Iranian Journal of Science
https://doi.org/10.1007/s40995-023-01517-4

RESEARCH PAPER

Empirical Likelihood Confidence Intervals for Lorenz Curve with Length-Biased Data

Mahdiyeh Vejdani[1] · Abdolhamid Rezaei Roknabadi[1] · Vahid Fakoor[1] · Sarah Jomhoori[2]

Received: 31 July 2021 / Accepted: 9 August 2023

Abstract
The Lorenz curve (LC) is the most fundamental and remarkable tool for processing the size distribution of income and wealth. The LC method is applied as a means to describe distributional consideration in economic analysis. On the other hand, the importance of the biased sampling problem has been well-recognized in statistics and econometrics. In this paper, the empirical likelihood (EL) procedure is proposed to make inferences about the LC in the length-biased setting. The limiting distribution of the EL-based log-likelihood ratio leads to a scaled Chi-square. This limiting distribution will be utilized to construct the EL ratio confidence interval for the LC. Another EL-based confidence interval is proposed by using the influence function method. Simulation studies are conducted to compare the performances of these EL-based confidence intervals with their counterparts in terms of coverage probability and average length. Real data analysis has been used to illustrate the theoretical results.

Keywords Confidence interval · Empirical likelihood · Length-biased data · Lorenz curve

Mathematics Subject Classification 62G05 · 62G20

1 Introduction

Recently, there is a wide interest in evaluating the inequality of incomes or wealth. In order to measure how desirable is a given distribution with respect to another one in terms of equality, it is important to determine on which basis one (income) distribution could be regarded as "more even" than another one. The most widely used tool to display and evaluate the inequality of either income or wealth is the Lorenz curve (LC). This measure, giving a graphical display of income or inequality in wealth, was developed by an American economist, Max Lorenz, in the year 1905. The graph indicates wealth or income, which is depicted on the vertical axis, against the population depicted on the horizontal axis. The LC generally co-occurs with a straight line having a slope one and illustrates the absolute balance in wealth or income distribution. The LC lies underneath and represents the real distribution. The level of unequal distribution increases when the LC drifts away from the baseline. Figure 1 gives a great description of the LC.

Interest in LC significantly increased around the 1970s when Atkinson (1970) and Gastwirth (1971) presented a quantitative measuring and inequality comparisons with the welfare economic implications of the LC. More contributions to the LC's analysis were made by Sen (1973), Jakobsson (1976), Kakwani (1977), Goldie (1977), and Marshall and Olkin (1979). Recent development has been made by Mosler (1994), Arnold (1990), and Lambert (2001), whose findings have led to numerous applications, particularly in reliability theory.

✉ Vahid Fakoor
fakoor@um.ac.ir

Mahdiyeh Vejdani
ias_2006_m@yahoo.com

Abdolhamid Rezaei Roknabadi
rezaei@um.ac.ir

Sarah Jomhoori
sjomhoori@birjand.ac.ir

[1] Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

[2] Department of Statistics, Faculty of Sciences, University of Birjand, Birjand, Iran

Published online: 07 September 2023

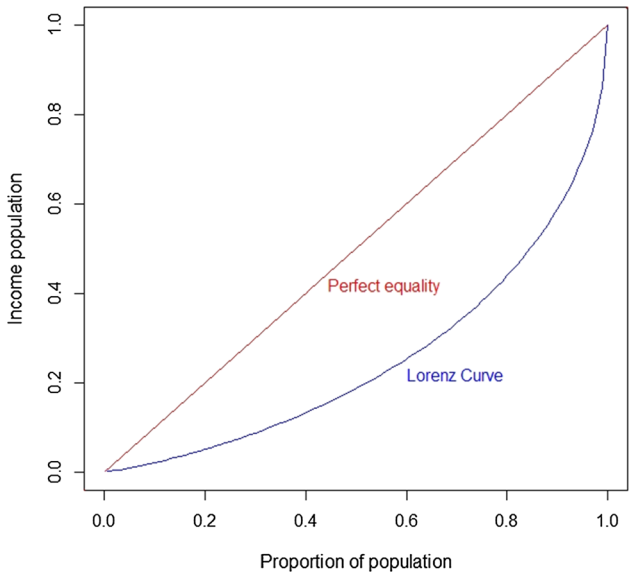

Fig. 1 Lorenz curve

The corresponding LC of a nonnegative random variable X, with a cumulative distribution function (cdf) F, is formulated as follows:

$$\rho(t) := \frac{1}{\mu}\int_0^{\xi_t} x dF(x), 0<t<1,$$

where $\mu = \int_0^{\infty} x dF(x) < \infty$ and $\xi_t = F^{-1}(t)$ is the tth quantile function of F. For a fixed $t \in (0,1)$, the Lorenz ordinate $\rho(t)$ is the ratio of the mean income of the lowest tth fraction of households and the mean income of total households.

However, the income distribution, F, is frequently unknown and must be estimated from the sample of income data. Let $X_1, \ldots, X_n$ be an independent and identically distributed (i.i.d.) sample from F. Then, the empirical estimate of the LC is

$$\hat{\rho}(t) = \frac{1}{\hat{\mu}}\int_0^{\hat{\xi}_t} x dF_n(x), \tag{1.1}$$

where $\hat{\mu}$ is the sample mean, F_n is the empirical distribution function of the data, and $\hat{\xi}_t$ is the tth sample quantile.

The asymptotic theory for empirical Lorenz processes $\{\sqrt{n}(\hat{\rho}(t) - \rho(t)), t \in [0,1]\}$ and what (Goldie 1977) calls concentration processes based on i.i.d random variables has been developed by many authors; see, for example, Gastwirth (1971), Gastwirth (1972), Kakwani and Podder (1973), Goldie (1977), Chandra and Singpurwalla (1978), Sendler (1982), Beach and Davidson (1983), Csörgő et al. (1986), and Zheng (2002). The Goldie concentration processes are, in fact, inverse Lorenz processes whose potential usefulness in econometrics suggests that they are at least as important as Lorenz processes themselves. However, when the population distribution F is skewed, and t falls in the tails of the LC, the existing normal approximation (NA) becomes unreliable unless the sample size is very large. In this situation, one may instead construct a confidence interval via the likelihood function. Likelihood-based inferences are known to have many optimal properties under some regularity conditions. The coverage probability of a likelihood interval is usually based on a Chi-square approximation to the distribution of the likelihood ratio statistic. However, the optimal properties naturally depend on the appropriateness of the parametric model and the precision of the Chi-square approximation.

The empirical likelihood (EL), introduced by Owen (1988, 1990), is a nonparametric methodology for constructing confidence regions and performing hypothesis tests. The EL method produces confidence regions whose shape and orientation are determined entirely by the data, and it possesses some advantages over other methods like the NA. Due to its simplicity and attractive properties, it has been widely applied in many areas, such as regression models (Chen and Van Keilegom 2009), quantile estimation (Chen and Hall 1993), the accelerated failure time model (Zhao 2011), and continuous scale diagnostic tests in the presence of verification bias (Wang and Qin 2013), to name only a few. The papers of particular importance have been done by Qin and Lawless (1994) and Hjort et al. (2009), which linked the concepts of EL, general estimating equations, and nuisance parameters. Since then, the EL has been applied to many different contexts. Belinga-Hill (2007) proposed an EL-based confidence interval for the generalized LC and compared it with the NA-based confidence interval. Qin et al. (2013) observed that most income data are skewed in economics studies, so they developed new EL-based methods to make inferences for the LC, like hybrid bootstrap and EL approach. Shi et al. (2020) proposed new nonparametric confidence intervals for the LC using the influence function-based EL method. They proved that the limiting distributions of the empirical log-likelihood ratio statistics are standard Chi-square distributions.

There are situations in which proper randomization cannot be achieved, and the observed sample is not representative of the population of interest. This biased sampling problem frequently appears since, in the real world, a truly random sampling is not easily achievable or practically feasible. The phenomenon of biased sampling was initially discovered and recognized by Wicksell (1925) in the field of anatomy. At that time, it was named the corpuscle problem, which later came to be known as length-biased sampling. Biased sampling problems occur in many research areas, including medicine, epidemiology and public health, social sciences, and economics. Meanwhile,

the length-biased sampling is one of the most naturally occurring types of biased sampling. Length-biased data are clearly encountered in applications of renewal processes, etiologic studies, genome-wide linkage studies, epidemiologic cohort studies, cancer prevention trials, and studies of the labor economy. Finding appropriate adjustments for the potential selection bias in analyzing length-biased data or, more generally, the biased sampling problems has been a long-standing statistical problem. The importance of biased sampling problem has also been well-recognized in econometrics, in the study of the concentration of income and wealth. When economists compare distributions of income or wealth, they may examine the extent to which the distributions are unequal or concentrated. This comparison may also be of interest when the variable is not wealth but time. To examine the concentration of income or wealth, it is customary to study the Lorenz curve and the Gini coefficient. The same tools apply when studying the concentration of unemployment durations and are, in fact, particularly analytically tractable for many common families of duration distributions. Lancaster (1990) deduced the length-biased distribution as the distribution of the total duration of an individual sampled from the stock in a constant population model. Heckman (1979) discussed sample selection bias as a specification error to utilize simple regression methods in econometrics. Nowell and Stanley (1991) addressed length-biased sampling in mall intercept surveys. Nowell et al. (1988) presented length-biased sampling in contingent valuation. Some applications of sample selection bias have appeared, most notably in wage comparison studies by Gronau (1974) and Hausman (1980).

