On the linearization of inequality indexes in the design-based framework

Sulla linearizzazione di indici di diseguaglianza nell'approccio da disegno

Lucio Barabesi and Giancarlo Diana and Pier Francesco Perri

Abstract Linearization methods are customarily adopted in sampling surveys to obtain approximated variance formulae for estimators of nonlinear functions of finite-population totals which can be usually rephrased in terms of statistical functionals. In the present paper, by considering Deville's (1999) approach stemming on the concept of design-based influence curve, we provide a general result for linearizing large families of inequality indexes. As an example, the achievement is applied to the Gini and the Amato indexes. We also discuss the case when income data are supposed to be collected by means of the randomized response technique.

Abstract *I metodi di linearizzazione sono tradizionalmente impiegati in indagini campionarie al fine di ottenere approssimazioni per la varianza di stimatori di funzioni non lineari di totali. Tali funzioni possono essere spesso riformulate in termini di funzionali statistici. Nel presente lavoro, adottando l'approccio basato sulle curve di influenza nel disegno campionario proposto da Deville (1999), si fornisce un risultato generale per linearizzare famiglie di indici di disuguaglianza. Il risultato viene esemplificato mediante l'applicazione agli indici di Gini e di Amato. Infine, si considera la situazione in cui la variabile reddito venga rilevata attraverso il metodo delle risposte casualizzate.*

Key words: Design-based inference, Substitution estimator, Variance estimation, Influence function, Gini index, Amato index, Randomized response technique

---

Lucio Barabesi
Department of Economics and Statistics, University of Siena, Piazza San Francesco 7, 53100, Siena (Italy), e-mail: lucio.barabesi@unisi.it

Giancarlo Diana
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova (Italy) e-mail: diana@stat.unipd.it

Pier Francesco Perri
Department of Economics, Statistics and Finance, University of Calabria, Via P. Bucci, 87036, Arcavacata di Rende (Italy) e-mail: pierfrancesco.perri@unical.it

1 Introduction

Sampling surveys frequently involve the estimation of target parameters which are nonlinear functions of population totals. Consequently, variance estimation is not a trivial matter and requires specific procedures. Methods for variance estimation can be classified according to two main approaches: (a) the resampling methods, and (b) the linearization methods. In this paper, we focus on the latter approach.

Under the usual design-based approach, let $U = \{1, \ldots, N\}$ be a fixed population of N identifiable individuals, and let y_i be the value of the study variable on the i-th individual. Moreover, let $\theta = \theta(y_1, \ldots, y_N)$ be the population parameter to be estimated on the basis of a random sample S of fixed size n selected from U with probability $P(S = s) = p(s)$. Let $\pi_i > 0$ and $\pi_{ij} > 0$ denote the first- and second-order inclusion probabilities, respectively. The linearization approach provides an approximation of the complex parameter θ and of its estimator $\widehat{\theta}$. The rationale underlying linearization consists in obtaining a linearized variable v_i for each observation y_i such that

$$\widehat{\theta} - \theta = \sum_{i \in S} \frac{v_i}{\pi_i} - \sum_{i \in U} v_i + R_p \ ,$$

where R_p is a (stochastically negligible) remainder term. Hence, the variance of the random variable $\sum_{i \in S} v_i / \pi_i$ may be used to approximate the variance of $\widehat{\theta}$. In practice, v_i is unknown and it is generally replaced by its sample counterpart $\widehat{v}_i$. Once the $\widehat{v}_i$'s are computed, variance estimation may be achieved by means of

$$\widehat{\mathrm{Var}}[\widehat{\theta}] = \sum_{i,j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \widehat{v}_i \widehat{v}_j \ .$$

Several techniques have been proposed to provide the $\widehat{v}_i$'s (see, e.g., Kovačević and Binder, 1997; Demnati and Rao, 2004). In the present paper, we consider Deville's (1999) method based on the concept of influence curve in the design-based approach. The next Section will be devoted to this linearization method when a large family of population functionals - which includes inequality indexes - is considered.

