Published in Journal of the Royal Statistical Society: Series A 176, issue 2, 521-540, 2013
which should be used for any reference to this work

Variance estimation of the Gini index: revisiting a result several times published

Matti Langel and Yves Tillé

University of Neuchâtel, Switzerland

[Received August 2011. Revised February 2012]

Summary. Since Corrado Gini suggested the index that bears his name as a way of measuring inequality, the computation of variance of the Gini index has been subject to numerous publications. We survey a large part of the literature related to the topic and show that the same results, as well as the same errors, have been republished several times, often with a clear lack of reference to previous work. Whereas existing literature on the subject is very fragmented, we regroup references from various fields and attempt to bring a wider view of the problem. Moreover, we try to explain how this situation occurred and the main issues that are involved when trying to perform inference on the Gini index, especially under complex sampling designs. The interest of several linearization methods is discussed and the contribution of recent references is evaluated. Also, a general result to linearize a quadratic form is given, allowing the approximation of variance to be computed in only a few lines of calculation. Finally, the relevance of the regression-based approach is evaluated and an empirical comparison is proposed.

Keywords: Gini; Inequality; Influence function; Linearization; Sampling; Variance estimation

1. Introduction

Despite the fact that the Gini index is the most widely used indicator of income inequality in a population, it is not a trivial measure to handle. To start with, its construction is not easy to understand for an uninitiated audience. Also, because the Gini index of a population is commonly estimated by means of a sample, its estimation should be completed by some knowledge of the accuracy of the point estimate, namely by reporting a variance estimator or standard error, allowing for the computation of a confidence interval. Since the Gini index is a non-linear function of interest, variance estimation is not straightforward, especially when data are collected by means of a complex sampling strategy. Thus, the computation of the sampling variance of the Gini index has prompted a great amount of research in statistics and economics.

Although the problem has now mostly been solved, the result has been republished many times in recent years. The original motivation of this paper is to understand why. For that purpose, we survey a large portion of the literature on the subject in a historical perspective. A thorough analysis of the evolution of the literature has stressed two important features: the main methodological issue in the computation of variance of the Gini index has sometimes been overlooked, and a serious lack of references to previous works is often witnessed.

One of the main contributions of this paper is to present and compare many different approaches to variance estimation of the Gini index. Additionally we give a general result on the linearization of a quadratic form which allows the initially intricate problem to become computationally quite simple. We also show why the regression-based methods, which have

Address for correspondence: Matti Langel, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland.
E-mail: matti.langel@unine.ch

attracted much attention recently, should be avoided. Finally, although publications on the topic have arisen from different fields, we try to propose a global clarification of the present state of the art on the subject.

In the next section, the Gini index and estimators are presented. A discussion on the evolution and issues regarding variance estimation of the index is proposed in Section 3. Section 4 is dedicated to linearization techniques, which encompass a variety of approaches for deriving a variance estimator. Many approaches are presented and a new result allowing for a fast linearization of the Gini index is suggested. The relevance of recent publications is also discussed. In Section 5, we examine the so-called regression approach which has prompted many recent research studies. Finally, a comparative numerical application is performed in Section 6. To show the shortcomings of the regression approach, an empirical study was conducted using the same data as in two recent references (Giles, 2004; Davidson, 2009). The paper ends with some concluding remarks.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://www.blackwellpublishing.com/rss`

2. The Gini index

2.1. Definitio and estimation in an infinit population

Like several other inequality measures, the Gini (1912, 1914, 1921) index is based on the cumulative income of a proportion of the poorer units. Let $f(y)$ be a probability density function of a positive continuous random variable Y that represents the income and $F(y)$ its cumulative distribution function. First define the Lorenz (1905) curve given by

$$L(\alpha) = \frac{\displaystyle\int_0^{F^{-1}(\alpha)} y\, f(y)\, \mathrm{d}y}{\displaystyle\int_0^{\infty} y\, f(y)\, \mathrm{d}y} = \frac{1}{\mu} \int_0^{\alpha} F^{-1}(u)\, \mathrm{d}u,$$

where $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$ and

$$\mu = \int_0^{\infty} y f(y)\, \mathrm{d}y.$$

The Gini index can be defined in several ways (see, for example, Xu (2004)):

$$G = 2 \int_0^1 \{\alpha - L(\alpha)\}\, \mathrm{d}\alpha = 1 - 2 \int_0^1 L(\alpha)\, \mathrm{d}\alpha = \frac{2}{\mu} \int_0^{\infty} y\, F(y)\, f(y)\, \mathrm{d}y - 1. \tag{1}$$

If $y_i, i = 1, \ldots, n$, is a sequence of positive random variables with the same probability density function $f(\cdot)$, the Gini index G can be estimated by

$$\hat{G} = \frac{2 \displaystyle\sum_{i=1}^{n} i y_{(i)}}{n \displaystyle\sum_{i=1}^{n} y_{(i)}} - \frac{n+1}{n}, \tag{2}$$

where the $y_{(i)}$ are the y_i ordered in increasing order. This expression can be found for example in Sen (1973) or Fei *et al.* (1978). A complete review of all the expressions of the Gini index that were proposed originally by Corrado Gini is found in Ceriani and Verme (2011). Note also that a

controversy exists on the use of $\hat{\tilde{G}} = n\hat{G}/(n-1)$ instead of $\hat{G}$. Indeed, $\hat{\tilde{G}}$ is less biased than $\hat{G}$, but the latter will be used hereafter because the construction of $\tilde{G}$ becomes difficult when the observations are weighted.

2.2. Definitio and estimation in a finit population

The Gini index is generally estimated by means of a sample survey. In survey sampling, we are interested in a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N from which a random sample S of size n is selected by means of a sampling design $p(s) = \Pr(S = s)$, for all $s \in U$. Let also $\pi_k = \Pr(k \in S)$ denote the inclusion probability of unit $k \in U$ and $d_k = 1/\pi_k$ the Horvitz and Thompson (1952) weights. In survey methodology, the observations are usually weighted. The sampling weight w_k that is associated with observation k can be equal to d_k or can be improved by a calibration technique (Deville and Särndal, 1992) or a non-response adjustment. Let $y_1, \ldots, y_k, \ldots, y_N$ denote the incomes of the units in the population. Let also N_k and n_k denote the rank of unit k in population U and in sample S respectively, with tied observations treated by using increasing integer ranks such that, for example, the series of observations $4, 6, 6, 8$ would be assigned ranks $1, 2, 3, 4$. To estimate totals

$$Y = \sum_{k \in U} y_k$$

and

$$N = \sum_{k \in U} 1,$$

we can use weighted estimators

$$\hat{Y} = \sum_{k \in S} w_k y_k$$

and

$$\hat{N} = \sum_{k \in S} w_k.$$

The total income of the αN poorest units is defined by

$$\tilde{Y}(\alpha) = \sum_{k \in U} y_k \, \mathbb{1}(y_k \leqslant Q_\alpha),$$

where Q_α is the α-quantile and $\mathbb{1}(A)$ is an indicator function equal to 1 if A is true and 0 otherwise. However, this definition is not very accurate because the quantiles can be defined in several ways when the cumulative distribution function is a step function (for a review, see Hyndman and Fan (1996)). We thus prefer to use the less ambiguous definition of the total income of the αN poorest units that was proposed in Langel and Tillé (2011) and given by

$$Y(\alpha) = \sum_{k \in U} y_k \, H(\alpha N - N_{k-1}),$$

where $H(\cdot)$ is the cumulative distribution function of a uniform [0,1] random variable

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leqslant x < 1, \\ 1 & \text{if } x \geqslant 1. \end{cases}$$

The function of interest $Y(\alpha)$ is then strictly increasing in α in $(0,1)$, which is not the case for $\tilde{Y}(\alpha)$. To estimate $Y(\alpha)$, we can use

$$\hat{Y}(\alpha) = \sum_{k \in S} w_k y_k \, H\left(\frac{\alpha \hat{N} - \hat{N}_{k-1}}{w_k}\right)$$

where $\hat{N}_k$ are the cumulative weights of the sampled units, i.e.

