Estimating Lorenz Curve of Income by Cubic Spline with Multiple Knots

Ying LIANG\(^1\), Han YANG\(^2\)*, Jia-xu ZHANG\(^2\), Qiang Li\(^2\) and Xue-zhang LIANG\(^2\)

\(^1\)Business School of Jilin University, Changchun, 130012, P.R. China
\(^2\)Mathematics School of Jilin University, Changchun, 130012, P.R. China

*Corresponding author

Keywords: Cubic spline with multiple knots, Linear programming, Aggregated Lorenz curve, Gini index.

Abstract. In this paper we propose a method for convex approximation of the Lorenz curve by cubic-spline with multiple knots. This method is applied to the China-2005 grouped data of urban and rural residents’ income, published in “China Statistical Yearbook-2006”. Then, we use the method proposed in this paper to estimate the Lorenz curves and the Gini index of urban and rural residents’ income in China, respectively. Finally we estimate the Lorenz curve of residents’ income of whole China (2005) by the aggregated approach.

Introduction

The Lorenz curve and the Gini index are important tools for analyzing the distribution of income and the degree of wealth distribution inequality. On the basis of the known grouped data of income, there are a variety of methods to approximate the Lorenz curve \[1-5\]. In this paper we present a method to approximate the Lorenz curve convexly by cubic-spline with multiple knots. We apply the method to China-2005 grouped data of urban and rural residents’ income, published in “China Statistical Yearbook-2006”\(^6\), to approximate urban and rural income Lorenz curves respectively. Furthermore, we obtain the Lorenz curve of whole China(2005) by the aggregation approach\(^7\).

The Lorenz curve is a monotonously increasing convex curve on \([0,1]\) and satisfies \(L(0) = 0, L(1) = 1\). In this paper, the income distribution function, its inversion and density function are respectively denoted as

\[ p = F(x), \]
\[ x = x(p) = F^{-1}(p), \]

and

\[ f(x) = F'(x) = \frac{dp}{dx}, \]

where \(x\) represents income and \(p\) represents cumulative population share, with \(x \in [0, \infty)\) and \(p \in [0,1]\).

Average income has the following expression:

\[ \mu = \int_{0}^{1} x(p)dp = \int_{0}^{\infty} xf(x)dx. \]
\[
L = L(p) = \frac{1}{\mu} \int_0^p x(q) dq = \frac{1}{\mu} \int_0^p f(t) dt
\]
and
\[
\frac{dL}{dp} = \frac{1}{\mu} x(p)
\]
Obviously, we have:
\[
x = x(p) = \mu L(p)
\]

The rest of this paper is organized as follows. In section 2, we introduce a new method of estimating Lorenz curve convexly by cubic-spline with multiple knots. In section 3, we briefly describes the aggregation formula. Section 4 provides the Lorenz curves of urban and rural area in China (2005) by the method proposed in section 2, and estimates the Lorenz curve of whole china (2005) by aggregation of the Lorenz curves of urban and rural area in China (2005). Concluding remarks is presented in section 5.

### A Kind of Approximation of Cubic Spline with Multiple Knots

Assume that \( \{x_i\}_{i=1}^n \) is a set of real nodes satisfying \( 0 = x_1 < \cdots < x_n = 1 \) and \( s(x) \) is continuous function on \([0,1]\). We call \( s(x) \) a cubic spline with multiple nodes \( \{x_i\}_{i=1}^n \) on \([0,1][8]\), if the function \( s(x) \) has a first-order continuous derivative on \([0,1]\) and equates to a real algebraic polynomial of degree \( \leq 3 \) in each interval \([x_j,x_{j+1}]\), \( j=1, \ldots, n-1 \).

Let \( L = L(p) \) be a given Lorenz curve, and \( \{(X_i,Y_i)\}_{i=1}^m \) be the grouped data of the Lorenz curve, which satisfy \( Y_i = L(X_i) \) \( (i=1,2,\ldots,m) \), with \( 0 = X_1 < X_2 < \cdots < X_m = 1 \) and \( 0 = Y_1 < Y_2 < \cdots < Y_m = 1 \).

