

FITTING THE GENERALIZED LAMBDA DISTRIBUTION TO INCOME DATA

Agostino Tarsitano

Key words: Distribution fitting, grouped data, quantile function.
COMPSTAT 2004 section: Model selection.

Abstract: This paper proposes the generalized lambda distribution (GLD) as a model for describing the distribution of income over a population. Performances of various methods of fitting the GLD to grouped income data are evaluated. Of the estimators considered it is concluded that the unweighted least squares regression on group means should be used.

1 Introduction

There has been an increased interest in describing the distributions of personal income for the last several decades. A number of monographs have been published in the area, including those by Dagum [1], Kleiber and Kotz [4]. The study of income distributions usually provide a mathematical description F for the cumulative distribution of incomes and use it to summarize in a small number of parameters the peculiarities one discovers in empirical distributions. Also, F can be employed to smooth out irregularities in the histogram of observed data and to compute summary measures that can be compared spatially and temporally.

A wide variety of functional forms have been considered as possible models for incomes. One approach is to view the income density function as the outcome of a stochastic process (*e.g.* the Champernowne model). A second approach exploits the connections between income and aptitudes (*e.g.* the lognormal model). Also, the model is derived from a differential equation designed to capture a stable structure of observed distributions of income (*e.g.* Singh-Maddala model). Another approach is the search of a flexible analytic form, which ensures a satisfactory goodness of fit (*e.g.* the generalized beta model). Other approaches can no doubt be suggested.

The generalized lambda distribution (GLD) is a flexible and manageable tool for modeling empirical and theoretical distributions. The GLD is primarily specified by the quantile function

$$X_p(p;\lambda) = \lambda_1 + \lambda_2^{-1}\left[p^{\lambda_3} - q^{\lambda_4}\right] \quad 0 \le p \le 1,\ q = 1-p;\ \lambda_2 \neq 0 \qquad (1)$$

Where λ_1 is a location parameter, λ_2 is a linear parameter related to (though not only to) the scale of X and λ_3, λ_4 are exponential parameters determining the shape of the quantile function. The following conditions are imposed:

$$If\ \lambda_2 \to \infty\ then\ \lambda_3,\ \lambda_4 > -\infty;\ If\ \lambda_3, \lambda_4 \to \infty\ then\ |\lambda_2| > 0 \qquad (2)$$

Although there is scarcely a need for another model to fit the distribution of income the flexibility and the adaptability offered by the GLD legitimate its advancement in this context.

The basic proposition of this paper is that personal income distributions can be adequately described by using the quantile function (1). The content of the paper is organized as follows: in Section 2 the properties of GLD are described and its analytical and statistical peculiarities are summarized. Section 3 contains a discussion of several estimation procedures in the case of grouped data paying special attention to the extension of these methods to a random variable defined by its quantile function. The goodness-of-fit statistics assessing their usefulness are also considered. The results of an application to a real data set are exposed in Section 4 providing information about the relative merits of the different estimation techniques.

2 Shape, moments, and Lorenz curve of the GLD

The support of the GLD random variable is bounded $(\lambda_1 - 1/\lambda_2, \lambda_1 + 1/\lambda_2)$ if $\lambda_3, \lambda_4 > 0$ and is the real line when $\lambda_3, \lambda_4 < 0$. Hence, the extremes of $X(p, \lambda)$ are finite or infinite according to the sign of the exponential parameter. Analytic expression for the cumulative distribution function $F(x, \lambda)$ is in general not available. However, the fact that the GLD is not invertible is not a serious drawback because the same is true for many popular models such as lognormal and generalized beta. The limiting form of (1) as λ_3 diverges to ∞ is the Pareto distribution.

The probability density function of a GLD random variable is defined by the density quantile function, that is the density expressed in terms of $p = F(x, \lambda)$

$$\frac{1}{\frac{dX(p;\lambda)}{dp}} = h\left[X(p;\lambda)\right] = \frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 q^{\lambda_4 - 1}} \tag{3}$$

If $\lambda_3 = \lambda_4$ then (3) is symmetric about the pole $X = \lambda_1$. When scale and location are changed we transform the variable $Y = a + bX$. The transformed distribution is another member of the GLD family with λ_1, λ_2 replaced by $a + b\lambda_1$ and $b\lambda_2$ respectively. Expression $h[X(p, \lambda)]$ represents a legitimate probability density function if and only if it is nonnegative and integrates to one. The latter condition follows directly from (3). A good summary of the regions in which the GLD is well defined is given in Karian and Dudewicvz [3].