$$\hat{N}_k = \sum_{l \in S} w_l \mathbb{1}(n_l \leqslant n_k), \tag{3}$$

and $\hat{N}_0 = 0$. Expression $Y(\alpha)$ is also strictly increasing in α in $(0,1)$. In a finite population, the Lorenz curve can then be defined by

$$L(\alpha) = Y(\alpha)/Y, \tag{4}$$

and can be estimated by

$$\hat{L}(\alpha) = \hat{Y}(\alpha)/\hat{Y}. \tag{5}$$

Accordingly, functions $L(\alpha)$ and $\hat{L}(\alpha)$ are also strictly increasing in $(0,1)$. If we use the definition of the Lorenz curve that is given in equation (4) and the Gini index as given in equation (1), we obtain, after some algebra, the following expressions of the Gini index for a finite population:

$$G = \frac{2}{NY} \sum_{k \in U} N_k y_k - \frac{N+1}{N} = \frac{\sum_{k \in U} \sum_{l \in U} |y_k - y_l|}{2NY}.$$

Finally, if we use the estimator of the Lorenz curve that is given in equation (5), we obtain an estimator of the Gini index for weighted observations:

$$\begin{aligned}
\hat{G} &= \frac{2}{\hat{N}\hat{Y}} \sum_{k \in S} w_k \hat{N}_k y_k - \left(1 + \frac{1}{\hat{N}\hat{Y}} \sum_{k \in S} w_k^2 y_k\right) \\
&= \frac{2 \sum_{k \in S} w_k \hat{N}_k y_k - \sum_{k \in S} w_k^2 y_k}{\hat{N}\hat{Y}} - 1 \qquad (6) \\
&= \frac{\sum_{k \in S} w_k y_k (2\hat{N}_k - \hat{N} - w_k)}{\hat{N}\hat{Y}} \qquad (7) \\
&= \frac{\sum_{k \in S} \sum_{l \in S} w_k w_l |y_k - y_l|}{2\hat{N}\hat{Y}}. \qquad (8)
\end{aligned}$$

$\hat{L}(\alpha)$ and $\hat{G}$ are generally biased. If, instead of equation (5), the Lorenz curve is estimated through an empirical distribution function (step function), the resulting expression of $\hat{G}$ may differ from that proposed above, depending on the type of step function that is used (see also Davidson (2009)). Our approach avoids this ambiguity.

3. Variance estimation: a result several times published

3.1. First results and evolution

The evolution of literature on the variance of the Gini index may be summarized as follows. Until the 1980s, the number of papers devoted to the subject was very limited. This period

of time is briefly discussed below. Research was then split into three main directions: survey sampling, robust statistics and economics.

The first calculation of variance relating to the Gini index was probably realized by Nair (1936) who computed the exact variance of the Gini mean difference, i.e. the numerator of expression (8) of the Gini index. The proposed expression of variance is nevertheless particularly cumbersome, even in the simplest case of unweighted observations that was presented by Nair. In a survey context, it is also always possible to compute the variance of a double sum like the Gini mean difference, but the computation requires inclusion probabilities up to the fourth order (see for example Ardilly and Tillé (2006)). Lomnicki (1952) and Glasser (1962) have approximated this expression and gave simpler variance estimators. The chronology between these three references is clear: Lomnicki (1952) refers to Nair (1936), whereas Glasser (1962) cites both. Finally, Sillitto (1969) also computed an expression for the variance of the Gini mean difference by using L-moments to approximate the quantile function by polynomials (see also Hosking (1990)).

Meanwhile, one of the first results for the variance of the full Gini index is attributed to Hoeffding (1948). After defining the notion of U-statistics based on earlier work by Halmos (1946) and expressing the variance of a U-statistic, Hoeffding showed that the Gini index is a function of two U-statistics and gave a result for the variance of the Gini index. In addition to giving a general form and an unbiased estimator for the variance of the Gini mean difference by using some results on U-statistics, Glasser (1962) proposed an approximation of the variance of the Gini index for simple random sampling without replacement from a finite population.

Expressions for an asymptotic variance of the Gini index have also been given by Sendler (1979) based on results of Shorack (1972) on functions of order statistics, by Cowell (1989) who generalized the U-statistic approach to weighted observations, as well as by Schechtman (1991) who combined the work of Hoeffding (1948) and the idea of the infinitesimal jackknife (Jaeckel, 1972). In related topics, Gastwirth (1972) computed lower and upper bounds for the estimated Gini index, whereas Beach and Davidson (1983) proposed inference on the Lorenz curve.

3.2. Fragmentation of the literature

Because the measure is widely used in practice (in official statistics or policy making for example), the question of sampling variance of the Gini index has interested scholars from partially unrelated domains. When analysing the whole corpus of literature, a clear separation between publications from the field of statistics and publications in economic journals is witnessed. As papers from one field are seldom cited in the other, it seems evident that researchers from these two fields do not necessarily read each other's work. Moreover, inside statistical literature, some contributions come from survey sampling whereas others come from robust statistics.

In the past 20 years, the gap between the economic and statistical literature on the Gini index seems to have become wider. The survey paper by Xu (2004) is a perfect illustration of this fragmentation. In this paper, which reviews literature on the Gini index across the 20th century, no references are made to survey sampling studies on the topic. Moreover, papers from robust statistics are not mentioned either. One could argue that this paper does not focus on inference *per se* but, in another paper, Xu (2007) proposed an introductory overview of inference for the Gini index, in which the literature from robust statistics and survey sampling of the last two decades is also absent.

The variety of existing methods to tackle the problem is another obstacle to a unified understanding of the latter. Indeed, at least three general categories of approach have been used: linearization (or asymptotic theory-based) methods, resampling methods and regression-based methods. Moreover, each of these categories encompass many different techniques.

In Section 4, several linearization techniques are presented. Linearization-based computation of variance of the Gini index has generated a large amount of literature and, thus, embodies the main focus of this paper. Furthermore, even after the problem of linearizing the Gini index had been solved, the result has been republished several times. The main reason why such a situation happened is probably the plurality of approaches that can be defined as linearization methods: direct computation of variance, use of estimating equations, computation of influence functions, the delta method or the Demnati and Rao (2004) method. These methods lead to the same result but are done by using different developments. Moreover, whereas some of these approaches have indeed been proposed directly for variance approximation, some were developed for other purposes such as robustness analysis.

The regression-based methods are described in Section 5. A jackknife approach is also discussed inside the regression section as well as in the empirical illustrations. Bootstrap methods are not treated in this paper but have also been applied to propose Gini index standard errors (Mills and Zandvakili, 1997; Kuan, 2000; Giorgi *et al.*, 2006). Finally, note that the problem has recently been addressed using the empirical likelihood method (Qin *et al.*, 2010; Peng, 2011).

3.3. The pitfall of the computation of the variance

There is certainly another, more methodological, reason why variance estimation for the Gini index has resulted in many publications. From today's point of view the problem seems quite simple; thus we show in Section 4 a result which gives an expression for the variance in only a few lines of calculation, but the complexity of the statistic has been a challenge for variance estimation in the past. In particular, some researchers who proposed straightforward or simplified solutions were in fact falling into a methodological pitfall. This pitfall can be described as follows: when we examine expression (6) of the estimated Gini index, we could believe that the numerator is composed of two simple sums and that an expression of the variance can be directly derived. Unfortunately, this reasoning is mistaken. Indeed, the quantity $\hat{N}_k$ that is defined in equation (3) is an estimator of N_k, the rank of unit k in the population, and it is random because its value depends on the sample selected. Thus, part of the variability of the estimated Gini index is due to $\hat{N}_k$, and this aspect must be taken into account. As shown by Sandström *et al.* (1985, 1988), the variance of the index is in fact considerably overestimated when this randomness is not taken care of.

