Metodološki zvezki, Vol. 1, No. 1, 2004, 109-118

Maximum Likelihood Estimation of Lorenz Curves using Alternative Parametric Model

Ibrahim M. Abdalla[1] and Mohamed Y. Hassan[2]

Abstract

In this paper the Lorenz curve proposed by Abdalla and Hassan is fitted to grouped income data of Abu-Dhabi Emirate family expenditure survey, 1997, using Maximum likelihood estimation method and assuming that income shares follow a Dirichlet distribution. Employing Abdalla and Hassan's together with some known parametric Lorenz models, estimates based on the maximum likelihood are compared with those based on non-linear least squares techniques. Given the nature of the distribution of income and the distinct characteristics of Abu-Dhabi Emirate, it is evident that the maximum likelihood estimation approach produces comparable parameter estimates to the non-linear least squares techniques, but higher standard errors and less goodness of fit. Under the two estimation techniques, the model proposed by Abdalla and Hassan performed well better than some well known parametric models in the literature.

1 Introduction

The Lorenz curve is a graphical representation, usually adopted to depict the distribution of income and wealth in a population. Horizontally, it displays the cumulative proportion of the population, say p, arranged in increasing order of income. Vertically, it measures the proportion of income, η, accruing to any particular fraction of the population thus arranged.

Direct estimation of the Lorenz curve is based on linear interpolation using empirical data presented in income group format. This representation assumes a

[1] College of Business and Economics, United Arab Emirates University, P. O. Box 17555; i.abdalla@uaeu.ac.ae

[2] College of Business and Economics, United Arab Emirates University, P. O. Box 17555; Myusuf@uaeu.ac.ae

complete uniformity of income distribution within each group. Consequently the estimated curve converges to the true Lorenz curve as the number of income groups increases. However, various models for parametric Lorenz curves have been suggested in the literature, including Kakwani (1980), Basmann *et al*. (1990), Ortega *et al*. (1991) and Chotikapanich (1993). Other models are based on well known income distribution models such as lognormal and gamma. Kakwani and Podder (1976) noted that Lorenz curves, driven from known density functions, hardly give reasonably good fit to actual data.

By means of goodness of fit tests, Gini ratios , and estimated income shares, it is evident that results of employing different models on different income data for different countries are in sharp contrast (Chotikapanich, 1993, Sarabia *et al*., 1999 and Cheong, 1999). Some models give a good approximation to the data over the middle of the distribution, but not well over the tails. The opposite happens when analyzing some other models. Arguably, these differences may be attributable to differences in the nature of income distributions and distinct characteristics of different countries.

The objective of this paper is to estimate the Lorenz curve for income data from the most recent Abu-Dhabi Emirate family expenditure survey, Ministry of Planning (1997), using the maximum likelihood approach suggested by Chotikapanich and Griffiths (2002) and compare estimates with those based on the non-linear least square techniques. Along with the model proposed by Abdalla and Hassan (2004), additional models are evaluated and corresponding Gini concentration ratios are assessed in terms of accuracy of fit.

The plan of this paper is as follows: In Section 2 some known parametric Lorenz curves (models), including the one proposed by Abdalla and Hassan (2004) are presented. In Sections 3 and 4 estimation methods are outlined. Goodness of fit and empirical results are discussed in Sections 5 and 6. Section 7 is dedicated to some concluding remarks.