We try to approximate the Lorenz curve \( L(p) \) by cubic spline \( s(x) \) with multiple knots \( \{x_i\}_{i=1}^n \) on \([0,1]\) satisfying \( \{X_i\}_{i=1}^m \subseteq \{x_i\}_{i=1}^n \) and \( m \leq n \) (We call the \( s(x) \) a spline representation of \( L(p) \)). Assume that \( x_{k_j} = X_j \) \( (j=1,\ldots,m) \) with \( 1 = k_1 < k_2 < \cdots < k_{m-1} < k_m = n \) and the corresponding values of spline function \( s(x) \) in the knots are \( \{y_j\}_{j=1}^m \) \( (0 \leq y_j \leq 1) \), and the first derivatives are \( \{y_j'\}_{j=1}^m \). Then, in each interval \([x_j,x_{j+1}]\) \( (j=0,1,\ldots,n-1) \), referring to cubic Hermite interpolation formula \([9]\), the spline function \( s(x) \) has the following expressions:

\[
\begin{align*}
    s_i(x) &= y_i (1 + \frac{2(x-x_i)}{h_i} (x-x_{i+1})^2 + y_i' (1 - \frac{2(x-x_{i+1})}{h_i}) (x-x_i) + y_i' (x-x_i) (x-x_{i+1})^2 + \\
    y_i' (x-x_{i+1}) (x-x_i)^2 \frac{1}{h_i^2},
\end{align*}
\]

and the second derivative of \( s_i(x) \) has the following expression:

\[
\begin{align*}
    s_i''(x) &= 6(2x-x_i-x_{i+1}) \frac{y_i-y_{i+1}}{h_i^3} + 2 \frac{y_i' (2x-x_i-2x_{i+1}) + y_i' (3x-2x_i-x_{i+1})}{h_i^2},
\end{align*}
\]
where \( h_i = x_{i+1} - x_i \), \( i = 0, \cdots, n - 1 \).
Because \( s_i''(x) \) is a linear function on \([x_i, x_{i+1}]\), \( s_i''(x) \geq 0 \) on \([x_i, x_{i+1}]\) if and only if \( s_i'(x_i) \geq 0 \) and \( s_i'(x_{i+1}) \geq 0 \). Therefore, \( s(x) \) is a monotone increasing and smooth convex curve on \([0,1]\) if and only if the following inequalities hold

\[
\begin{align*}
\frac{2}{3}y_i + \frac{1}{3}y_{ri+1} + \frac{1}{h_i}y_i - \frac{1}{h_i}y_{ri+1} & \leq 0 (1 \leq i \leq n-1) \\
\frac{1}{3}y_i - \frac{2}{3}y_{ri+1} - \frac{1}{h_i}y_i + \frac{1}{h_i}y_{ri+1} & \leq 0 (1 \leq i \leq n-1)
\end{align*}
\]

(2.3)

Then, the formula of Gini index corresponding to \( s(x) \) is

\[
G = 1 - 2 \int_0^1 s(p) dp
\]

(2.4)

where

\[
\int_0^1 s(p) dp = \frac{1}{2} \sum_{i=1}^{n} h_i (y_i + y_{ri+1}) + \frac{1}{12} \sum_{i=1}^{n} h_i^2 (y_{ri+1} + y_i)
\]

(2.5)

Obviously, as definite integral in formula (2.5) reaches the minimum value, the Gini index reaches the maximum value. Assume

\[
C = \frac{1}{2} \sum_{i=1}^{n} h_i (y_i + y_{ri+1}) + \frac{1}{12} \sum_{i=1}^{n} h_i^2 (y_{ri+1} + y_i)
\]

(2.6)

The problem of estimating Lorenz curve of income by using cubic-spline with multiple knots can be transformed to the question of solving the linear programming as following:

Minimize the objective function \( C \) subject to the following constraint:

\[
\begin{align*}
\frac{2}{3}y_i + \frac{1}{3}y_{ri+1} + \frac{1}{h_i}y_i - \frac{1}{h_i}y_{ri+1} & \leq 0 (i = 1, 2, \cdots, n - 1) \\
y_n = Y_n (i = 1, 2, \cdots, m)
\end{align*}
\]

(2.7)

We solve the above linear programming problem with the simplex (Big-M) method[10] to obtain \( y_i \) and \( y_{ri} (i = 1, 2, \cdots, n) \). Furthermore, we can obtain the Lorenz curve by using the Hermite interpolation formula (2.1) in each interval \([x_i, x_{i+1}]\).