One could imagine some situations where the rank is known in the population, e.g. when the sample is selected from a register. This question has been discussed by Deville (1996). If so, computing the variance of the Gini index amounts to expressing the variance of a ratio. The estimator of the Gini index nevertheless has a smaller variance when an estimator of the rank is used rather than the true population rank. Although this overestimation may at first seem surprising, it can be easily illustrated. Suppose that an unexpected number of high incomes are selected in the sample. The true population rank N_k of selected unit k will then have a tendency to be underestimated by $\hat{N}_k$. If, in contrast, very few high incomes are selected, $\hat{N}_k$ will have a tendency to overestimate the rank of unit k. The sampling distribution of $\Sigma_{k\in S} w_k \hat{N}_k y_k$ becomes less scattered than that of the respective expression computed with the true population rank, namely $\Sigma_{k\in S} w_k N_k y_k$, resulting in a smaller variance for the Gini index that is estimated by using the former sum.

4. Linearization techniques

4.1. Rationale behind linearization

Linearization combines a range of techniques used to calculate the approximated variance of

a non-linear statistic. It consists in approximating a non-linear or complex statistic (here, the Gini index) by a sum of terms, i.e. finding a linearized variable z_k such that

$$\hat{G} - G \approx \sum_{k \in S} w_k z_k - \sum_{k \in U} z_k.$$

Next, the variance of $\hat{G}$ is simply approximated by the variance of the estimated total

$$\hat{Z} = \sum_{k \in S} w_k z_k.$$

Nevertheless, the z_ks often depend on population parameters that must be estimated. By estimating these parameters, we can obtain $\hat{z}_k$ an estimator of z_k and thus construct an estimator of variance by plugging $\hat{z}_k$ into the expression of the variance of a total corresponding to the given sampling design. Thus, the method is applicable to any sampling design, provided that an expression for the variance of the total is known. For instance, if the sampling design is simple without replacement, with fixed sample size n, we have the estimator

$$\widehat{\text{var}}_{\text{lin}}(\hat{G}) = \frac{N-n}{Nn(n-1)} \sum_{k \in S} (\hat{z}_k - \bar{\hat{z}})^2, \tag{9}$$

where

$$\bar{\hat{z}} = \frac{1}{n} \sum_{k \in S} \hat{z}_k.$$

For details on the asymptotic framework validating linearization, one can relate to Isaki and Fuller (1982), Deville and Särndal (1992) and Deville (1999). Without using this terminology, Glasser (1962) already pointed out that the linearized variable given by

$$u_k = 2 \sum_{l \in U} |y_k - y_l| \tag{10}$$

can be used to approximate the variance of the Gini mean difference by plugging it into the estimator of variance of a total in place of the interest variable. For instance, if the sampling design is simple without replacement with fixed sample size n, we have the approximation of variance

$$\text{Avar}(\hat{A}) = N^2 \frac{N-n}{Nn(n-1)} \sum_{k \in U} (u_k - \bar{u})^2,$$

where

$$\hat{A} = \sum_{k \in S} \sum_{l \in S} |y_k - y_l|$$

and

$$\bar{u} = \frac{1}{N} \sum_{k \in U} u_k.$$

4.2. Taylor series expansion

In practice, there are several ways to calculate this linearized variable. For smooth functions of totals, one can linearize by performing a Taylor series expansion with respect to these totals (Woodruff, 1971). In the survey sampling literature, Nygård and Sandström (1981, 1985, 1989) and especially Sandström *et al.* (1985, 1988) are usually considered as seminal works on the variance of the Gini index. Sandström *et al.* (1985, 1988) discussed four variance estimators for

the Gini index: first in simple random sampling; then in unequal probability sampling. Furthermore, Sandström *et al.* (1988) showed links between their work and those of Glasser (1962) or Sendler (1979). Two of the estimators that were presented in Sandström *et al.* (1985, 1988) are based on a first-order Taylor series approximation and are thus probably the first explicit tentative linearization of the Gini index. However, the results have been shown to be only partially satisfactory because the Gini index cannot be expressed as a smooth function of totals.

For the first estimator (the *ratio* estimator) that was presented by Sandström *et al*. (1985, 1988) the Gini index is simply considered to be a ratio of totals. Thus, the classical first-order Taylor series approximation of a ratio is used to linearize the index. They noted that this estimator does not take into account the fact that the ranks depend on the other units in the sample (namely, that $\hat{N}_k$ is random) and that these ranks should be estimated in some way. These conclusions are in line with the pitfall that was discussed in Section 3.3. Indeed, the ratio estimator is unsatisfactory because it can only be constructed if the Gini index is mistakenly interpreted as a function of totals.

The second estimator (which is hereafter denoted the *Taylor* estimator) that was proposed by Sandström *et al.* (1985, 1988) is also based on Taylor series but this time it takes the randomness of the ranks into account. However, the expression is cumbersome in the simple random sampling case and becomes virtually inapplicable in the probability sampling framework because it requires joint inclusion probabilities up to order 4.

In opposition to the first two design-based estimators, the third proposed estimator is what Sandström *et al*. (1985, 1988) call a *model-based* estimator, denoting that the observations are realizations of independent, identically distributed random variables which form a fixed, given, sample. This estimator is related to the U-statistics approach, and, under simple random sampling, reduces to the variance estimator that was proposed by Sendler (1979). The last estimator discussed is a straightforward jackknife estimator.

In Sandström *et al.* (1985), the simulation studies under simple random sampling show that the first ratio estimator, which is the only estimator to treat the ranks as constants, greatly overestimates the variance. This should act as an indication that handling the random ranks issue is crucial. Other estimators show good results. In the probability sampling case (Sandström *et al.*, 1988), the ratio estimator is unsurprisingly also ineffective and the Taylor estimator is, as noted previously, not applicable. Thus, in that situation, only the jackknife and the model-based estimators were considered satisfactory by them. Sandström *et al.* (1988) also noted that the Taylor estimator is similar to that of Glasser (1962) and that, when both sample and finite population sizes become large, the Taylor and model-based estimators are equal.

4.3. Influenc functions

When the function of interest is, like the Gini index, not a function of totals, a more radical way of computing a linearized variable involves computing the influence function that was initially proposed by Hampel (1974) and Hampel *et al.* (1985). The influence function was first proposed to study the robustness of an estimator but can also be used to approximate the variance. The influence function of the Gini index seems to have been computed for the first time by Monti (1991) and next by Cowell and Victoria-Feser (1996, 2003) but was used to make the Gini estimator robust rather than to estimate the variance. The influence function of the Gini index that was given by Monti (1991) can be rewritten

$$z_k = \frac{1}{NY}\{2N_k(y_k - \bar{Y}_k) + Y - Ny_k - G(Y + y_k N)\}, \tag{11}$$

where

$$\bar{Y}_k = \frac{\sum_{l \in U} y_l \mathbb{1}(N_l \leqslant N_k)}{N_k}.$$

The result that was obtained by Monti (1991) has been overlooked by survey statisticians, probably because the role of the influence function as a tool for variance estimation was not well established.

We shall show in the next sections that this linearized variable can be computed in only a few lines of calculation by means of different methods. This linearized variable can be estimated by

$$\hat{z}_k = \frac{1}{\hat{N}\hat{Y}}\{2\hat{N}_k(y_k - \hat{\bar{Y}}_k) + \hat{Y} - \hat{N}y_k - \hat{G}(\hat{Y} + y_k\hat{N})\}, \tag{12}$$

where

$$\hat{\bar{Y}}_k = \frac{\sum_{l \in S} w_l y_l \mathbb{1}(\hat{N}_l \leqslant \hat{N}_k)}{\hat{N}_k}.$$

4.4. Estimating equations

Another method that is used to compute a linearized variable for the Gini index is to express it as the solution of an estimating equation (Binder and Kovacevic, 1995; Kovacevic and Binder, 1997). Using the estimating equation methodology (for details on this approach, see for example Binder and Patak (1994)), a linearized variable for the Gini index is derived. The result is equal to that of expression (11) and can be estimated by equation (12).

4.5. Deville approach

Deville (1996) used Sandström *et al.* (1985, 1988) as a starting point for the linearization of the Gini index. He identified clearly the source of the overestimation that was obtained by Sandström *et al.* (1985, 1988) when the randomness of the ranks was neglected. Later, Deville (1999) proposed a modified version of the influence function to compute a linearized variable for sampling from a finite population. Unfortunately, in Deville (1999), a term is missing in the final expression of the linearized variable of the Gini index.

