Economic Modelling 27 (2010) 1429–1435

Contents lists available at ScienceDirect

Economic Modelling

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

Empirical likelihood confidence intervals for the Gini measure of income inequality

Yongsong Qin [a], J.N.K. Rao [b,*], Changbao Wu [c]

[a] Department of Mathematics, Guangxi Normal University, Guilin, Guanxi, 541004, China
[b] School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6
[c] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1

ARTICLE INFO

Keywords:
Bootstrap percentile
Bootstrap-t
Confidence interval
Coverage probability
Gini coefficient
Normal approximation

ABSTRACT

Gini coefficient is among the most popular and widely used measures of income inequality in economic studies, with various extensions and applications in finance and other related areas. This paper studies confidence intervals on the Gini coefficient for simple random samples, using normal approximation, bootstrap percentile, bootstrap-t and the empirical likelihood method. Through both theory and simulation studies it is shown that the intervals based on normal or bootstrap approximation are less satisfactory for samples of small or moderate size than the bootstrap-calibrated empirical likelihood ratio confidence intervals which perform well for all sample sizes. Results for stratified random sampling are also presented.

1. Introduction

Income inequality has long been an active research area in economic studies. Among various measures of income inequality proposed in the statistical and economic literature, the Gini coefficient, G, is probably the most popular and widely used measure. It was originated from Gini's mean difference (Gini 1912, 1936), and is closely related to the Lorenz curve, the popular measure for the size distribution of income and wealth. Lorenz curves are also widely used in economic analysis (Kakwani, 1977).

Let $F(y) = P(Y \le y)$ be the cumulative distribution function of a nonnegative continuous random variable Y. We will refer to Y as the income variable. Let X and Y be two independent random variables following the same distribution $F(y)$. The Gini mean difference is then defined as

$$D = E|X-Y| = \int_0^{+\infty}\int_0^{+\infty}|x-y|dF(x)dF(y).$$

The value of D is the average absolute difference of incomes of two randomly selected individuals and hence reflects the income inequality in the population. Noting that $0 \le D \le 2\mu$, where $\mu = E(Y) = \int_0^{+\infty} ydF(y)$ is the population mean income, the Gini coefficient, G, is defined as the normalized mean difference, i.e., $G = D/(2\mu) \in [0,\ 1]$, which can be equivalently written as (David, 1968)

$$G = \frac{1}{\mu}\int_0^{+\infty}\{2F(y)-1\}ydF(y). \tag{1}$$

The Gini coefficient is also closely related to another popular measure of income inequality, the Lorenz curve (Lorenz, 1905; Sendler, 1979). Let $F^{-1}(t) = \inf\{\xi : F(\xi) \ge t\}$ for $t \in [0,\ 1]$. The Lorenz curve based on the income distribution $F(\cdot)$ is then defined as

$$L(\alpha, F) = \frac{1}{\mu}\int_0^{\alpha} F^{-1}(t)dt = \frac{1}{\mu}\int_0^{F^{-1}(\alpha)} xdF(x)$$

for $\alpha \in [0,\ 1]$. The Gini coefficient G is equal to twice the area between a 45-degree line and the Lorenz curve, i.e., $G = 2\left\{0.5 - \int_0^1 L(\alpha, F)d\alpha\right\}$.

There exists an extensive literature on the Gini measure of income inequality. In addition to various applications and extensions in economic studies, statistical investigations focused largely on variance estimation; see, for instance, Glasser (1962), Sandström et al. (1985, 1988), Yitzhaki (1991), Karagiannis and Kovacevic (2000), among others. In particular, Yitzhaki (1991) calculated jackknife variance estimators of the plug-in moment estimator, $\hat{G}$, of G, under simple random sampling and stratified random sampling. However, confidence intervals for the Gini coefficient have not been studied by previous authors, with the exception of Sandström et al. (1988) where 95% normal approximation confidence intervals based on three variance estimators were briefly mentioned.