Next, we introduce the methods for selecting urban and rural spline node \( \{x_i\}_{i=1}^{n} \) respectively. Assume that there are \( m_i \) sets of the given grouped data of urban residents’ income, spline nodes \( \{x_i\}_{i=1}^{n} \) of estimating Lorenz curve for urban area can be expressed as following:
\[ x_i = X_i, \quad i = 1, 2, 3. \]
\[ x_4 = (X_3 + X_4) / 2 \]
\[ x_{k(i-1)+j+4} = X_{i+j} + \frac{(j-1)}{k} (X_{i+j} - X_{i+j-1}) \quad i = 1, 2, 3; \quad j = 1, 2, \ldots, k. \]  
(2.8)
\[ x_{2i+3k+3} = X_{i+6}, i = 1, 2, 3. \]
\[ x_{3k+m+2i-5} = (X_{m+i-3} + X_{m+i-2}) / 2, i = 1, 2. \]

where \( h_i = x_{i+1} - x_i \) (0 \( \leq i \leq n-1 \)), then the number of spline nodes is \( n = 3k + m_1 \) with \( m_1 = 9 \) and \( k = 4 \) in section 4.

Assume that there are \( m_2 \) sets of the given grouped data of rural residents’ income, and rural spline nodes \( \{x_i\}_{i=1}^{n} \) of estimating Lorenz curve for rural area can be expressed as following:

\[ x_{(i-1) + k + j} = X_i + (j-1)(x_{i+1} - x_i) / k, \quad i = 1, 2, \ldots, m_2 - 1; \quad j = 1, 2, \ldots, k. \]
\[ x_m = X_m \]  
(2.9)

where \( h_i = x_{i+1} - x_i \) (0 \( \leq i \leq n-1 \)), then the number of spline nodes is \( n = (m_2 - 1)k + 1 \) with \( m_2 = 5 \) and \( k = 3 \) in section 4.

**Aggregation Formulas for Lorenz Curve**

Assume that the population of urban and rural area are \( P_1 \) and \( P_2 \) respectively, the total population is \( P \). The ratio of urban and rural population to nation are represented as \( \lambda = P_1 / P \) and \( \hat{\lambda} = P_2 / P \) respectively, obviously, \( \lambda + \hat{\lambda} = 1 \). Let \( \mu_1 \) be the average of urban residents’ income and \( \mu_2 \) be the average of rural residents’ income.

Distribution function, density function and Lorenz curve of urban income are defined as,

\[ p_1 = F_1(x), \quad f_1(x) = F_1'(x), \quad \text{and} \quad L_1 = L_1(p_1). \]  
(3.1)

Distribution function, density function and Lorenz curve of rural income are defined as,

\[ p_2 = F_2(x), \quad f_2(x) = F_2'(x), \quad \text{and} \quad L_2 = L_2(p_2). \]  
(3.2)

And nationwide Lorenz curve is defined as,

\[ L = L(p). \]  
(3.3)

Then, the aggregation formula\(^7\) is as following:

\[ \begin{align*}
   p &= \hat{\lambda} p_1 + \lambda p_2 \\
   x &= L_1(p_1) \mu_1 = L_2(p_2) \mu_2 = \mu L(p) \\
   L(p) &= \frac{1}{\mu} \left[ \hat{\lambda} \mu_1 L_1(p_1) + \lambda \mu_2 L_2(p_2) \right] 
\end{align*} \]  
(3.4)

where \( \mu \) represents the aggregation average income, with
\[ \mu = \lambda_1 \mu_1 + \lambda_2 \mu_2. \]  

(3.5)

Based on the spline representations of Lorenz curve \( L_1(p_1) \) and \( L_2(p_2) \), we can obtain the values of \( L(p) \) by aggregation formula (3.4) for given \( p (0 \leq p \leq 1) \).

