Estimation of the Lorenz Curve and Concentration Ratio
Shi-Tao Yeh, Consultant, Malvern, PA, USA

ABSTRACT

The concentration ratio is a measure of the inequalities of the size distribution. It has been used for studying market share concentration, city size distribution, income distribution inequality, environmental pollution concentration, ..., etc. Such inequalities can be recorded in the form of a Lorenz curve. The diagonal straight line shows what a distribution of complete equality in cumulative percentage would look like, so the extent to which the Lorenz curve deviates from this line gives an indication of relative concentration.

The Gini coefficient, or concentration ratio, provides a summary measure of the extent to which the Lorenz curve deviated from the equitarian line.

This paper discusses the evaluation of functional forms of the Lorenz curve and its mathematical properties. The SAS modules with SAS code for estimating the Lorenz curve and the concentration ratio are provided.

The last section of this paper deals with real-world application.

INTRODUCTION

The concentration ratio is a measure of the inequalities of the size distribution. It has been used for studying market share concentration, city size distribution, income distribution inequality, environmental pollution concentration, ..., etc. Such inequalities can be recorded in the form of a Lorenz curve, as in Figure 1. The diagonal straight line shows what a distribution of complete equality in cumulative percentage would look like, so the extent to which the Lorenz curve deviates from this line gives an indication of relative concentration. For example, the diagonal line shows how one might expect 50% of total population to be accounted for by 50% of the total income, whereas in fact 50% of the lower income from the total population are accounted for by only 23% of total income, as the Lorenz curve indicates in Figure 1.

The Gini coefficient, crosshatched lines area in Figure 1, provides a summary measure of the extent to which the Lorenz curve deviated from the equitarian line. It indicates the extent of the crosshatched lines area in the figure by dividing the area of crosshatched lines by the area below the line of equality. The value of the Gini coefficient ranges from 0 (complete equality) to 1 (complete inequality).

This paper discusses the evaluation of functional forms of the Lorenz curve and its mathematical properties. The SAS modules with SAS code for estimating the Lorenz curve and the concentration ratio are provided.

The last section of this paper deals with real-world application.
In recent years, there has been some development on deriving the functional forms for the Lorenz curve. The functional form

\[ y = f(x) \]

represents the Lorenz curve if it satisfies the following properties:

(i) \( f(0) = 0; \)
(ii) \( f(1) = 1; \)
(iii) \( f'(x) > 0, \text{ for } 0 < x < 1; \)
(iv) \( f''(x) > 0, \text{ for } 0 < x < 1; \)
(v) \( f(x) < x, \text{ for } 0 < x < 1; \)
(vi) \( 0 \leq \int f(x) \, dx \leq 1/2. \)

There are two proposed functional forms that satisfy all of the properties. One is proposed by Rache, Gaffney, Koo, and Obst [2] as follows:

\[ y = \left[ 1 - (1-x)^\alpha \right]^{1/\beta} \quad (1) \]

\[ \text{where } 0 < \alpha < 1, 0 < \beta < 1; \]

the other, suggested by Gupta [1], is the following:

\[ y = (x-1)^A \quad \text{with } A > 1. \quad (2) \]

Functional form (1) has been criticized with its nonlinearity in the parameter, which makes estimation of the parameters by the linear least squares method impossible [1]. Functional form (2) is intrinsically linear; nonlinear with respect to the variables, but linear with respect to the parameters to be estimated. These two functional forms has been used for parameter estimation. The results from these two forms are very similar. The NLIN procedure in SAS/STAT® provides a powerful tool that the estimation of nonlinear parameters becomes relatively simple and easy.

**MODULE FOR PROCESSING INPUT DATA**

The data from Table 1 are used throughout the paper for demonstration purpose.

The following SAS® macro module will process input data and convert raw data into data with sorted cumulative percentage.

<table>
<thead>
<tr>
<th>County</th>
<th>1970</th>
<th>1980</th>
<th>1990</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philadelphia</td>
<td>1,949,996</td>
<td>1,688,210</td>
<td>1,585,577</td>
</tr>
<tr>
<td>Montgomery</td>
<td>624,080</td>
<td>643,621</td>
<td>678,111</td>
</tr>
<tr>
<td>Chester</td>
<td>277,746</td>
<td>316,660</td>
<td>376,396</td>
</tr>
<tr>
<td>Delaware</td>
<td>603,456</td>
<td>555,007</td>
<td>547,651</td>
</tr>
<tr>
<td>Bucks</td>
<td>416,278</td>
<td>479,211</td>
<td>541,174</td>
</tr>
<tr>
<td>Camden</td>
<td>456,291</td>
<td>471,500</td>
<td>502,824</td>
</tr>
<tr>
<td>Burlington</td>
<td>323,132</td>
<td>362,542</td>
<td>395,066</td>
</tr>
<tr>
<td>Gloucester</td>
<td>172,681</td>
<td>199,917</td>
<td>230,082</td>
</tr>
<tr>
<td>New Castle</td>
<td>385,856</td>
<td>398,115</td>
<td>441,946</td>
</tr>
</tbody>
</table>

