Semi-Parametric Estimation of Lorenz Curve and Gini-index in an Exponential Distribution

P.P. Chandrasekhara Pillai[*1], G. Rajesh[2] and E.I. Abdul-Sathar[3]

[*1,2]Department of Statistics, DB Pampa College, Parumala, India
[3]Department of Statistics, University of Kerala,Thiruvananthapuram, India
Corresponding Author: chandranpbhavan@gmail.com

Received: September 5, 2013/ Revised: February 12, 2014 / Accepted: March 03, 2014

ABSTRACT

In this article, we estimate Lorenz curve and Gini-index for exponential distribution using the semi-parametric method. We derived the estimates using least square method based on the survival function. The estimation is also carried out under various censoring schemes. The performances of the estimators in terms of MSE are compared via a Monte Carlo simulation study.

Keywords: Lorenz curve, Gini-index, Least square estimation, Kaplan-Meier estimator, Type-I censoring, Type-II censoring, Interval censoring.

1. Introduction

The Lorenz curve and Gini-index play a central role in the analysis of income data and the evaluation of welfare judgments. Also these have been extensively used in the study of inequality of distributions. Let X be a non-negative random variable with distribution function F. The Lorenz curve corresponding to the random variable *X*, denoted by L(*p*), is defined (Gastwirth, 1971) by the formula

$$L(p)\frac{1}{\mu}\int_0^p F^{-1}(x)dx, 0 \le p \le 1, \qquad (1.1)$$

where

$$F^{-1}(x) = \sup\{y: F(y) \le x\}$$

We assume that the mean μ is finite and positive. If the distribution which is being studied is the income of a certain population, then L(*p*) denotes the fraction of the total income received by the 100*p*%of the population which has the lowest income. The Gini index is twice the area between the Lorenz curve and the `line of equal angle'(also known as `line of equal distribution'). The Gini-index G is defined as

$$G = 1-2\int_0^1 L(p)dp \tag{1.2}$$

For the applications of Lorenz curve and Gini-index we refer to Moothathu (1990) and Bhattacharjee (1993) and the references therein. These measures have also been found applications in reliability theory. For example, see Chandra and Singpurwalla (1981).

Iyengar (1960) showed that the maximum likelihood estimator of the Gini-index for the normal-logarithmic model is asymptotically normal and Moothathu (1985a and 1985b) obtains the distribution of the maximum likelihood estimators of the Lorenz curve and the Gini-index of the Pareto and exponential distribution, respectively. He also obtains uniformly minimum variance unbiased estimators which are strongly consistent and asymptotically normal, for the lognormal distribution (Moothathu, 1989) and the Pareto distribution. (Moothathu, 1990). Sathar *et al*. (2005) discussed the problem of estimation of Lorenz curve and Gini-index for the Pareto distribution in the Bayesian frame-work.

The outline of the paper is as follows. In Section 2, the general least square procedure of estimating the parameter of an exponential distribution is introduced. Then using its survival function, least square estimators of the Lorenz Curve and Gini-index of exponential distribution under various censoring schemes are obtained. Results of a Monte Carlo simulation study conducted to evaluate the performance of these estimators in terms of mean squared error (MSE) are provided in Section 3. A brief summary of the results are given in Section 4.

2. Estimation of Lorenz curve and Gini-index

In this Section, we obtained the estimates of Lorenz curve and Gini-index based on a sample $\underline{x} = (x_1, x_{2,\ldots,} x_n)$ from an exponential density using various methods. The estimation under various censoring schemes using least square method is discussed in this section.

