Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03701-2 OPEN

A regression method for estimating Gini index by decile

Xiaobo Shen[1] & Pingsheng Dai[2]✉

Based on the three-parametric Lorenz curve proposed by Kakwani (1980), this paper builds a multiple linear regression model to estimate the parameters by the weighted least square method named the regression method. Using the Lorenz curve, the Gini index and its variance are then calculated. Compared to the error minimization technique, the regression method has a better performance in estimating the Gini index using Kakwani (1980)'s Lorenz curve and a dataset of sixteen economies from the United Nation University-World Income Inequality Database (UNU-WIID). The results also suggest that the regression method has an advantage when estimating the Gini index and fitting the income shares by decile for the medium and higher inequality economies. We find that the three-parametric Lorenz curve has a better performance than the double-parametric Lorenz curve, and the double-parametric Lorenz curve is superior to the single-parametric Lorenz curve, judged by the RMSE of the actual Gini index and the estimated ones.

[1] School of Economics, Xiamen University, Xiamen, China. [2] School of Finance and Economics, Jimei University, Xiamen, China. ✉email: daips@jmu.edu.cn

Introduction

Recently, periodic reports of certain summary statistics on income or wealth distribution have become quite common (Sitthiyot and Holasut, 2021). International Labor Organization's ILOSTAT, United Nations Development Programme's Human Development Report (UNDP-HDR), the United Nations University-World Income Inequality Database (UNU-WIID), the World Bank's Poverty and Inequality Platform (PIP), and the World Income Database (WID) are the largest cross-country databases that provide grouped data. By using the grouped data on income or wealth, Gini index could be estimated (1) by assuming a statistic distribution of income, such as lognormal, Beta-II, Generalized Pareto, or mixture distribution (McDonald, 1984; Chotikapanich et al., 1997, 2007; Blanchet et al., 2022) and (2) by specifying a parametric function form for Lorenz curve (Paul and Shanker, 2020; Sitthiyot and Holasut, 2021).

Fitting Lorenz curve is more convenient than fitting income distribution because the cumulative population share and the corresponding cumulative income share form the points of the Lorenz curve. Numerous studies have suggested a variety of parametric functional forms to estimate directly the Lorenz curve. There are single-parametric Lorenz curve (Kakwani and Podder, 1973; Aggarwal, 1984; Chotikapanich, 1993; Paul and Shankar, 2020), double-parametric Lorenz curve (Rasche et al., 1980; Ortega et al., 1991; Sitthiyot and Holasut, 2021), three-parametric Lorenz curve (Kakwani, 1980; Sarabia et al., 1999). Chotikapanich and Griffiths (2002) have suggested estimating parameter(s) of the Lorenz curve using the Maximum Likelihood (ML) method assuming that each income share is subject to the joint Dirichlet distribution, while Jorda et al., (2021) have proposed estimating parameter(s) of the Lorenz curve utilizing the error minimization technique. Sitthiyot and Holasut (2021) introduce a simple and straightforward method for estimating the Lorenz curve using three indicators, namely, the Gini index, the income share of the bottom, and that of the top, which is associated with a specific functional form based on the weighted average of the exponential function and the functional form implied by Pareto distribution.

This study proposes a new approach named the regression method for estimating the Gini index by decile based on a specified functional form suggested by Kakwani (1980). The new approach builds a linear regression model to estimate three parameters, and can get better estimates of Gini index compared to the ML method and the error minimization technique for the same Lorenz curve. The new approach can easily obtain an estimation of the Lorenz curve and corresponding Gini index, and get an estimation of the variance of the Gini index utilizing popular computer programs such as EVIEWS and STATA. In addition to, the new approach allows negative income or wealth share values, which are not allowed in the Beta-II density function[1] of Chotikapanich et al. (2007) and the Gamma function[2] in the maximum likelihood of Chotikapanich and Griffiths (2002).

We demonstrate how to estimate the Gini index by using the new approach based on a dataset of the income shares of sixteen economies, which differ in the level of income inequality, economic, sociological, and regional backgrounds. We also compare the performance of our new method to that of the methods suggested by Sitthiyot and Holasut, 2021 and Sarabia et al. (1999), which aim to estimate the Gini index and fit the income shares by using the error minimization technique and decile data.

Methods

Kakwani (1980) suggests a functional form for fitting the Lorenz curve as follows:

$$L(x; a, p, q) = x - ax^p(1-x)^q;\ a > 0,\ 0 < p \leq 1,\ 0 < q \leq 1 \quad (1)$$

where, x is the cumulative population share, 0≤ x≤1. When fitting function of the Lorenz curve, the functional form in model (1) is more consistent with the actual data (Cheong, 2002; Tanak et al., 2018; Sitthiyot and Holasut, 2021).

The advantage of the function $L(x)$ is that it is applicable to negative income or wealth share. This can deal with the Sarabia et al. (1999)'s criticism that Eq.(1) violates $L'\ (0^+) \geq 0$. Actually, the wealth share of the poorest 10% is often negative in many economies; in this case, we only request that the function $L(x)$ meets the conditions: $L(x) \geq 0$, $L'(x) \geq 0$ and $L''(x) \geq 0$ in the right-handed area within the interval [0,1].

