Journal of Econometrics 150 (2009) 30–40

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

# Reliable inference for the Gini index

Russell Davidson [*]

*Department of Economics and CIREQ, McGill University, Montréal, Québec, Canada H3A 2T7*
*GREQAM, Centre de la Vieille Charité, 2 Rue de la Charité, 13236 Marseille cedex 02, France*

ARTICLE INFO

*Article history:*
Received 25 January 2008
Received in revised form
4 October 2008
Accepted 21 November 2008
Available online 20 February 2009

*JEL classification:*
C12
C13
C81
D31
I32

*Keywords:*
Gini index
Delta method
Asymptotic inference
Jackknife
Bootstrap

ABSTRACT

Although attention has been given to obtaining reliable standard errors for the plug-in estimator of the Gini index, all standard errors suggested until now are either complicated or quite unreliable. An approximation is derived for the estimator by which it is expressed as a sum of IID random variables. This approximation allows us to develop a reliable standard error that is simple to compute. A simple but effective bias correction is also derived. The quality of inference based on the approximation is checked in a number of simulation experiments, and is found to be very good unless the tail of the underlying distribution is heavy. Bootstrap methods are presented which alleviate this problem except in cases in which the variance is very large or fails to exist. Similar methods can be used to find reliable standard errors of other indices which are not simply linear functionals of the distribution function, such as Sen's poverty index and its modification known as the Sen–Shorrocks–Thon index.

## 1. Introduction

Some attention has been given recently to the standard error of a Gini index estimated by a plug-in estimator with no distributional assumptions. Quite a number of techniques for computing an asymptotically valid standard error have been proposed, of varying degrees of complexity or computational intensiveness. Sandström et al. (1988) discuss estimation of the Gini coefficient with arbitrary probability sampling design, and then propose three ways to compute a standard error. The first is from a complicated analytic formula, the second is based on the jackknife, and the third is discarded as "quite useless".

More recently, Bishop et al. (1997) have given a discussion of the variance of the Gini index in the context of estimating Sen's index of poverty; their approach is based on U-statistics, as is also that of Xu (2007). Ogwang (2000) provided a method for computing the Gini index by an OLS regression, and discussed how to use this regression to simplify the computation of the jackknife standard error. Then Giles (2004) claimed that the OLS standard error from this regression could be used directly in order to compute the standard error of the Gini index itself. See also the reply by Ogwang (2004).

Subsequently, Modarres and Gastwirth (2006) struck a cautionary note on the use of Giles's approach, showing by simulation that the standard errors it produces are quite inaccurate. They recommended a return to the complex or computationally intensive methods used previously, and, in their replies, Ogwang (2006) and Giles (2006) did not fundamentally disagree with the criticism. More recently still, Bhattacharya (2007) has developed techniques of asymptotic inference for Lorenz curves and the Gini index with stratified and clustered survey data. These techniques are based on sample empirical process theory and the functional delta method, and they lead to a formula for the variance of an estimated Gini index, which is however not at all easy to implement.

This paper shows how to compute an asymptotically correct standard error for an estimated Gini index, based on a reasonably simple formula that is very easy to compute. The proposed standard error is based on the delta method, but makes no use of empirical process theory. The approach also provides a simple and effective bias correction for the estimate of the index. The methods used can be extended to other commonly used indices, including Sen's (1976) poverty index, and the modification of it proposed

[*] Corresponding address: Department of Economics and CIREQ, McGill University, Montréal, Québec, Canada H3A 2T7. Tel.: +1 514 398 4835; fax: +1 514 398 4938.

*E-mail addresses:* russell.davidson@mcgill.ca, russell.davidson@univmed.fr.

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.11.004

by Shorrocks (1995), often referred to as the Sen–Shorrocks–Thon (SST) index.

In Section 2, we review some well-known properties of the Gini index, and give an expression for the Gini index of a sample. This is then related to the regression proposed by Ogwang (2000). Then, in Section 3, an asymptotic approximation for the usual plug-in estimator of the index is derived. This approximation shows that the estimator is asymptotically normal, since it takes the form of a sum of IID random variables. In Section 4, inference based on the estimate is investigated. The asymptotic variance is easily found from the approximation, and it is shown how it can easily be estimated from the sample. Bias is studied next, and a simple bias correction proposed. Section 5 considers the jackknife as an alternative way of doing bias correction and variance estimation. It is found that the jackknife does not give reliable inference. The bootstrap is discussed briefly in Section 6. Unlike the jackknife, the bootstrap can yield reasonably reliable inference. Section 7 provides simulation evidence that bears out the main conclusions of the paper, and reveals their limitations when used with heavy-tailed distributions. The empirical study given in Giles (2004) is redone in Section 8 so as to make clear how the methods of this paper differ from those used by Giles. In Section 9, the methods of the paper are used to find the asymptotic variance of Sen's (1976) poverty index and the SST variant. Section 10 concludes.