To define the influence function, Deville used a measure M with unit mass for each point of the population. According to Deville's definition, the measure M is positive, discrete, with a total mass N whereas the total mass is equal to 1 for the measure used in the influence function proposed by Hampel (1974). A function of interest can be presented as a functional $T(M)$ that associates for each measure a real number or a vector. For instance, a total Y can be written

$$Y = \int y \, \mathrm{d}M = \sum_{k \in U} y_k.$$

Besides, we also suppose that the functionals considered are homogeneous in the sense that there always exists a real number α such that

$$T(tM) = t^{\alpha} T(M), \qquad \text{for all } t \in \mathbb{R}_+^*.$$

Coefficient α is called the degree of the functional $T(M)$. The measure M is estimated by a measure $\hat{M}$ that has a mass equal to w_k for each point x_k of sample S. The plug-in estimator of a functional $T(M)$ is simply $T(\hat{M})$. For instance, the estimator of a total is given by

$$\int y \, \mathrm{d}\hat{M} = \sum_{k \in S} w_k y_k.$$

Deville's influence function is defined by

$$\mathrm{IT}(M,x)=\lim_{t\to 0}\frac{T(M+t\delta_x)-T(M)}{t},$$

when this limit exists, where δ_x is the Dirac measure at point x. This influence function is the Gâteaux differential in the direction of the Dirac mass at point x. Deville (1999) showed that this influence function $z_k=\mathrm{IT}(M,x_k)$ is a linearized variable of $T(\hat{M})$ in the sense that it allows for the approximation of the interest function:

$$\frac{T(\hat{M})-T(M)}{N^{\alpha}}\approx\frac{1}{N^{\alpha}}\left(\sum_{k\in S}w_k z_k-\sum_{k\in U}z_k\right).$$

The approximation of the variance of $T(\hat{M})$ is obtained by computing the variance of the weighted sum of z_k on the sample:

$$\mathrm{Avar}\{T(\hat{M})\}=\mathrm{var}\left(\sum_{k\in S}w_k z_k\right).$$

The influence function generally depends on population parameters that are unknown and can simply be estimated by replacing the latter by their plug-in estimators. Thereby, we obtain $\hat{z}_k$, the estimator of the linearized variable z_k. Computation of influence functions follows the rules of differential calculus. Deville (1999) has shown, among other properties, the following results.

Result 1. If $T(M)=\Sigma_{k\in U}\,y_k=\int y\,\mathrm{d}M(y),$ then the influence function is $\mathrm{IT}(M,x_k)=y_k$.

Result 2. Let $\mathbf{S}$ be a functional of $\mathbb{R}^q$ and $\mathbf{T}_{\boldsymbol{\lambda}}$ a family of functionals that depend of $\boldsymbol{\lambda}\in\mathbb{R}^q$; then

$$I(\mathbf{T}_{\mathbf{S}})=I(\mathbf{T}_{\boldsymbol{\lambda}=\mathbf{S}})+\left.\frac{\partial\mathbf{T}}{\partial\boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\mathbf{S}}I\mathbf{S}.$$

Result 2 shows that the computation of the influence function can also be realized by steps. We propose hereafter an additional result that enables us to compute the linearized variable of a double sum (e.g. quadratic form) directly such that

$$S=\sum_{k\in U}\sum_{l\in U}\phi(y_k,y_l).$$

Result 3. If

$$S(M)=\int\int\phi(x,y)\,\mathrm{d}M(x)\,\mathrm{d}M(y),$$

where $\phi(\cdot,\cdot)$ is a function from $\mathbb{R}^2$ in $\mathbb{R}$, then

$$\mathrm{IS}(M,\xi)=\int\phi(x,\xi)\,\mathrm{d}M(x)+\int\phi(\xi,y)\,\mathrm{d}M(y).$$

The proof is given in Appendix A.

If $\phi(x,y)=\phi(y,x)$ for all x and y then the influence function can simply be written as

$$\mathrm{IS}(M,\xi)=2\int\phi(x,\xi)\,\mathrm{d}M(x).$$

Result 3 allows for a fast computation of the linearized variable of the Gini index. The latter can be written

$$G=A/2NY,$$

where A is a double sum:

$$A = \sum_{k \in U} \sum_{l \in U} |y_k - y_l|.$$

By using result 3, we obtain a linearized variable of A immediately:

$$\mathrm{IA}(M, x_k) = 2 \sum_{l \in U} |y_k - y_l| = 2\{2N_k(y_k - \bar{Y}_k) + Y - Ny_k\},$$

which is the result that was presented in expression (10) obtained by Glasser (1962) by using a different method. Now, if we apply the technique of linearization by steps as well as result 1 for totals Y and N, we obtain a linearized variable of the Gini index

$$\begin{aligned} z_k = \mathrm{IG}(M, x_k) &= \frac{\mathrm{IA}(M, x_k)}{2NY} - \frac{A}{2N^2Y}\mathrm{IN}(M, x_k) - \frac{A}{2NY^2}\mathrm{IY}(M, x_k) \\ &= \frac{1}{NY}\left\{\frac{\mathrm{IA}(M, x_k)}{2} - G(Y + y_k N)\right\} \\ &= \frac{1}{NY}\{2N_k(y_k - \bar{Y}_k) + Y - Ny_k - G(Y + y_k N)\}, \end{aligned}$$

which is the result that was proposed in Monti (1991) and can be estimated by equation (12).

4.6. Demnati and Rao approach

A fast technique to obtain a direct linearized variable consists in computing the Deville influence function, not on the measure M but on the estimated measure $\hat{M}$. We then obtain

$$\mathrm{IT}(\hat{M}, x_k) = \lim_{t \to 0} \frac{T(\hat{M} + t\delta_x) - T(\hat{M})}{t}.$$

Measure $\hat{M}$ has a mass equal to w_k for each point x_k of the sample. If we refer to the definition of the derivative, we can note that a simple way to obtain a linearized variable is to differentiate the estimate with respect to w_k:

$$\mathrm{IT}(\hat{M}, y_k) = \frac{\partial T(\widehat{M})}{\partial w_k}.$$

The computation of a simple derivative with respect to the weights was advocated by Demnati and Rao (2004) to compute the linearized variable of a function of totals. This method also enables us to compute a linearized variable for any function of interest whose observations are weighted by w_k. By computing the derivative of expression (6) with respect to w_k, we obtain, in a couple of lines, the estimator of the linearized variable that is given in expression (12). The result is exactly the same as that obtained by Deville's method.

4.7. Graf approach

Recently Graf (2011) has proposed another way of computing the linearized variable by applying a Taylor series expansion with respect to the indicator variables I_k, where for all $k \in U$

$$I_k = \begin{cases} 1 & \text{if } k \in S, \\ 0 & \text{if } k \notin S, \end{cases}$$

determines the presence of unit k in the sample. The Graf method is coherent because the expansion is done with respect to the only source of randomness in the estimator. In sampling from a finite population, the estimator of the Gini index $\hat{G}$ can be written as a function of these

indicator variables: $\hat{G} = Q(I_1, \ldots, I_k, \ldots, I_N)$. By using a Taylor series expansion, we can then write

$$Q(I_1, \ldots, I_k, \ldots, I_N) \approx Q(\pi_1, \ldots, \pi_k, \ldots, \pi_N) + \sum_{k \in U} (I_k - \pi_k) Q'_k, \tag{13}$$

where $Q'_k(\cdot)$ is the partial derivative of $Q(I_1, \ldots, I_k, \ldots, I_N)$ with respect to I_k. Since

$$Q(\pi_k, \ldots, \pi_k, \ldots, \pi_N) = G,$$

we can then compute a linearized variable as

$$z_k = I_k Q'_k.$$

From expression (13), we obtain

$$\hat{G} - G \approx \sum_{k \in S} \frac{z_k}{\pi_k} - \sum_{k \in U} z_k.$$

If we use the Horvitz–Thompson weights $d_k = 1/\pi_k$, expression (6) can be written

$$\hat{G} = \frac{\sum_{k \in U} \sum_{l \in U} d_k d_l |y_k - y_l| I_k I_l}{2 \sum_{k \in U} d_k I_k \sum_{k \in U} d_k I_k y_k}.$$

The computation of $I_k Q'_k$ directly gives expression (12). Note that, in all the proposed linearization methods, if the weights are not fixed but depend on the I_ks (e.g. if they result from a calibration procedure) they must be differentiated as well. The advantage of the Graf approach, however, is that the effect of the calibration procedure on the estimator is directly accounted for when the weights are differentiated with respect to I_k.