2.1 Least square estimator based on survival function

Afify (2003) derived the least square estimator using the regression of survival function on the observations, for the Pareto distribution. In this set-up, we use this approach to derive another estimator for the Lorenz Curve and Gini-index, for the exponential distribution with pdf

$$\int(\chi, \sigma, \theta) = \tfrac{1}{\sigma}\exp\left(-\tfrac{1}{\sigma}(x-\theta)\right), x > \theta > 0, \sigma > 0, \tag{2.1}$$

and survival function

$$S(x_i) = P(X \geq x_i) = \exp\left(-\tfrac{1}{\sigma}(x_i - \theta)\right), i = 1,2,.....n. \quad (2.2)$$

The Lorenz curve and the Gini-index for (2.1) simplifies respectively to

$$L(p) = p + \sigma(\theta + \sigma)^{-1}(1-p)\log(1-p),\ 0 \leq p \leq 1, \quad (2.3)$$

and

$$G = \tfrac{\sigma}{2}(\theta + \sigma)^{-1}. \quad (2.4)$$

Taking logarithms on both sides of (2.2) we get

$$\text{Log } S(xi) = -\tfrac{1}{\sigma}(x_i - \theta), i = 1,2,.....n. \quad (2.5)$$

or

$$x_i = \theta - \sigma \log S(x_i), i = 1,2,....n. \quad (2.6)$$

Equation (2.6) can be written in the form, $Y_i = AX_i + b$, where $A = -\sigma, Y_i = x_i, X_i = \log S(x_i)$ and $b = \theta$. By least square procedure, the estimator of A is

$$\hat{A} = \frac{\sum_{i=1}^{n} Y_i - nb}{\sum_{i-1}^{n} X_I} = \frac{\sum_{i-1}^{n} x_i - n\theta}{\sum_{i=1}^{n} \log S(x_i)}, \quad (2.7)$$

where

$$\hat{\sigma}_{ls} = -\frac{\sum_{i-1}^{n} x_i - n\theta}{\sum_{i-1}^{n} \log S(x_i)}. \quad (2.8)$$

An estimate of the survival function $S(x_i)$ is $1 - \hat{F}(x_{i:n} : \theta))$ where $x_{i:n}$ is the i^{th} order statistic and $\hat{F}(x_{i:n} : \theta) = \frac{i}{n}$, the empirical distribution function. In order to avoid log (0), D'Agostino and Stephans (1986) suggested that, $\hat{F}(x_{i:n} : \theta)$ can be approximated by, $\frac{i-c}{n-2c+1,}$ i = 1, 2,….. n where $0 \leq c \leq 1$, generally. In this paper, we take the three popular values for c, viz c = 0, 0.3 and 0.5. Then (2.8) becomes

$$\hat{\sigma}_{ls} = -\frac{\sum_{i-1}^{n} x_i - n\theta}{\sum_{i-1}^{n} \log \frac{n+1-c-i}{n+1-2c}} \ . \tag{2.9}$$

From (2.7) we obtain the estimate of L and G respectively as

$$\hat{L}_{ls} = p + \hat{\sigma}_{ls}(\theta + \hat{\sigma}_{ls})^{-1}(1-p)\log(1-p), \tag{2.10}$$

and

$$\hat{G}_{ls} = \frac{\hat{\sigma}_{ls}}{2}(\theta + \hat{\sigma}_{ls})^{-1}. \tag{2.11}$$

From (2.9), we obtain the estimate of L and G respectively as

$$\hat{L}_{ls} = p + \hat{\sigma}_{cs}(\theta + \hat{\sigma}_{cs})^{-1}(1-p)\log(1-p), \tag{2.12}$$

and

$$\hat{G}_{ls} = \frac{\hat{\sigma}_{ls}}{2}(\theta + \hat{\sigma}_{ls})^{-1}. \tag{2.13}$$

2.2 Estimation using type-II censored sample

Censored data are commonly encountered in practical applications to income and wealth distributions, for several reasons such as confidentiality or convenience. Sometimes the data may have been censored by the data providers. This usually affects extreme values, the problem of `top-coding' and `bottom-coding'. Some high income observations may have been removed from the sample because of concern for confidentiality. Sometimes extreme values have been modified because it was inconvenient to store large numbers in the data set. Therefore, here we consider the censored sample situation.