The ordinates of the density quantile function at the extremes of the range of variation are $(\lambda_2/\lambda_4, \lambda_2/\lambda_3)$ if $\lambda_3, \lambda_4 \geq 1$ and zero for $\lambda_3, \lambda_4 < 1$. The parameters λ_3 and λ_4 determine the type of tails of the GLD (provided that the sign of λ_2 ensures that (3) is a valid density function). For example, if $\lambda_3, \lambda_4 > 0$ then (3) has increasingly peakedness and short tails; if $\lambda_3, \lambda_4 < 0$ the tails have increasingly heaviness. The density tends to zero both as p goes to 0 and as p goes to 1 if, respectively, $\lambda_3 < 1$ and $\lambda_4 < 1$. On the other hand, if $\lambda_4 \geq 1(\lambda_3 \geq 1)$ then the density has truncated left (right) tail. The density (3) is unimodal if $\lambda_3, \lambda_4 > 2$, if $0 < \lambda_3, \lambda_4 < 1$ or if $0 < \lambda_3, \lambda_4 < 0$. It is zeromodal if $1 < \lambda_3, \lambda_4 < 2$. The arithmetic mean and the median of a GLD are

$$\mu = \lambda_1 + \lambda_2^{-1}\left\lfloor \frac{1}{(\lambda_3+1)} - \frac{1}{(\lambda_4+1)} \right\rfloor;\ M_e = \lambda_1 + \lambda_2^{-1}\left(0.5^{\lambda_3} - 0.5^{\lambda_4}\right) \quad (4)$$

Consider the linear transformation $Z = X - \lambda_1$. Then

$$E(Z^i) = \sum_{j=0}^{i} \binom{j}{i} (-1)^j \lambda_2^{-i} B(\lambda_3(i-j)+1, \lambda_4 j + 1);\ i = 1, 2, \cdots \quad (5)$$

Where $B(x, y)$ denotes the complete beta function. The i-th moment of the GLD exists if and only if $\min(\lambda_3, \lambda_4) > -i^{-1}$. Since $Z - E(Z) = X - E(X)$ the central moments of X coincide with the central moments of Z. The degree of skewness can be measured by

$$\begin{aligned}\frac{\mu - M_e}{S_{Me}} &= b(\lambda) \quad (6)\\ &= \frac{(\lambda_4+1)\left[1-(\lambda_3+1)0.5^{\lambda_3}\right] - (\lambda_3+1)\left[1-(\lambda_4+1)0.5^{\lambda_4}\right]}{(\lambda_4+1)\left[1-0.5^{\lambda_3}\right] + (\lambda_3+1)\left[1-0.5^{\lambda_4}\right]}\end{aligned}$$

where S_{Me} is the mean deviation about the median. From (6) it easily checked that (3) has a positive skewness if $\lambda_3 < \lambda_4$. The practical advantage of using $X(p, \lambda)$ instead of $F(x, \lambda)$ depends on having the $X(p, \lambda)$ in closed form. First, the Lorenz curve and other characteristics are handled simply.

$$L(p; \lambda) = \mu^{-1}\left\{\lambda_1 p + \lambda_2^{-1}\left[(\lambda_3+1)^{-1} p^{\lambda_3+1} + (\lambda_4+1)^{-1}\left(q^{\lambda_4+1} - 1\right)\right]\right\} \quad (7)$$

The condition $\lambda_2\lambda_3\lambda_4 \geq 0$ suffices to ensure the convexity of the Lorenz curve as long as the mean exists and $h[X(p, \lambda)]$ is a valid density function. Sarabia [7] used this model to define a hierarchy of Lorenz curves. Maddala and Singh [5] employed a version of (7) obtaining good results in terms of fitting. The use of (7) can be done analytically and not numerically. For instance, the Lorenz orderings can be obtained by a direct comparison of involved curves.

Second, several measures of inequality can be written as $\int J(p)X(p, \lambda)dp$ with $\int J(p)dp = 0$ where $J(.)$ is a monotone weight function. The following formulae express three well-known measures of income inequality.