2 Parametric Lorenz Curves

Suppose that population units, individuals or families, are ordered according to increasing income. Thus income distribution data can take the form (p_i, η_i) for each of T income groups, where p_i is the cumulative proportions of population units associated with group i, with $p_T = 1$, and η_i is the corresponding cumulative proportions of income, with $\eta_T = 1$. Dropping the subscripts and utilizing these observations, a parametric Lorenz curve is given as $\eta = L(p;\theta)$, where θ is a vector of unknown parameters. A function is a Lorenz curve if it satisfies the following conditions:

$(i)\quad L(0;\theta)=0,$

$(ii)\quad L(1;\theta)=1,$

(iii) $L(p;\theta)$ is twice continuously differentiable monotone increasing function; therefore: $L'(p;\theta)\geq 0,\quad L''(p;\theta)\geq 0.$

Abdalla and Hassan (2004) suggest the model defined by

$$L(p)=p^{\alpha}[1-(1-p)^{\delta}e^{\beta p}]\qquad \alpha\geq 0,\ 0\leq\beta\leq\delta\leq 1 \tag{2.1}$$

The Gini concentration ratio (R_G) associated with the different parametric models are calculated based on the definition

$$R_G=1-2\int_0^1 L(p;\theta)dp \tag{2.2}$$

In particular, the Gini ratio associated with Abdalla and Hassan model is reduced to the following closed form

$$\begin{aligned} R_G &= 1-2\int_0^1 p^{\alpha}[1-(1-p)^{\delta}e^{\beta p}dp \\ &= \frac{\alpha-1}{\alpha+1}+2\sum_{j=0}^{\infty}\frac{\Gamma(\alpha+j+1)\Gamma(\delta+1)\beta^{j}}{\Gamma(j+1)\Gamma(\alpha+\delta+j+2)} \end{aligned} \tag{2.3}$$

Using Abu-Dhabi Emirate income data set, 1997, the performance of the parametric model suggested by Abdalla and Hassan, L_{AH}, is compared with different alternatives. The first alternative is the model L_C introduced by Chotikapanich (1993). It is a simple one parameter model:

$$L_C(p;\alpha)=\frac{e^{\alpha p}-1}{e^{\alpha}-1}\qquad \alpha>0 \tag{2.4}$$

Ortega *et al.* (1991) suggested the model L_O which is a special case of Abdalla and Hassan model L_{AH} when $\beta=0$;

$$L_O(p;\alpha,\delta)=p^{\alpha}[1-(1-p)^{\delta}]\qquad \alpha\geq 0,\quad 0<\delta\leq 1 \tag{2.5}$$

When $\beta=0$ and $\alpha=0$ L_{AH}, is reduced to $L_P(p;\delta)=1-(1-p)^{\delta}$, which originates from the Pareto distribution. This curve is a special case of Rasche's *et al*. (1980), two parameter model, L_R, with $\alpha=1$

$$L_R(p;\delta,\alpha)=[1-(1-p)^{\delta}]^{\alpha}\qquad \alpha\geq 1,\quad 0<\delta\leq 1 \tag{2.6}$$

Kakwani (1980) introduced the beta model;

$$L_K(p;\alpha,\delta,\beta)=p-\alpha p^{\delta}(1-p)^{\beta}\qquad \alpha>0,\ 0<\delta\leq 1,\quad 0<\beta\leq 1 \tag{2.7}$$

Kakwani’s proposal is considered as the best performer compared to a number of different models; see for example Chotikapanich and Griffith (2002), Cheong (1999) and Sarabia *et al*. (1999).

3 Maximum Likelihood Estimation

Chotikapanich and Griffiths (2002) proposed a maximum likelihood estimation methodology based on assuming income proportions (from grouped data) as realizations of a Dirichlet distribution. They argued that Lorenz curve estimation based on linear or non-linear least squares techniques defy the reality that error terms associated with observations on cumulative proportions are neither independent nor normally distributed. Alternatively, conditional on the population proportions p_i, income shares $q_i = \eta_i - \eta_{i-1}$ are considered as random variables with means $E(q_i) = E(\eta_i) - E(\eta_{i-1}) = L(p_i;\theta) - L(p_{i-1};\theta)$. The vector $q = (q_1, q_2, \ldots, q_T)'$ is assumed to be generated from a Dirichlet distribution with density

