-spline estimation for the Gini index and section 5 presents the empirical studies.

2 Nonparametric model-assisted estimation of totals

Consider a finite population U of N elements labeled $k = 1, \dots, N$. Let y_k (resp. z_k), the value of the study (resp. auxiliary) variable $\mathcal{Y}$ (resp. $\mathcal{Z}$) for the kth population element. The values $z_1, \dots, z_N$ are assumed to be known for the entire population (i.e., complete information). In this section, the parameter to estimate is the finite population total $t_y = \sum_U y_k$. A sample s is selected from U according to a sampling design $p(\cdot)$ of fixed size n. Many approaches can be used to take into account auxiliary information $\mathcal{Z}$ and thus improve on the Horvitz-Thompson estimator $\hat{t}_{y,HT} = \sum_s y_k/\pi_k$. Note that $\pi_k = Pr(k \in s) > 0$ are the first-order inclusion probabilities. The goal is to derive a weighted linear estimator $\hat{t}_{wy} = \sum_s w_{ks} y_k$ of t_y, such that the sample weights w_{ks} do not depend on the study variable values y_k but include the values z_k, for all $k \in U$. Among the different methods for deriving the w_{ks}, we focus on the model-assisted approach. The construction of the model-assisted (MA) class of estimators $\hat{t}_{wy}$ is based on a superpopulation model ξ:

$$(1) \quad \xi : \quad y_k = f(z_k) + \varepsilon_k$$

where the ε_k are independent random variables with mean zero and variance $v(z_k)$. Recently, Breidt & Opsomer (2000) proposed local linear estimators and Breidt *et al.* (2005) and Goga (2005) used nonparametric spline regression for estimating the total t_y. Let $\hat{f}_{y,k}$ be the estimator of $f(z_k)$ obtained using one of the three nonparametric methods mentioned above. The nonparametric generalized difference estimator of the finite population total is:

$$(2) \quad \hat{t}_{y,\text{diff}} = \sum_s \frac{y_k - \hat{f}_{y,k}}{\pi_k} + \sum_U \hat{f}_{y,k}.$$

This estimator is still design unbiased but it is asymptotically model unbiased because nonparametric estimators $\hat{f}_{y,k}$ are always biased for f_k. The estimators $\hat{f}_{y,k}$ are usually obtained by a least square method (weighted, penalized or ordinary) and in general, we write

$$(3) \quad \hat{f}_{y,k} = \mathbf{G}'_k \mathbf{y}_U, \quad \text{for all } k \in U$$

where the vector $\mathbf{G}_k$ depends on the population values z_k, for all $k \in U$ but does not depend on $\mathcal{Y}$. As in the parametric case, we estimate $\hat{f}_{y,k}$ by $\tilde{f}_{y,k}$ using the sampling design,

$$(4) \quad \tilde{f}_{y,k} = \widehat{\mathbf{G}}'_{ks} \mathbf{y}_s, \quad \text{for all } k \in U$$

where $\widehat{\mathbf{G}}'_{ks}$ is a design-based estimator of $\mathbf{G}'_k$ and $\mathbf{y}_s = (y_k)_{k \in s}$ is the vector of sample values of $\mathcal{Y}$. Plugging $\tilde{f}_{y,k}$ into (2) yields the following nonparametric model-assisted estimator for t_y,

$$(5) \quad \hat{t}_{y,np} = \sum_s \frac{y_k - \tilde{f}_{y,k}}{\pi_k} + \sum_U \tilde{f}_{y,k}.$$

Nonparametric model-assisted estimators (NMA) can be written as weighted sums of the sampled observations

$$(6) \quad \hat{t}_{y,np} = \sum_s w_{ks} y_k$$

where the weights depend only on the sample and on the auxiliary information. The expression of w_{ks} depends on the nonparametric method chosen, as discussed in Breidt and Opsomer (2000), Breidt *et al.* (2005) and Goga (2005). Under mild hypothesis, the variance of $\hat{t}_{y,np}$ may be approximated by the variance of $\hat{t}_{y,\text{diff}}$. This result states that all the NMA estimators are bias robust, regardless of whether the model is valid. Besides, they bring an improvement over parametric methods and the Horvitz-Thompson estimator when the relation between $\mathcal{Y}$ and $\mathcal{Z}$ is not linear. In the latter, the residuals $y_k - \hat{f}_{y,k}$ will be smaller than under a parametric smoother, which explains the diminution of the design variance of NMA estimators.