**A Practical Application**

The method described in section 2 is applied to the grouped data of urban and rural area in China (2005) published in "China Statistical Yearbook (2006)"\(^{[6]}\), we obtain the Lorenz curve of urban and rural area. Furthermore we get the Lorenz curve of whole China (2005) by aggregation approach. In the process of calculation, income of urban residents adopts disposable income, rural residents’ income adopts net income. The average income of urban residents is \( \mu_1 = 10483 \), the average income of rural residents is \( \mu_2 = 3254.9 \), and \( \lambda_1 = 0.4299, \lambda_2 = 0.5701 \). The grouped data of urban and rural China are as following Tables 1 and 2. The calculation results are as following Table 3 to 9.

**Table 1.** The grouped data of residents’ income in urban China (2005).

<table>
<thead>
<tr>
<th>X</th>
<th>0.0000</th>
<th>0.0559</th>
<th>0.1115</th>
<th>0.2205</th>
<th>0.4311</th>
<th>0.6319</th>
<th>0.8201</th>
<th>0.9115</th>
<th>1.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0.0000</td>
<td>0.0131</td>
<td>0.0329</td>
<td>0.0830</td>
<td>0.2160</td>
<td>0.3897</td>
<td>0.6140</td>
<td>0.7605</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Table 2.** The grouped data of residents’ income in rural China (2005).

<table>
<thead>
<tr>
<th>X</th>
<th>0.0000</th>
<th>0.2258</th>
<th>0.4384</th>
<th>0.6397</th>
<th>0.8292</th>
<th>1.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0.0000</td>
<td>0.0721</td>
<td>0.2017</td>
<td>0.3742</td>
<td>0.6042</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Table 3.** The values of Lorenz curve of urban China (2005) and the corresponding first derivatives.

<table>
<thead>
<tr>
<th>x</th>
<th>0.0000</th>
<th>0.0559</th>
<th>0.1115</th>
<th>0.1660</th>
<th>0.2205</th>
<th>0.2732</th>
<th>0.3258</th>
<th>0.3784</th>
<th>0.4310</th>
<th>0.4813</th>
<th>0.5315</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0.0000</td>
<td>0.0131</td>
<td>0.0329</td>
<td>0.0579</td>
<td>0.0830</td>
<td>0.1072</td>
<td>0.1321</td>
<td>0.1709</td>
<td>0.2160</td>
<td>0.2590</td>
<td>0.3020</td>
</tr>
<tr>
<td>y’</td>
<td>0.0943</td>
<td>0.3044</td>
<td>0.4596</td>
<td>0.4596</td>
<td>0.4596</td>
<td>0.4596</td>
<td>0.4991</td>
<td>0.8564</td>
<td>0.8564</td>
<td>0.8564</td>
<td>0.8564</td>
</tr>
</tbody>
</table>

**Table 4.** The values of Lorenz curve of urban China (2005) and the corresponding first derivatives.

<table>
<thead>
<tr>
<th>x</th>
<th>0.0581</th>
<th>0.6319</th>
<th>0.6792</th>
<th>0.7265</th>
<th>0.7737</th>
<th>0.8210</th>
<th>0.8862</th>
<th>0.9115</th>
<th>0.9557</th>
<th>1.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0.3457</td>
<td>0.3897</td>
<td>0.4312</td>
<td>0.4726</td>
<td>0.5375</td>
<td>0.6140</td>
<td>0.6873</td>
<td>0.7605</td>
<td>0.8321</td>
<td>1.0000</td>
</tr>
<tr>
<td>y’</td>
<td>0.8771</td>
<td>0.8771</td>
<td>0.8771</td>
<td>0.8771</td>
<td>1.6188</td>
<td>1.6188</td>
<td>1.6188</td>
<td>1.6188</td>
<td>1.6188</td>
<td>8.1434</td>
</tr>
</tbody>
</table>