The derivation process of its second-order derivative is as follows:

Let $f(x) = x^p(1\text{-}x)^q$ namely $L(x) = x\text{-}f(x)$, so that

$$f'(x) = \frac{d(e^{\ln f})}{dx} = \frac{d(e^{p\ln x + q\ln(1-x)})}{dx} = e^{\ln f}\left(\frac{p}{x} - \frac{q}{1-x}\right)$$

Then

$$f''(x) = e^{\ln f}\left[\left(\frac{p}{x} - \frac{q}{1-x}\right)^2 - \left(\frac{p}{x^2} + \frac{q}{(1-x)^2}\right)\right]$$
$$= e^{\ln f}\frac{(p+q)(p+q-1)x^2 - 2p(p+q-1)x + p^2 - p}{x^2(1-x)^2}$$

Let $g(x) = (p+q)(p+q-1)x^2 - 2p(p+q-1)x + p^2 - p$. We can obtain the discriminant of the root nature of $g(x)$ as follows

$$\Delta = 4p^2(p+q-1)^2 - 4p(p-1)(p+q)(p+q-1)$$
$$= 4pq(p+q-1)$$

When $p+q \leq 1$, $p > 0$, $q > 0$, then $\Delta < 0$, indicating g(x) ≤0, which means $f''(x) \leq 0$,so that $L''(x) \geq 0$ ($a > 0$). That is, the curve is convex.When $p+q > 1$, $p > 0$, $q > 0$, then $\Delta > 0$, thus we can obtain two roots of $g(x)$, x_1 and x_2 as follows

$$x_1 = \frac{p - \sqrt{\frac{pq}{p+q-1}}}{p+q} \leq 0 \Longleftrightarrow (p+q)(p-1) \leq 0 \Longleftrightarrow 0 < p \leq 1$$

$$x_2 = \frac{p + \sqrt{\frac{pq}{p+q-1}}}{p+q} \geq 1 \Longleftrightarrow (1-q)(p+q) \geq 0 \Longleftrightarrow 0 < q \leq 1$$

So that

(1) when $p+q \leq 1$, $p > 0$, $q > 0$, we have $L''(x) \geq 0$, $x \in [0,1]$.

(2) when $p+q > 1$, $0 < p \leq 1$, $0 < q \leq 1$, we have $L''(x) \geq 0$, $x \in [0,1]$.

(3) when $p > 1$, $0 < q \leq 1$, we have $x_1 < 1$, $x_2 > 1$, we have $L''(x) \geq 0$, $x \in [x_1,1]$.

Furthermore, $L'(x) = 1\text{-}af'(x)$, when $x \to 1^-$, we have $L'(x) > 0$, because

$$\lim_{x\to 1^-} f'(x) = \lim_{x\to 1^-} \frac{p(1-x) - qx}{x^{1-p}(1-x)^{1-q}} < 0 \Rightarrow \lim_{x\to 1^-} L'(x) > 0$$

It means $L(x)$ is convex, increasing in the right-handed area within the interval [0, 1].

When analyzing the condition $L''(x) \geq 0$, we find that under $p+q > 1$, the condition $p > 0$ must hold. So Eq. (1) can be expressed as follow:

$$L(x; a, p, q) = x - ax^p(1-x)^q;\quad a > 0,\quad 0 < q \leq 1,\quad p > 0 \quad (2)$$

Then, the $L(x)$ satisfies with $L(x) \geq 0$, $L'(x) \geq 0$ and $L''(x) \geq 0$ in the right-handed area within the interval [0,1]. For example, in Fig. 1, the Lorenz curve passes through the point (0.53, −0.0021), which is formed from fitting the generalized Pareto curves for the United States (Blanchet et al., 2022). Although model (2) cannot

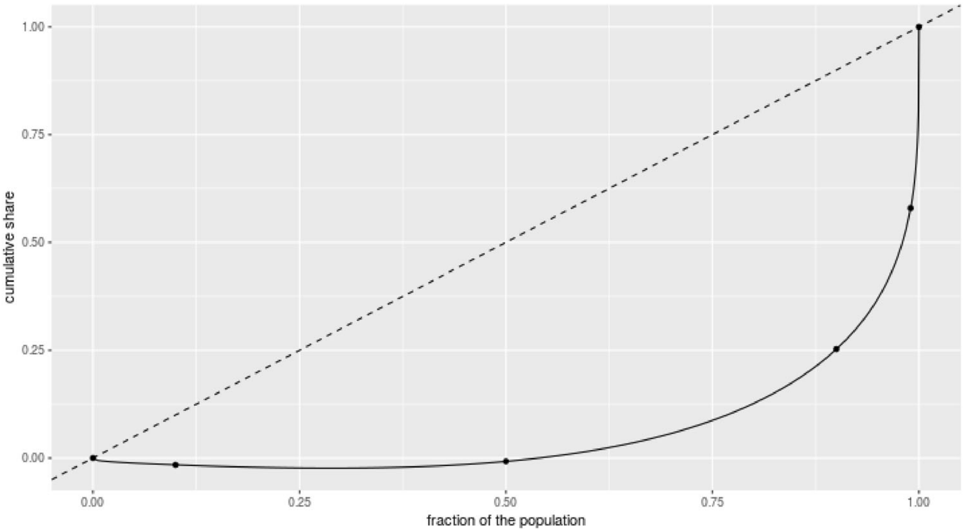

Fig. 1 The Lorenz curve of the capital of the USA in 2010. https://wid.world/gpinter/ (file2).

meet the properties of the classic Lorenz curve, i.e. $L(x) \geq 0$, $L'(x) \geq 0$, and $L''(x) \geq 0$ in the interval [0, 1], we would like to call model (2) a Lorenz curve, more precisely, a fitting function of a Lorenz curve.