$$f(q\,|\,\alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_T)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_T)} q_1^{\alpha_1 - 1} q_2^{\alpha_2 - 1} \cdots q_T^{\alpha_T - 1} \tag{3.1}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_T)'$ are the parameters of the Dirichlet density and $\Gamma(\cdot)$ is the gamma function. Setting α as a function of the curve parameters, $\alpha_i = \lambda[L(p_i;\theta) - L(p_{i-1};\theta)]$, a probability density function for q with mean $E(q_i)$ is given by

$$f(q\,|\,\phi) = \Gamma(\lambda) \prod_{i=1}^{T} \frac{q_i^{\lambda[L(p_i;\theta) - L(p_{i-1};\theta)] - 1}}{\Gamma(\lambda[L(p_i;\theta) - L(p_{i-1};\theta)])} \tag{3.2}$$

where $\phi = (\theta', \lambda)'$ and λ is an additional parameter. This leads to the possibility of getting maximum likelihood estimate for ϕ which is based on maximizing the likelihood function,

$$\begin{aligned} \log[f(q\,|\,\phi)] = \log\Gamma(\lambda) + \sum_{i=1}^{T}(\lambda[L(p_i;\theta) - L(p_{i-1};\theta)] - 1) \times \log q_i \\ - \sum_{i=11}^{T} \log\Gamma(\lambda[L(p_i;\theta) - L(p_{i-1};\theta)]) \end{aligned} \tag{3.3}$$

4 Non-linear Least Squares Estimation

Non-linear least squares estimators are defined by the estimators which minimize the sum of the squared differences between the predicted and observed values. For a particular Lorenz curve $L(p;\theta)$ the minimization is associated with the expression

$$\sum_{i=1}^{T}(\eta_i - L(p_i;\theta))^2.$$

Least squares estimates are usually obtained by direct numerical search procedures implemented in many statistical computer packages.

5 Goodness of Fit

Goodness of fit of alternative models discussed in Section 2 is based on comparing values of the information inaccuracy measure as shown in Theil (1967), and Chotikapanich and Griffiths (2002), which is given by

$$I = \sum_{i=1}^{T} q_i \, log(\frac{q_i}{\hat{q}_i}), \tag{5.1}$$

where $\hat{q}_i$ is the predicted income shares from an estimated model.

Smaller values of the index I depicted in (5.1) indicate better fits. Moreover, the index I can be used as a Deviance measure of fit, Agresti (1996). A value of zero indicates a perfect model fit. For each model, the value of I is compared with χ^2 with $T-k$ degrees of freedom, where T denotes the number of income groups and k is the number of parameters in the model. This shows some sort of a penalty when too many parameters are included in the model.

Gastwirth (1972) criterion for the Gini ratio is also used as another goodness of fit measure.

Assessment of the reliability of the two techniques, maximum likelihood (ML) and non-linear least squares (NL) is based on comparing parameter estimates and associated standard errors. The standard errors for the ML and the NL estimates are estimated by utilizing the Hessian matrix generated from a numerical maximization/minimization routine as shown in the “Introduction to R” manual, pages 66-67, R Development Team (2000).

6 Empirical Results

Incomes of 3206 families that constitute Abu-Dhabi Emirate surveyed sample are summarized by grouping the data into $T = 12$ intervals. Income share q_i for each interval (income group) together with cumulative proportion of income η_i and cumulative proportion of families p_i are calculated, $i = 1,\ldots,T$. These values are used in fitting the models outlined in Section 2 using ML and NL estimation techniques.

Table 1: Lorenz curve: Maximum likelihood (ML) and non-linear least squares (NL) parameter estimates, and related standard errors (*s.e.*) and Gini ratios (R_G).