2 Linearization via influence curve

Let $M = \sum_{i \in U} \delta_{y_i}$ be the discrete measure on $\mathbb{R}$ which allocates a unit mass on each y_i and where δ_y represents the Dirac mass at y. In Deville's approach, it is assumed that the target parameter θ can be written as a functional F with respect to M, namely $\theta = F(M)$. Under this set-up, if $\widehat{M} = \sum_{i \in S} \delta_{y_i} / \pi_i$ denotes the empirical measure corresponding to M, a substitution estimator of θ is given by $\widehat{\theta} = F(\widehat{M})$. If F is homogeneous of degree α, under broad assumptions, Deville (1999) has proven the linearization

$$\sqrt{n}N^{-\alpha}(F(\widehat{M})-F(M)) = \sqrt{n}N^{-\alpha}\int \mathrm{IF}_F(u;M)\mathrm{d}(\widehat{M}-M)(u)+o_p(1)\ ,$$

where

$$\mathrm{IF}_F(u;M) = \lim_{t\to 0}\frac{1}{t}\left(F(M+t\delta_u)-F(M)\right)$$

represents the influence function in the design-based approach. The influence curve plays a central role in Deville's approach and - in particular - in the variance estimation. In fact, it provides the linearized variable in the sense that $v_i = \mathrm{IF}_F(y_i;M)$.

We now introduce a rule for dealing with the influence function in the presence of complex functionals. In this setting, let us consider a functional which may be expressed as

$$F(M) = \int \psi_y(L_y(M))\,\mathrm{d}M(y)\ , \tag{1}$$

where $L_y(M) = (L_{1,y}(M),\ldots,L_{k,y}(M))^{\mathrm{T}}$ is a vector of further functionals (eventually) indexed by y and $\psi_y : \mathbb{R}^k \mapsto \mathbb{R}$ is a function family assumed to be differentiable and regularly indexed by y. For instance, the inequality measures commonly considered in practice are members of the functional family F, or may be expressed at most as $\varphi(F(M)) = (\varphi\circ F)(M)$ where $\varphi : \mathbb{R}\mapsto\mathbb{R}$ is a smooth function. In the next Section some illustrative examples will be provided. In order to obtain the linearization of the functional F defined in expression (1), the following result is useful.

Proposition 1. *Let F be the functional defined in (1). If $L_{j,y}$ is Fréchet differentiable for each j, the influence function of F is given by*

$$IF_F(u;M) = \psi_u(L_u(M)) + \int \nabla\psi_y(L_y(M))^T IF_{L_y}(u;M)\,dM(y)\ ,$$

where $IF_{L_y}(u;M) = (IF_{L_{1,y}}(u;M),\ldots,IF_{L_{k,y}}(u;M))^T$.

The formal proof of Proposition 1 is given by Barabesi et al. (2014a).

Hence, Proposition 1 provides a simple rule for obtaining the influence function corresponding to the functional (1). Finally, it should be remarked that the influence function of $(\varphi\circ F)$ promptly follows as

$$\mathrm{IF}_{\varphi\circ F}(u;M) = \varphi'(F)\mathrm{IF}_F(u;M)\ ,$$

by assuming that φ be differentiable.

3 Application to some inequality indexes

The result stated in Proposition 1 turns out to be particularly useful for the linearization of inequality measures. Indeed, it can be easily shown that the Gini concentration index, the Amato index, the families of Generalized Entropy and Atkinson indexes and the Zenga index (2007) may be expressed as functionals of type (1). For

illustrative purposes, we solely discuss the achievement of the influence function for the Gini index and the Amato index.

The finite-population version of the Gini index (see, e.g., Berger, 2008) may be expressed as the functional

$$G(M) = \int \frac{2yH_y(M)}{N(M)T(M)}\,\mathrm{d}M(y) - 1\ ,$$

where

$$N(M) = \int \mathrm{d}M(x)\ ,\ T(M) = \int x\,\mathrm{d}M(x)$$

respectively represent the population size and the population total rephrased as functionals. In addition, by assuming that I_B is the usual indicator function of a set B, in the following we also assume that

$$H_y(M) = \int I_{[x,\infty[}(y)\,\mathrm{d}M(x)\ ,\ K_y(M) = \int xI_{[y,\infty[}(x)\,\mathrm{d}M(x)\ .$$