## 2. Properties of the Gini index

The classical definition of the Gini index of inequality is twice the area between the 45°-line and the Lorenz curve. If we denote by $F$ the cumulative distribution function (CDF) of the incomes under study, the Lorenz curve is defined implicitly by the equation

$$L\left(F(x)\right) = \frac{1}{\mu}\int_0^x y\mathrm{d}F(y), \tag{1}$$

where $\mu \equiv \int_0^\infty y\,\mathrm{d}F(y)$ is expected income. It is assumed that there are no negative incomes. The function $L$ is increasing and convex, and maps the $[0, 1]$ interval into itself. Twice the area between the graph of $L$ and the 45°-line is then

$$G = 1 - 2\int_0^1 L(y)\mathrm{d}y. \tag{2}$$

Using the definition (1) in (2), we find that

$$G = 1 - 2\int_0^\infty L\left(F(x)\right)\mathrm{d}F(x) = 1 - \frac{2}{\mu}\int_0^\infty\int_0^x y\mathrm{d}F(y)\mathrm{d}F(x).$$

Then, on interchanging the order of integration and simplifying, we obtain

$$\begin{aligned} G &= 1 - \frac{2}{\mu}\int_0^\infty y\int_y^\infty \mathrm{d}F(x)\mathrm{d}F(y) \\ &= 1 - \frac{2}{\mu}\int_0^\infty y\left(1 - F(y)\right)\mathrm{d}F(y) \\ &= 1 + \frac{2}{\mu}\int_0^\infty yF(y)\mathrm{d}F(y) - 2 = \frac{2}{\mu}\int_0^\infty yF(y)\mathrm{d}F(y) - 1. \end{aligned} \tag{3}$$

The last expression above corresponds to a result cited in Modarres and Gastwirth (2006) according to which $G$ is $2/\mu$ times the covariance of $Y$ and $F(Y)$, where $Y$ denotes the random variable "income" of which the CDF is $F$. There are of course numerous other ways of expressing the index $G$, but (3) is most convenient for present purposes. See Appendix A for further discussion of this point.

Suppose now that an IID sample of size $n$ is drawn randomly from the population, and let its empirical distribution function (EDF) be denoted as $\hat{F}$. The natural plug-in estimator of $G$ is then $\hat{G}$, defined as

$$\hat{G} = \frac{2}{\hat{\mu}}\int_0^\infty y\hat{F}(y)\mathrm{d}\hat{F}(y) - 1. \tag{4}$$

Evaluating $\hat{G}$ using (4) reveals an ambiguity: different answers are obtained if the EDF is defined to be right- or left-continuous. The ambiguity can be resolved by splitting the difference, or by noting that we can write

$$\begin{aligned} \hat{G} &= \frac{1}{\hat{\mu}}\int_0^\infty y\mathrm{d}\left(\hat{F}(y)\right)^2 - 1 \\ &= \frac{1}{\hat{\mu}}\sum_{i=1}^n y_{(i)}\left(\left(\frac{i}{n}\right)^2 - \left(\frac{i-1}{n}\right)^2\right) - 1 \\ &= \frac{2}{\hat{\mu}n^2}\sum_{i=1}^n y_{(i)}\left(i - \frac{1}{2}\right) - 1. \end{aligned} \tag{5}$$

Here the $y_{(i)}, i = 1, \ldots, n$, are the order statistics. The definition (5) has the advantage over alternative possibilities that, when $y_{(i)} = \hat{\mu}$ for every $i$, $\hat{G} = 0$.

In order to compute $\hat{G}$ itself, Ogwang (2000) suggested the use of the regression

$$i = \theta + u_i, \quad i = 1, \ldots, n, \tag{6}$$

estimated by weighted least squares under the assumption that the variance of $u_i$ is proportional to $1/y_{(i)}$. The parameter estimate $\hat{\theta}$ is then

$$\hat{\theta} = \left(\sum_{i=1}^n y_i\right)^{-1}\sum_{i=1}^n iy_{(i)}.$$

It is easy to check that $\hat{G}$, as given by (5), is equal to $2\hat{\theta}/n - 1 - 1/n$. Giles (2004) reformulated the weighted regression as

$$i\sqrt{y_{(i)}} = \theta\sqrt{y_{(i)}} + v_i, \quad i = 1, \ldots, n, \tag{7}$$

now to be estimated by OLS. His proposal was then simply to use the OLS standard error, multiplied by $2/n$, as the standard error of $\hat{G}$. As pointed out by Modarres and Gastwirth (2006), however, the fact that the order statistics are correlated means that the OLS standard error may be unreliable.

## 3. An asymptotic expression for the Gini index

Standard arguments show that the estimator (4) is consistent under weak regularity conditions. Among these, we require the existence of the second moment of the distribution characterised by $F$. This is not quite enough, as the class of admissible CDFs $F$ must be further restricted so as to avoid the Bahadur–Savage problem; see Bahadur and Savage (1956). Asymptotic normality calls for a little more regularity, but not a great deal. In this section, we examine the quantity $n^{1/2}(\hat{G} - G)$ that should be asymptotically normal under the required regularity, and derive the variance of its limiting distribution as $n \to \infty$.

Let

$$I \equiv \int_0^\infty yF(y)\mathrm{d}F(y) \quad \text{and} \quad \hat{I} \equiv \int_0^\infty y\hat{F}(y)\mathrm{d}\hat{F}(y). \tag{8}$$

Notice that the integral defining $I$ exists if we assume that the first moment of $F$ exists, since $F(y)$ is bounded above by 1. Then we have

$$\begin{aligned} n^{1/2}(\hat{G} - G) &= n^{1/2}\left(\frac{2\hat{I}}{\hat{\mu}} - \frac{2I}{\mu}\right) = n^{1/2}\frac{2}{\mu\hat{\mu}}(\mu\hat{I} - \hat{\mu}I) \\ &= \frac{2}{\mu\hat{\mu}}\left(\mu n^{1/2}(\hat{I} - I) - In^{1/2}(\hat{\mu} - \mu)\right). \end{aligned} \tag{9}$$