Suppose that n items, X1,X2, …. Xn which follows exponential distribution, are put on a test. Let $\{X_i\}$ be a sequence of independent and identically distributed (iid) random variables with pdf (2.1). Assuming that $\underline{x} = (x_{(2)},...x_{(r)})$ are the first *r* type –II right censored sample from a sequence$\{X_{(i)}\}$ of iid exponential distribution with pdf (2.1) are observed. The test is terminated when a pre-assigned number of items, say r(< n) have failed. The lifetimes of these first r failed items say $\underline{x} = (x_{(2)},...x_{(r)})$ are observed. In this case, we derived the least square estimator for Lorenz Curve and Gini-index, by replacing the empirical survival function Sn(x) with the Kaplan-Meier (Kaplan and Meier, 1958) estimator

$$S_r(x) = \prod_i \ (1 - \frac{1}{n-i+1})^{di}, \tag{2.14}$$

where

$$d_i = \begin{cases} 1 & if\ x_i < x_r \\ 2 & if\ x_i < x_r \end{cases}. \tag{2.15}$$

This gives

$$\hat{\sigma}_{rcs} = -\frac{\sum_{i=1}^{n} xi - n\theta}{\sum_{i-1}^{n} \log(1 - \frac{1}{n-i+1})}. \tag{2.16}$$

The estimate of Lorenz Curve and Gini-index under type –II censored samples are given respectively as

$$\hat{L}_{rcs} = p + \hat{\sigma}_{rcs}(\theta + \hat{\sigma}_{rcs})^{-1}(1-p)\log(1-p), \tag{2.17}$$

and

$$\hat{G}_{rcs} = \frac{\hat{\sigma}_{rcs}}{2}(\theta + \hat{\sigma}_{rcs})^{-1}. \tag{2.18}$$

2.3 Estimation using type-I censored sample

In this section, we propose a semi parametric estimator for the Lorenz Curve and Gini-index of the exponential distribution (2.1), when the data is censored at a pre determined time T. Suppose n components with exponential life times are put on test and we observe the number of components failed at each time point *x, x + k, x + 2k, ...* up to the time T. Define, n_j as the number of components still functioning at the *time* $x_j = x+jk,\ j = 0, 1, 2, \ldots\ x_j \leq T$ and $d - j$ as the number of components whose failure occurs in the time interval $(x_j\text{-}1, x_j)$.Then the Kaplan-Meier estimator of the survival function S(x) (Jan *et al.*, 2005) for a given x is

$$\mathrm{S*(x)} = \prod_{j:x_{(j)} \leq x} \left(\frac{nj - dj}{nj} \right) \tag{2.19}$$

for $x < x_1$; $S*(x) = 1$. Then by least square procedure as in Section2.1, we have

$$\hat{\sigma}_{tcs} = -\frac{\sum_{i=1}^{n} xi - n\theta}{\sum_{i=1}^{n}\sum_{j:x_j \le x} log\left(1-\frac{n_j}{d_j}\right)}, \tag{2.20}$$

where m_j as the number of components still functioning at the time y, $y_j = y+jk$, and d_j as the number of components whose failure occurs in the time interval $(y_j\text{-}1, y_j)$. This leads to the estimate of Lorenz Curve and Gini-index as

$$\hat{L}_{tcs} = p + \hat{\sigma}_{tcs}(\theta + \hat{\sigma}_{tcs})^{-1}(1-p)log(1-p), \tag{2.21}$$

and

$$\hat{G}_{tcs} = \frac{\hat{\sigma}_{tcs}}{2}(\theta + \hat{\sigma}_{tcs})^{-1} \tag{2.22}$$

.

2.4 Estimation under interval censoring

Interval censored data arises naturally in some applications. For Example, when a failure time T cannot be observed, but can only be determined to lie in an interval obtained from a sequence of examination times. Jan *et al.* (2005) proposed weighted estimate of the survival function of the exponential distribution under interval censoring. In the present section we extent the result to obtain weighted estimate of Lorenz Curve and Gini-index under interval censoring.