Let X be a nonnegative random variable with an unknown cdf F. If the probability of selecting an item is proportional to its length, then the recorded item will be distributed according to the following cdf:

$$G(y) = \frac{1}{\mu}\int_0^y x dF(x), \quad y \geq 0, \tag{1.2}$$

where $\mu = \int_0^\infty x dF(x)$ is the mean of X and G is called the length-biased version of F. It can be easily deduced from (1.2) that

$$F(x) = \mu \int_0^x y^{-1} dG(y), \quad x \geq 0. \tag{1.3}$$

There is a natural connection between the LC and length-biased sampling distribution. The curve is actually the plot of $F(x)$ versus its length-biased version, that is, $G(x)$. It should be pointed out that μ is the average income and that $\int_0^{\xi_t} x dF(x) = E[XI(X \leq \xi_t)]$ is the average income for those households with income less than ξ_t.

Accordingly, providing confidence intervals for the LC in the presence of length-biased sampling is of special interest. In this paper, the EL is used for the construction of confidence intervals for the LC. It is proved that the EL ratio admits a limiting scaled Chi-square distribution with one degree of freedom. Another EL-based confidence interval is derived from the influence function technique.

The rest of this article is classified as follows. In Sect. 2, the EL procedure will be proposed to construct two different confidence intervals in the length-biased setting. In Sect. 3, the NA and bootstrap-based confidence intervals are provided for the Lorenz ordinates. Simulation studies for comparing the performances of different methods and real data applications are reported in Sect. 4. Proofs of the theorems are presented in Appendix.

2 EL-Based Confidence Intervals for LC

In order to apply the EL approach for length-biased data, it is necessary to restrict our attention to the estimation of the biased distribution function G instead of the unbiased distribution F. In the length-biased setting, instead of observing a random sample from F, the observations $Y_1, \ldots, Y_n$ are obtained randomly from the distribution G. Apparently, the corresponding empirical distribution function of G is

$$G_n(y) = \frac{1}{n}\sum_{i=1}^n I(Y_i \leq y), \tag{2.1}$$

where $I(A)$ is the indicator function of the event A. According to Eq. (1.3), the plug-in estimator of F, proposed by Cox (1969), is given by

$$F_n(x) = \mu_n \int_0^x y^{-1} dG_n(y), \tag{2.2}$$

where $\mu_n^{-1} = \int_0^\infty y^{-1} dG_n(y)$.

Let $Y_{(1)}, \ldots, Y_{(n)}$ be the order statistics corresponding to $Y_1, \ldots, Y_n$. The sample estimator of ξ_t, proposed by Sen (1984), is then taken as

$$\hat{\xi}_{n,t} = Y_{(k_n)}, \tag{2.3}$$

where the random integer k_n is suitably chosen from

$$k_n = \max\left\{k : \sum_{i=1}^k Y_{(i)}^{-1} \leq t \sum_{i=1}^n Y_{(i)}^{-1}\right\}. \tag{2.4}$$

Remark 1 When t is close to zero and the sample size n is small, the inequality in (2.4) may not hold even for $k = 1$. To overcome this difficulty, Sen (1984) proposed to choose $k_n = 1$ whenever $Y_{(1)}^{-1} > t\sum_{i=1}^n Y_{(i)}^{-1}$. Thus, k_n is a positive

integer-valued random variable and $P(1 \le k_n \le n) = 1$ for every $n \ge 1$.

In what follows, two different EL-based confidence intervals for the LC will be given. To achieve the main theorems, some regularity conditions are assumed as follows:

A1.
$E(Y^{-2}) < \infty$.
A2.
F has a continuous probability density function (pdf) f in some neighborhood of ξ_t and $0 < \xi_t f(\xi_t) < \infty$.

Remark 2 Assumption A1 is used to prove Lemmas 1 and 3 and is also needed to derive main theorems to ensure finiteness of $\sigma_1^2(t_0)$. This moment condition is common and not unlikely. It has also been used by Sen (1984) to derive NA of the sample quantile. Assumption A2 has already been used by Sen (1984) to derive weak convergence of the sample quantile. This result is used to derive Lemmas 2 and 3 in Appendix.

2.1 EL

The fundamental principle of the EL method is based on obtaining the EL ratio statistics via the Lagrange multiplier method under the specified restrictions. According to Eq. (2.1), for a fixed value of $t \in (0,1)$, the restriction

$$E\big(I(Y \le \xi_t) - \rho(t)\big) = 0, \tag{2.5}$$

can be expressed as the following estimating equation:

$$U(\rho(t)) := \frac{1}{n}\sum_{i=1}^{n}\Big(I(Y_i \le \xi_t) - \rho(t)\Big) = 0. \tag{2.6}$$

Let $\mathbf{p} = (p_1, \ldots, p_n)'$ be a probability vector for which $p_i \ge 0, i = 1, \ldots, n$, and $\sum_{i=1}^{n} p_i = 1$. Define

$$D_i(t) := I(Y_i \le \xi_t) - \rho(t), \quad i = 1, \ldots, n.$$

Given Eq. (2.6), the EL for the LC function, $\rho(t)$, could be defined as follows:

$$L(\rho(t)) := \sup_{\mathbf{p}}\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i D_i(t) = 0\right\}. \tag{2.7}$$

Substituting the unknown population quantile ξ_t, with its sample estimator $\hat{\xi}_{n,t}$, the profile EL for $\rho(t)$ is obtained as

$$\mathcal{L}(\rho(t)) := \sup_{\mathbf{p}}\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i D_{ni}(t) = 0\right\},$$

where

$$D_{ni}(t) := I(Y_i \le \hat{\xi}_{n,t}) - \rho(t), \quad i = 1, \ldots, n.$$

Applying the Lagrange multiplier method, the likelihood will be maximized at

$$p_i = \{n(1 + \lambda(t) D_{ni}(t))\}^{-1}, \quad i = 1, \ldots, n,$$

where $\lambda(t)$ will be obtained from

$$\frac{1}{n}\sum_{i=1}^{n}\frac{D_{ni}(t)}{1 + \lambda(t) D_{ni}(t)} = 0. \tag{2.8}$$

Subjecting to the conditions $\sum_{i=1}^{n} p_i = 1$ and $p_i \ge 0, i = 1, \ldots, n$, the product of the elements of $\mathbf{p}$ attains its maximum, which is n^{-n} at $p_i = n^{-1}$. Thus, the profile EL ratio for $\rho(t)$ can be defined as

$$r(\rho(t)) := \prod_{i=1}^{n}(np_i) = \prod_{i=1}^{n}\{1 + \lambda(t) D_{ni}(t)\}^{-1}. \tag{2.9}$$

Finally, the profile empirical log-likelihood ratio of $\rho(t)$ is

$$\begin{aligned}\mathcal{R}(\rho(t)) &:= -2\log r(\rho(t)) \\ &= 2\sum_{i=1}^{n}\log\{1 + \lambda(t) D_{ni}(t)\}.\end{aligned} \tag{2.10}$$

The following theorem gives the limiting distribution of $\mathcal{R}(\rho(t))$. This result can be used to construct confidence intervals for the Lorenz ordinates.

Theorem 1 *Under the stated regularity conditions and for a fixed value of $t = t_0 \in (0,1)$, as n goes to infinity, $\mathcal{R}(\rho(t_0))$ tends to a scaled Chi-square random variable with one degree of freedom, that is,*

$$\kappa \mathcal{R}(\rho(t_0)) \xrightarrow{\mathcal{D}} \chi_1^2,$$

where $\xrightarrow{\mathcal{D}}$ is used to denote convergence in the distribution and

$$\kappa = \frac{\sigma_2^2(t_0)}{\sigma_1^2(t_0)}$$

is the scale constant in which

$$\begin{aligned}\sigma_1^2(t_0) :=& \xi_{t_0}^2\left((1 - t_0)^2\int_0^{\xi_{t_0}} y^{-2} dG(y) + t_0^2\int_{\xi_{t_0}}^{\infty} y^{-2} dG(y)\right) \\ &+ \rho(t_0)\big(1 - \rho(t_0)\big) + 2\xi_{t_0}\frac{t_0(t_0 - 1)}{\mu}\end{aligned} \tag{2.11}$$

and

$$\sigma_2^2(t_0) := \rho(t_0)\big(1 - \rho(t_0)\big). \tag{2.12}$$

Theorem 1 can be used to present a confidence interval for $\rho(t)$ at a fixed time t_0 for $0<t_0<1$. First, it is necessary to estimate σ_1^2 and σ_2^2, consistently. Hence, the following plug-in estimators of σ_1^2 and σ_2^2 are used

$$\hat{\sigma}_1^2(t_0) := \hat{\xi}_{n,t_0}^2 \left((1-t_0)^2 \int_0^{\hat{\xi}_{n,t_0}} y^{-2} dG_n(y) + t_0^2 \int_{\hat{\xi}_{n,t_0}}^{\infty} y^{-2} dG_n(y) \right) + \rho_n(t_0)\big(1-\rho_n(t_0)\big) + 2\hat{\xi}_{n,t_0} \frac{t_0(t_0-1)}{\mu_n} \tag{2.13}$$

and

$$\hat{\sigma}_2^2(t_0) := \rho_n(t_0)\big(1-\rho_n(t_0)\big), \tag{2.14}$$

where G_n, μ_n, and $\hat{\xi}_{n,t}$ are given by (2.1)–(2.3), respectively, and

$$\rho_n(t_0) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le \hat{\xi}_{n,t_0}). \tag{2.15}$$

Therefore, an asymptotic $100(1-\alpha)\%$ confidence interval for $\rho(t)$ at a fixed time $t = t_0 \in (0,1)$ can be obtained from the following equation:

$$I_{EL1} = \left\{ \rho(t_0) : \frac{\hat{\sigma}_2^2(t_0)}{\hat{\sigma}_1^2(t_0)} \mathcal{R}(\rho(t_0)) \le \chi_{1,\alpha}^2 \right\}, \tag{2.16}$$

where $\chi_{1,\alpha}^2$ is the upper α-quantile of Chi-square distribution with one degree of freedom.