Then rearranging Eq. (1) and taking the natural logarithm, we can get the following equation:

$$\log(x-y) = \log a + p\log x + q\log(1-x) + \varepsilon \tag{3}$$

where $y = L(x)$, ε is an error term with zero mean value.

Let $\beta_0 = \log(a)$, the model (3) can be expressed as the following matrix form:

$$\mathbf{Y} = \begin{pmatrix} \log(x_1 - y_1) \\ \log(x_2 - y_2) \\ \vdots \\ \log(x_9 - y_9) \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & \log x_1 & \log(1-x_1) \\ 1 & \log x_2 & \log(1-x_2) \\ \vdots & \vdots & \vdots \\ 1 & \log x_9 & \log(1-x_9) \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ p \\ q \end{pmatrix}$$

we can obtain the parameters estimates of the model using the least square method:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},\ \mathbf{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

where $Var(\varepsilon) = \sigma2$. Therefore, we estimate not only the parameters of the Lorenz curve, but also the covariance matrix of the parameters. Using the estimated parameters, we calculate the Gini coefficient and its variance according to the following formulas:

$$G = 2aBeta(p+1, q+1),\ Var(G) = \left(\frac{\partial G}{\partial \boldsymbol{\beta}}\right)' Var(\boldsymbol{\beta})\frac{\partial G}{\partial \boldsymbol{\beta}} \tag{4}$$

According to Kakwani and Podder (1976), we can obtain the following partial derivatives:

$$\frac{\partial G}{\partial \beta_0} = G,\quad a = e^{\beta_0}$$
$$\frac{\partial G}{\partial p} = [\Psi(1+p) - \Psi(2+p+q)]G, \Psi(1+p)$$
$$-\Psi(2+p+q) = \sum_{k=0}^{+\infty}\left(\frac{1}{2+p+q+k} - \frac{1}{1+p+k}\right)$$
$$\frac{\partial G}{\partial q} = [\Psi(1+q) - \Psi(2+p+q)]G$$

Because both x and y are increasing, this can lead to heteroscedasticity of ε. To avoid the problem, we can exploit the weighted least square method to estimate the parameters of Eq.(3) using $1/x_i$ ($i = 1, 2,\ldots, n$) as the weights:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\Lambda\mathbf{X})^{-1}\mathbf{X}'\Lambda\mathbf{Y},\quad \mathbf{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\Lambda\mathbf{X})^{-1}\sigma^2$$

where $\Lambda = \mathrm{diag}(1/\mathrm{x}_1, 1/\mathrm{x}_2, \ldots, 1/\mathrm{x}_\mathrm{n})$, $\mathrm{Var}(\varepsilon\varepsilon') = \Lambda^{-1}\sigma^2$.

The total income or wealth is usually greater than 0, and the average income or wealth is also required to be greater than 0. The income Gini coefficient is equal to the Gini mean difference divided by the mean income, the Gini mean difference is non-negative, so the Gini coefficient is always non-negative.

The UNU-WIID database has data on the income shares by decile and the Gini index for the economies in the world. These economies differ significantly in the degree of income inequality according to the UNU-WIID database. For example, Belgium (BEL), Czechia (CZE), Slovakia (SVK), and Iceland (ISL) have a lower Gini index ranging from 0.2 to 0.3; while Panama (PAN), Brazil (BRA), South Africa (ZAF), and Hong Kong (HKG) have higher Gini indices greater than 0.5. We classify the economies in the UNU-WIID database into four groups according to these economies' Gini index. The economies in the first group has a Gini index less than 0.3, the economies in the second group with a Gini index ranging from 0.3 to 0.4, the economies in the third group with a Gini index ranging from 0.4 to 0.5, and the economies in the last group with a Gini index greater than 0.5.

To demonstrate the regression method for estimating Gini index by decile, we choose sixteen economies (see Table 1) from the above four groups and utilize the data on the income shares of these sixteen economies during the period of 2016 to 2020.

As suggested by Dagum (1977), a good parametric functional form for the Lorenz curve should be able to characterize income distributions of different countries, regions, socioeconomic groups in different periods. We can use some goodness-of-fit statistics, such as coefficient of determination (R^2), mean square error (MSE), and mean absolute error (MAE), to gauge how close the estimated income shares are to the actual observations (Chotikapanich, 1993; Cheong, 2002; Tanak et al., 2018; Paul and Shankar, 2020; Sitthiyot and Holasut, 2021).

Fitting Lorenz curves using income shares by decile, we can estimate parameter(s) of the Lorenz curve based on the curve

Table 1 Estimating the Gini index by the error minimization technique and the regression method using the data on the income shares by decile as published in the UNU-WIID.