4.8. Other recent publications

In survey sampling, Dell *et al.* (2002) and Osier (2006, 2009) also showed results based on the influence function in the sense of Deville. Both derived a linearized variable for various inequality and poverty measures including the Gini index as well as application to survey data. Although the whole technique and computation rules were attributed to Deville (1999) in Dell *et al.* (2002) and Osier (2006, 2009), the result for the Gini index was not presented as a problem already solved.

In the work of Cowell and Victoria-Feser (2003), the influence function result by Monti (1991) is presented and reference to Deville (1999) is made to show that influence functions can also be used for variance estimation, not only to study robustness. Finally, the computation of the influence function and the approximation of the variance of the Gini index by linearization are presented as well-known results in the works of Berger (2008) and Lesage (2009): the former referring to Kovacevic and Binder (1997); the latter citing Deville (1999).

Outside survey statistics, the same result by using different methodologies related to influence functions and linearization has also been republished at least three times in recent years (Bhattacharya, 2007; Davidson, 2009; Barrett and Donald, 2009). The common point of these papers is an obvious lack of references to prior papers from robust and survey statistics, nor to seminal papers like Hoeffding (1948) or Glasser (1962). The fact is that these papers all propose a short 'historical' introduction on the state of the art of the problem, in which the whole literature presented in Sections 3 and 4 is absent. They are briefly discussed below.

Bhattacharya (2007) has proposed asymptotic inference for the Gini index by using an asymp-

totic framework based on a generalized method of moments and empirical process theory. An influence function for the Gini index is derived and the method is extended to clustered and stratified sampling. Although the outcome is similar to that in previous references, the development and theoretical basis are different. Bhattacharya (2007) made no reference to previous research studies on the influence function of the Gini index. Instead Cowell (1989), Zheng (2001) or Bishop *et al.* (1997) and an earlier draft of Barrett and Donald (2009) were cited. Moreover, Bhattacharya noted that existing literature on inference for inequality measures has almost always assumed simple random sampling, whereas we have shown that many references in the survey sampling literature have tackled the issue of complex sampling designs.

In Davidson (2009), the delta method is used to produce an asymptotically valid expression for the variance of the Gini index. The paper starts with a brief history of the subject. Reference is made to recent papers (Bishop *et al.*, 1997; Xu, 2007) on the U-statistic approach instead of older papers like Hoeffding (1948), as well as references from the regression-based approach. From survey sampling literature, only Sandström *et al.* (1988) is cited. The proposed variance estimator for the Gini index is constructed via the delta method and the result that was obtained is equivalent to earlier suggestions in Monti (1991), Kovacevic and Binder (1997), Deville (1999) or Cowell and Victoria-Feser (1996). The numerical illustration applied in Davidson (2009) is of interest and is discussed in Section 6.

Barrett and Donald (2009) worked on a similar corpus of literature to that of Davidson (2009). Their work is said to take grounds in results by Cowell (1989), Bishop *et al.* (1997), Xu (2007), Zitikis and Gastwirth (2002) and Zitikis (2003). Barrett and Donald (2009) used influence functions to derive asymptotic variance expressions for generalized Gini indices (the 'E-Gini' and 'S-Gini' indices). Cowell and Victoria-Feser (1996) is referred to as one of the first applications of the influence function in econometrics. The main contribution of the paper is that the influence function is derived for a whole class of Gini indices and not only for the classical Gini index. However, references to previous literature on the influence function of the latter and its use for variance estimation were again nearly completely omitted.

5. Regression-based variance estimation

5.1. Ogwang's jackknife

A different way of expressing the Gini index has prompted a new wave of publications. In fact, it has been advocated already since the 1980s that the Gini index can be expressed by means of a covariance (Anand, 1983; Lerman and Yitzhaki, 1984; Shalit, 1985). With $F(y)$ denoting the cumulative distribution of incomes y, and $\mu(y)$ the mean income, the Gini index can thus be written as

$$G = \frac{2\,\mathrm{cov}\{y, F(y)\}}{\mu(y)}. \tag{14}$$

This idea has been exploited by Ogwang (2000) to derive a fast algorithm for the computation of jackknife estimates of the variance of the Gini index. Ogwang (2000) showed that, using equation (14) and sorting incomes in non-decreasing order, the Gini index estimated by means of a sample of size n is a function of a regression coefficient such that

$$\hat{G} = \frac{n^2 - 1}{6n}\frac{\hat{\beta}}{\bar{y}},$$

where $\bar{y} = n^{-1}\Sigma_{i=1}^{n} y_i$, $\hat{\beta}$ is the ordinary least square estimator of β from the regression model $y_i = \alpha + \beta i + u_i$, for $i = 1, \ldots, n$, and the error terms u_i are assumed to be homoscedastic with mean 0 and variance σ^2. Equivalently, the Gini index can be estimated by

$$\hat{G} = \frac{2}{n}\hat{\theta} - 1 - \frac{1}{n}, \tag{15}$$

where $\hat{\theta} = \Sigma_{i=1}^{n} i y_i / \Sigma_{i=1}^{n} y_i$ is the weighted least square estimator of θ in the regression model $i = \theta + u_i$ with heteroscedastic error u_i of variance σ^2 / y_i. Note that equation (15) is equivalent to equation (2). The algorithm that was proposed by Ogwang is fast because it avoids reranking incomes at each step, i.e. every time that an observation is dropped from the sample. The Ogwang jackknife variance estimator is

$$\widehat{\mathrm{var}}_{\mathrm{Oj}}(\hat{G}) = \frac{n-1}{n} \sum_{k=1}^{n} \left(\widehat{G^{(k)}} - \frac{1}{n} \sum_{l=1}^{n} \widehat{G^{(l)}} \right)^2, \tag{16}$$

where $\widehat{G^{(k)}}$ is the estimated Gini index of the remaining $n-1$ observations after deletion of unit k,

$$\widehat{G^{(k)}} = \hat{G} + \frac{2}{\sum_{i=1}^{n} y_i - y_k} \left\{ \frac{y_k \hat{\theta}}{n} + \frac{\sum_{i=1}^{n} i y_i}{n(n-1)} - \frac{\sum_{i=1}^{n} y_i - \sum_{i=1}^{k} y_i + k y_k}{n-1} \right\} - \frac{1}{n(n-1)}, \tag{17}$$

and can be computed as an n-component vector in only one pass through the data. It is crucial to note that $\widehat{G^{(k)}}$ takes into account the fact that the rank of y_i changes when observation k is dropped from the sample and $y_i \geqslant y_k$. Computing $\widehat{G^{(k)}}$ is thus equivalent to applying equation (15) successively to each sample of size $n-1$. The standard error $\mathrm{SE}_{\mathrm{Oj}}(\hat{G})$ is obtained simply by

$$\mathrm{SE}_{\mathrm{Oj}}(\hat{G}) = \sqrt{\widehat{\mathrm{var}}_{\mathrm{Oj}}(\hat{G})}. \tag{18}$$

Note that a jackknife estimator had already been applied with satisfactory results by Sandström *et al.* (1985, 1988) in their simulation study, as well as by others (for a review see Berger (2008)). In their approach, the Gini index is recomputed on the $n-1$ remaining observations for each jackknife subsample by using expression (7). The result is no different from that of Ogwang (2000), but the contribution of the latter is that, by using equation (17), computation becomes far less intensive because all $\widehat{G^{(k)}}$s can be computed at once while still taking the rank changes into account. This is an important feature when the sample size becomes large. A similar simplification was also proposed by Karagiannis and Kovacevic (2000).