Gini

$$\mu^{-1}\left\{\lambda_1 - \mu + 2\lambda_2^{-1}\left[(\lambda_3+1)^{-1} - (\lambda_4+1)^{-1}(\lambda_4+2)^{-1}\right]\right\} \quad (8)$$

Bonferroni

$$\mu^{-1}\left\{\mu-\lambda_1+\lambda_2^{-1}\left[(\lambda_4+1)^{-1}(\lambda+\psi(\lambda_4+2))-(\lambda_3+1)^{-2}\right]\right\} \tag{9}$$

Pietra-Ricci

$$\mu^{-1}\left\{(\mu-\lambda_1)p_\mu+\lambda_2^{-1}\left[(\lambda_3+1)^{-1}p_\mu^{\lambda_3+1}+(\lambda_4+1)^{-1}q_\mu^{\lambda_4+1}\right]\right\} \tag{10}$$

Where γ is the Eulero's constant and $\psi(.)$ is the digamma function.

Finally, the expected value of the i-th order statistic exists in closed form for each i

$$E(X_{i:n})=\lambda_1+\lambda_2^{-1}\left[\frac{B(n+1,\lambda_3)}{B(i,\lambda_3)}-\frac{B(n+1,\lambda_4)}{B(n-i+1,\lambda_4)}\right];\ i=1,\cdots,n \tag{11}$$

3 Parameter estimation

Suppose that n ordered incomes have been grouped (preserving the ordering) into k intervals where the boundaries are $(L_i,U_i],\ i=1,2,\cdots,k$. The number of values in the i-th interval is n_i with $\Sigma n_i=n$. The mean income is $m_i, f_i=n_i/n$ denotes the relative frequency, N_i and p_i are, respectively, the cumulative absolute and relative frequency of incomes not exceeding X_i. Clearly, the grouping scheme may significantly affect the parameter estimation and the variance of estimators. For instance, if the observations cluster significantly around particular values producing multimodal distributions, no GLD can give an acceptable agreement with this behavior.

Karian and Dudewicz [3, p. 155] considered the following system

$$S1: r_3=\frac{A_1-A_2}{A_3-A_1};\ r_4=\frac{A_4-A_5}{A_3-A_2};\ S2:\ r_2=\lambda_2^{-1}(A_3-A_2);\ r_1=\lambda_1+\lambda_2^{-1}A_1 \tag{12}$$

Where $A_i=(\alpha_i)^{\lambda_3}-(1-\alpha_i)^{\lambda_4},\ i=1,2,\cdots,5;\ \alpha_2<\alpha_1,\ \alpha_2<\alpha_3,\ \alpha_5<\alpha_4;\ \alpha_I$ is an observed percent point and r_i is its sample counterpart. The subsystem formed by the first two equations is free of $\lambda_1 and \lambda_2$. Now, given a solution $(\lambda_3,\lambda_4) of S1$, one can rapidly determine the best companion choice for (λ_1,λ_2) by solving the linear system $S2$. The roots of $S1$ can be obtained by a Newton method. This, however, should be preceded both by a trial and error search over the relevant range values of (λ_3,λ_4) and a direct search like the Nelder-Mead simplex algorithm to establish a reasonable starting point.

The method of quantiles has the advantage of being operative without the necessity of knowing every measurement. Moreover, the outliers are given less weight than in the moment estimates; in fact, (12) can be still be applied when the moments do not exist. The choice of α, however, involves an inherent arbitrariness. If the alfa's favor the central part of the distribution, then the X_i's around the mode are efficiently estimated, but at the cost of underestimating higher incomes. If the alfa's were selected in the tails then the most frequent incomes would be neglected. Karian and Dudewicz [3, p. 158] suggest: $\alpha=(0.5,0.1,0.9,0.75,0.25)$ which is quite unsatisfactory for income distributions that are typically skewed to the right. The estimates determined by equating four percentage points seem to be a valid alternative to (12). However, all the $C_{k-1,4}$ combinations should be investigated (supposing that at least one of the non linear four equations systems will give permissible values) to establish an optimal choice. The difficulties of applying this method for large k are such that it would probably be better to abandon it.

The method of moments has been advocated because of its widespread use in practice. The first step is the solution, following closely that of $S1$, of a nonlinear system that depends solely on (λ_3, λ_4)

$$\gamma_1 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{x}}{s} \right)^3 \quad \gamma_2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{x}}{s} \right)^4 \tag{13}$$

Once the best values for $(\lambda_3,\ \lambda_4)$ have been attained, the values of $(\lambda_1,\ \lambda_2)$ are given by $\lambda_2 = \pm(b - a^2)^{0.5}/s,\ \lambda_1 = -a/\lambda_3,\ a = (1+\lambda_3)^{-1} - (1+\lambda_4)^{-1},\ b = (1+2\lambda_3)^{-1} - (1+2\lambda_4)^{-1} - 2B(1+\lambda_3, 1+\lambda_4),\ \min(\lambda_3, \lambda_4) \geq -0.25$.