2.2 Influence Function-Based EL

According to Theorem 1, the limiting distribution of the log-likelihood ratio is a scaled Chi-square. When applying this method to construct EL-based confidence interval for LC, one needs to estimate the unknown $\sigma_1^2(t)$ and $\sigma_2^2(t)$. According to the proof of Lemma 1 (see Appendix), the influence function

$$g(Y_i, \rho(t)) = \frac{\xi_t}{Y_i}\big(t - I(Y_i \le \xi_t)\big) + \big(I(Y_i \le \xi_t) - \rho(t)\big)$$

has zero expectation. Hence, another EL for $\rho(t)$ can be defined based on this influence function as follows:

$$L_{IF}(\rho(t)) := \sup_{\mathbf{w}} \left\{ \prod_{i=1}^{n} w_i : \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i g(Y_i, \rho(t)) = 0 \right\}, \tag{2.17}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)'$ is a probability vector satisfying $\sum_{i=1}^{n} w_i = 1$ and $w_i \ge 0$ for all i. Substituting g with its estimator, that is,

$$\hat{g}(Y_i, \rho(t)) = \frac{\hat{\xi}_{n,t}}{Y_i}\big(t - I(Y_i \le \hat{\xi}_{n,t})\big) + I(Y_i \le \hat{\xi}_{n,t}) - \rho(t),$$

we derive the following influence function-based profile EL,

$$\mathcal{L}_{IF}(\rho(t)) := \sup_{\mathbf{w}} \left\{ \prod_{i=1}^{n} w_i : \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i \hat{g}(Y_i, \rho(t)) = 0 \right\}. \tag{2.18}$$

The Lagrange multiplier method results in

$$w_i = \{n(1+\gamma(t)\hat{g}(Y_i, \rho(t))\}^{-1}, \quad i = 1, \ldots, n,$$

where $\gamma(t)$ will be obtained from

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{g}(Y_i, \rho(t))}{1+\gamma(t)\hat{g}(Y_i, \rho(t))} = 0. \tag{2.19}$$

Consequently, the corresponding influence function-based profile empirical log-likelihood ratio is

$$\mathcal{R}_{IF}(\rho(t)) := 2\sum_{i=1}^{n} \log\{1+\gamma(t)\hat{g}(Y_i, \rho(t)\}. \tag{2.20}$$

The limiting distribution of $\mathcal{R}_{IF}(\rho(t))$ is given in the following theorem.

Theorem 2 *Under the stated regularity conditions and for a fixed value of $t = t_0 \in (0,1)$, as n goes to infinity, $\mathcal{R}_{IF}(\rho(t_0))$ tends to a Chi-square random variable with one degree of freedom, that is,*

$$\mathcal{R}_{IF}(\rho(t_0)) \xrightarrow{\mathcal{D}} \chi_1^2.$$

According to Theorem 2, there is no need to estimate any scale parameter. Hence, an asymptotic $100(1-\alpha)\%$ confidence interval for $\rho(t)$ at a fixed time $t = t_0 \in (0,1)$ can be obtained from the following equation:

$$I_{EL2} = \left\{ \rho(t_0) : \mathcal{R}_{IF}(\rho(t_0)) \le \chi_{1,\alpha}^2 \right\}. \tag{2.21}$$

3 NA-Based and Bootstrap-Based Confidence Intervals

One of the old-fashioned ideas to find confidence intervals straightaway without considering auxiliary information is the NA method. Lemma 1 in Appendix allows us to construct NA-based confidence intervals for $\rho(t)$. It is obvious that $\sigma_1^2(t)$ can be consistently estimated by its plug-in estimator $\hat{\sigma}_1^2(t)$. Consequently, an asymptotic $100(1-\alpha)\%$

confidence interval for $\rho(t)$ at a fixed time $t = t_0 \in (0,1)$ based on the stated NA is given by

$$I_{NA1} = \left(\rho_n(t_0) - z_{\alpha/2}\frac{\hat{\sigma}_1(t_0)}{\sqrt{n}} \quad \rho_n(t_0) + z_{\alpha/2}\frac{\hat{\sigma}_1(t_0)}{\sqrt{n}}\right),$$

where z_p is the pth quantile of the standard normal distribution and ρ_n was given previously in (2.15).

According to the part (iii) of Lemma 4, one can also propose another NA-based $100(1-\alpha)\%$ confidence interval for $\rho(t)$ at a fixed time $t = t_0 \in (0,1)$, which is given by

$$I_{NA2} = \left(\tilde{\rho}_n(t_0) - z_{\alpha/2}\frac{\hat{\sigma}_1(t_0)}{\sqrt{n}} \quad \tilde{\rho}_n(t_0) + z_{\alpha/2}\frac{\hat{\sigma}_1(t_0)}{\sqrt{n}}\right),$$

where

$$\tilde{\rho}_n(t_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\xi}_{n,t_0}}{Y_i}\left(t - I(Y_i \le \hat{\xi}_{n,t_0})\right) + \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le \hat{\xi}_{n,t_0}).$$

The NA-based and EL-based confidence intervals may have disadvantages. First, the asymptotic normality holds with large sample size. Second, the variance estimation efficiency also depends on the sample size. When the sample size is not large enough, these confidence intervals may perform poorly, especially when the variance estimation efficiency also depends on the sample size. To overcome this problem, the useful method of the bootstrap could be effective. There are several ways to construct bootstrap confidence intervals that vary in ease of calculation and accuracy. The simplest is the normal interval. In order to obtain normal bootstrap-based confidence intervals, we construct the sequence of B, bootstrap estimator

$$\rho_n^b(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i^b \le \hat{\xi}_{n,t}^b), \quad b = 1,2,\ldots,B,$$

where $Y_1^b,\ldots,Y_n^b$ are bootstrap re-samples from original data and $\hat{\xi}_{n,t}^b$ is the tth sample quantile estimator, introduced by (2.3), in the bootstrap re-sample. The bootstrap variance of $\rho_n(t)$ is estimated by

$$V^b = \frac{1}{B-1}\sum_{b=1}^{B}\left(\rho_n^b(t) - \bar{\rho}_n(t)\right)^2,$$

where

$$\bar{\rho}_n(t) = \frac{1}{B}\sum_{b=1}^{B}\rho_n^b(t).$$

As $B \longrightarrow \infty$, by the law of large numbers,

$$V^b \xrightarrow{a.s.} Var(\rho_n(t)).$$

[see, for example, Wasserman (2006, p. 30)].

The $100(1-\alpha)\%$ bootstrap confidence interval is therefore constructed by

$$I_{Boot} = \left(\bar{\rho}_n(t) - z_{\alpha/2}\sqrt{V^b}, \bar{\rho}_n(t) + z_{\alpha/2}\sqrt{V^b}\right).$$

4 Numerical Results

4.1 Simulation Study

The design of the simulation study begins with the choice of the underlying distributions.[1] As income analysis is a natural framework in which inequality measures are needed, we will focus our attention on the inverse gamma distribution, which was used by Vinci (1921) for his income distribution applications. The pdf is given by

$$f(x) = \frac{\beta^p}{\Gamma(p)}x^{-p-1}e^{-\beta/x}, \quad x > 0,$$

where $p, \beta > 0$. The stated Assumptions A1 and A2 are satisfied for this distribution. In order to investigate the performances of the proposed confidence intervals, we have conducted a Monte Carlo simulation study from an Inverse gamma distribution with $p = 4$ and $\beta = 1$, to derive the coverage probability and the average length of each candidate. The number of replications for each sample size $n = 50, 100, 400$ and the number of bootstrap re-samples are $N = 10000$ and $B = 400$, respectively. The results are presented in Table 1. These results are summarized as follows:

- For any sample sizes and various values of t, $EL1$, and $NA1$ confidence intervals are not significantly different from each other.
- $NA2$ performs better than $NA1$ in terms of coverage probabilities. Among others, this confidence interval has the highest coverage probability for small values of t.
- As n increases, the superiority of $EL2$ over the rivals is determined.
- For small values of n and when t falls in the lower tail (next to the origin), $EL2$ has better performance in the sense of minimum average length.
- As t increases, $EL2$ has uniformly higher coverage probability than $EL1$ and NA. However, it produces wider confidence intervals for small values of n.
- The performance of bootstrap confidence intervals is not satisfactory for small sample sizes.
- For large sample sizes, all the methods work well.

[1] The statistical language R is used to perform simulation study. R codes are available on https://github.com/sarah1360/EL-inervals-for-LC-for-LB-data.