In this case, we have $L_y(M) = (H_y(M), N(M), T(M))^{\mathrm{T}}$, while

$$\psi_y(L_y(M)) = \frac{2yH_y(M)}{N(M)T(M)}$$

and $\varphi(F) = F - 1$. Thus, with a slight abuse in notation, i.e. by suppressing the argument of the functionals for the sake of simplicity, it holds that

$$\nabla\psi_y(L_y(M)) = \frac{2y}{NT}\left(1, -\frac{H_y}{N}, -\frac{H_y}{T}\right)^{\mathrm{T}}$$

and $\mathrm{IF}_{L_y}(u;M) = (I_{[u,\infty[}(y), 1, u)^{\mathrm{T}}$. Hence, by applying Proposition 1, after some algebra it follows that

$$\mathrm{IF}_G(u;M) = \frac{2}{NT}\left(uH_u + K_u\right) - (G+1)\left(\frac{1}{N} + \frac{u}{T}\right)\ .$$

The Amato index has recently received renewed interest for its properties by Arnold (2012). The influence function for the Amato index is not available in literature. To this aim, on the basis of the continuous-population expression of the Amato index (Arnold, 2012), the finite-population counterpart of this inequality measure may be given as the functional

$$A(M) = \int \sqrt{\frac{1}{N(M)^2} + \frac{y^2}{T(M)^2}}\,\mathrm{d}M(y)\ .$$

Hence, in this case $L_y(M) = (N(M), T(M))^{\mathrm{T}}$, while

$$\psi_y(L_y(M)) = \sqrt{\frac{1}{N(M)^2} + \frac{y^2}{T(M)^2}}$$

and, trivially, $\varphi(F) = F$. By adopting the same notational simplification as above, and by assuming that $\mu = T/N$ denotes the population mean, it holds that

$$\nabla\psi_y(L_y(M)) = \frac{T}{\sqrt{\mu^2+y^2}}\left(-\frac{1}{N^3}, -\frac{y^2}{T^3}\right)^{\mathrm{T}}$$

and $\mathrm{IF}_{L_y}(u;M) = (1,u)^{\mathrm{T}}$. Hence, by applying Proposition 1, it turns out that

$$\mathrm{IF}_A(u;M) = \frac{1}{T}\sqrt{\mu^2+u^2} - \frac{\mu}{N^2}\int \frac{1}{\sqrt{\mu^2+y^2}}\,\mathrm{d}M(y) - \frac{u}{T^2}\int \frac{y^2}{\sqrt{\mu^2+y^2}}\,\mathrm{d}M(y)\,.$$

4 Collecting income data via randomized response theory

When dealing with inequality indexes based on income data, it should be realized that income is notoriously considered a sensitive character to be surveyed, in the sense that people are reluctant to disclose it - mostly, in the case of income from self-employment, property and financial assets. Consequently, this issue may result in seriously-biased estimates of inequality indicators. To alleviate this problem, the respondent cooperation has to be increased. The randomized response technique can be used to achieve this aim (see, e.g, Barabesi et al., 2013 and the references therein). For quantitative data, the idea behind the technique is to perturb the true response y_i for ensuring confidentiality to the respondents. Let us suppose that the randomization procedure proposed by Greenberg et al. (1971) is adopted. Under this protocol, the i-th individual reports her/his true income value y_i with probability q, or she/he generates a random variate from a (suitable) absolutely-continuous random variable (r.v.) X with probability $(1-q)$. Note that $q = 1$ leads to direct questioning. In order to obtain the influence function when the randomization stage is added, let the r.v. Z_i represent the answer of the i-th individual and let us consider the measure M_{R} on $\mathbb{R}$ given by

$$M_{\mathrm{R}} = \sum_{i\in U}(q\delta_{y_i} + (1-q)\lambda_X)\,,$$

where λ_X represents the law of the r.v. X with expectation μ_X. In such a case, the empirical measure corresponding to M_{R} is given by $\widehat{M}_{\mathrm{R}} = \sum_{i\in S}\delta_{Z_i}/\pi_i$. By suitably reformulating the original functional, i.e. by determining the further functional $F_{\mathrm{R}}(M_{\mathrm{R}}) = F(M)$, the substitution estimator for $F_{\mathrm{R}}(M_{\mathrm{R}})$ is given by $F_{\mathrm{R}}(\widehat{M}_{\mathrm{R}})$ and the influence function may be in turn defined as

$$\mathrm{IF}_{F_{\mathrm{R}}}(u;M_{\mathrm{R}}) = \lim_{t\to 0}\frac{1}{t}\left(F_{\mathrm{R}}(M_{\mathrm{R}} + t\delta_u) - F_{\mathrm{R}}(M_{\mathrm{R}})\right)\,.$$