Suppose n components with exponential life times are put on test and let data set consists of n observations $x_1; x_2; \ldots; x_n$. It there is no censoring, x_i is the failure time of the component. But in this case we consider the observation times $x_1; x_2; \ldots; x_n$ includes the censoring time also. That is, for some of the x_j it is only known that the equipment j was still function at x_j and disappeared (either failed or censored) after time x_j. Let r be the number of distinct failure times or censored times and $x_{(1)} < \ldots < x_{(r)}$ be the ordered failure times. Define n_j as the number of equipment functioning just before $x_{(j)}$ including the equipment about to fail. Define d_j as the number of equipment which has failed and c_j as the number of equipment censored at time $x_{(j)}$. Then the weight w_j at time $x_{(j)}$ is defined as (Jan et al., 2005) $w_j = \frac{n_j - c_j}{n_j}$. Here $w_j = 1$ if there is no censoring, and $w_j < 1$, if there is censoring at $x_{(j)}$. Then the weighted Kaplan-Meier estimator of the survival function $S_{(x)}$ is

$$S*(x)=\begin{cases}1, \text{ for } x=0\\ \prod_{j:x_{(j)}\le x} w_j\left(\frac{n_j-d_j}{n_j}\right), \text{ for } x_{(j)},\\ 0, \text{ for } x>x_n\end{cases} \tag{2.23}$$

which gives

$$\hat{\sigma}_{ics}=-\frac{\sum_{i=1}^{n}x_i-n\theta}{\sum_{i=1}^{n}\sum_{j:x_j\le x}\left[\log w_j+\log\left(1-\frac{n_j}{d_j}\right)\right]} \tag{2.24}$$

.

Then the estimate of Lorenz Curve and Gini-index turns out to be

$$\hat{L}_{ics}=p+\hat{\sigma}_{ics}(\theta+\hat{\sigma}_{ics})^{-1}(1-p)log(1-p), \tag{2.25}$$

and

$$\hat{G}_{ics}=\frac{\hat{\sigma}_{ics}}{2}(\theta+\hat{\sigma}_{ics})^{-1} \tag{2.26}$$

.

3. Simulation results

In this Section we present the results of a simulation study. In order to assess the performance of the estimators of the Lorenz Curve and Gini-index, we perform a simulation study of 2000 samples of sizes n = 20; 30; 50 and 100 generated for different values of the parameters specified in the table. We present the simulation results concerning the mean and mean square errors of all these estimators. The computational results for the means and MSEs (in parentheses) are displayed in Tables 1-3. The means and mean squared errors of the estimators under least square method are presented in Table 1, and censored sampling are presented in Tables 2-3. It is to be noted that the MSE of the estimators become smaller as the sample size increases.

Table 1: Means and MSEs (in parentheses) of the estimates of Lorenz Curve and Gini-index under least square method

		σ = 2.5	σ =2.2	σ =2.5	σ =3.5
True Lorenz		0.07106	0.06786	0.06366	0.05330
n = 20	$\hat{L}_{ls}$	0.04032	0.04207	0.04284	0.02476
		(0.0223)	(0.0324)	(0.0239)	(0.0221)
	$\hat{G}_{ls}$	0.25032	0.26207	0.28284	0.32476
		(0.0223)	(0.0324)	(0.0239)	(0.0221)
n = 30	$\hat{L}_{ls}$	0.05621	0.05608	0.05602	0.03603
		(0.0123)	(0.0214)	(0.0117)	(0.0113)
	$\hat{G}_{ls}$	0.25621	0.27608	0.29602	0.33603
		(0.0123)	(0.0214)	(0.0117)	(0.0113)
n = 50	$\hat{L}_{ls}$	0.06227	0.06725	0.06228	0.04326
		(0.0011)	(0.0017)	(0.0034)	(0.0017)
	$\hat{G}_{ls}$	0.26227	0.28725	0.30228	0.34326
		(0.0011)	(0.0017)	(0.0034)	(0.0017)
n = 100	$\hat{L}_{ls}$	0.07521	0.0609	0.06912	0.05231
		(0.0009)	(0.0001)	(0.0005)	(0.0004)
	$\hat{G}_{ls}$	0.28521	0.2909	0.31912	0.35231
		(0.0009)	(0.0001)	(0.0005)	(0.0004)