Table 1 Simulation results of 95% level confidence intervals from Invgamma(4,1)

t	n	Coverage accuracy					Average length				
		EL1	EL2	NA1	NA2	Boot	EL1	EL2	NA1	NA2	Boot
	50	0.9957	0.8113	0.9973	0.9978	0.8111	0.0719	0.0206	0.0714	0.0714	0.0692
0.1	100	0.9799	0.8579	0.9876	0.9925	0.8526	0.0381	0.0165	0.0379	0.0379	0.0366
	400	0.9548	0.9220	0.9609	0.9810	0.8741	0.0123	0.0089	0.0123	0.0123	0.0119
	50	0.9754	0.8934	0.9843	0.9930	0.8977	0.0737	0.0406	0.0738	0.0738	0.0781
0.2	100	0.9468	0.9123	0.9618	0.9825	0.8716	0.0428	0.0298	0.0428	0.0428	0.0442
	400	0.9422	0.9429	0.9472	0.9670	0.9014	0.0166	0.0149	0.0166	0.0166	0.0168
	50	0.9397	0.9166	0.9506	0.9764	0.8845	0.0777	0.0562	0.0778	0.0778	0.0879
0.3	100	0.9291	0.9314	0.9401	0.9715	0.8813	0.0481	0.0403	0.0481	0.0481	0.0520
	400	0.9437	0.9474	0.9453	0.9578	0.9174	0.0209	0.0200	0.0209	0.0209	0.0215
	50	0.9114	0.9359	0.9161	0.9682	0.8857	0.0818	0.0691	0.0819	0.0820	0.0972
0.4	100	0.9098	0.9399	0.9172	0.9563	0.8942	0.0532	0.0490	0.0533	0.0533	0.0594
	400	0.9397	0.9484	0.9429	0.9535	0.9284	0.0247	0.0243	0.0247	0.0247	0.0256
	50	0.9066	0.9432	0.9081	0.9593	0.8949	0.0850	0.0796	0.0852	0.0852	0.1055
0.5	100	0.9063	0.9456	0.9086	0.9512	0.9040	0.0578	0.0562	0.0579	0.0579	0.0658
	400	0.9447	0.9473	0.9446	0.9497	0.9322	0.0280	0.0279	0.0280	0.0280	0.0291
	50	0.8321	0.9450	0.8363	0.9353	0.8950	0.0863	0.0872	0.0864	0.0865	0.1121
0.6	100	0.9079	0.9470	0.9083	0.9416	0.9064	0.0611	0.0617	0.0612	0.0612	0.0709
	400	0.9358	0.9461	0.9368	0.9454	0.9366	0.0306	0.0307	0.0306	0.0306	0.0319
	50	0.8032	0.9469	0.8031	0.9210	0.8953	0.0843	0.0919	0.0844	0.0844	0.1158
0.7	100	0.8848	0.9466	0.8848	0.9315	0.9108	0.0623	0.0649	0.0624	0.0624	0.0739
	400	0.9392	0.9455	0.9392	0.9417	0.9417	0.0321	0.0324	0.0321	0.0321	0.0336
	50	0.7842	0.9456	0.7825	0.8394	0.8884	0.0752	0.0920	0.0754	0.0754	0.1153
0.8	100	0.8863	0.9475	0.8857	0.9078	0.9091	0.0597	0.0651	0.0597	0.0597	0.0739
	400	0.9301	0.9477	0.9286	0.9417	0.9389	0.0319	0.0325	0.0319	0.0319	0.0337
	50	0.5974	0.9403	0.5943	0.6155	0.8693	0.0467	0.0831	0.0468	0.0468	0.1059
0.9	100	0.8139	0.9452	0.8079	0.8620	0.8951	0.0482	0.0588	0.0482	0.0482	0.0676
	400	0.9226	0.9469	0.9226	0.9298	0.9321	0.0282	0.0294	0.0281	0.0282	0.0306

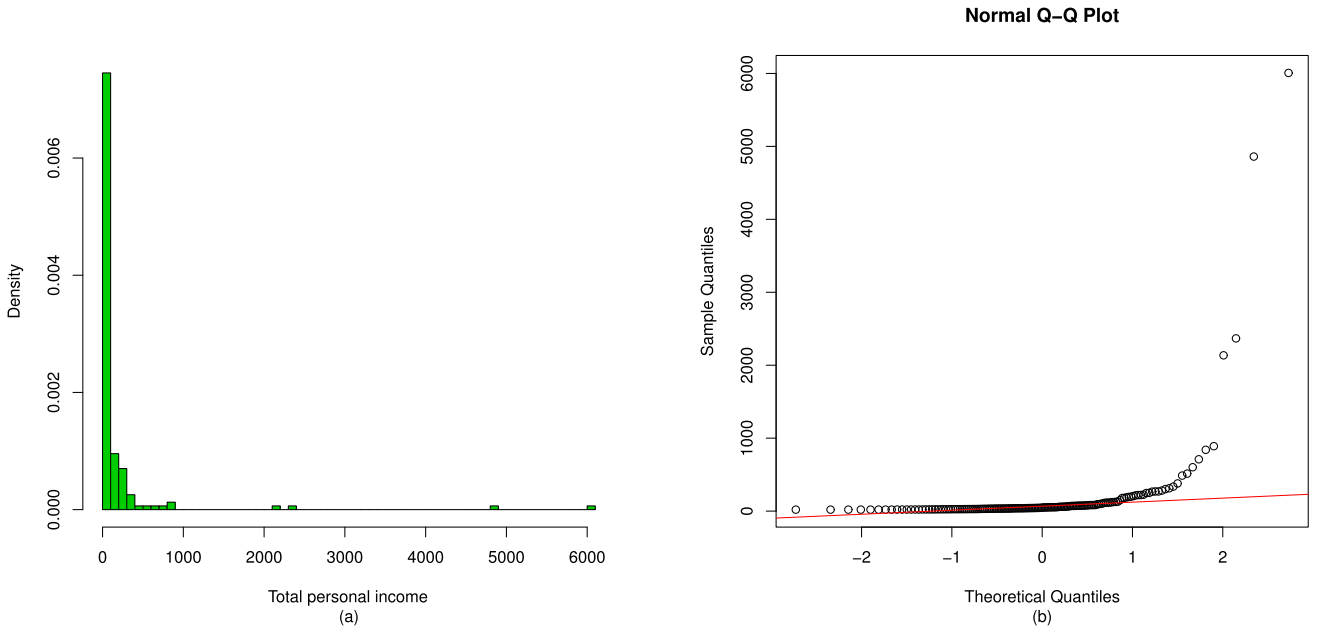

Fig. 2 **a** Histogram and **b** quantile-quantile normal plot of Texas counties data

4.2 Real Data Application

For illustration purposes, we consider data to be referred to as the *Texas counties data* consists of 157 observations. Each observation represents the total personal income accruing to the population of someone of 254 counties in Texas in 1969. The 157 observations included in the present data set represent all the Texas counties in which the total personal income exceeds $\$20,000,000$. This kind of cross-section results in left truncation. Here, the truncation variable is defined as the total personal income, which is less than $\$20,000,000$. When the personal income population size may be assumed to be constant, a suitable modelization is given by the length-bias model (1.2); see (Lancaster 1990).

The histogram of these lifetimes is depicted in Fig. 2a, while the Q-Q normal plot is given in Fig. 2b. Summary statistics are provided in Table 2. The histogram shows that these data are highly skewed to the right. From the Q-Q plot, an inflection with a different slope to the right can be detected. According to these results, the distribution of the data is apparently unimodal and far from being Gaussian. There exist some outliers in the observations. It would be hard enough to come up with a parametric family that fits the data and allows both skewness and kurtosis to vary freely. To avoid having to specify a parametric family for the data, the proposed nonparametric methods are used to construct confidence intervals for LC. Five rival 95% confidence intervals are provided in Table 3 for different values of t. The enlarged graph for $100*$ (lower/upper confidence bounds$-\rho_n(t)$) is depicted in Fig. 3. It can be concluded from Fig. 4 that the proposed intervals have a short length for any fixed values of t. In general, the smaller the t value, the shorter the confidence interval. From the graph, it can be concluded that EL1 and NA intervals have almost the same lengths. The EL confidence intervals are more stable than bootstrap ones, and EL2 has generally shorter interval length.

Table 2 Texas counties data: descriptive statistics

Mean	195.05
SD	662.48
Min	20.20
Max	6007.10
Skewness	6.92
Kurtosis	51.62
Jarque–Bera (test statistics)	4.6782
Jarque–Bera (p-value)	0.09642
Number of observations	157

5 Conclusion

In this paper, the EL method was proposed and used to construct confidence intervals for the Lorenz ordinate in the case of length-biased sampling. Our simulation results showed that the influence function-based EL confidence intervals have good coverage probabilities in the upper tails of the Lorenz curve. Compared with other rivals, they give the shortest interval length over small values of t. EL1 and NA intervals have almost the same performances in terms of the average length and coverage probabilities. Generally, the main advantage of EL, relative to the bootstrap, stems from its use of a likelihood function. Not only does EL provide data-determined shapes for confidence intervals, but it can also easily incorporate known constraints on parameters and adjust for biased sampling schemes. Indeed it should be mentioned that sometimes, it can be computationally challenging to optimize the likelihood ratio over some nuisance parameters. The optimization problem is time-consuming, and this is the main shortcoming of EL relative to the bootstrap.

One might be interested in simultaneous confidence bands. To construct these bands, we have to establish the weak convergence of the EL ratio-based stochastic process.