Year	Economy	True Gini	L_k				K_{RE}				
			a	*p*	*q*	ΔG_1	*a*	*p*	*q*	ΔG_2	95% conf. interval
2020	BEL	0.2540	0.4803	0.8569	0.6109	−0.0004	0.4621	0.8284	0.5905	0.0000	0.2540 ± 0.0010
2019	CZE	0.2402	0.4595	0.8852	0.5985	−0.0004	0.4674	0.8979	0.6074	−0.0002	0.2402 ± 0.0010
2018	SVK	0.2097	0.4254	0.8130	0.7227	−0.0005	0.4352	0.8293	0.7351	−0.0008	0.2097 ± 0.0010
2017	ISL	0.2487	0.4195	0.8245	0.5167	0.0001	0.4130	0.8126	0.5087	−0.0006	0.2487 ± 0.0020
2020	AUT	0.3802	0.7689	0.9478	0.6088	0.0002	0.7775	0.9559	0.6146	0.0005	0.3802 ± 0.0012
2019	ALB	0.3430	0.7068	0.9388	0.6285	−0.0024	0.7248	0.9574	0.6418	−0.0018	0.3430 ± 0.0022
2018	DEU	0.3159	0.5739	0.9054	0.5338	−0.0006	0.5819	0.9161	0.5410	−0.0002	0.3159 ± 0.0012
2017	RUS	0.3670	0.7405	0.9625	0.5950	−0.0004	0.7501	0.9721	0.6018	−0.0001	0.3670 ± 0.0012
2020	FRA	0.4230	0.8002	0.9725	0.5261	−0.0006	0.8090	0.9808	0.5318	−0.0002	0.4230 ± 0.0014
2019	COL	0.4810	0.8418	0.9567	0.4610	−0.0001	0.8464	0.9608	0.4638	0.0002	0.4810 ± 0.0014
2018	USA	0.4709	0.9127	0.9881	0.5389	−0.0003	0.9207	0.9946	0.5434	0.0000	0.4709 ± 0.0010
2017	MYS	0.4107	0.7748	0.9850	0.5142	−0.0011	0.7842	0.9941	0.5205	−0.0009	0.4107 ± 0.0008
2019	PAN	0.5150	0.9531	1.0157	0.4816	0.00053	0.9558	1.0179	0.4831	0.00046	0.5150 ± 0.0008
2018	BRA	0.5400	0.8917	0.9880	0.3872	−0.0027	0.8996	0.9949	0.3917	−0.0024	0.5400 ± 0.0024
2017	ZAF	0.6170	1.1084	1.0821	0.4076	0.0020	1.1387	1.1022	0.4215	0.0021	0.6170 ± 0.0067
2016	HKG	0.5390	0.9184	0.9764	0.4274	−0.0038	0.9164	0.9747	0.4263	−0.0028	0.5390 ± 0.0008
RMSE	All					0.0015				0.0012	
	I					0.0004				0.0005	
	II					0.0012				0.0009	
	III					0.0006				0.0005	
	IV					0.0025				0.0021	

The L_K denotes fitting the Kakwani (1980)'s Lorenz curve using the error minimization technique, and the K_{RE} means fitting the Kakwani (1980)'s Lorenz curve using the regression method. I, II, III, and IV stand for the low, medium, higher, and high inequality groups of sixteen economies, respectively.

fitting technique and the method of minimizing the sum of squared errors:

$$\hat{\theta} = \min_{\theta} \sum_{i=1}^{N} [y_i - L(x_i, \theta)]^2, \ \theta' = (a, p, q) \tag{5}$$

From the view of fitting the Lorenz curve, though the goodness-of-fit of our regression method in most cases is smaller than that of the minimization technique of Eq.(5), it has better performance in estimating Gini index. We introduce the root mean squared error (RMSE) in Eq. (6) to measure the difference between the estimated and the actual Gini indices.

$$RMSE = \sqrt{\frac{1}{k} \sum_{i=1}^{k} \left(\hat{G}_i - G_i\right)^2} \tag{6}$$

Results and discussion

Table 1 reports the estimated parameters a, p and q for the Lorenz curve of Eq.(2) for the sixteen economies using the error minimization technique and the regression method. We can substitute the estimated values of parameters a, p, and q into Eq. (4) to obtain the estimated Gini index. The ΔG_1 and ΔG_2 in Table 1 are the differences between the estimated and the actual Gini indices based on the two estimation methods, respectively. It can be seen that mostly the ΔG_2 is smaller than the ΔG_1 for the sixteen economies, indicating a better performance of the regression method. Furthermore, we can see that the regression method has a lower RMSE compared to the minimization technique for the full sample estimation. This conclusion also holds for the estimations for the medium, higher, and high groups. This also conveys that the regression method is superior to the minimization technique. We also note that in some cases the estimated value of parameter p is greater than one under both methods, this provides the justification of our adjustment of the range of parameter p in Eq. (2). Table 1 also gives the standard deviation of the estimated Gini index employing the regression method.

Comparison of goodness-of-fit by two methods. We then use the fitted Lorenz curves to calculate the values of the (cumulative) income shares by decile for the sixteen economies under the two methods, and compare the estimated income shares with the observed income shares. Table 2 reports the values of goodness-of-fit statistics, i.e., information inaccuracy measure (IIM), R^2, MSE, MAE, and maximum absolute error (MAS). All the values of goodness-of-fit measures suggest that there is no significant difference between the estimated (cumulative) income shares and the observed income shares for each economy. We find that the MSE of the fitted Lorenz curve by the error minimization technique is less than the MSE by the regression method for all the sampled economies. In most cases, this conclusion holds in terms of the statistics MAS, MAE, and IIM. Therefore, the error minimization technique has a better performance than the regression method.