Lorenz Function	Estimates (*s.e.*)	I	R_G
Chotikapanich	$\hat{\alpha} = 2.9646$ (0.3497)	0.03252	0.4341
ML	$\hat{\alpha} = 3.4876$ (0.1323)	0.03615	0.4896
Ortega *et al.*	$\hat{\alpha} = 0.3650$ (0.1797) $\hat{\delta} = 0.4835$ (0.0524)	0.00129	0.4370
ML	$\hat{\alpha} = 0.4543$ (0.0255) $\hat{\delta} = 0.4787$ (0.0045)	0.00115	0.4585^*
Rasche *et al.*	$\hat{\delta} = 0.5383$ (0.0679) $\hat{\alpha} = 1.3067$ (0.1586)	0.00077	0.4427^*
ML	$\hat{\delta} = 0.5577$ (0.0051) $\hat{\alpha} = 1.3862$ (0.0148)	0.00048	0.4564^*
Kakwani	$\hat{\alpha} = 0.8040$ (0.1400) $\hat{\delta} = 1.0189$ (0.0933) $\hat{\beta} = 0.4558$ (0.0919)	0.00125	0.4442^*
ML	$\hat{\alpha} = 0.8168$ (0.0223) $\hat{\delta} = 1.0205$ (0.0250) $\hat{\beta} = 0.4484$ (0.0127)	0.00115	0.4544^*
Abdalla and Hassan	$\hat{\alpha} = 0.0476$ (0.3738) $\hat{\delta} = 0.5798$ (0.1098) $\hat{\beta} = 0.3154$ (0.3008)	0.00076	0.4454^*
ML	$\hat{\alpha} = 0.1447$ (0.0575) $\hat{\delta} = 0.5515$ (0.0111) $\hat{\beta} = 0.2466$ (0.0383)	0.00031	0.4536^*

* Satisfy Gastwirth criterion.

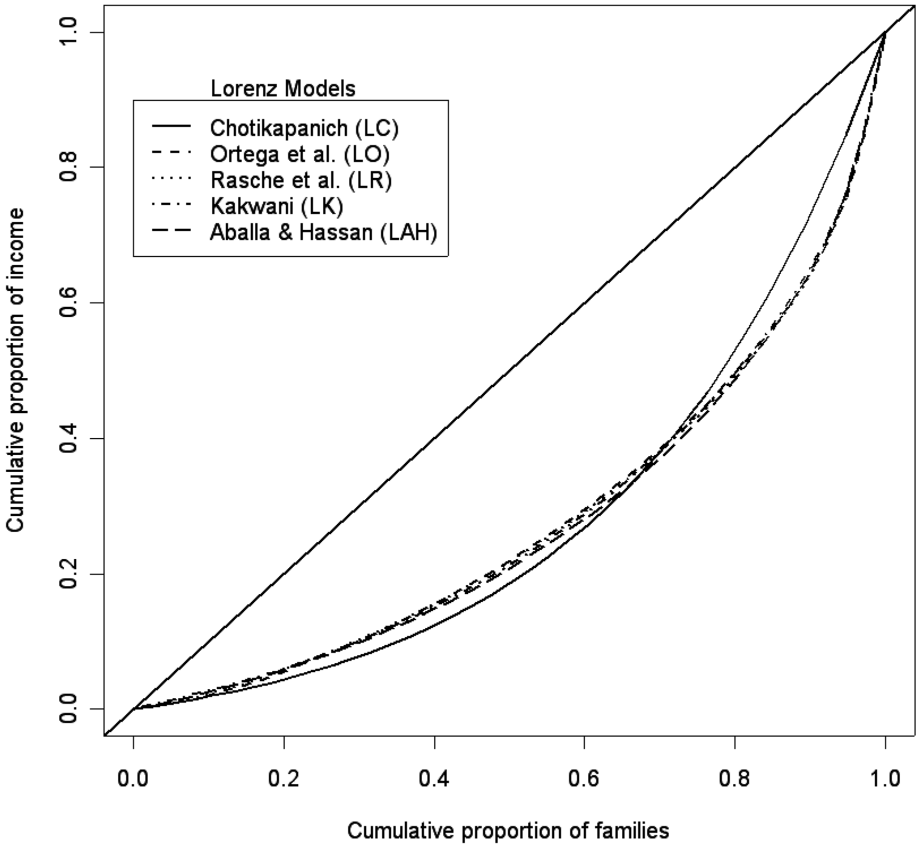

Figure 1: Lorenz curves for Abu-Dhabi Emirate income data, 1997.