3 Nonparametric model-assisted estimation of complex parameters

Let us consider the estimation of some nonlinear parameters Φ by taking into account complete auxiliary information $\mathcal{Z}$. Examples of nonlinear parameter of interest Φ are the ratio, the empirical distribution function or the Gini coefficient. A parameter Φ may depend on one or several variables of interest but we consider a single auxiliary variable $\mathcal{Z}$. As such, we aim to provide a general method for estimating Φ using $\mathcal{Z}$ by considering the functional approach introduced by Deville (1999). The methodology consists in writing Φ as a functional T of a discrete and finite measure $M = \sum_U \delta_{y_k}$ such that there is unity mass on each point y_k, $k \in U$ and zero mass elsewhere, $\Phi = T(M)$. A substitution estimator of Φ is a functional T of a random measure $\hat{M}$ that takes into account the sampling weights w_{ks}. Deville (1999) suggests using the Horvitz-Thompson weights $w_{ks} = 1/\pi_k$ or more generally, calibration weights.

We suggest a simple method that consists in using the nonparametric weights w_{ks} provided by (6) and defining $\widehat{M}_{np} = \sum_s w_{ks}\delta_{y_k}$ and the nonparametric substitution estimator $\widehat{\Phi}_{np} = T(\widehat{M}_{np})$. The weights w_{ks} derived in (6) to estimate the total t_y can be used to estimate other finite population totals. It can also be used to estimate any nonlinear parameter of interest Φ as soon as it can be expressed as a functional of M.

Let u_k be the linearized variables of Φ, for all $k \in U$ and let ξ' be the nonparametric model for the linearized variable

$$\xi' : \quad u_k = g(z_k) + \eta_k$$

where g is supposed to be a smooth function. An estimator of g is obtained by using the same nonparametric method employed for estimating f from the model ξ. This means that the same vectors $\mathbf{G}_k$ and $\hat{\mathbf{G}}_{ks}$ from (3) and (4) are used to derive estimators of g. More precisely, let us denote $\hat{g}_{u,k} = \mathbf{G}_k'\mathbf{u}_U$ as the estimator of $g(z_k)$ under the model ξ', where $\mathbf{u}_U = (u_k)_{k\in U}$ and $\tilde{g}_{u,k} = \widehat{\mathbf{G}}_{ks}'\mathbf{u}_s$, where $\mathbf{u}_s = (u_k)_{k\in s}$ is the sample restriction of $\mathbf{u}_U$. Unlike the linear case, $\tilde{g}_{u,k}$ is not an estimate of $\hat{g}_{u,k}$ since the sample linearized variable vector $\mathbf{u}_s$ is not known. Plugging $\hat{g}_{u,k}$ (resp. $\tilde{g}_{u,k}$) into (2) (resp. (5)) yields the nonparametric difference estimator $\hat{t}_{u,\mathrm{diff}}$ (resp. the NMA estimator $\hat{t}_{u,np}$),

$$(7) \quad \hat{t}_{u,\mathrm{diff}} = \sum_s \frac{u_k - \hat{g}_{u,k}}{\pi_k} + \sum_U \hat{g}_{u,k}$$

$$(8) \quad \hat{t}_{u,np} = \sum_s \frac{u_k - \tilde{g}_{u,k}}{\pi_k} + \sum_U \tilde{g}_{u,k}$$

The following theorem (Goga and Ruiz-Gazen, 2011) shows that the nonparametric estimator $\widehat{\Phi}_{np}$ is approximated by the nonparametric difference estimator for the population total of the linearized variable.

Theorem 1 *We suppose that the parameter $\Phi = T(M)$ has degree α, that is $T(rM) = r^\alpha T(M)$ and $\lim_{N\to\infty} N^{-\alpha}T(M) < \infty$. Under general assumptions (Goga and Ruiz-Gazen, 2011), the nonparametric substitution estimator $\widehat{\Phi}_{np}$ fulfills*

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

Furthermore, if the asymptotic distribution of $\sqrt{n}N^{-\alpha}\left(\hat{t}_{u,\mathit{diff}} - t_u\right)$ is normal with mean zero and asymptotic variance

$$\frac{n}{N^{2\alpha}} \sum_U \sum_U \Delta_{kl} \frac{u_k - \hat{g}_{u,k}}{\pi_k} \frac{u_l - \hat{g}_{u,l}}{\pi_l}$$

then the asymptotic distribution of $\sqrt{n}N^{-\alpha}\left(\widehat{\Phi}_{np} - \Phi\right)$ is normal with mean zero and the same asymptotic variance.

Variance estimation and other properties of $\widehat{\Phi}_{np}$ are extensively discussed in Goga and Ruiz-Gazen (2011).