Comparison of the estimated Gini index using different Lorenz curves. There are many different functional forms of

Table 2 Goodness-of-fit of the fitted Lorenz curves by the error minimization technique and the regression method using the data on the income shares by decile for the sixteen economies.

year	Economy	R^2		MSE		MAE		MAS		IIM	
		L_K	K_{RE}	L_K	K_{RE}	L_K	K_{RE}	L_K	K_{RE}	L_K	K_{RE}
2020	BEL	1.0000	1.0000	1.4E−06	2.3E−06	0.0010	0.0011	0.0017	0.0035	5.1E−05	1.2E−04
2019	CZE	1.0000	1.0000	1.7E−07	3.2E−07	0.0004	0.0005	0.0007	0.0012	6.4E−06	1.3E−05
2018	SVK	1.0000	1.0000	2.8E−07	4.9E−07	0.0004	0.0005	0.0008	0.0014	9.6E−06	2.2E−05
2017	ISL	1.0000	1.0000	2.1E−07	3.5E−07	0.0004	0.0004	0.0007	0.0014	8.3E−06	1.7E−05
2020	AUT	1.0000	1.0000	2.5E−07	4.0E−07	0.0004	0.0005	0.0008	0.0012	1.1E−05	1.7E−05
2019	ALB	1.0000	1.0000	7.6E−07	1.4E−06	0.0007	0.0009	0.0016	0.0023	3.0E−05	7.6E−05
2018	DEU	1.0000	1.0000	3.3E−07	5.0E−07	0.0005	0.0006	0.0010	0.0015	1.3E−05	2.6E−05
2017	RUS	1.0000	1.0000	3.1E−07	4.9E−07	0.0005	0.0006	0.0009	0.0014	1.4E−05	3.0E−05
2020	FRA	1.0000	1.0000	6.4E−07	8.2E−07	0.0007	0.0007	0.0013	0.0020	3.9E−05	6.8E−05
2019	COL	1.0000	1.0000	1.9E−07	2.4E−07	0.0004	0.0004	0.0006	0.0009	1.4E−05	2.5E−05
2018	USA	1.0000	1.0000	2.9E−07	4.2E−07	0.0004	0.0005	0.0009	0.0015	2.7E−05	5.5E−05
2017	MYS	1.0000	1.0000	5.4E−07	7.4E−07	0.0006	0.0007	0.0012	0.0019	3.2E−05	5.5E−05
2019	PAN	1.0000	1.0000	1.6E−07	1.8E−07	0.00034	0.00033	0.0007	0.0008	1.9E−05	2.4E−05
2018	BRA	1.0000	1.0000	2.1E−06	2.3E−06	0.0012	0.0012	0.0028	0.0026	2.4E−04	3.0E−04
2017	ZAF	1.0000	1.0000	1.4E−05	1.6E−05	0.0032	0.0034	0.0057	0.0075	1.4E−03	2.0E−03
2016	HKG	1.0000	1.0000	1.7E−07	1.9E−07	0.0003	0.0003	0.0009	0.0010	1.6E−05	1.4E−05

The L_K denotes fitting the Kakwani (1980)'s Lorenz curve using the error minimization technique, and the K_{RE} means fitting the Kakwani (1980)'s Lorenz curve using the regression method.

Lorenz curve, of which the most common forms are as the following:

$$\text{Paul}-\text{Shankar}: L_1(x;r) = x[e^{-r(1-e^x)} - 1]/[e^{-r(1-e)} - 1],\ r > 0$$
$$\text{Aggarwal}: L_2(x;r) = [(1-r)^2 x]/[(1+r)^2 - 4rx], 0 < r < 1$$
$$\text{Chotikapanich}: L_3(x;r) = (e^{rx} - 1)/(e^r - 1), r > 0$$
$$\text{Pareto}: L_4(x;r) = 1 - (1-x)^{1/r}, r > 1$$
$$\text{Kakwani}-\text{Podder}: L_5(x;r) = xe^{-r(1-x)}, r > 0$$
$$\text{Rasche et al.}: L_6(x;q,r) = [1-(1-x)^q]^r, 0 < q \le 1,\ r \ge 1$$
$$\text{Ortega et al.}: L_7(x;q,r) = x^q[1-(1-x)^r], q \ge 0, 0 < r \le 1$$
$$\text{Sitthiyot}-\text{Holasut}: L_8(x;q,r) = (1-r)x^q + r[1-(1-x)^{1/q}], q \ge 1, 0 \le r \le 1$$
$$\text{Sarabia et al.}: L_9(x;q,r,s) = x^q[1-(1-x)^r]^s, q \ge 0, 0 < r \le 1, s \ge 1$$

The first five specifications of the above Lorenz curve are single-parametric, the last one is three-parametric, and the rest are double-parametric.

The minimization technique are applicable to all the above specifications of Lorenz curve, while the regression method can only be applied to the Kakwani's Lorenz curve (1980). We fit the above-mentioned nine specifications using the error minimization technique, and fit the specification of the Kakwani's Lorenz curve using the regression method based on the dataset of the sampled economies. Table 3 reports the estimated results. Column (4) presents the estimated result of the Kakwani's Lorenz curve using the regression method, and columns (5)-(13) present the estimated results of the above nine specifications using the error minimization technique, respectively.

We find that the performance of the regression method is better than that of the error minimization technique under the single-parametric and the double-parametric specifications, while it is poorer than the error minimization method under the three-parametric specification presented in column (13). In addition, we find the regression method is always better than the error minimization technique under any specifications for the medium and the higher groups, since the RMSE of the regression method is smaller than the RMSE of the error minimization method under any specifications.