5.2. *Direct regression*

Later, Giles (2004) suggested a direct analytical variance estimator for the Gini index based on regression theory. Indeed, as expression (15) shows, $\hat{G}$ is a function of a regression coefficient, and an estimator of the variance of the Gini index can thus be derived directly from the variance of the regression coefficient by

$$\widehat{\mathrm{var}}_{\mathrm{reg}}(\hat{G}) = \frac{4}{n^2} \widehat{\mathrm{var}}(\hat{\theta}), \tag{19}$$

and, accordingly, the standard error by

$$\mathrm{SE}_{\mathrm{reg}}(\hat{G}) = \frac{2}{n} \mathrm{SE}(\hat{\theta}),$$

where $\mathrm{SE}(\hat{\theta}) = \sqrt{\widehat{\mathrm{var}}(\hat{\theta})}$ can be obtained from any basic statistical package handling ordinary least squares or weighted least squares. Thereby, Giles advocated that using the jackknife method to derive a variance estimator is unnecessary. We argue that this idea is a deep methodologi-

cal error and that the variance estimator in equation (19) must not be used. Giles (2004) has been followed by a discussion concerning the violation of the model assumptions leading to a substantial overestimation of the variance (Ogwang, 2004, 2006; Giles, 2006; Modarres and Gastwirth, 2006). Above all, Modarres and Gastwirth (2006) argued that, because the y_is are ordered, they are dependent, and therefore the independence assumption for the error terms in the regression model does not hold. This dependence is totally ignored in Giles's proposal. In a classical regression, the independent variable is supposed to be non-random, which absolutely does not correspond to the problem of the variance of the Gini index. The solution of Giles (2004) is thus wrong since he committed the same error as in the first ratio estimator of Sandström *et al.* (1985, 1988). In other words, he fell into the pitfall that was described in Section 3.3. The discussion that follows Giles (2004) (see Ogwang (2006), Modarres and Gastwirth (2006) and Giles (2006)) is a little vain since this problem had already been identified by Deville (1996, 1999). Indeed, as discussed earlier in this paper, the major issue in leading reliable inference for the Gini index is that the population rank N_k of unit k must be estimated by means of the sample. Also, the model proposed does not correspond to the way that the data have been produced or could be modelled, because it assumes that the independent variable is not random. Moreover, as pointed out by Berger (2008), the regression approach takes no account of the sampling design.

6. Empirical comparisons

Giles (2004) proposed an empirical illustration of the regression-based method by using data that are available from Heston *et al.* (1995) for 133 countries. The variable of interest is, for each country, the real consumption *per capita* in constant US dollars (international prices based on year 1985). The Gini index and its standard error were computed for four different time periods (1970, 1975, 1980 and 1985). In this numerical application, Giles (2004) compared $\mathrm{SE}_{\mathrm{reg}}(\hat{G})$ with a jackknife variance estimator which Giles incorrectly attributed to Ogwang (2000) and showed that both variance estimators behave similarly. However, we shall show hereafter that the jackknife estimator that was computed by Giles is by no means equal to $\mathrm{SE}_{\mathrm{Oj}}(\hat{G})$ of expression (18). Recently, Davidson (2009) used the same data set to compare a linearization variance estimator based on the delta method with both Giles's and Ogwang's approaches. Note that the linearization estimator that was proposed in Davidson (2009) is equivalent to expression (9), the only difference being the way that the cumulative distribution function is estimated in $\hat{z}_k$. Because we have witnessed only a negligible effect in the empirical applications that are proposed later in this paper, we simply consider both of them under the term linearization variance estimator and denoted $\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{G})$. We also define

$$\mathrm{SE}_{\mathrm{lin}}(\hat{G}) = \sqrt{\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{G})},$$

because standard errors, rather than variances, were reported in both Giles (2004) and Davidson (2009).

The results of these empirical illustrations are confusing and the conclusions unclear. Indeed, Giles (2004), who in addition also conducted a simulation, argued that the jackknife approach produces biased estimates in samples smaller than 5000 observations, where the regression approach is more reliable, and that both methods tend to converge as the sample size increases. Thus, the use of the regression method was advocated by Giles because it is computationally much simpler and allows for some robustness testing.

Davidson (2009) argued in contrast that both methods might be unreliable. When reproducing the empirical application of Giles, he obtained the same numerical results as the latter and both approaches are shown to yield much larger variance estimators than the linearization method.

When analysing the results, however, Davidson (2009) did not clearly discard the jackknife and regression estimators. No clear comment is made on the quality of the three estimators and on the discrepancies between them.

Hereafter, we show that the regression approach is conceptually wrong and that, when used correctly as suggested by Ogwang (2000), the jackknife provides reliable estimates in addition to being computationally straightforward. Both Giles (2004) and Davidson (2009) used the same method for computing the jackknife variance estimator and, obviously, the estimator is not that proposed in equation (16). The jackknife approach that was used in Giles (2004) and Davidson (2009) is hereafter denoted the Davidson–Giles jackknife DGj and the resulting variance estimator is

$$\widehat{\mathrm{var}}_{\mathrm{DGj}}(\hat{G}) = \frac{n-1}{n} \sum_{k=1}^{n} \left(\widehat{\widetilde{G^{(k)}}} - \frac{1}{n} \sum_{l=1}^{n} \widehat{\widetilde{G^{(k)}}} \right)^2, \tag{20}$$

where $\widehat{\widetilde{G^{(k)}}}$ is the Gini index computed by using expression (15) directly on the $n-1$ values without taking into account the effect of dropping unit k on the ranks of the other sampled units (i.e. without recomputing $\hat{\theta}$ each time that a unit is dropped). Likewise, the standard error is

$$\mathrm{SE}_{\mathrm{DGj}}(\hat{G}) = \sqrt{\widehat{\mathrm{var}}_{\mathrm{DGj}}(\hat{G})}.$$

Within the procedure of Davidson and Giles, the rank of observation y_i is kept constant among all the n jackknife subsamples, whereas in fact, as discussed in Section 3.3, the ranks depend on the sample and should thus be recalculated within each jackknife subsample. This issue is taken into account in equation (16) but not in equation (20). In our empirical study below, we show that equation (20) leads to a heavy overestimation of the variance, as was the case with the direct regression approach. Indeed, using this version of the jackknife is nothing other than falling again into the pitfall that was described in Section 3.3. In addition to the fact that the jackknife estimator is computed incorrectly, these empirical illustrations have another major issue: because the set-up is not that of a repeated random-sampling simulation method, the estimates obtained cannot be compared with a Monte Carlo variance which would mimic the true value. Thus, it is not possible to assess the genuine reliability of the variance estimators proposed.

To take these issues into account, we first repeat the exact empirical set-up that was proposed by Giles (2004) and Davidson (2009) and add a fourth estimator, the Ogwang jackknife estimator $\mathrm{SE}_{\mathrm{Oj}}(\hat{G})$. Subsequently, we use the same data to perform a Monte Carlo simulation study using two different sample sizes. In both numerical illustrations, only data for the year 1970 are used.

Table 1 summarizes the results of the first replication. The figures of the previous studies are reported for comparison. Note that the results we produced are exactly the same as in Davidson (2009) and that they differ very slightly from those of Giles (2004), probably because the data were processed differently to create the desired real consumption *per capita* variable. These results show that the Ogwang jackknife algorithm gives a variance estimator that is very close to the linearization estimator and very different from the other two estimators, showing that taking the changes of the ranks into account in the jackknife procedure has a massive effect on variance estimation.