Table 1 reports point estimates and standard errors for the parameters and the Gini ratio, R_G , associated with each model together with the values of information inaccuracy, *I*, as goodness of fit measures. It is evident from Table 1 that the estimates of the Lorenz curve parameters associated with ML estimation are slightly different and have higher standard errors compared to estimates associated with NL estimation. Gini ratios associated with ML estimation are noticeably smaller compared to those associated with NL estimation. Using I as a Deviance measure of fit and compare it with the value of χ^2 with $T-k$ degrees of freedom (see Section 5) , there is no evidence of model lack of fit. The p-value (the right tail probability) associated with each model is 1. However, the model L_{AH} proposed by Abdalla and Hassan is associated with the smallest values of I compared to the other models. This indicates that L_{AH} provides better fit compared to the other models.

Table 2: Actual and Estimated Income Shares (based on maximum likelihood (ML) and non-linear least squares (NL) methods) for the U.A.E. Income Groups[*]

Income	Actual	Chotikapanich	Ortega	Rasche *et al.*	Kakwani	Abdalla &
			et al.			Hassan
(000DHs)		L_C	L_O	L_R	L_K	L_{AH}
Less than 3	0.30	0.37	0.27	0.31	0.57	0.50
ML	0.30	0.25	0.19	0.23	0.55	0.39
3 – 6	5.20	4.06	5.46	5.54	5.38	5.48
ML	5.20	2.95	4.73	4.92	5.18	5.28
6 – 9	10.50	9.13	11.27	10.97	10.25	10.21
ML	10.50	7.38	10.68	10.59	9.98	10.21
9 – 12	10.60	11.44	11.17	10.84	10.80	10.58
ML	10.60	10.18	11.02	10.83	10.61	10.54
12 – 15	9.60	11.99	9.89	9.68	10.00	9.85
ML	9.60	11.43	9.94	9.83	9.89	9.76
15 – 18	7.20	9.48	7.15	7.07	7.40	7.37
ML	7.20	9.46	7.26	7.24	7.35	7.29
18 – 21	6.30	8.47	6.17	6.15	6.44	6.49
ML	6.30	8.73	6.29	6.33	6.42	6.41
21 – 24	4.50	6.01	4.34	4.37	4.55	4.64
ML	4.50	6.32	4.44	4.50	4.55	4.58
24 – 27	5.00	6.18	4.52	4.58	4.74	4.86
ML	5.00	6.60	4.63	4.72	4.74	4.81
27 – 30	4.50	5.35	4.03	4.11	4.21	4.36
ML	4.50	5.80	4.13	4.24	4.23	4.31
30 –33	3.70	4.27	3.35	3.43	3.48	3.63
ML	3.70	4.68	3.43	3.54	3.50	3.60
33 and over	32.70	23.24	32.39	32.95	32.18	32.03
ML	32.70	26.20	33.27	33.04	33.00	32.83

[*] Underlined numbers represent estimated income shares that are closest to the actual

Estimated Gastwirth bounds corresponding to the 12 income groups have been calculated. This is based on the idea that any fitted curve whose related Gini ratio falls outside the lower and upper bounds should be declared to fit the data inadequately. Thus, using Abu-Dhabi Emirate data, a lower bound of 0.4414 and an upper bound of 0.4600 are produced. Based on the results reported in Table 1, it is evident that the Gini ratios associated with the NL estimation of Ortega *et al.*, Rasche, Kakawani and Abdalla and Hassan models are contained within Gastwirth bounds, whereas the one estimated using Chotikapanich model violates the criterion. Those associated with ML estimation of Chotikapanich and Ortega *et al*. models also violates Gastwirth criterion.