We also find that the three-parametric Lorenz curve has a better performance than the double-parametric one when the error minimization technique is used to fit Lorenz curve, since the RMSE under the three-parametric specification is smaller than the RMSE of the double-parametric one. Similarly, the double-parametric specification is better than the single-parametric one. The results from Table 3 also suggest that the L_8 proposed by Sitthiyot and Holasut (2021) has the best performance in the three double-parametric Lorenz curves, and the L_2 proposed by Aggarwal (1984) has the best performance in the five single-parametric Lorenz curves. Sitthiyot and Holasut (2023) find that the estimated Gini index has a lower bound of 0.4180 for the Lorenz curve L_1 under the condition $r > 0$. In order to better fit the above-mentioned nine Lorenz curves, we relax the constraints on the parameters of those Lorenz curves when we fit them using the error minimization technique. For example, for the Lorenz curve L_1, we allow the parameter r to vary between negative infinity and positive infinity, i.e. $-\infty < r < \infty$, and for the Lorenz curve L_9, we allow the following looser parameter constaints: $-\infty < q < \infty$, $0 < r \le 1$, and $s > 0$.

Comparison of the estimated income shares between L_{SH} and K_{RE}. Given the poorer performance of the regression method that fits the Lorenz curve in Table 2, we examine the relative performance of the regression method to the error minimization technique in fitting the income shares by decile. We compare the estimated income shares of the regression method under Kakwani's Lorenz curve to those of the error minimization technique under the Lorenz curve L_8, since the specification L_8 has the best performance in all the single-parametric and the double parametric specifications. When estimating the income shares, we choose four countries from the sixteen economies, namely, Belgium, Albania (ALB), the United States (USA), and South Africa, which have significant differences in the level of inequality. For instance, the Gini index of Belgium is 0.2540 in 2019, Albania 0.343 in 2019, the USA 0.4709 in 2018, and South Africa 0.6170 in 2017.

Table 4 reports the estimated income shares for these four countries. Panel A in Table 4 presents the actual income shares and the estimated income shares using two methods for Belgium. Panel B, Panel C, and Panel D display the actual and the estimated income shares of Albania, the USA, and South Africa,

Table 3 Estimating the Gini index by the regression method and the error minimization technique with the different Lorenz curves using the data on the income shares by decile as published in the UNU-WIID.

year	Economy	True Gini	K_{RE}	The Lorenz curves								
				L_1	L_2	L_3	L_4	L_5	L_6	L_7	$L_8(L_{SH})$	L_9
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
2020	BEL	0.2540	0.2546	0.1981	0.2622	0.2521	0.3008	0.2484	0.2533	0.2535	0.2532	0.2541
2019	CZE	0.2402	0.2404	0.1761	0.2434	0.2340	0.2807	0.2305	0.2406	0.2408	0.2405	0.2402
2018	SVK	0.2097	0.2097	0.1593	0.2311	0.2242	0.2645	0.2215	0.2095	0.2096	0.2094	0.2095
2017	ISL	0.2487	0.2495	0.1828	0.2500	0.2401	0.2884	0.2365	0.2471	0.2474	0.2470	0.2491
2020	AUT	0.3802	0.3803	0.3503	0.3968	0.3784	0.4456	0.3727	0.3806	0.3814	0.3802	0.3800
2019	ALB	0.3430	0.3448	0.3101	0.3607	0.3445	0.4070	0.3392	0.3458	0.3465	0.3456	0.3449
2018	DEU	0.3159	0.3154	0.2622	0.3168	0.3019	0.3614	0.2969	0.3150	0.3156	0.3149	0.3151
2017	RUS	0.3670	0.3672	0.3309	0.3773	0.3590	0.4251	0.3534	0.3684	0.3693	0.3682	0.3670
2020	FRA	0.4230	0.4228	0.3886	0.4288	0.4063	0.4800	0.4001	0.4234	0.4246	0.4228	0.4222
2019	COL	0.4810	0.4819	0.4510	0.4872	0.4614	0.5414	0.4553	0.4801	0.4820	0.4790	0.4814
2018	USA	0.4709	0.4709	0.4469	0.4838	0.4600	0.5376	0.4541	0.4707	0.4723	0.4696	0.4705
2017	MYS	0.4107	0.4109	0.3711	0.4107	0.3882	0.4611	0.3819	0.4122	0.4134	0.4119	0.4103
2019	PAN	0.5150	0.5129	0.4864	0.5166	0.4898	0.5712	0.4841	0.5133	0.5155	0.5119	0.5123
2018	BRA	0.5400	0.5424	0.5094	0.5378	0.5076	0.5928	0.5022	0.5412	0.5438	0.5396	0.5413
2017	ZAF	0.6170	0.6198	0.5995	0.6185	0.5916	0.6744	0.5894	0.6229	0.6266	0.6202	0.6171
2016	HKG	0.5390	0.5385	0.5108	0.5417	0.5139	0.5980	0.5085	0.5358	0.5384	0.5340	0.5383
RMSE	All		0.0012	0.0413	0.0097	0.0174	0.0553	0.0215	0.0020	0.0030	0.0020	0.0010
	I		0.0005	0.0594	0.0116	0.0090	0.0458	0.0101	0.0009	0.0008	0.0009	0.0002
	II		0.0009	0.0393	0.0132	0.0082	0.0588	0.0124	0.0017	0.0022	0.0015	0.0010
	III		0.00049	0.0325	0.0077	0.0180	0.0589	0.0240	0.0009	0.0018	0.0013	0.0005
	IV		0.0021	0.0267	0.0021	0.0272	0.0564	0.0319	0.0035	0.0052	0.0034	0.0015

The K_{RE} means fitting the Kakwani (1980)'s Lorenz curve using the regression method, the L_i ($i = 1,\ldots,9$) denotes fitting the i_{th} Lorenz curve using the error minimization technique, especially, the $L_8(L_{SH})$ denotes fitting the Lorenz curve proposed by Sitthiyot and Holasut (2021). I, II, III, and IV stand for the low, medium, higher, and high inequality groups of sixteen economies, respectively.