Table 3 95% confidence intervals for Lorenz ordinates of Texas counties data

t	EL1	EL2	NA1	NA2	Boot
0.1	(0.0363, 0.0689)	(0.0406, 0.0496)	(0.0346, 0.0673)	(0.0287, 0.0614)	(0.0377, 0.0717)
0.2	(0.0828, 0.1094)	(0.0839, 0.1015)	(0.0823, 0.1088)	(0.0794, 0.1060)	(0.0854, 0.1153)
0.3	(0.1322, 0.1616)	(0.1317, 0.1582)	(0.1318, 0.1612)	(0.1303, 0.1597)	(0.1332, 0.1699)
0.4	(0.1860, 0.2224)	(0.1855, 0.2215)	(0.1856, 0.2220)	(0.1854, 0.2218)	(0.1899, 0.2349)
0.5	(0.2504, 0.2983)	(0.2451, 0.2901)	(0.2499, 0.2978)	(0.2439, 0.2918)	(0.2520, 0.3035)
0.6	(0.3168, 0.3718)	(0.3139, 0.3688)	(0.3164, 0.3715)	(0.3143, 0.3694)	(0.3206, 0.3805)
0.7	(0.3950, 0.4589)	(0.3927, 0.4568)	(0.3948, 0.4587)	(0.3937, 0.4576)	(0.3975, 0.4659)
0.8	(0.4947, 0.5624)	(0.4868, 0.5569)	(0.4948, 0.5625)	(0.4896, 0.5573)	(0.4952, 0.5666)
0.9	(0.6268, 0.6969)	(0.6169, 0.6915)	(0.6273, 0.6975)	(0.6216, 0.6918)	(0.6262, 0.7010)

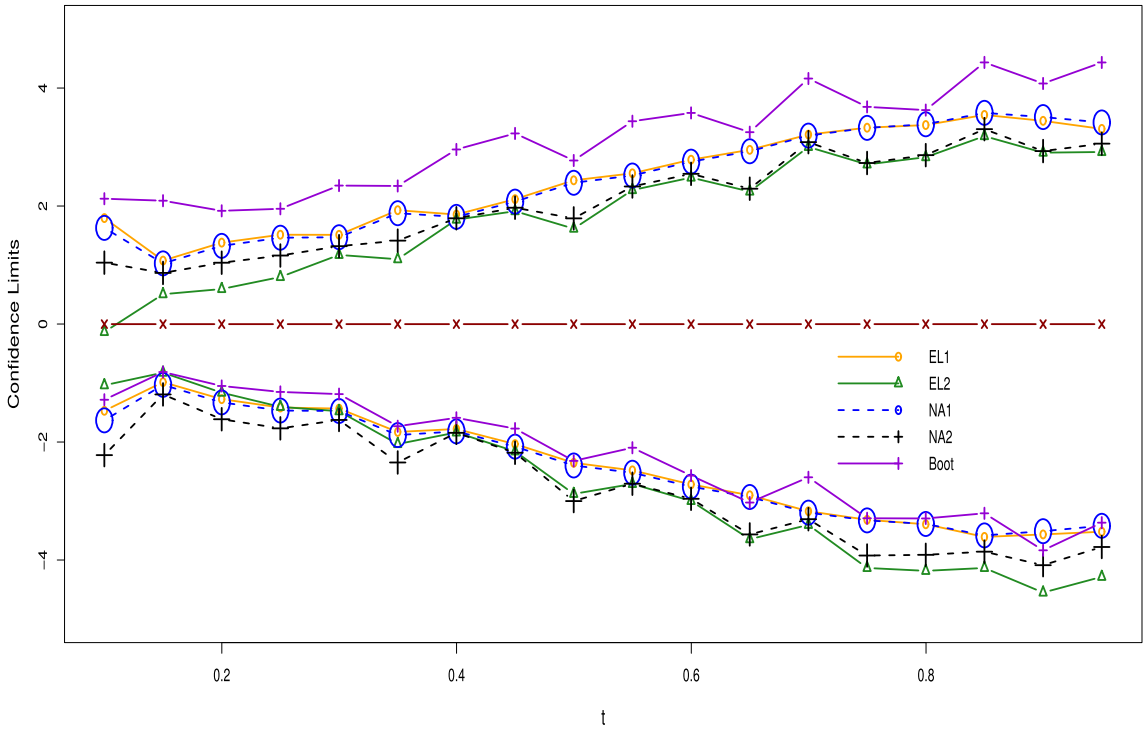

Fig. 3 100 * (95% confidence limits-$\rho_n(t)$) for the Lorenz ordinates of Texas counties data

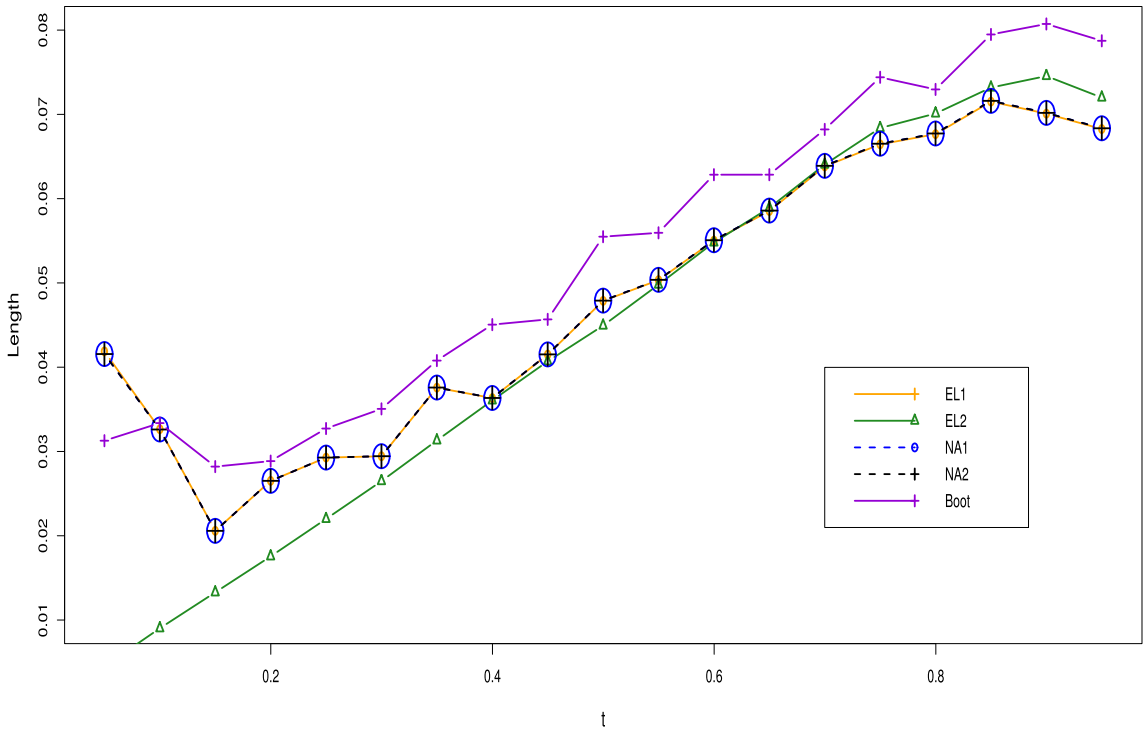

Fig. 4 Length of 95% confidence intervals for the Lorenz ordinates of Texas counties data

We leave this issue as our future work since it is a topic of a different research project.

Appendix

Lemma 1 *Suppose that the stated conditions of Theorem 1 are satisfied. Then, for a fixed $t = t_0 \in (0,1)$, as $n \longrightarrow \infty$, it holds that*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{ni}(t_0) \xrightarrow{\mathcal{D}} N\big(0, \sigma_1^2(t_0)\big), \tag{6.1}$$

where $\sigma_1^2(t_0)$ is given in (2.11).

Proof Let $\mathcal{G}$ be the space of all distribution functions. The line segment in $\mathcal{G}$ joining F and F_n consists of the set of distribution functions, $\{(1-v)F + vF_n, 0 \le v \le 1\}$, also written as $\{F + v(F_n - F), 0 \le v \le 1\}$. Consider the functional $T(F) = \int_0^{\xi_{t_0}} x dF(x)$ and define $T(F_v) = T\big(F + v(F_n - F)\big)$. The first-order Gâteaux differential of T at F in the direction of F_n is

$$\begin{aligned}
&\frac{d}{d\lambda} T(F_v)\Big|_{v=0^+} \\
&= \xi_{t_0}\big(t_0 - F_n(\xi_{t_0})\big) + \left(\frac{\mu_n}{n}\sum_{i=1}^{n} I(Y_i \le \xi_{t_0}) - T(F)\right) \\
&= \xi_{t_0}\left(t_0 - \frac{\mu_n}{n}\sum_{i=1}^{n}\frac{1}{Y_i} I(Y_i \le \xi_{t_0})\right) \\
&\quad + \left(\frac{\mu_n}{n}\sum_{i=1}^{n} I(Y_i \le \xi_{t_0}) - T(F)\right).
\end{aligned} \tag{6.2}$$