4 B-spline estimation of the Gini index

Spline functions have many attractive properties, and they are often used in practice because of their good numerical features and their easy implementation. Consider the superpopulation model ξ given by (1) where f is a smooth function. We suppose without loss of generality that all z_k have been normalized and lie in $[0,1]$. The set of spline functions of order m, $m \geq 2$ with K interiors knots $0 = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = 1$ is the set of C^{m-2} continuously differentiable functions on $[0,1]$. Note that these functions are piecewise polynomials of degree $m-1$ on the intervals between knots. For each fixed set of knots, $S_{K,m}$ is a linear space of functions of dimension $q = K + m$. A basis for this linear space is provided by B-spline functions $B_1, \ldots, B_q$ (Dierckx, 1993).
Let $\boldsymbol{B}_U$ be the $N \times q$ matrix having the vectors $\boldsymbol{b'}(z_k) = (B_1(z_k), \ldots, B_q(z_k))$, $k \in U$, as rows and $\boldsymbol{B}_s$ the sample restriction. Let $\boldsymbol{\Pi}_s$ be the $n \times n$ diagonal matrix with π_k, $k \in s$, along the diagonal. The design-based estimators of f are $\tilde{f}_{y,k} = \widehat{\mathbf{G}}'_{ks}\boldsymbol{y}_s$ where $\widehat{\mathbf{G}}'_{ks} = \boldsymbol{b'}(z_k)(\boldsymbol{B}'_s\boldsymbol{\Pi}_s^{-1}\boldsymbol{B}_s)^{-1}\boldsymbol{B}'_s\boldsymbol{\Pi}_s^{-1}$ and the B-spline NMA estimator of t_y is as follows

$$(9) \quad \hat{t}_{BS,y} \;=\; \sum_s \frac{y_k - \tilde{f}_{y,k}}{\pi_k} + \sum_U \tilde{f}_{y,k}.$$

The B-spline functions have the attractive property that $\sum_{j=1}^{q} B_j(x) = 1$ for all $x \in [0,1]$. As a consequence (Goga, 2005), $\hat{t}_{BS,y}$ is equal to the finite population total of the prediction $\tilde{f}_{y,k}$, $\hat{t}_{BS,y} = \sum_s w_{ks} y_k$ where

$$(10) \quad w_{ks} = \frac{1}{\pi_k}\left(\sum_U \boldsymbol{b'}(z_i)\right)\left(\sum_s \frac{\boldsymbol{b}(z_i)\boldsymbol{b'}(z_i)}{\pi_i}\right)^{-1} \boldsymbol{b}(z_k).$$

Consider now the complex parameter given by the Gini index,

$$\text{Gini} = \frac{\sum_U y_k\,(2F(y_k) - 1)}{t_y} = \frac{\int (2F(y) - 1) y dM(y)}{\int y dM(y)}$$

where $F(y) = \int \mathbf{1}_{\{\xi \leq y\}} dM(\xi) / \int dM(y) = \sum_U \mathbf{1}_{\{y_k \leq y\}}/N$ is the empirical distribution function. The nonparametric estimator for Gini is obtained by simply replacing M with $\widehat{M}_{np}$. Hence,

$$\widehat{\text{Gini}}_{np} = \frac{\sum_s w_{ks}(2\hat{F}_{np}(y_k) - 1)y_k}{\sum_s w_{ks} y_k}$$

where $\hat{F}_{np}(y) = \dfrac{\int \mathbf{1}_{\{\xi \leq y\}} d\hat{M}_{np}(\xi)}{\int d\hat{M}_{np}(y)} = \dfrac{\sum_s w_{ks}\mathbf{1}_{\{y_k \leq y\}}}{\sum_s w_{ks}}$ with w_{ks} given by (10). The linearized variable u_k of Gini is given by

$$u_k = 2F(y_k)\frac{y_k - \bar{y}_{k,<}}{t_y} - y_k \frac{\text{Gini} + 1}{t_y} + \frac{1 - \text{Gini}}{N}$$

where $\bar{y}_{k,<}$ denotes the mean of the y_j lower than y_k. We can remark from the above expression that the linearized variable u_k has a rather complicated expression. For the French Labour Force data used in the empirical study, the linearized variable of the Gini index computed for the wage variable exhibits a clearly nonlinear relationship versus the auxiliary variable given by the wage variable from the year before. This is why, fitting a nonparametric model on the linearized variable u_k is strongly justified.