Based on maximum likelihood estimates, the comparison between the competing Lorenz models depicted in Figure 1 suggests that, optically, four of the curves provide almost identical fit. The only one which is distinguishable is the one-parameter L_C model. This is also confirmed by the largest value I = 0.03.

Moreover, actual and estimated income shares for different models and income groups employing Abu-Dhabi data are displayed in Table 2. It can be noted that no single model consistently outperforms the others. However, the model proposed by Abdalla and Hassan, L_{AH}, has the maximum number of the closest estimates to actual shares, using either estimation techniques.

7 Concluding remarks

Empirical analysis conducted in this paper compared two estimation techniques for Lorenz curves, namely the maximum likelihood and non-linear least squares, in terms of goodness of fit and reliability of parameter estimates. Findings and reported results based on Abu-Dhabi Emirate income data, a country with relatively middle inequality, suggest that non-linear least squares techniques provide better and more reliable fit compared to the maximum likelihood method. This contradicts conclusions reported by Chotikapanich and Griffith (2002). The Lorenz model proposed by Abdalla and Hassan performed well better than some known models in the literature. Generally, as confirmed previously by Chotikapanich and Griffith, the estimation of Gini ratios does not depend on the estimation technique or Lorenz model specification. The results reported for Abu-Dhabi data do not automatically carry over to other data sets and other models; none the less, they are meaningful for the specific context. It would be useful to study the time trend of the indices, in order to evaluate and forecast the tendencies of income distribution. This, however, depends on availability of periodic and deeper surveys.

References

[1] Abdalla, I.M. and Hassan, M.Y. (2004): Fitting income distribution using a new parametric Lorenz curve. *Pakistan Journal of Statistics*, to appear.

[2] Agesti, A. (1996): *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

[3] Basmann, R.L., Hayes, K.J., Slottje, D.J., and Johnson, J.D. (1990): A general functional form for approximating the Lorenz curve. *Journal of Econometrics*, **43**, 77-90.

[4] Cheong, K.S. (1999): A comparison of alternative functional forms for parametric estimation of the Lorenz curve. *Working Paper No. 99-2R*, Department of Economics, University of Hawaii at Manoa.

[5] Chotikapanich, D. (1993): A comparison of alternative functional forms for the Lorenz curve. *Economic Letters*, **41**, 129-138.

[6] Chotikapanich, D. and Griffith, W.E. (2002): Estimating Lorenz curves using a Dirichlet distribution, *Journal of Business and Economic Statistics,* **20**, 290-295.

[7] Gastwirth, J.L. (1972): The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics,* **54**, 306-316.

[8] Kakwani, N.C. (1980): On a class of poverty measures. *Econometrica*, **48**, 437-446.

[9] Kakwani, N.C. and Podder, N. (1976): Efficient estimation of the Lorenz curve and associated inequality measures from grouped data. *Econometrica*, **44**, 137-148.

[10] Ministry of Planning, Statistics Department (1997): *Abu-Dhabi Emirate Family Expenditure Survey*. Abu-Dhabi, United Arab Emirates.

[11] Ortega, P.G., Martin, A., Fernandez, M., Lodoux, M., and Garcia, A. (1991): A new functional form for estimating Lorenz curves. *Review of Income and Wealth*, **37**, 447-452.

[12] R Development Team (2000): *An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics*. Version 1.2.0.

[13] Ray, D. (1998): *Development Economics*. New Jersey, Princeton: Princeton University Press.

[14] Rasche, R.H., Gaffney, J., Koo, A., and Obst, N. (1980): Functional forms for estimating Lorenz curve. *Econometrica*, **48**, 1061-1062.

[15] Sarabia, J.M., Castillo, E., and Slottje, D.J. (1999): An ordered family of Lorenz curves. *Journal of Econometrics*, **91**, 43-60.

[16] Theil, H. (1967): *Economics and Information Theory*. Amsterdam: North-Holland.