We then proceed to a Monte Carlo simulation set-up on the same data. 10 000 simple random samples of size $n = 30$ are drawn without replacement from the full data of $N = 133$ countries. On each sample, the Gini index $\hat{G}$ as well as the four competing variance estimators are computed. The Monte Carlo standard error of the Gini index, which is denoted $\mathrm{SE}_{\mathrm{sim}}(\hat{G})$, is then com-

Table 1. Standard errors of $\hat{G}$ by using different approaches (data from Heston *et al.* (1995), year 1970) and comparing results obtained in Giles (2004), Davidson (2009) and in our replication

Method	$SE_{\mathrm{reg}}(\hat{G})$, *regression approach*	$SE_{\mathrm{DGj}}(\hat{G})$, *Davidson–Giles jackknife*	$SE_{\mathrm{lin}}(\hat{G})$, *linearization approach*	$SE_{\mathrm{Oj}}(\hat{G})$, *Ogwang jackknife*
Giles (2004)	0.0417	0.0481		
Davidson (2009)	0.0418	0.0478	0.0173	
Our replication	0.0418	0.0478	0.0173	0.0176

Table 2. Simulation results: standard errors of $\hat{G}$ by using various approaches†

Method		*Standard errors for the following values of n:*	
		$n = 30$	$n = 100$
Regression approach	$E_{\mathrm{sim}}[\mathrm{SE}_{\mathrm{reg}}(\hat{G})]$	0.0896	0.0483
Davidson–Giles jackknife	$E_{\mathrm{sim}}[\mathrm{SE}_{\mathrm{DGj}}(\hat{G})]$	0.1066	0.0554
Linearization approach	$E_{\mathrm{sim}}[\mathrm{SE}_{\mathrm{lin}}(\hat{G})]$	0.0333	0.0100
Ogwang jackknife	$E_{\mathrm{sim}}[\mathrm{SE}_{\mathrm{Oj}}(\hat{G})]$	0.0355	0.0102
Monte Carlo estimator	$\mathrm{SE}_{\mathrm{sim}}(\hat{G})$	0.0337	0.0101

†Data from Heston *et al.* (1995), year 1970; 10000 replicates; simple random sampling without replacement.

puted on the 10000 estimators of G obtained. The value of $\mathrm{SE}_{\mathrm{sim}}(\hat{G})$ is the benchmark to which the four estimators can be compared. The exact same simulation set-up was also conducted with a larger sample size of $n = 100$. In Table 2, the Monte Carlo expected value $E_{\mathrm{sim}}[\mathrm{SE}(\hat{G})]$ for each method is reported. The right-hand column contains the Monte Carlo value which approximates the true value for each sample size. Results show that the linearization technique as well as the jackknife provide reliable estimates for the variance of the Gini index. Indeed, both methods give estimators that are close to the Monte Carlo estimator. Unlike what has been reported previously, the jackknife procedure that was proposed by Ogwang (2000) is valid if applied correctly. On the contrary, the two methods that do not account for the rank issue yield unsatisfactory results. As suggested earlier in the paper, the variance is greatly overestimated when these approaches are used. In our simulations, for example, the variance that is estimated by the regression approach is overestimated by a factor of more than 20 with a sample size of $n = 100$. We thus point out that the main problem with the regression-based variance estimation is that the rank is not identified as a major source of randomness, leading to overestimation. The method should therefore be avoided.

7. Conclusion

As this paper tries to illustrate, the computation of variance of the Gini index has been subject to many publications. For numerous reasons, the same results have been republished several

times. The segmentation of the different research fields and the multiplicity of methods that can be applied to obtain an approximation of variance are probably the main causes of this phenomenon. Indeed, an examination of the references in the publications clearly shows the lack of communication between statisticians, economists and survey statisticians. As a consequence, the same 'mistakes' have been reproduced when tackling the problem, an example being the recent regression-based variance estimator that was suggested by Giles (2004).

In addition to giving a small contribution (the linearization of a double sum) which tends to simplify the problem, we try to propose a global picture of the state of the art regarding inference for the Gini index. We hope that the paper clarifies the situation to a much larger extent than previous survey papers on the Gini index have. We also hope that it will provide researchers from different fields with an update on what is done elsewhere as well as a comprehensive survey of the literature. Finally, we hope that it will enlighten the reader on the interesting issues regarding inference on a non-linear statistic and on the various possible approaches leading to the result.

Acknowledgements

The authors thank the Joint Editor and three referees for their insightful comments and suggestions. This research is supported by grant 200021-121604 of the Swiss National Science Foundation.

Appendix A: Proof of result 3

For direct computation of the linearized variable of the double sum (e.g. quadratic form)

$$S = \sum_{k \in U} \sum_{l \in U} \phi(y_k, y_l),$$

let

$$\begin{aligned}
C(t) &= \frac{1}{t}\{S(M + t\delta_\xi) - S(M)\} \\
&= \frac{1}{t}\left\{\int\int \phi(x,y)\,\mathrm{d}(M + t\delta_\xi)(x)\,\mathrm{d}(M + t\delta_\xi)(y) - \int\int \phi(x,y)\,\mathrm{d}M(x)\,\mathrm{d}M(y)\right\} \\
&= \frac{1}{t}\int\left\{\int \phi(x,y)\,\mathrm{d}(M + t\delta_\xi)(x) - \int \phi(x,y)\,\mathrm{d}M(x)\right\}\mathrm{d}M(y) \\
&\quad + \frac{1}{t}\int\int \phi(x,y)\,\mathrm{d}(M + t\delta_\xi)(x)\,\mathrm{d}(t\delta_\xi)(y) \\
&= \frac{1}{t}\left\{\int\int \phi(x,y)\,\mathrm{d}M(y)\,\mathrm{d}(M + t\delta_\xi)(x) - \int\int \phi(x,y)\,\mathrm{d}M(y)\,\mathrm{d}M(x)\right\} \\
&\quad + \int \phi(x,\xi)\,\mathrm{d}(M + t\delta_\xi)(x).
\end{aligned}$$

We obtain

$$\mathrm{IS}(M,\xi) = \lim_{t \to 0} C(t) = \int \phi(\xi, y)\,\mathrm{d}M(y) + \int \phi(x,\xi)\,\mathrm{d}M(x).$$

References

Anand, S. (1983) *Inequality and Poverty in Malaysia: Measurement and Decomposition*. New York: Oxford University Press.

Ardilly, P. and Tillé, Y. (2006) *Sampling Methods: Exercises and Solutions*. New York: Springer.