The method of moments is inadequate. Its use is restricted to distributions possessing their first four moments, but the heavy tail usually observed in empirical income distributions does not support such a premise. Furthermore, when the available data are grouped, a correction for grouping should be considered and if L_1 and/or U_k were left unspecified, the moments cannot be estimated without making arbitrary assumptions. On the other hand (11) is cryptic: the GLD density is symmetric for $\lambda_3 = \lambda_4$ but $\gamma_1 = 0$ even if $\lambda_3 \neq \lambda_4$ and it is far from clear which characteristic is being measured by γ_2 in skewed distributions. Finally, for some data sets, the iterative process might converge to (λ_3, λ_4) for which GLD has no finite moments. The method of quantiles and the method of moments do not appear to be very convenient for income data at the present. The ordinary least squares estimates of λ can be obtained by minimizing

$$\begin{aligned}
S(\lambda) &= \sum_{i=1}^{k} \left[y_i - \lambda_1 - \beta_2 g_i(\lambda_3, \lambda_4) \right]^2 f_i; \quad \beta_2 = \lambda_2^{-1} \\
M1 &: \ y_i = U_i;\ g_i = p_i^{\lambda_3} - q_i^{\lambda_4};\ i = 1, 2, \cdots, k-1 \\
M2 &: \ y_i = U_i;\ g_i = \frac{B(n+1, \lambda_3)}{B(N_i, \lambda_3)} - \frac{B(n+1, \lambda_4)}{B(n - N_i + 1, \lambda_4)}, \\
&\quad\ i = 1, 2, \cdots, k-1 \\
M3 &: \ y_i = m_i;\ g_i = \frac{p_i^{\lambda_3+1} - p_{i-1}^{\lambda_3+1}}{f_i(\lambda_3+1)} + \frac{q_i^{\lambda_4+1} - q_{i-1}^{\lambda_4+1}}{f_i(\lambda_4+1)}, i = 1, 2, \cdots, k
\end{aligned}$$

$M1$ defines the estimators that minimize the sum of squared differences between predicted and observed quantiles. $M2$, based on (9), is an extension to grouped data of the method proposed by Oztürk and Dale [6]. $M3$ suggests itself because of the importance of the group means for measuring income inequality. This new approach is more demanding since it requires knowledge of the mean of each income group, but has the advantage of using more information than the other methods. Since λ_1 and λ_2 are in linear form, they can be replaced by their least squares estimates given (λ_3, λ_4)

$$\begin{aligned} \hat{\lambda_1} &= \bar{y} - \hat{\lambda_2}^{-1}\bar{g}. \\ \hat{\lambda_2} &= \frac{1}{\hat{\beta_2}} = \frac{\sum_{i=1}^{k}(g_i - \bar{g})^2 f_i}{\sum_{i=1}^{k}(y_i - \bar{y})(g_i - \bar{g}) f_i} \\ &\Rightarrow S(\lambda_3, \lambda_4) = (1 - r_{yg}^2)\sum_{i=1}^{k}(y_i - \bar{y})^2 f_i \end{aligned} \tag{14}$$

Where r_{yg} is the correlation coefficient between y and g and r_yg does not depend on λ_1, β_2. Therefore, the pair (λ_3, λ_4) that minimizes $[1 - (r_{yg})^2]$ also minimizes $S(\lambda_3, \lambda_4)$. It should be remarked that $S(\lambda_3, \lambda_4)$ in (14), like $S1$ and (12), can have multiple solutions or no solution for some data sets. Even when a solution exists, the numerical procedure devoted to its search may fail to find it because of convergence failure. Moreover, the observed y_i will not have equal variance nor will they be uncorrelated. Since this drawback is, at least in theory, serious further studies (e.g. in the line of generalized least squares) are needed to assess the effectiveness of GLD for income data.

4 Parameter estimation

Gastwirth [2] gives an income distribution in ten classes. The Gini index for the entire sample is 0.4014 and the crude bounds within which the index must lie are $(0.3883, 0.4083)$. Table 1 reveals the relative merit of five distinct estimators of λ.