Using (6.2) and applying the Taylor expansion of statistical functionals introduced by Mises (1947), we can write

$$\begin{aligned}
T(F_n) - T(F) =& \xi_{t_0}\left(t_0 - \frac{\mu_n}{n}\sum_{i=1}^{n}\frac{1}{Y_i} I(Y_i \le \xi_{t_0})\right) \\
&+ \left(\frac{\mu_n}{n}\sum_{i=1}^{n} I(Y_i \le \xi_{t_0}) - T(F)\right) \\
&+ o_p(n^{-1/2}),
\end{aligned}$$

where

$$T(F_n) = \frac{\mu_n}{n}\sum_{i=1}^{n} I(Y_i \le \hat{\xi}_{n,t_0}).$$

Hence, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{ni}(t_0) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(I(Y_i \le \hat{\xi}_{n,t_0}) - \rho(t_0)\big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\xi_{t_0}}{Y_i}\big(t_0 - I(Y_i \le \xi_{t_0})\big) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(I(Y_i \le \xi_{t_0}) - \rho(t_0)\big) + o_p(1).
\end{aligned} \tag{6.3}$$

Applying Central limit theorem for the i.i.d. random variables

$$\begin{aligned}
g(Y_i, \rho(t_0)) =& \frac{\xi_{t_0}}{Y_i}\big(t_0 - I(Y_i \le \xi_{t_0})\big) \\
&+ I(Y_i \le \xi_{t_0}) - \rho(t_0), \quad i = 1, 2, \ldots,
\end{aligned} \tag{6.4}$$

gives that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{ni}(t_0) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \sigma_1^2(t_0)\big),$$

where $\sigma_1^2(t_0)$ is given in (2.11). □

Lemma 2 *Suppose that the stated conditions of Theorem 1 are satisfied. Then, for a fixed $t = t_0 \in (0,1)$, as $n \longrightarrow \infty$, the following statements are satisfied*:

(i) $\big|I(Y_i \le \hat{\xi}_{n,t_0}) - I(Y_i \le \xi_{t_0})\big| \xrightarrow{\mathcal{P}} 0$,

(ii) $\frac{1}{n}\sum_{i=1}^{n} D_{ni}^2(t_0) \xrightarrow{\mathcal{P}} \sigma_2^2(t_0)$,

where $\xrightarrow{\mathcal{P}}$ denotes the convergence in probability and $\sigma_2^2(t_0)$ is given in (2.12).

Proof

(i) According to (3.10) of Sen (1984), for every $\eta > 0$, there exists an integer n_0, such that for every (fixed) $c > 0$ and $n \ge n_0$,

$$P(\hat{\xi}_{n,t_0} \in J_n) \ge 1 - \eta,$$

where $J_n = \{y : \xi_{t_0} - \frac{c}{\sqrt{n}} \le y \le \xi_{t_0} + \frac{c}{\sqrt{n}}\}$. Hence, we conclude that

$$P(|\hat{\xi}_{n,t_0} - \xi_{t_0}| > \frac{c}{\sqrt{n}}) < \eta, \qquad \text{for every } n \ge n_0.$$

Let $0 < \epsilon < 1$, using Markov's inequality, we can write

$$
\begin{aligned}
&P\big(\big|I(Y_i \le \hat{\xi}_{n,t_0}) - I(Y_i \le \xi_{t_0})\big| > \epsilon\big) \\
&\quad \le \frac{E\big|I(Y_i \le \hat{\xi}_{n,t_0}) - I(Y_i \le \xi_{t_0})\big|}{\epsilon} \\
&\quad = \frac{1}{\epsilon}\Big(P(\hat{\xi}_{n,t_0} < Y_i \le \xi_{t_0}) + P(\xi_{t_0} < Y_i \le \hat{\xi}_{n,t_0})\Big) \\
&\quad = \frac{P(\hat{\xi}_{n,t_0} - \xi_{t_0} < Y_i - \xi_{t_0} \le 0)}{\epsilon} \\
&\qquad + \frac{P(0 < Y_i - \xi_{t_0} \le \hat{\xi}_{n,t_0} - \xi_{t_0})}{\epsilon} \\
&\quad \le \frac{P(|\hat{\xi}_{n,t_0} - \xi_{t_0}| > |Y_i - \xi_{t_0}|)}{\epsilon} \\
&\quad = \int_0^{\infty} \frac{P(|\hat{\xi}_{n,t_0} - \xi_{t_0}| > |y - \xi_{t_0}|)}{\epsilon} dG(y) \\
&\quad = \int_{J_n} \frac{P(|\hat{\xi}_{n,t_0} - \xi_{t_0}| > |y - \xi_{t_0}|)}{\epsilon} dG(y) \\
&\qquad + \int_{J_n'} \frac{P(|\hat{\xi}_{n,t_0} - \xi_{t_0}| > |y - \xi_{t_0}|)}{\epsilon} dG(y) \\
&\quad =: A_{n1} + A_{n2},
\end{aligned}
\tag{6.5}
$$

where $J_n' = \{y : y < \xi_{t_0} - \frac{c}{\sqrt{n}}\} \cup \{y : y > \xi_{t_0} + \frac{c}{\sqrt{n}}\}$. To get rid of A_{n1}, we can write

$$
\begin{aligned}
A_{n1} &\le \frac{1}{\epsilon} \int_{J_n} dG(y) \\
&= \frac{G(\xi_{t_0} + \frac{c}{\sqrt{n}}) - G(\xi_{t_0} - \frac{c}{\sqrt{n}})}{\epsilon} \\
&= \frac{2c}{\epsilon\sqrt{n}} \left(\frac{G(\xi_{t_0} + \frac{c}{\sqrt{n}}) - G(\xi_{t_0} - \frac{c}{\sqrt{n}})}{\frac{2c}{\sqrt{n}}} \right).
\end{aligned}
$$

Obviously, we have

$$
\lim_{n\to\infty} \frac{G(\xi_{t_0} + \frac{c}{\sqrt{n}}) - G(\xi_{t_0} - \frac{c}{\sqrt{n}})}{\frac{2c}{\sqrt{n}}} = g(\xi_{t_0}),
$$

where $g(\cdot)$ is the corresponding density function of $G(\cdot)$. Hence, for every $\epsilon' > 0$, there exists an integer n_1, such that for all values of $n \ge n_1$,

$$
A_{n1} < \epsilon'. \tag{6.6}
$$

For the second term in (6.5), we have

$$
\begin{aligned}
A_{n2} &\le \int_{J_n'} \frac{P(|\hat{\xi}_{n,t_0} - \xi_{t_0}| > \frac{c}{\sqrt{n}})}{\epsilon} dG(y) \\
&< \frac{\eta}{\epsilon}, \qquad \text{for every } n \ge n_0.
\end{aligned}
\tag{6.7}
$$

Since ϵ' and η are positive arbitrary values, we choose $\eta = \frac{\epsilon\delta}{2}$ and $\epsilon' = \frac{\delta}{2}$ and hence, we conclude from (6.5), (6.6) and (6.7) that for every $n \ge \max\{n_0, n_1\}$,

$$
P\big(\big|I(Y_i \le \hat{\xi}_{n,t_0}) - I(Y_i \le \xi_{t_0})\big| > \epsilon\big) < \delta, \quad i = 1, 2, \ldots \tag{6.8}
$$

(ii) Applying the law of large numbers for the sequence $\{D_i^2(t_0), i = 1, 2, \ldots\}$ of i.i.d. random variables, we have

$$
\frac{1}{n}\sum_{i=1}^{n} D_i^2(t_0) \xrightarrow{\mathcal{P}} E(D_i^2(t_0)),
$$

where

$$
E(D_i^2(t_0)) = \rho(t_0)\big(1 - \rho(t_0)\big).
$$

Hence,

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} D_{ni}^2(t_0) - \frac{1}{n}\sum_{i=1}^{n} D_i^2(t_0)\right| \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n} \big|D_{ni}(t_0) + D_i(t_0)\big|\big|D_{ni}(t_0) - D_i(t_0)\big| \\
&\quad \le \frac{2}{n}\sum_{i=1}^{n} \big|I(Y_i \le \hat{\xi}_{n,t_0}) - I(Y_i \le \xi_{t_0})\big| \\
&\quad = o_p(1),
\end{aligned}
$$

where the last convergence follows from part (*i*) and this completes the proof. □

Proof of Theorem 1 Since $E(D_{ni}^2(t_0)) < \infty$, for a fixed value of $t = t_0 \in (0, 1)$, Lemma 2 (*ii*) jointly with a similar argument as in Owen (2001) results that

$$
\max_{1 \le i \le n} |D_{ni}(t_0)| = o_p(n^{1/2}), \tag{6.9}
$$

and also,

$$
\frac{1}{n}\sum_{i=1}^{n} \big|D_{ni}(t_0)\big|^3 = o_p(n^{1/2}). \tag{6.10}
$$

Thus, combining (6.9) and (6.10), we can prove that

$$
\big|\lambda(t_0)\big| = O_p(n^{-1/2}). \tag{6.11}
$$

Using Taylor expansion of (2.20), we have

$$
\begin{aligned}
\mathcal{R}(\rho(t_0)) &= 2\sum_{i=1}^{n} \log\{1 + \lambda(t_0) D_{ni}(t_0)\} \\
&= 2\sum_{i=1}^{n} \left(\lambda(t_0) D_{ni}(t_0) - \frac{\lambda^2(t_0) D_{ni}^2(t_0)}{2}\right) \\
&\quad + Rem(t_0),
\end{aligned}
\tag{6.12}
$$

where

$$Rem(t_0) = 2\sum_{i=1}^{n}\sum_{k=3}^{\infty}(-1)^{k-1}\left(\frac{\lambda^k(t_0)D_{ni}^k(t_0)}{k}\right).$$