Table 4 The comparison of the estimated income shares of four countries by decile between the error minimization technique with L_8 and the regression method.

Decile	Panel A BEL(2020)			Panel B ALB (2019)			Panel C USA (2018)			Panel D ZAF (2017)		
	Actual	L_{SH}	K_{RE}	Actual	L_{SH}	K_{RE}	Actual	L_{SH}	K_{RE}	Actual	L_{SH}	K_{RE}
D1	0.0390	0.0407	0.0355	0.0230	0.0268	0.0253	0.0105	0.0178	0.0120	0.0064	0.0133	0.0139
D2	0.0560	0.0548	0.0577	0.0410	0.0387	0.0402	0.0243	0.0235	0.0235	0.0141	0.0149	0.0102
D3	0.0660	0.0662	0.0687	0.0540	0.0514	0.0525	0.0365	0.0328	0.0355	0.0209	0.0181	0.0160
D4	0.0770	0.0767	0.0781	0.0660	0.0647	0.0649	0.0491	0.0453	0.0486	0.0283	0.0242	0.0255
D5	0.0880	0.0869	0.0872	0.0780	0.0788	0.0780	0.0630	0.0610	0.0633	0.0391	0.0347	0.0384
D6	0.0980	0.0973	0.0966	0.0920	0.0937	0.0924	0.0802	0.0800	0.0804	0.0528	0.0515	0.0553
D7	0.1080	0.1084	0.1073	0.1080	0.1100	0.1090	0.1004	0.1029	0.1010	0.0732	0.0766	0.0780
D8	0.1200	0.1216	0.1204	0.1280	0.1290	0.1295	0.1278	0.1311	0.1281	0.1077	0.1136	0.1109
D9	0.1390	0.1401	0.1398	0.1610	0.1547	0.1589	0.1711	0.1699	0.1703	0.1757	0.1709	0.1677
D10	0.2090	0.2074	0.2087	0.2490	0.2522	0.2495	0.3370	0.3357	0.3372	0.4818	0.4822	0.4841
	MSE	1.1E−6	2.3E−6	MSE	5.0E−6	1.4E−6	MSE	1.3E−5	4.2E−7	MSE	1.9E−5	1.6E−5
	MAE	0.0004	0.0006	MAE	0.0011	0.0005	MAE	0.0011	0.0003	MAE	0.0014	0.0017
	MAS	0.0007	0.0016	MAS	0.0028	0.0009	MAS	0.0024	0.0006	MAS	0.0025	0.0036
	IIM	3.6E−5	1.2E−4	IIM	2.6E−4	7.6E-5	IIM	9.8E−4	5.5E-5	IIM	0.0015	0.0020
K-S	D	0.1	0.1	D	0.1	0.1	D	0.1	0.1	D	0.1	0.1
test	*p*-value	1	1	*p*-value	1	1	*p*-value	1	1	*p*-value	1	1

The K_{RE} means fitting the Kakwani (1980)'s Lorenz curve using the regression method, and the L_{SH} denotes fitting the Lorenz curve proposed by Sitthiyot and Holasut (2021) using the error minimization technique. The K-S test is the Kolmogorov–Smirnov test.

respectively. The lower half of Table 4 reports the K-S test and the goodness-of-fit statistics such as IIM, MSE, MAE, and MAS.

The results of the K-S test suggest that there is no significant differences between the actual income shares and the estimated income shares for each country. Furthermore, judged by the MSE, MAE, MAS, and IIM, the regression method has a better performance than the error minimization technique in estimating income shares for Albania and the USA. In contrast, the error minimization technique has a better performance than the regression method for Belgium and South Africa.

Conclusions

Estimating Gini index with the income shares by decile attracts considerable attentions of researchers. Because the cumulative population shares and the corresponding cumulative income shares by decile form the points of Lorenz curve, fitting Lorenz curve is more convenient than fitting income distribution function. Based on the Lorenz curve suggested by Kakwani (1980), we propose a new approach named the regression method to estimate parameters of this Lorenz curve and calculate the Gini index for the sample economies.

We build a linear multiple regression equation and obtain the estimated values of three parameters of the Kakwani (1980)'s Lorenz curve using the latest data on the income shares by decile from the UNU-WIID database. We then calculate Gini index based on the Beta function, which is easily calculated utilizing some popular computer programs such as EWIEWS, STATA, MATHLAB, and R. We also provide a method to estimate the variance of Gini index.

The results suggest that the regression method has a better performance than the error minimization technique when fitting the Lorenz curve proposed by Kakwani (1980). We extend the range of the parameter p in Eq. (2) by replacing $0 < p \leq 1$ with $p > 0$. In the same time, the Lorenz curve proposed by Kakwani (1980) allows for negative income and wealth values. This is very useful for research using survey data with negative values.