Source: CB Commercial Real Estate Group

**Table 1. Population in Tri-State Area**
This macro function %cinput can be used anywhere in the SAS program. For example, a SAS data set F1 contains data of Table 1 with variable names: COUNTY $, Y1970, Y1980, and Y1990. The following SAS statements will produce a data set F2, with variables: OBS, Y70, X, Y80, and Y90, shown in Figure 2.

```sas
%macro cinput(fl, var1, x1, var2, off);
data fl;
set &fl;
ct = 1;
proc sort; by ct &var1;
proc summary;
class ct;
var &var1;
output out=a1 sum=s1 n=n1;
data a1;
set a1;
ct = 1;
if _type_ = 0;
keep ct s1 n1;
data &off;
merge a1 fl;
by ct;
pct = &var1/s1;
retain &var2 0;
&var2 = &var2 + pct;
ctx = ct / n1;
retain &x1 0;
&x1 = &x1 + ctx;
keep &x1 &var2;
%mend cinput;

%cinput(f1,y1970,0,y70,0)
%cinput(f1,y1980,x,y80,02)
%cinput(f1,y1990,x,y90,03)
run;
data f2;
set o1; set o2; set o3;
```

Figure 2 SAS Macro Code for Input Data Processing

The arguments of this macro function are:

- **fl**: data set name which contains the variables to be processed,
- **var1**: variable name in fl data set for which to be processed,
- **x1**: computed values on equalitarian line,
- **var2**: variable name for processed var1 values,
- **off**: output data set name.

To invoke this function, write its macro function name %cinput and then the arguments of data set name fl, variable name var1, name for computed values on equalitarian line x1, variable name var2 for processed var1 values, and output data set name off, enclosed in parentheses.

---

**MODULE FOR PARAMETER ESTIMATION**

The NLIN procedure in SAS/STAT is employed for the Lorenz curve parameter estimation. Figure 3 shows a SAS macro estimation module.

```sas
%macro nlreg(x,y,a,b,p);
proc nlin data=&f
maxiter=30
converge=.00001;
parms &a=0.5
&b=0.6;
bounds 0<&a<1;
0<&b<1;
model &y=(1-(1-&a)**&a)**(&y);
output out=&p parms=&a &b;
%mend nlreg;
```

Figure 3 Data Set F2

Figure 4 Parameter Estimation Module
The macro function %nlreq accepts users supplied arguments of:

- \( f \): data set name,
- \( x \): values on equalitarian line,
- \( y \): target variable for analysis,
- \( aI, bI \): estimates of parameters \( a \) and \( b \) in Lorenz curve (1),
- \( pl \): output data set name.

After executing the macro function %nlreq, you can use the following macro function to process output file.

```
%macro fone(pI,aI);
  data &pI;
  set &pI;
  proc sort; by &aI;
  data &pI;
  set &pI;
  by &aI;
  if first.&aI;
  %mend fone;
```

The complete macro code are shown in the following figure.

```
%macro cr(pI,aI,bI,year);
  data &pI;
  set &pI;
  a = &aI;
  b = &bI;
  aI = 1/&aI;
  bI = (1/&bI) + 1;
  cI = aI + bI;
  cr = 1.0 - (2.0/a) * (gamma(aI)) * (gamma(bI)/gamma(cI));
  year = &year;
  keep year cr a b;
  %mend cr;
```

### APPLICATION

The Tri-state population data in Table 1 show the population changes from 1970 to 1990. One may ask the question of equity in the population size distribution in the study area. The following SAS statements are used to produce Table 2.

```
%nlreg(x,y70,a0,b0,p0)
%fone(p0,a0)
%cr(p0,a0,b0,1970)
%nlreg(x,y80,a1,b1,p1)
%fone(p1,a1)
%cr(p1,a1,b1,1980)
%nlreg(x,y90,a2,b2,p2)
%fone(p2,a2)
%cr(p2,a2,b2,1990)
```

Here, the SAS function of GAMMA is used for computing the beta distribution. The macro function %cr is designed for computing concentration ratio. The arguments of this function are

- \( pI \): data set name which is the output file of macro function %nlreg,
- \( aI, bI \): estimates of parameters, 
- \( year \): the year which the data are represented.

Substituting variables

\[ u = 1 - (1 - x)^\alpha, \]

this is equal to:

\[ CR = 1.0 - 2.0 \left( \int \frac{1}{a} \left(1 - \frac{u}{b} \right)^{\alpha - 1} du \right) \]

\[ = 1.0 - (2.0/a) \cdot B(1/\alpha, 1/\beta + 1) \]

where \( B \) represents the beta distribution.

SAS function of GAMMA is used for computing the beta distribution. The macro function %cr is designed for computing concentration ratio. The arguments of this function are

- data name of output file,
- \( aI, bI \): estimates of parameters \( a \) and \( b \) in Lorenz curve (1),
- \( pl \): data set name which is the output file of macro function %nlreg.

The complete macro code are shown in the following figure.
Population Concentration Trend in Tri-State

<table>
<thead>
<tr>
<th>year</th>
<th>concentration ratio</th>
<th>estimate of a</th>
<th>estimate of b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>0.38806</td>
<td>0.50352</td>
<td>0.89967</td>
</tr>
<tr>
<td>1980</td>
<td>0.33115</td>
<td>0.55668</td>
<td>0.91887</td>
</tr>
<tr>
<td>1990</td>
<td>0.28947</td>
<td>0.59915</td>
<td>0.93148</td>
</tr>
</tbody>
</table>

Table 2 Population Concentration Trend in Tri-State Area

Table 2 presents the estimated parameters of the Lorenz curves along with the different inequality measures. Figure 3 exhibits the corresponding Lorenz curves. It shows a trend of equity in the population size distribution in the Tri-State area.

REFERENCES


SAS, and SAS/STAT are registered trademarks of SAS Institute Inc., Cary NC, USA