Since the α_i have not been reached, the r_i's in (12) were computed by using linear interpolation on the given values ($Q1$) and the closest observed quantiles ($Q2$). It is easily seen that the quantile estimates depend markedly on the particular choice of percentage points. The moments have been calculated by assuming that all incomes in the i-th interval equal the average income m_i whereas, the solutions of (14), were obtained by using the Nelder-Mead simplex procedure. According to the SSE there is a sufficiently close agreement between observed and estimated percentiles with the exception of the two methods based on quantiles. As a general result, the fit of GLD is reasonable good in the middle part, but is poor in describing both the upper and the lower tails. The best performance has been obtained by $M2$ with $M1$ close competitor. $M3$ has an unduly bad fit in the last class. The Chi-squared criterion confirms the ranking of the six techniques determined by SSE. However, only the method of moments and the method of least squares on group means were able to provide an estimated Gini index (reported in

U_i	P_i	m_i		Q1	Q2	Mom.	M1	M2	M3
1	0.048235	0.54141		2.11	0.87	1.17	1.30	1.00	1.13
2	0.130757	1.46363		3.68	3.92	2.15	2.10	1.97	2.08
3	0.202900	2.44572		4.51	5.35	3.02	2.84	2.85	2.94
4	0.271913	3.43890		5.16	6.10	3.85	3.61	3.72	3.78
5	0.338056	4.43732		5.74	6.49	4.66	4.39	4.59	4.62
6	0.414029	5.40118		6.37	6.76	5.61	5.37	5.64	5.62
7	0.492491	6.39292		7.04	7.00	6.61	6.49	6.79	6.69
10	0.706509	8.30464		9.19	8.35	9.89	10.41	10.40	9.98
15	0.897600	11.90433		12.73	11.92	16.88	16.60	14.87	15.56
□	1.000000	22.26150							
			λ_1	14.67930	6.77021	13.78170	0.45451	0.45486	17.46992
			λ_2	0.08218	0.11396	0.07589	0.05036	0.05025	0.05923
			λ_3	25.81397	4.92769	9.29542	3.96E-08	3.40E-08	20.59420
			λ_4	0.68402	8.03131	0.89215	0.56603	0.56498	0.66202
			SSE	1.52479	3.59048	0.71697	0.06914	0.06907	0.09283
			χ^2	0.23735	0.53782	0.03390	0.01530	0.01520	0.01331
			G	0.28616	0.25153	0.39533	0.36650	0.36666	0.40026

$$SSE=\sum_{i=1}^{k-1}\left(U_i-\hat{U}_i\right)^2 f_i$$

$$\chi^2=\sum_{i=1}^{k-1}\left(\pi_i-\pi_{i-1}-f_i\right)^2 f_i$$

$$\hat{\lambda}_2\left(U_i-\hat{\lambda}_1\right)=\pi_i^{\hat{\lambda}_3}-\left(1-\pi_i\right)^{\hat{\lambda}_4}$$

Table 1: Observed and estimated quantiles of income data.

the last row of Table 1) lying inside the prescribed bounds. In this sense $M3$ carries the gold medal.

References

[1] Dagum C. (1990). *Generation and properties of income distribution functions.* In C. Dagum, M. Zenga (eds.), Income and Wealth Distribution, Inequality and Poverty. Springer-Verlag Berlin Heidelberg, 1 − 17.

[2] Gastwirth J.L. (1971). *The estimation of the Lorenz curve and Gini index.* The Review of Economics and Statistics **54**, 306 − 316.

[3] Karian Z.A., Dudewicz E.J. (2000). *Fitting statistical distributions.* The Generalized lambda distribution and gereralized bootstrap methods. CRC, Boca Raton (FL).

[4] Kleiber C., Kotz S. (2003). *Statistical size distributions in economics and actuarial sciences.* John Wiley &, Sons, New York.

[5] Maddala G.S., Singh A.K. (1977). *A flexible functional form for Lorenz curves.* Economie Appliquée, **30**, 481 − 486.

[6] Oztürk A., Dale R.F. (1985). *Least squares estimation of the parameters of the generalized lambda distributions.* Technometrics, **27**, 81 − 84.

[7] Sarabia J. M. (1996). *A hierarchy of Lorenz curves based on generalized Tukey's lambda distribution.* Econometric Reviews, **16**, 305 − 320.

Address: A. Tarsitano, Dipartimento di Economia e Statistica. Universita della Calabria, 87030 Arcavacata di Rende (Cs). Italy

E-mail: `agotar@unical.it`