Applying (6.10) and (6.11), it can be seen that

$$\begin{aligned}|Rem(t_0)| &\le C\sum_{i=1}^{n}|\lambda(t_0)D_{ni}(t_0)|^3\\ &\le C|\lambda(t_0)|^3\sum_{i=1}^{n}|D_{ni}(t_0)|^3\\ &= o_p(1).\end{aligned} \tag{6.13}$$

Recalling (2.8), it follows that

$$\begin{aligned}0 &= \sum_{i=1}^{n}\frac{D_{ni}(t_0)}{1+\lambda(t)D_{ni}(t_0)}\\ &= \sum_{i=1}^{n}D_{ni}(t_0)\left[1-\lambda(t_0)D_{ni}(t_0)+\frac{\lambda^2(t)D_{ni}^2(t_0)}{1+\lambda(t)D_{ni}(t)}\right]\\ &= \sum_{i=1}^{n}D_{ni}(t_0)-\lambda(t_0)\left(\sum_{i=1}^{n}D_{ni}^2(t_0)\right)\\ &\quad+\sum_{i=1}^{n}\frac{\lambda^2(t_0)D_{ni}^3(t_0)}{1+\lambda(t_0)D_{ni}(t_0)}.\end{aligned} \tag{6.14}$$

Applying (6.9) and (6.11) in (6.14) results in

$$\lambda(t_0) = \frac{\sum_{i=1}^{n}D_{ni}(t_0)}{\sum_{i=1}^{n}D_{ni}^2(t_0)} + o_p(n^{-1/2}). \tag{6.15}$$

Once more by recalling (2.8), we have

$$\begin{aligned}0 &= \sum_{i=1}^{n}\frac{\lambda(t_0)D_{ni}(t_0)}{1+\lambda(t_0)D_{ni}(t_0)}\\ &= \sum_{i=1}^{n}\left(\lambda(t_0)D_{ni}(t_0)\right)-\sum_{i=1}^{n}\left(\lambda^2(t_0)D_{ni}^2(t_0)\right)\\ &\quad+\sum_{i=1}^{n}\frac{\lambda^3(t_0)D_{ni}^3(t_0)}{1+\lambda(t_0)D_{ni}(t_0)}.\end{aligned}$$

Moreover, from (6.9) and (6.11), we can write

$$\sum_{i=1}^{n}\frac{\lambda^3(t_0)D_{ni}^3(t_0)}{1+\lambda(t_0)D_{ni}(t_0)} = o_p(n^{-1/2}).$$

Hence,

$$\sum_{i=1}^{n}\left(\lambda^2(t_0)D_{ni}^2(t_0)\right) = \sum_{i=1}^{n}\left(\lambda(t_0)D_{ni}(t_0)\right)+o_p(1). \tag{6.16}$$

Substituting (6.15) and (6.16) into (6.12), and using Lemmas 1 and 2, as $n \longrightarrow \infty$, we conclude that

$$\begin{aligned}\kappa\mathcal{R}(\rho(t_0))) &= \frac{\sigma_2^2(t_0)}{\sigma_1^2(t_0)}\sum_{i=1}^{n}\left(\lambda^2(t_0)D_{ni}^2(t_0)\right)+o_p(1)\\ &= \frac{\sigma_2^2(t_0)}{\sigma_1^2(t_0)}\frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_{ni}(t_0)\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}D_{ni}^2(t_0)\right)}+o_p(1)\\ &= \chi_1^2+o_p(1),\end{aligned}$$

which completes the proof. □

The following two lemmas are needed to prove Theorem 2.

Lemma 3 *Under the stated conditions of Theorem 1, we have*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(Y_i,\rho(t_0))-g(Y_i,\rho(t_0)))^2 = o_p(1).$$

Proof We can write

$$\hat{g}(Y_i,\rho(t_0))-g(Y_i,\rho(t_0)) =: C_{1i}+C_{2i}+C_{3i}, \tag{6.17}$$

where

$$\begin{aligned}C_{1i} &= \frac{t_0(\hat{\xi}_{n,t_0}-\xi_{t_0})}{Y_i},\\ C_{2i} &= I(Y_i\le\hat{\xi}_{n,t_0})-I(Y_i\le\xi_{t_0}),\\ C_{3i} &= \frac{\xi_{t_0}}{Y_i}I(Y_i\le\xi_{t_0})-\frac{\hat{\xi}_{n,t_0}}{Y_i}I(Y_i\le\hat{\xi}_{n,t_0}).\end{aligned}$$

Applying the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2), a, b, c \in \mathbb{R}$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}(Y_i,\rho(t_0))-g(Y_i,\rho(t_0))\right)^2 &\le \frac{3}{n}\sum_{i=1}^{n}C_{1i}^2\\ &\quad+\frac{3}{n}\sum_{i=1}^{n}C_{2i}^2+\frac{3}{n}\sum_{i=1}^{n}C_{3i}^2.\end{aligned} \tag{6.18}$$

Since $E(Y^{-2})<\infty$, using the law of large numbers and the weak consistency of $\hat{\xi}_{n,t_0}$, which was studied by Sen (1984), we have

$$\frac{1}{n}\sum_{i=1}^{n}C_{1i}^2 = t_0^2\left(\hat{\xi}_{n,t_0}-\xi_{t_0}\right)^2\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i^2} = o_p(1). \tag{6.19}$$

To calculate the second term in the right side of (6.18), first we compute the expectation of $C_{2i}^2, i = 1, 2, \ldots, n$. Choose $\epsilon > 0$ and n so large that (6.8) holds, then, we have

$$\begin{aligned}E(C_{2i}^2) &= E\left(C_{2i}^2I(|C_{2i}|\le\epsilon)\right)+E\left(C_{2i}^2I(|C_{2i}|>\epsilon)\right)\\ &= P(|C_{2i}|>\epsilon)<\delta.\end{aligned}$$

According to the Markov's inequality, we have for $x > 0$,

$$P\Big(\frac{1}{n}\sum_{i=1}^{n} C_{2i}^2 > x\Big) \le \frac{1}{n}\sum_{i=1}^{n}\frac{E(C_{2i}^2)}{x}.$$

Hence, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n} C_{2i}^2 = o_p(1).$$

To get rid of the third term in (6.18), we decompose C_{3i} into

$$\begin{aligned} C_{3i} =& (\xi_{t_0} - \hat{\xi}_{n,t_0})\frac{I(Y_i \le \xi_{t_0})}{Y_i} \\ &+ \frac{1}{Y_i}(\hat{\xi}_{n,t_0} - \xi_{t_0})\big(I(Y_i \le \xi_{t_0}) - I(Y_i \le \hat{\xi}_{n,t_0})\big) \\ &+ \frac{\xi_{t_0}}{Y_i}\big(I(Y_i \le \xi_{t_0}) - I(Y_i \le \hat{\xi}_{n,t_0})\big) \\ =:& I_{1i} + I_{2i} + I_{3i}. \end{aligned}$$

Using similar arguments, we obtain simply that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} I_{i1}^2 =& o_p(1), \\ \frac{1}{n}\sum_{i=1}^{n} I_{i2}^2 =& o_p(1). \end{aligned}$$

We have also

$$\frac{1}{n}\sum_{i=1}^{n} I_{i3}^2 \le \frac{2\xi_{t_0}^2}{n}\sum_{i=1}^{n}\frac{\big|I(Y_i \le \xi_{t_0}) - I(Y_i \le \hat{\xi}_{n,t_0})\big|}{Y_i^2} = o_p(1),$$

and analogously,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} C_{3i}^2 \le& \frac{3}{n}\sum_{i=1}^{n} I_{1i}^2 + \frac{3}{n}\sum_{i=1}^{n} I_{2i}^2 + \frac{3}{n}\sum_{i=1}^{n} I_{3i}^2 \\ =& o_p(1). \end{aligned} \tag{6.20}$$

The result follows from (6.18)–(6.20). □

Lemma 4 *Under the stated conditions of Theorem 1, the following statements are satisfied*:

(i) $\max_{1\le i\le n}|\hat{g}(Y_i, \rho(t_0))| = o_p(n^{1/2})$.
(ii) $\frac{1}{n}\sum_{i=1}^{n}\hat{g}^2(Y_i, \rho(t_0)) = \sigma_1^2(t_0) + o_p(1)$.
(iii) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}(Y_i, \rho(t_0)) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \sigma_1^2(t_0)\big)$,

where $\sigma_1^2(t_0)$ is given in (2.11).