We also analyze the effects of the different forms of Lorenz curves on the estimated Gini index. Using nine popular Lorenz curves, we find that the three-parametric Lorenz curves have a better performance than the double-parametric curves judged by the RMSE, and the double-parametric specifications are better than the single-parametric. Furthermore, we find that the three-parametric Lorenz curve proposed by Sarabia et al. (1999) has the best performance among the nine Lorenz curves. In addition, we find the regression method has the best performance for the economies with medium and higher levels of inequality, while the error minimization technique is the best for the economies with low and high levels of inequality under the Lorenz curve proposed by Sarabia et al. (1999).

Among the double-parametric Lorenz curves, the Lorenz curve suggested by Sitthiyot and Holasut (2021) has the best performance in terms of the estimated Gini index. Therefore, we compare the performance of the regression method to that of the error minimization technique in estimating the income shares by decile. The results show that the regression method is better than the error minimization technique for the economies with medium and higher inequality, while the minimization technique is better than the regression method for the economies with low and high inequality.

Data availability

The datasets generated and/or analyzed during the current study can be accessed from the United Nations University World Institute for Development Economics Research (UNU-WIDER) website (http://www4.wider.unu.edu/).

Received: 22 November 2023; Accepted: 30 August 2024;
Published online: 17 September 2024

Notes

[1] The density function of Beta-II distribution in Chotikapanich et al. (2007) is as following:$f(x) = \frac{x^{p-1}}{b^p B(p,q)(1+x/b)^{p+q}}$, $x > 0$ where, $B(\cdot)$ is the Beta function, b, p, and q are all positive parameters.

[2] The dirichlet distribution function in Chotikapanich and Griffiths (2002) is as following:$f(q|\alpha) = \frac{\Gamma(\alpha_1}{} + \alpha_2 + \cdots + \alpha_M)\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha\ M)q_1^{\alpha_1-1}q_2^{\alpha}$ $2 - 1\cdots q_M^{\alpha_M-1}$, $\alpha_i = \lambda[L(p_i;\theta) - L(p_{i-1};\theta)]$ here, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the Gamma function, λ is an additional unknown parameter. Correspondingly, the likelihood function of the above distribution function is as following:$\log[f(q|\theta)] = \log\Gamma(\lambda) + \sum_{i=1}^{M}\{\lambda[L(p_i,\theta) - L(p_{i-1},\theta)] - 1\} \times \log q_i - \sum_{i=1}^{M}\log\Gamma\{\lambda[L(p_i,\theta) - L(p_{i-1},\theta)]\}$.

References

Aggarwal V (1984) On optimum aggregation of income distribution data. Sankhya B 46:343–355

Blanchet T, Piketty T, Fournier J (2022) Generalized Pareto curves: theory and applications. Rev Income Wealth 1:263–288

Chotikapanich D, Valenzuela MR, Prasada Rao DS (1997) Global and regional inequality in the distribution of income: Estimation with limited/incomplete data. Empir Econ 20:533–546

Chotikapanich D, Griffiths WE, Rao DSP (2007) Estimating and combining national income distributions using limited data. J Bus Econ Stat 25:97–109

Chotikapanich D (1993) A comparison of alternative functional forms for the Lorenz curve. Econ Lett 41:21–29

Chotikapanich D, Griffiths WE (2002) Estimating Lorenz curves using a dirichlet distribution. J Bus Econ Stat 20:290–295

Cheong KS (2002) An empirical comparison of alternative functional forms for the Lorenz curve. Appl Econ Lett 9:171–176

Dagum C (1977) A new model of personal income distribution: Specification and estimation. In: Chotikapanich D (Ed) Modeling income distributions and Lorenz curves. Economic studies in equality, social exclusion and well-being, vol 5. Springer, New York, pp. 3–25

Jorda V, Sarabia JM, Jantti M (2021) Inequality measurement with grouped data: parametric and non-parametric methods. J R Stat Soc Ser A 184:964–984

Kakwani NC (1980) On a class of poverty measures. Econometrica 48:437–446

Kakwani NC, Podder N (1973) On the estimation of Lorenz curves from grouped observations. Int Econ Rev 14:278–292

Kakwani NC, Podder N (1976) Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations. Econometrica 44:137–148

McDonald JB (1984) Some generalized functions for the size distribution of income. Econometrica 52:647–663

Ortega P, Martn G, Fernndez A, Ladoux M, Garca A (1991) A new functional form for estimating Lorenz curves. Rev Income Wealth 37:447–452

Paul S, Shankar S (2020) An alternative single parameter functional form for Lorenz curve. Empir Econ 59:1393–1402

Rasche RH, Gaffney J, Koo A, Obst N (1980) Function forms for estimating the Lorenz curve. Econometrica 48:1061–1062

Sarabia JM, Castillo E, Slottje D (1999) An ordered family of Lorenz curves. J Econ 91:43–60

Sitthiyot T, Holasut K (2021) A simple method for estimating the Lorenz curve. Hum Soc Sci Commun. 8:268

Sitthiyot T, Holasut K (2023) An investigation of the performance of parametric functional forms for the Lorenz curve. PLoS ONE 18(6):e0287546

Tanak AK, Mohtashami Borzadaran GR, Ahmadi J (2018) New functional forms of Lorenz curves by maximizing Tsallis entropy of income share function under the constraint on generalized Gini index. Phys A 511:280–288

Author Contributions

Conceptualization: Pingsheng Dai. Formal analysis: Xiaobo Shen, Pingsheng Dai. Methodology: Pingsheng Dai. Validation: Xiaobo Shen. Writing - original draft: Pingsheng Dai. Writing - review & editing: Pingsheng Dai, Xiaobo Shen.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Pingsheng Dai.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.