 ISBN-1391-4987

Table 2: Means and MSEs (in parentheses) of the estimates of Lorenz Curve under various censored schemes

		σ = 2.5	σ =2.2	σ =2.5	σ =3.5
True Lorenz		0.063662	0.07870	0.076613	0.08860
n = 20	$\hat{L}_{rcs}$	0.060318	0.043243	0.051594	0.057525
		(0.000618)	(0.000944)	(0.000767)	(0.000657)
	$\hat{L}_{tcs}$	0.060323	0.043251	0.051600	0.057531
		(0.000616)	(0.000943)	(0.000767)	(0.000656)
	$\hat{L}_{ics}$	0.062046	0.045214	0.053544	0.059302
		(0.000611)	(0.000833)	(0.000663)	(0.000570)
n = 30	$\hat{L}_{rcs}$	0.068741	0.050665	0.059226	0.065252
		(0.000408)	(0.000504)	(0.000418)	(0.000317)
	$\hat{L}_{tcs}$	0.068743	0.050670	0.059230	0.065254
		(0.000410)	(0.000503)	(0.000418)	(0.000317)
	$\hat{L}_{ics}$	0.069664	0.052021	0.060366	0.066263
		(0.000412)	(0.000445)	(0.000373)	(0.000282)
n = 50	$\hat{L}_{rcs}$	0.075529	0.051460	0.065219	0.071857
		(0.000209)	(0.000472)	(0.000199)	(0.000125)
	$\hat{L}_{tcs}$	0.075529	0.051463	0.065221	0.071857
		(0.000209)	(0.000472)	(0.000199)	(0.000125)
	$\hat{L}_{ics}$	0.075972	0.052278	0.065813	0.072357
		(0.000222)	(0.000438)	(0.000183)	(0.000114)
n = 100	$\hat{L}_{rcs}$	0.078886	0.056867	0.068669	0.076014
		(0.000113)	(0.000259)	(0.000104)	(0.000048)
	$\hat{L}_{tcs}$	0.078886	0.056868	0.068669	0.076015
		(0.000118)	(0.000104)	(0.000104)	(0.000048)
	$\hat{L}_{ics}$	0.079075	0.057237	0.68945	0.76230
		(0.000320)	(0.000247)	(0.000099)	(0.000045)

Table 3: Means and MSEs (in parentheses) of the estimates of Gini-index under various censored schemes