Proof

(i) Since $g(Y_i, \rho(t_0)), i = 1, \ldots, n$ are i.i.d. mean zero random variables with variance $\sigma_1^2(t_0)$, we have

$$\max_{1\le i\le n}|g(Y_i, \rho(t_0))| = o_p(n^{1/2}).$$

Hence,

$$\begin{aligned} &\max_{1\le i\le n}|\hat{g}(Y_i, \rho(t_0))| \\ &\le \max_{1\le i\le n}|\hat{g}(Y_i, \rho(t_0)) - g(Y_i, \rho(t_0))| \\ &\quad + \max_{1\le i\le n}|g(Y_i, \rho(t_0))| \\ &= o_p(n^{1/2}). \end{aligned}$$

(ii) Similar to the proof of Lemma 3, we can prove that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \rho(t_0))\big(\hat{g}(Y_i, \rho(t_0)) \\ &\quad - g(Y_i, \rho(t_0))\big) = o_p(1). \end{aligned} \tag{6.21}$$

Using the law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n} g^2(Y_i, \rho(t_0)) = \sigma_1^2(t_0) + o_p(1). \tag{6.22}$$

Consequently, we obtain from (6.21) and (6.22) that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\hat{g}^2(Y_i, \rho(t_0)) \\ &= \frac{1}{n}\sum_{i=1}^{n}\big(\hat{g}(Y_i, \rho(t_0)) - g(Y_i, \rho(t_0))\big)^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} g^2(Y_i, \rho(t_0)) \\ &\quad + \frac{2}{n}\sum_{i=1}^{n} g(Y_i, \rho(t_0))\big(\hat{g}(Y_i, \rho(t_0)) - g(Y_i, \rho(t_0))\big) \\ &= \sigma_1^2(t_0) + o_p(1). \end{aligned}$$

(iii) Considering definition $\hat{g}$, we have

$$\begin{aligned} &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}(Y_i, \rho(t_0)) \\ &= \hat{\xi}_{n,t_0}\sqrt{n}\left(\frac{t_0}{\mu_n} - H_n(\hat{\xi}_{n,t_0})\right) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{ni}(t_0), \end{aligned} \tag{6.23}$$

where

$$H_n(\hat{\xi}_{n,t_0}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}I(Y_i \le \hat{\xi}_{n,t_0}).$$

According to Sen (1984), (p. 64, Line 1), we have

$$\mu_n H_n(\hat{\xi}_{n,t_0}) = t_0 + O_p(1/n).$$

Applying the law of large numbers for μ_n and using the weak consistency of $\hat{\xi}_{n,t_0}$, we obtain

$$\hat{\xi}_{n,t_0}\sqrt{n}\left(\frac{t_0}{\mu_n} - H_n(\hat{\xi}_{n,t_0})\right) = o_p(1). \tag{6.24}$$

Hence, the result follows from (6.23), (6.24) and Lemma 1. □

Proof of Theorem 2 Lemma 4(i) together with a similar argument given in the proof of Theorem 1 gives

$$|\gamma(t_0)| = O_p(n^{-1/2}). \tag{6.25}$$

Taylor expansion results in

$$\begin{aligned}\mathcal{R}_{IF}(\rho(t_0)) &= 2\sum_{i=1}^{n}\left(\gamma(t_0)\hat{g}(Y_i,\rho(t_0)) - \frac{1}{2}\gamma^2(t_0)\hat{g}^2(Y_i,\rho(t_0))\right) \\ &\quad + rem(t_0),\end{aligned} \tag{6.26}$$

where

$$|rem(t_0)| = o_p(1).$$

Recalling (2.19), we have

$$\begin{aligned}0 &= \sum_{i=1}^{n}\frac{\hat{g}(Y_i,\rho(t_0))}{1+\gamma(t_0)\hat{g}(Y_i,\rho(t_0))} \\ &= \sum_{i=1}^{n}\hat{g}(Y_i,\rho(t_0)) - \gamma(t_0)\sum_{i=1}^{n}\hat{g}^2(Y_i,\rho(t_0)) \\ &\quad + \sum_{i=1}^{n}\frac{\gamma^2(t_0)\hat{g}^3(Y_i,\rho(t_0))}{1+\gamma(t_0)\hat{g}(Y_i,\rho(t_0))}.\end{aligned} \tag{6.27}$$

Applying Lemma 4(i) and (6.25) in (6.27) results in

$$\gamma(t_0) = \frac{\sum_{i=1}^{n}\hat{g}(Y_i,\rho(t_0))}{\sum_{i=1}^{n}\hat{g}^2(Y_i,\rho(t_0))} + o_p(n^{-1/2}). \tag{6.28}$$

Moreover, we have

$$\begin{aligned}&\sum_{i=1}^{n}\gamma(t_0)\hat{g}(Y_i,\rho(t_0)) \\ &\quad = \sum_{i=1}^{n}\gamma^2(t_0)\hat{g}^2(Y_i,\rho(t_0)) + o_p(1).\end{aligned} \tag{6.29}$$

Substituting (6.28) and (6.29) into (6.26), and using Lemma 4, as $n \longrightarrow \infty$, we conclude that

$$\begin{aligned}\mathcal{R}(\rho(t_0))) &= \sum_{i=1}^{n}\left(\gamma^2(t_0)\hat{g}^2(Y_i,\rho(t_0))\right) + o_p(1) \\ &= \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}(Y_i,\rho(t_0))\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{g}^2(Y_i,\rho(t_0))\right)} + o_p(1) \\ &= \chi_1^2 + o_p(1),\end{aligned}$$

which completes the proof. □

Acknowledgements The authors are grateful to the referees and the Editor-in-Chief for careful reading and numerous suggestions which greatly improved this paper.

Author Contributions On behalf of all authors, the corresponding author states that there is no *contributions*.

Funding On behalf of all authors, the corresponding author states that no *funding* was received.

Availability of Data and Materials On behalf of all authors, the corresponding author states that there is no *availability of data and material*.

Code Availability On behalf of all authors, the corresponding author states that there is no *code availability*.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Arnold BC (1990) The Lorenz order and the effects of taxation policies. Bull Econ Res 42:249–264

Atkinson AB (1970) On the measurement of inequality. Econ Theory 2:244–263

Beach CM, Davidson R (1983) Distribution-free statistical inference with Lorenz curves and income shares. Rev Econ Stud 50:723–735

Belinga-Hill N (2007) Empirical likelihood confidence intervals for generalized Lorenz curve. Master thesis at Georgia State University, Atlanta

Chandra M, Singpurwalla ND (1978) The Gini index, the Lorenz curve, and the total time on test transform. George Washington University Serial T-368

Chen SX, Hall P (1993) Smoothed empirical likelihood confidence intervals for quantiles. Ann Stat 21:1166–1181

Chen SX, Van Keilegom I (2009) A goodness-of-fit test for parametric and semiparametric models in multiresponse regression. Bernoulli 15:955–976

Cox DR (1969) Some sampling problems in technology. In: Johnson NL, Smith H (eds) New developments in survey sampling. Wiley, New York, pp 506–527

Csörgő M, Csörgő S, Horváth L (1986) An asymptotic theory for empirical reliability and concentration process. Lecture Notes in Statistics, vol 33. Springer, New York

Gastwirth JL (1971) A general definition of the Lorenz curve. Econometrica 39:1037–1039

Gastwirth JL (1972) The estimation of the Lorenz curve and Gini index. Rev Econ Stat 54:306–316

Goldie CM (1977) Convergence theorems for empirical Lorenz curves and their inverses. Adv Appl Probab 9:765–791

Gronau R (1974) Wage comparisons—a selectivity bias. J Polit Econ 82:1119–43

Hausman JA (1980) The effect of wages, taxes, and fixed costs on women's labor force participation. J Public Econ 14:161–94

Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47:153–161

Hjort NL, McKeague IW, Van Keilegom I (2009) Extending the scope of empirical likelihood. Ann Stat 37:1079–1111

Jakobsson U (1976) On the measurement of the degree of progression. Public Econ 5:161–168
Kakwani NC (1977) Applications of Lorenz curves in economic analysis. Econometrica 45:719–727
Kakwani NC, Podder N (1973) On the estimation of Lorenz curves from grouped observations. Int Econ Rev 14:278–292
Lambert PJ (2001) The distribution and redistribution of income, 3rd edn. Manchester University Press
Lancaster T (1990) The econometric analysis of transition data. Cambridge University Press, Cambridge
Marshall AW, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic Press, Orlando
Mises RV (1947) On the asymptotic distribution of differentiable statistical functions. Ann Math Stat 18(3):309–348
Mosler K (1994) Majorization in economic disparity measures. Linear Algebra Appl 199:91–114
Nowell C, Stanley LR (1991) Length-biased sampling in mall intercept surveys. J Mark Res 28:475–479
Nowell C, Evans MA, McDonald L (1988) Length-biased sampling in contingent valuation studies. Land Econ 64:367–371
Owen A (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75:237–249
Owen A (1990) Empirical likelihood ratio confidence regions. Ann Stat 18(1):90–120
Owen A (2001) Empirical likelihood. Chapman & Hall/CRC, New York
Qin J, Lawless JF (1994) Empirical likelihood and general estimating equations. Ann Stat 22:300–325
Qin GS, Yang BY, Belinga-Hill NE (2013) Empirical likelihood-based inferences for the Lorenz curve. Ann Inst Stat Math 65:1–21
Sen A (1973) On economic inequality. Oxford University Press Inc., New York
Sen PK (1984) On asymptotic representations for reduced quantiles in sampling from a length-biased distribution. Calcutta Stat Assoc Bull 33:59–68
Sendler W (1982) On functionals of order statistics. Metrika 29:19–54
Shi Y, Liu B, Qin G (2020) Influence function-based empirical likelihood and generalized confidence intervals for the Lorenz curve. Stat Methods Appl 29:427–446
Vinci F (1921) Nuovi contributi allo studio della distribuzione dei redditi. G Degl Econ Riv Stat 61:365–369
Wang B, Qin G (2013) Empirical likelihood confidence regions for the evaluation of continuous scale diagnostic tests in the presence of verification bias. Can J Stat 41:398–420
Wasserman L (2006) All of nonparametric statistics. Springer, New York
Wicksell SD (1925) The corpuscle problem: a mathematical study of a biometric problem. Biometrika 17:84–99
Zhao Y (2011) Empirical likelihood inference for the accelerated failure time model. Stat Probab Lett 81:603–610
Zheng B (2002) Testing Lorenz curves with non-simple random samples. Econometrica 70:1235–1243