The Gini index is a parameter of degree zero, so $\alpha = 0$. The nonparametric estimator of the Gini index may be approximated by

$$\widehat{\text{Gini}}_{np} - \text{Gini} = \hat{t}_{u,diff} - t_u + o_p(n^{-1/2})$$

where $\hat{t}_{u,\text{diff}}$ is given by (7) and $\hat{g}_{u,k} = \boldsymbol{b'}(z_k)(\boldsymbol{B}'_U\boldsymbol{B}_U)^{-1}\boldsymbol{B}'_U\boldsymbol{u}_U$. The variance of $\widehat{\text{Gini}}_{np}$ may be approximated by the variance of $\hat{t}_{u,diff}$ which is equal to

$$\sum_U\sum_U \Delta_{kl}\frac{u_k - \hat{g}_{u,k}}{\pi_k}\frac{u_l - \hat{g}_{u,l}}{\pi_l}.$$

5 Empirical results

In this section, we consider two data sets, with one study variable and one auxiliary variable. The first data set is from the French Labour Force surveys of 1999 and 2000; it consists in the yearly wages of 22,741 wage-earners who were sampled in both years. The second data set consists of television audience measurements (i.e., the amount of television viewed in minutes) of 6,658 persons for a particular channel during two consecutive Mondays in September 2010. These data are confidential and are from the French audience measurement company Médiamétrie. Both data sets are considered the finite populations of interest.

The employment data set is used in order to compare asymptotic variances of several estimators, including the non-parametric estimators we propose for different complex parameters of interest. We estimate the mean, the median, the Gini index and the poverty rate for the wages in 2000 using the wages in 1999 as auxiliary information. The poverty rate is the proportion of persons whose wages are below the threshold of 60% of the median wage. For each parameter (in column), the scatter plots in Figure 1 show the relationship between the linearized variable and the auxiliary variable. For the mean, the linearized variable is the study variable itself and it is clear from the first plot that a linear model fits this relationship well. However, this is no longer the case for complex parameters such as the median, the Gini index or the poverty rate.

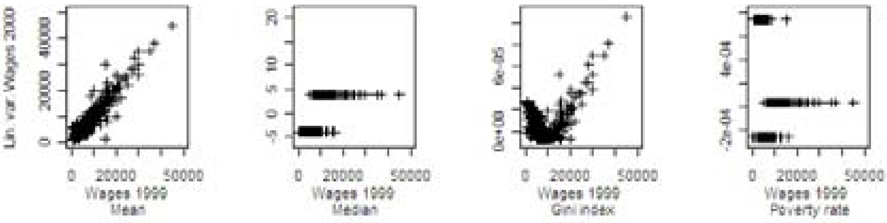

Figure 1: Linearized variables versus the wages in 1999

We focus on a simple random sampling without replacement and consider several estimators for each parameter such as the Horvitz-Thompson estimator, the poststratified estimator with six strata bounded at the empirical quantiles for 1999 wages, the GREG estimator, which takes into account the 1999 wages as auxiliary information using a simple linear model, the calibrated estimator proposed by Harms and Duchesne (2006), a multivariate GREG estimator that incorporates various auxiliary variables, including the constant and finally, the B-spline estimators, which take into account the wages from 1999 as auxiliary information by using a nonparametric model with $K = 5$ knots located at the quantiles of the empirical distribution for wages from 1999 and for different orders $m = 2, 3, 4$. During the oral presentation, the comparison results will be presented. For all parameters, results are very stable for different B-spline orders, and almost all the results favor the B-spline estimators

We use the second data set for simulation studies to investigate the finite-sample performance of the proposed nonparametric estimators. Recently, the Médiamétrie company focus on the estimation of Gini concentration measures for different television channels together with confidence intervals, taking into account past auxiliary information. In the present study, we focus on one particular channel. We focus on the estimation of the Gini index for the audience viewing duration on a given Monday by taking into account the audience viewing duration of the previous Monday for the same channel. We look at the finite-sample properties of the proposed estimators. These data are quite challenging because they contain many zeros and ties. Simulation results concerning relative bias, ratio of root mean squared errors and coverage probabilities will be presented during the presentation. Again, the results are significantly better for the B-spline approach.

REFERENCES (RÉFERENCES)

Berger, Y. G. and Skinner, C. J. (2003), Variance estimation for a low income proportion, Applied Statistics, **52**, 457-468.

Breidt, F. J. and Opsomer, J. (2000), Local Polynomial Regression Estimators in Survey Sampling, The Annals of Statistics, **28**, 1026-1053.

Breidt, F. J., Claeskens G. and Opsomer, J. (2005), Model-assisted estimation for complex surveys using penalised splines, Biometrika, **92**, 831-846.

Deville, J. C. (1999), Variance estimation for complex statistics and estimators: linearization and residual techniques, Survey Methodology, **25**, 193-203.

Dierckx, P. (1993), Curves and Surface Fitting with Splines, Clarendon Press, Oxford, United Kingdom.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles, Survey Methodology, **32**, 37-52.

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, A. (2011), Nonparametrics for complex parameters estimation in finite population, submitted.