This paper presents confidence intervals on the Gini coefficient, G, using normal and bootstrap approximations and empirical likelihood (EL) based methods. We first consider the case of independent and identically distributed (*iid*) samples (or simple random samples when the sampling fraction is negligible), and then extend the results to stratified random sampling. In Section 2, we establish the asymptotic normality of the point estimator, $\hat{G}$, of G and construct confidence intervals on G based on the normal approximation. Confidence intervals on G based on the bootstrap percentile and the bootstrap-t methods are also given. In Section 3, the limiting distribution of the EL ratio statistic is established and the EL ratio confidence intervals are presented. A bootstrap-

[*] Corresponding author. Tel.: +1 613 520 2600x2167.
E-mail address: jrao@math.carleton.ca (J.N.K. Rao).

0264-9993/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.econmod.2010.07.015

calibrated EL confidence interval on G is also presented. Results of a limited simulation study on the finite sample performance of the proposed confidence intervals are reported in Section 4. Extensions to stratified random sampling are outlined in Section 5. Proofs of theorems are relegated to Appendix A.

2. Normal and bootstrap approximation confidence intervals

Let $\{y_1,\cdots,y_n\}$ be an *iid* sample from $F(y)$ and $F_n(u)=n^{-1}\sum_{j=1}^{n} I(y_j\leq u)$ be the empirical distribution function based on the sample, where $I(\cdot)$ denotes the indicator function. Noting that $G=E[\{2F(Y)-1\}Y]/E(Y)$, a simple plug-in moment estimator of G is given by

$$\hat{G} = \frac{1}{\hat{\mu}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left[\{2F_n(y_i)-1\}y_i\right], \tag{2}$$

where $\hat{\mu} = \bar{y}$ is the sample mean. Let $h(u_1,u_2)=I(u_2\leq u_1)u_1+I(u_1\leq u_2)u_2$. For $u_1\geq 0$, let

$$h_1(u_1) = Eh(u_1,Y) = u_1F(u_1) + \int_{u_1}^{\infty} ydF(y). \tag{3}$$

We have the following result on the asymptotic normality of $\hat{G}$.

Theorem 1. *Suppose that* $0<E(Y^2)<\infty$. *Then, as* $n\to\infty$,

$$\sqrt{n}\left(\hat{G}-G\right)\xrightarrow{d} N\left(0,\sigma_1^2\right),$$

where $\sigma_1^2=\mu^{-2}Var\{2h_1(Y)-(G+1)Y\}$ *and* $\xrightarrow{d}$ *denotes convergence in distribution.*

Using this result, a $(1-\alpha)$-level normal approximation confidence interval on G is given by

$$\left(\hat{G}-z_{\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}},\quad \hat{G}+z_{\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}}\right), \tag{4}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the standard normal distribution and

$$\hat{\sigma}_1^2 = \frac{1}{\hat{\mu}^2}\cdot\frac{1}{n-1}\sum_{i=1}^{n}\left(u_{1i}-\bar{u}_1\right)^2 \tag{5}$$

with

$$u_{1i} = 2\hat{h}_1(y_i)-\left(\hat{G}+1\right)y_i,\quad \bar{u}_1 = \frac{1}{n}\sum_{i=1}^{n} u_{1i} \tag{6}$$

and

$$\hat{h}_1(u) = uF_n(u) + \frac{1}{n}\sum_{j=1}^{n} y_jI\left(y_j\geq u\right). \tag{7}$$

The symmetric interval (Eq. (4)) has asymptotically correct coverage probability for large samples. For small samples, however, the normal interval (Eq. (4)) tends to have under-coverage problems, as observed from the simulation results reported in Section 4. In addition, the tail error rates of this interval also tend to be unbalanced, due to skewness of income distributions.