		σ = 1.5	σ =0.75	σ =0.48	σ =0.39
True Lorenz		0.25000	0.16666	0.12121	0.103317
n = 20	$\hat{G}_{rcs}$	0.250007	0.415864	0.345515	0.280050
		(0.000817)	(0.065380)	(0.052313)	(0.033912)
	$\hat{G}_{tcs}$	0.250236	0.416829	0.346202	0.280484
		(0.000823)	(0.065887)	(0.052646)	(0.034080)
	$\hat{G}_{ics}$	0.236545	0.394478	0.327240	0.264769
		(0.000919)	(0.055098)	(0.044229)	(0.028488)
n = 30	$\hat{G}_{rcs}$	0.196218	0.358446	0.287541	0.209092
		(0.003266)	(0.040811)	(0.028089)	(0.012697)
	$\hat{G}_{tcs}$	0.196328	0.358939	0.287828	0.209222
		(0.003255)	(0.041026)	(0.028190)	(0.012727)
	$\hat{G}_{ics}$	0.189473	0.345426	0.277228	0.201728
		(0.004011)	(0.035705)	(0.024714)	(0.011092)
n = 50	$\hat{G}_{rcs}$	0.189571	0.345869	0.277483	0.201844
		(0.004000)	(0.035887)	(0.024799)	(0.011117)
	$\hat{G}_{tcs}$	0.153241	0.328175	0.224546	0.175755
		(0.009769)	(0.027566)	(0.011234)	(0.006064)
	$\hat{G}_{ics}$	0.153276	0.328396	0.224634	0.175803
		(0.009764)	(0.027643)	(0.011254)	(0.006072)
n = 100	$\hat{G}_{rcs}$	0.150133	0.321253	0.219965	0.172186
		(0.010363)	(0.025308)	(0.010294)	(0.005523)
	$\hat{G}_{tcs}$	0.127559	0.291718	0.196472	0.146123
		(0.015106)	(0.016399)	(0.005285)	(0.002244)
	$\hat{G}_{ics}$	0.127569	0.291794	0.196501	0.146138
		(0.015104)	(0.016419)	(0.005289)	(0.002245)

4. Conclusion

The present paper proposes semi-parametric approaches to estimate the Lorenz curve and Gini-index of exponential distribution. Comparisons are made between the different estimators based on a simulation study, and it is to be noted that the MSE of the estimators become smaller as the sample size increases.

Acknowledgment

The authors are grateful to the editor and the referees for their valuable comments on an earlier version of the paper.

References

1. Afify, E.E. (2003). Estimation of Parameters for Pareto Distribution. InterStat, http:// interstat.statjournals.net/ YEAR/2003/ articles/ 0302004.pdf.
2. Bhattacharjee, M.C. (1993). How rich are the rich? Modeling affluence and inequality via reliability theory. *Sankhyā*, **55**,1-26.
3. Chandra, M. and Singpurwalla, N.D. (1981). Relationships between some notions which are common to reliability theory and economics. *Mathematics of Operations Research*, **6**, 113-121, DOI: 10.1287/moor.6.1.113
4. D'Agostino, R.B. and Stephens, M.A. (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker, Inc.
5. Gastwirth, J.L. (1971). A general definition of the Lorenz curve. *Econometrica*, **39**, 1037-1039, DOI: 10.2307/1909675
6. Iyengar, N.S. (1960). On the standard error of the Lorenz concentration ratio. *Sankhyā*, **22**, 371-378.
7. Jan, B., Ali Shah, W.S., Shah, S. and Qudir, F.M. (2005). Weighted Kaplan-Meier estimation of survival function in heavy censoring. *Pakistan Journal of Statistics*, **21(1)**, 55-63.
8. Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of American Statistical Association*, **53**, 457-481, DOI: 10.1080/01621459.1958.10501452
9. Moothathu, T.S.K. (1985a). Sampling distributions of Lorenz curve and Gini index of the Pareto distribution. *Sankhyā*, **47**, 247-258.
10. Moothathu, T.S.K. (1985b). Distributions of maximum likelihood estimators of the Lorenz curve and Gini index of exponential distribution. *Annals of the Institute of Statistical Mathematics*, **37**, 473-479, DOI: 10.1007/BF02481115
11. Moothathu, T.S.K. (1989). On unbiased estimation of Gini index and Yntema-pietra index of lognormal distribution and their variances. *Communications in Statistics - Theory and Methods*, **18**, 661-672, DOI: 10.1080/03610928908829925

12. Moothathu, T.S.K. (1990). The best estimator and strongly consistent asymptotically normal unbiased estimator of Lorenz curve, Gini index and theil entropy index of the Pareto distribution, *Sankhyā*, **52**, 115-127.

13. Sathar, A.E.I., Jeevanand, E.S. and Nair, K.R.M. (2005). Bayesian estimation of Lorenz curve, Gini-index and variance of logarithms in a in a Pareto distribution. *Statistica*, **65(2)**, 193-205.