From the previous definition of the influence function, a result equivalent to Proposition 1 holds. As an example, in the case of the Gini concentration index, Barabesi et al. (2014b) shown that

$$G_{\mathrm{R}}(M_{\mathrm{R}}) = \frac{2}{q^2}\int \frac{yH_y(M_{\mathrm{R}})}{N_{\mathrm{R}}(M_{\mathrm{R}})T_{\mathrm{R}}(M_{\mathrm{R}})}\,\mathrm{d}M_{\mathrm{R}}(y) - \frac{2(1-q)}{q^2}\,\frac{C_{\mathrm{R}}(M_{\mathrm{R}})}{T_{\mathrm{R}}(M_{\mathrm{R}})} + \frac{2(1-q)^2\gamma_X}{q^2}\,\frac{N_{\mathrm{R}}(M_{\mathrm{R}})}{T_{\mathrm{R}}(M_{\mathrm{R}})} - 1\,,$$

where

$$N_{\mathrm{R}}(M_{\mathrm{R}}) = \int \mathrm{d}M_{\mathrm{R}}(y)\,,\; T_{\mathrm{R}}(M_{\mathrm{R}}) = \frac{1}{q}\int y\,\mathrm{d}M_{\mathrm{R}}(y) - \frac{1-q}{q}\,\mu_X N_{\mathrm{R}}(M_{\mathrm{R}})$$

and

$$C_{\mathrm{R}}(M_{\mathrm{R}}) = \int (yH_y(\lambda_X) + K_y(\lambda_X))\,\mathrm{d}M_{\mathrm{R}}(y)\,,$$

while $\gamma_X = \int K_y(\lambda_X)\mathrm{d}\lambda_X(y)$. Hence, with the usual slight abuse in notation, it reads

$$\mathrm{IF}_{G_{\mathrm{R}}}(u;M) = \frac{2}{q^2 N_{\mathrm{R}} T_{\mathrm{R}}}\,(uH_u(M_{\mathrm{R}}) + K_u(M_{\mathrm{R}})) - \frac{2(1-q)}{q^2 T_{\mathrm{R}}}\,(uH_u(\lambda_X) + K_u(\lambda_X)) + \frac{2(1-q)C_{\mathrm{R}}}{q^2 N_{\mathrm{R}} T_{\mathrm{R}}} + \frac{4(1-q)^2\gamma_X}{q^2 T_{\mathrm{R}}} - (G_{\mathrm{R}}+1)\left(\frac{1}{N_{\mathrm{R}}} + \frac{u}{qT_{\mathrm{R}}} - \frac{(1-q)\mu_X}{qT_{\mathrm{R}}}\right).$$

Acknowledgements Work supported by project PRIN-2012F42NS8:"Household wealth and youth unemployment: new survey methods to meet current challenges".

References

1. Arnold, B.C.: On the Amato inequality index. Stat Probabil Lett, **82**, 1504–1506 (2012)
2. Barabesi, L., Diana, G., Perri, P.F.: Design-based distribution function estimation for stigmatized populations. Metrika, **76**, 919–935 (2013)
3. Barabesi, L., Diana, G., Perri, P.F.: A functional derivative useful for the linearization of inequality indexes in the design-based framework. Available via http://arxiv.org/abs/1402.3478 (2014a)
4. Barabesi, L., Diana, G., Perri, P.F.: Gini index estimation in randomized response surveys. AStA-Adv Stat Anal, forthcoming (2014b)
5. Berger, Y.G: A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. J Off Stat, **24**, 541–555 (2008)
6. Demnati, A., Rao, J.N.K.: Linearization variance estimators for surveys data (with discussion). Surv Methodol, **30**, 17-26 (2004)
7. Deville, J.C.: Variance estimation for complex statistics and estimators: linearization and residual techniques. Surv Methodol, **25**, 193–203 (1999)
8. Greenberg, B.G., Kubler, R.R., Abernathy, J.R., Horvitz, D.G.: Applications of the randomized response technique in obtaining quantitative data. J Am Stat Assoc, **66**, 243–250 (1971)
9. Kovačević, M., Binder, D.A.: Variance estimation for measures of income inequality and polarization. The estimating equations approach. J Off Stat, **13**, 41-58 (1997)
10. Zenga, M.: Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. Statistica & Applicazioni, **4**, 3-27 (2007)