The normal approximation can be replaced by bootstrap procedures. A $(1-\alpha)$-level confidence interval based on the bootstrap percentile of $\hat{G}-G$ is given by

$$\left(\hat{G}-P_{1-\alpha/2},\quad \hat{G}-P_{\alpha/2}\right), \tag{8}$$

where P_α is the 100αth percentile of the sampling distribution of $\hat{G}^*-\hat{G}$, and $\hat{G}^*$ is the estimator of G calculated based on a bootstrap sample $\{y_1^*,\cdots,y_n^*\}$ taken from the original sample $\{y_1,\cdots,y_n\}$ by simple random sampling with replacement. The percentile P_α can be obtained through Monte Carlo approximations by drawing a large number of bootstrap samples. Let $\hat{G}^*(b)$ be the estimate of G computed from the bth bootstrap sample $\{y_1^*(b),\cdots,y_n^*(b)\}$, $b=1,\cdots,B$ and let $\hat{G}^*[1]\leq\cdots\leq\hat{G}^*[B]$ be the ordered sequence of the $\hat{G}^*(b)$s. Then $P_\alpha\doteq\hat{G}^*[\alpha B]-\hat{G}$.

The bootstrap-t confidence interval on G is constructed as

$$\left(\hat{G}-T_{1-\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}},\quad \hat{G}-T_{\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}}\right), \tag{9}$$

where T_α is the 100αth percentile of the sampling distribution of $\left(\hat{G}^*-\hat{G}\right)/\left(\hat{\sigma}_1^*/\sqrt{n}\right)$, and $\hat{G}^*$ and $\hat{\sigma}_1^*/\sqrt{n}$ are the estimator of G and the associated standard error based on a bootstrap sample $\{y_1^*,\ldots,y_n^*\}$. Once again, T_α can be obtained through Monte Carlo approximations.

3. Empirical likelihood ratio confidence intervals

The empirical likelihood (EL) method is a nonparametric approach and is particularly suitable to handle inferential problems involving skewed distributions. EL confidence intervals, obtained from profiling the empirical likelihood ratio statistic, are range respecting and transformation invariant. The shape and orientation of the EL intervals are determined by the data (Owen, 2001), unlike the normal approximation and bootstrap intervals. The log-EL ratio statistic for $\theta=G$ is given by

$$R(\theta) = \sum_{i=1}^{n}\log\{n\tilde{p}_i(\theta)\}, \tag{10}$$

where $\tilde{p}_1(\theta),\cdots,\tilde{p}_n(\theta)$ maximize the log-EL function $l(\mathbf{p}) = \sum_{i=1}^{n} log(p_i)$ subject to the following set of constraints:

$$p_i>0,\ \sum_{i=1}^{n} p_i = 1 \ \text{and}\ \sum_{i=1}^{n} p_i[\{2F_n(y_i)-1\}y_i-\theta y_i] = 0. \tag{11}$$

The last constraint in Eq. (11) is induced by the estimating equation $E[\{2F(Y)-1\}Y-\theta Y]=0$ which defines the parameter $\theta=G$, with the unknown distribution function $F(\cdot)$ replaced by the empirical distribution function $F_n(\cdot)$.

Let $Z(y_i,\theta)=\{2F_n(y_i)-1\}y_i-\theta y_i, i=1,\cdots,n$. It can be shown, by using the Lagrange multiplier method, that

$$R(\theta) = -\sum_{i=1}^{n}\log\{1+\lambda Z(y_i,\theta)\},$$

where λ is the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z(y_i,\theta)}{1+\lambda Z(y_i,\theta)} = 0.$$

Theorem 2 establishes the asymptotic distribution of the log-EL ratio statistic $R(\theta)$.

Theorem 2. *Suppose that* $0<E(Y^3)<\infty$. *Then, as* $n\to\infty$,

$$-2R(\theta)\xrightarrow{d}\frac{\sigma_3^2}{\sigma_2^2}\chi^2(1),$$

where $\sigma_2^2=Var\{2YF(Y)-(\theta+1)Y\}$, $\sigma_3^2=Var\{2h_1(Y)-(\theta+1)Y\}$, *and* $h_1(\cdot)$ *is defined in Eq. (3).*

Using this result, a $(1-\alpha)$-level EL ratio confidence interval on G can be constructed as

$$\left\{\theta\,|\,-2R(\theta)\leq\hat{k}^{-1}\chi_\alpha^2(1)\right\}, \tag{12}$$