Barrett, G. F. and Donald, S. G. (2009) Statistical inference with generalized Gini indices of inequality, poverty, and welfare. *J. Bus. Econ. Statist.*, **27**, 1–17.
Beach, C. M. and Davidson, R. (1983) Distribution-free statistical inference with Lorenz curves and income shares. *Rev. Econ. Stud.*, **50**, 723–735.
Berger, Y. G. (2008) A note on asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *J. Off. Statist.*, **24**, 541–555.
Bhattacharya, D. (2007) Inference on inequality from household survey data. *J. Econmetr.*, **137**, 674–707.
Binder, D. A. and Kovacevic, M. S. (1995) Estimating some measures of income inequality from survey data: an application of the estimating equation approach. *Surv. Methodol.*, **21**, 137–145.
Binder, D. A. and Patak, Z. (1994) Use of estimating functions for estimation from complex surveys. *J. Am. Statist. Ass.*, **89**, 1035–1043.
Bishop, J. A., Formby, J. and Zheng, B. (1997) Statistical inference and the Sen index of poverty. *Int. Econ. Rev.*, **38**, 381–387.
Ceriani, L. and Verme, P. (2011) The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J. Econ. Ineq.*, to be published, doi 10.1007/s10888-011-9188-x.
Cowell, F. A. (1989) Sampling variance and decomposable inequality measures. *J. Econmetr.*, **42**, 27–41.
Cowell, F. A. and Victoria-Feser, M.-P. (1996) Robustness properties of inequality measures. *Econometrica*, **64**, 77–101.
Cowell, F. A. and Victoria-Feser, M.-P. (2003) Distribution-free inference for welfare indices under complete and incomplete information. *J. Econ. Ineq.*, **1**, 191–219.
Davidson, R. (2009) Reliable inference for the Gini index. *J. Econmetr.*, **150**, 30–40.
Dell, F., d'Haultfoeuille, X., Février, P. and Massé, E. (2002) Mise en oeuvre du calcul de variance par linéarisation. *Actes Journ. Methodol. Statist.*, 73–104.
Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data (with discussion). *Surv. Methodol.*, **30**, 17–34.
Deville, J.-C. (1996) Estimation de la variance du coefficient de Gini estimé par sondage. *Actes Journ. Methodol. Statist.*, **69–70–71**, 269–288.
Deville, J.-C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Surv. Methodol.*, **25**, 193–204.
Deville, J.-C. and Särndal, C.-E. (1992) Calibration estimators in survey sampling. *J. Am. Statist. Ass.*, **87**, 376–382.
Fei, J. C. H., Ranis, G. and Kuo, S. W. Y. (1978) Growth and the family distribution of income by factor components. *Q. J. Econ.*, **92**, 17–53.
Gastwirth, J. L. (1972) The estimation of the Lorenz curve and Gini index. *Rev. Econ. Statist.*, **54**, 306–316.
Giles, D. E. A. (2004) Calculating a standard error for the Gini coefficient: some further results. *Oxf. Bull. Econ. Statist.*, **66**, 425–433.
Giles, D. E. A. (2006) A cautionary note on estimating the standard error of the Gini index of inequality: comment. *Oxf. Bull. Econ. Statist.*, **68**, 395–396.
Gini, C. (1912) *Variabilità e Mutabilità*. Bologna: Tipografia di Paolo Cuppin.
Gini, C. (1914) Sulla misura della concertrazione e della variabilità dei caratteri. *Atti R. Ist. Ven. Sci. Lett. Arti*, **73**, 1203–1248.
Gini, C. (1921) Measurement of inequality and incomes. *Econ. J.*, **31**, 124–126.
Giorgi, G. M., Palmitesta, P. and Provasi, C. (2006) Asymptotic and bootstrap inference for the generalized gini indices. *Metron*, **64**, 107–124.
Glasser, G. (1962) Variance formulas for the mean difference and coefficient of concentration. *J. Am. Statist. Ass.*, **57**, 648–654.
Graf, M. (2011) Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis: Theory and Applications* (eds V. Pawlowsky-Glahn and A. Buccianti). Chichester: Wiley.
Halmos, P. R. (1946) The theory of unbiased estimation. *Ann. Math. Statist.*, **17**, 34–43.
Hampel, F. R. (1974) The influence curve and its role in robust estimation. *J. Am. Statist. Ass.*, **69**, 383–393.
Hampel, F. R., Ronchetti, E., Rousseuw, P. J. and Stahel, W. (1985) *Robust Statistics: the Approach based on the Influence Function*. New York: Wiley.
Heston, A., Summers, R. and Aten, B. (1995) Penn World Table. *Technical Report.* Center for International Comparisons of Production, Income and Prices, University of Pennsylvania, Philadelphia.
Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, **19**, 293–325.
Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.
Hosking, J. R. M. (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Statist. Soc.* B, **52**, 105–124.
Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages. *Am. Statistn*, **50**, 361–365.
Isaki, C. T. and Fuller, W. A. (1982) Survey design under a regression population model. *J. Am. Statist. Ass.*, **77**, 89–96.

Jaeckel, L. (1972) The infinitesimal jackknife. *Memorandum MM 72-1215-11*. Bell Laboratories, Murray Hill.
Karagiannis, E. and Kovacevic, M. (2000) A method to calculate the jackknife variance estimator for the Gini coefficient. *Oxf. Bull. Econ. Statist.*, **62**, 119–122.
Kovacevic, M. S. and Binder, D. A. (1997) Variance estimation for measures of income inequality and polarization—the estimating equations approach. *J. Off. Statist.*, **13**, 41–58.
Kuan, X. (2000) Inference for generalized Gini indices using the iterated bootstrap method. *J. Bus. Econ. Statist.*, **18**, 223–227.
Langel, M. and Tillé, Y. (2011) Statistical inference for the quintile share ratio. *J. Statist. Planng Inf.*, **141**, 2976–2985.
Lerman, R. I. and Yitzhaki, S. (1984) A note on the calculation and interpretation of the Gini index. *Econ. Lett.*, **15**, 363–368.
Lesage, E. (2009) Calage non linéaire. *Actes Xème Journ. Methodol. Statist*.
Lomnicki, Z. A. (1952) The standard error of Gini's mean difference. *Ann. Math. Statist.*, **23**, 635–637.
Lorenz, M. O. (1905) Methods of measuring the concentration of wealth. *Publ. Am. Statist. Ass.*, **9**, 209–219.
Mills, J. A. and Zandvakili, S. (1997) Statistical inference via bootstrapping for measures of inequality. *J. Appl. Econmetr.*, **12**, 133–150.
Modarres, R. and Gastwirth, J. L. (2006) A cautionary note on estimating the standard error of the Gini index of inequality. *Oxf. Bull. Econ. Statist.*, **68**, 385–390.
Monti, A. C. (1991) The study of the Gini concentration ratio by means of the influence function. *Statistica*, **51**, 561–577.
Nair, U. S. (1936) The standard error of Gini's mean difference. *Biometrika*, **28**, 428–436.
Nygård, F. and Sandström, A. (1981) *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell.
Nygård, F. and Sandström, A. (1985) The estimation of the Gini and the entropy inequality parameters in finite populations. *J. Off. Statist.*, **1**, 399–412.
Nygård, F. and Sandström, A. (1989) Income inequality measures based on sample surveys. *J. Econmetr.*, **42**, 81–95.
Ogwang, T. (2000) A convenient method of computing the Gini index and its standard error. *Oxf. Bull. Econ. Statist.*, **62**, 123–129.
Ogwang, T. (2004) Calculating a standard error for the Gini coefficient: some further results: reply. *Oxf. Bull. Econ. Statist.*, **66**, 435–437.
Ogwang, T. (2006) A cautionary note on estimating the standard error of the Gini index of inequality: comment. *Oxf. Bull. Econ. Statist.*, **68**, 391–393.
Osier, G. (2006) Variance estimation: the linearization approach applied by Eurostat to the 2004 SILC operation. In *Proc. Methodological Wrkshp European Statistics on Income and Living Conditions*. Helsinki: Eurostat and Statistics Finland.
Osier, G. (2009) Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Surv. Res. Meth.*, **3**, 167–195.
Peng, L. (2011) Empirical likelihood methods for the gini index. *Aust. New Zeal. J. Statist.*, **53**, 131–139.
Qin, Y., Rao, J. and Wu, C. (2010) Empirical likelihood confidence intervals for the gini measure of income inequality. *Econ. Modllng*, **27**, 1429–1435.
Sandström, A., Wretman, J. H. and Walden, B. (1985) Variance estimators of the Gini coefficient: simple random sampling. *Metron*, **43**, 41–70.
Sandström, A., Wretman, J. H. and Walden, B. (1988) Variance estimators of the Gini coefficient: probability sampling. *J. Bus. Econ. Statist.*, **6**, 113–120.
Schechtman, E. (1991) On estimating the asymptotic variance of a function of U-statistics. *Am. Statistn*, **45**, 103–106.
Sen, A. K. (1973) *On Economic Inequality*. Oxford: Clarendon.
Sendler, W. (1979) On statistical inference in concentration measurement. *Metrika*, **26**, 109–122.
Shalit, H. (1985) Calculating the Gini index of inequality for individual data. *Oxf. Bull. Econ. Statist.*, **47**, 185–189.
Shorack, G. (1972) Functions of order statistics. *Ann. Math. Statist.*, **43**, 412–427.
Sillitto, G. P. (1969) Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample. *Biometrika*, **56**, 641–650.
Woodruff, R. S. (1971) A simple method for approximating the variance of a complicated estimate. *J. Am. Statist. Ass.*, **66**, 411–414.
Xu, K. (2004) How has the literature on Gini's index evolved in the past 80 years? *Working Paper*. Department of Economics, Dalhousie University, Halifax.
Xu, K. (2007) U-statistics and their asymptotic results for some inequality and poverty measures. *Econmetr. Rev.*, **26**, 567–577.
Zheng, B. (2001) Statistical inference for poverty measures with relative poverty lines. *J. Econmetr.*, **101**, 337–356.
Zitikis, R. (2003) Asymptotic estimation of the E-Gini index. *Econmetr. Theor.*, **19**, 587–601.
Zitikis, R. and Gastwirth, J. L. (2002) The asymptotic distribution of the S-Gini index. *Aust. New Zeal. J. Statist.*, **44**, 439–446.