Figure 1. Estimation of Lorenz curve for urban China (2005).
Table 5. The values of Lorenz curve of rural China (2005) and the corresponding first derivatives.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0.0000</th>
<th>0.0749</th>
<th>0.1499</th>
<th>0.2248</th>
<th>0.2960</th>
<th>0.3672</th>
<th>0.4384</th>
<th>0.5055</th>
<th>0.5726</th>
<th>0.6397</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0266</td>
<td>0.0721</td>
<td>0.1153</td>
<td>0.1585</td>
<td>0.2017</td>
<td>0.2424</td>
<td>0.2950</td>
<td>0.3742</td>
</tr>
<tr>
<td>$y'$</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.6067</td>
<td>0.6067</td>
<td>0.6067</td>
<td>0.6067</td>
<td>0.6067</td>
<td>0.6067</td>
<td>1.1362</td>
<td>1.2032</td>
</tr>
</tbody>
</table>

Table 6. The values of Lorenz curve of rural China (2005) and the corresponding first derivatives.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0.7029</th>
<th>0.7660</th>
<th>0.8292</th>
<th>0.8861</th>
<th>0.9431</th>
<th>1.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>0.4502</td>
<td>0.5262</td>
<td>0.6022</td>
<td>0.6707</td>
<td>0.7392</td>
<td>1.0000</td>
</tr>
<tr>
<td>$y'$</td>
<td>1.2032</td>
<td>1.2032</td>
<td>1.2032</td>
<td>1.2032</td>
<td>1.2032</td>
<td>11.3360</td>
</tr>
</tbody>
</table>

Figure 2. Estimation of Lorenz curve for rural China (2005).

Table 7. The values of Lorenz curve of whole China (2005) and the corresponding first derivatives.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0.0000</th>
<th>0.1000</th>
<th>0.2000</th>
<th>0.3000</th>
<th>0.4000</th>
<th>0.5000</th>
<th>0.6000</th>
<th>0.7000</th>
<th>0.8000</th>
<th>0.9000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>0.0000</td>
<td>0.0118</td>
<td>0.0428</td>
<td>0.0739</td>
<td>0.1241</td>
<td>0.1856</td>
<td>0.2492</td>
<td>0.3300</td>
<td>0.4709</td>
<td>0.6476</td>
</tr>
<tr>
<td>$y'$</td>
<td>0.0000</td>
<td>0.3102</td>
<td>0.3102</td>
<td>0.3156</td>
<td>0.6146</td>
<td>0.6151</td>
<td>0.7575</td>
<td>1.2533</td>
<td>1.4456</td>
<td>2.6680</td>
</tr>
</tbody>
</table>

Table 8. The values of Lorenz curve of whole China (2005) and the corresponding first derivatives.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0.9100</th>
<th>0.93</th>
<th>0.94</th>
<th>0.9500</th>
<th>0.9600</th>
<th>0.9700</th>
<th>0.9800</th>
<th>0.9900</th>
<th>1.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>0.6743</td>
<td>0.7010</td>
<td>0.7277</td>
<td>0.7544</td>
<td>0.7810</td>
<td>0.8077</td>
<td>0.8344</td>
<td>0.8660</td>
<td>0.9121</td>
</tr>
<tr>
<td>$y'$</td>
<td>2.6680</td>
<td>2.6680</td>
<td>2.6680</td>
<td>2.6680</td>
<td>2.6680</td>
<td>2.6680</td>
<td>2.6680</td>
<td>2.6933</td>
<td>3.7791</td>
</tr>
</tbody>
</table>

Figure 3. Estimation of Lorenz curve of whole China (2005).
Table 9. The estimations of rural, urban and nationwide Gini index in China (2005).

<table>
<thead>
<tr>
<th>year</th>
<th>Urban Gini index</th>
<th>Rural Gini index</th>
<th>nationwide Gini index</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>0.3468</td>
<td>0.3911</td>
<td>0.4843</td>
</tr>
</tbody>
</table>

Conclusion

From the above computation results, we can see that Lorenz curve obtained by cubic-spline with multiple knots not only keeps monotonicity and convexity of the Lorenz curve, but also has better approximation performance. Table 9 shows that nationwide Gini index of China (2005) calculated by this paper is very close to 0.485, reported by National Statistical Bureau in 2014. Obviously, the approach proposed in the paper is feasible and effective. The results reported in tables 7 and 8 have some reference value

Acknowledgement

This research was financially supported by the Chinese National Science Foundation (No.11271041), Jilin University Philosophy and Social Science Fund(2011QY093), and China Scholarship Council (CSC).

References