Uses of parametric modeling of income for aggregate income indicators from EU-SILC data and AMELIA synthetic universe

Monique Graf and Desislava Nedyalkova
Swiss Federal Statistical Office, Statistical Methods Unit

Swiss Statistics Meeting, Fribourg 2011
Outline

The AMELI Project

Income and indicators in the EU-SILC context

Parametric estimation

Examples of use

Variance of ML estimates

Simulation setup

Some results

Discussion

Uses of parametric estimation:
The AMELI Project

"Advanced Methodology for Laeken Indicators" is a research project on the statistical methodology for social inclusion indicators, under the 7th framework programme of the European Union.

- Aim(s): Improved estimation; visualization; simulations.
- Database: EU-SILC (social and living conditions) survey.
- Main emphasis on monetary indicators,
- but should be transposable to other indicator areas.
- Delivery of open source codes in R.

Uses of parametric estimation: The AMELI Project
The AMELI Team

Germany
University of Trier Ralf Münnich (Co-ordinator)
German Federal Statistical Office Oliver Bode

Switzerland
Univ. of Applied Sciences NW Switzerland Beat Hulliger
Swiss Federal Statistical Office Monique Graf

Austria
Vienna University of Technology Matthias Templ
Statistics Austria Thomas Burg

Finland
University of Helsinki Risto Lehtonen
Statistics Finland Timo Alanko

Slovenia
Statistical Office of the Republic of Slovenia Rudi Seljak

Estonia
Statistics Estonia Kaja Sõstra
Equivalized income

- Main income variable in the EU-SILC survey:
  \[
  \text{Equivalized income} = \frac{\text{total gross household income}}{\text{household equivalized size}}.
  \]

- Household equivalized size = weighted sum of the no of adults and children in the household.

All household members have:

- the same equivalized income = the household equivalized income.
- the same sampling weight.
Indicators of poverty and social exclusion

- The at-risk-of-poverty rate ($ARPR$): proportion of incomes $< 60\%$ of the median income $=$ proportion of poors.
- The relative median poverty gap ($RMPG$): relative difference between the median income of the poors and $60\%$ of the median income.
- The income quintile share ratio ($QSR$): total income of the upper quintile to the total income of the lower quintile.
- The Gini index ($GINI$).
These indicators are computed within the EU-SILC survey at the country level and for several subgroups. One of the main objectives of the AMELI project is to review the estimation methods of these indicators in different contexts.

This presentation is about parametric estimation.
Parametric estimation: why?

1. to study the properties of estimators of indicators;
2. to stabilize estimation;
3. to get insight into the relationships between the characteristics of the theoretical distribution and a set of indicators, e.g. by sensitivity plots;
4. to deduce the whole distribution from known empirical indicators, when the raw data are not available.
Context

- Fisk (1961) introduces a new skewed distribution as a competitor to the lognormal.
- Dagum (1975, 1977) and Singh - Maddala (1976) extend Fisk’s distribution by adding a parameter in two different ways.
- McDonald (1984), among others, defines the Generalized Beta distribution of the second kind (GB2) that encompasses Fisk’s, Dagum’s and Singh - Maddala’s distributions.

Empirical studies on income, see Table B2 of Kleiber and Kotz (2003) and Jenkins (1997), tend to show that the GB2 outperforms other 4-parameter distributions.
Generalized beta distribution of the second kind
\( GB2(a, b, p, q) \)

- Density:

\[
f(x; a, b, p, q) \sim \frac{(x/b)^{ap-1}}{(1 + (x/b)^a)^{p+q}},
\]

where

- \( b > 0 \) is a scale parameter,
- \( p > 0, q > 0 \) and \( a > 0 \) are shape parameters.

- Distribution function \( F(x; a, b, p, q) \):
  no closed form, but easily computable in R.

- Moment of order \( k \):
  moments exist for \( -ap < k < aq \).

Dagum: \( q = 1 \); SM: \( p = 1 \); Fisk: \( p = q = 1 \).
Indicators of poverty and social exclusion under the GB2

Under the GB2, indicators are functions of the 3 scale parameters $a, p, q$.

In the R package GB2 (Graf and Nedyalkova 2010), we find:

- density, distribution and quantile function, random GB2 generator;
- functional expressions of the indicators, moments, moments of log;
- different types of estimation methods (weighted maximum likelihood, NLS fit from indicators);
- graphical procedures (distribution plots, likelihood plots, sensitivity plots).
Log-likelihood function

Pseudo log-likelihood function (see e.g. Skinner et al, 1989):

\[ \ell_m(\theta) = \sum_{i=1}^{m} w_i n_i \log f(x_i; \theta), \]

- \( m \) - number of households,
- \( n_i \) - number of persons in household \( i \),
- \( w_i \) - the sampling weights,
- \( \theta = (a, b, p, q)^T \) - vector of parameters,
- \( f(x_i; \theta) \) is the GB2 density.

Log-likelihood equations:

\[ \ell'_m(\theta) = \frac{\partial \ell_m(\theta)}{\partial \theta} = 0 \Rightarrow \hat{\theta}_m. \]
Example: Swiss data 2007

Cumulative Distribution plot
- ML full
- NLS fit

Density plot
- Kernel
- GB2 density

Uses of parametric estimation: Examples of use
Sensitivity plot for ARPR

At risk of poverty rate

a = 2.7

a = 4.9

a = 7

a = 9.2

Uses of parametric estimation: Examples of use
Simulation study of direct indicators

Once parameters have been estimated, one can generate GB2 deviates that give a realistic view of the income distribution. We can:

- compare the distribution of direct estimators with the theoretical value;
- illustrate the clustering effect (same income for all household members);
- investigate the normality of the indicator’s estimates.
Relative errors (%) in 1000 simulations (s. size 13000)

GB2 parameters: $a = 4.964$, $p = 0.654$, $q = 0.79$

Uses of parametric estimation: Examples of use
Relative errors (%) in 1000 simulations (6000 HH, s.size 13000)

GB2 parameters: a = 4.964 p = 0.654 q = 0.79

Uses of parametric estimation: Examples of use
Distribution of empirical estimates, 6000 HH, s.size=13000

- **arpr**
- **rmob**
- **qsr**
- **gini**

Uses of parametric estimation: Examples of use
Variance of GB2 ML estimates

\[ \theta = (a, b, p, q) \quad \ell'_m(\theta) = 0: \text{pseudo-likelihood eq.} \]

- **Variance of the GB2 parameters**
  
  "Sandwich variance estimator" (Huber (1967), Freedman (2006) and Pfeffermann & Sverchkov (2003)):

\[
\hat{\text{Var}}(\hat{\theta}_m) \approx [\ell''_m(\hat{\theta}_m)]^{-1} \hat{V}_{sc}(\hat{\theta}_m)[\ell''_m(\hat{\theta}_m)]^{-1}
\]

\[ \hat{V}_{sc} - \text{design-based variance of the scores } \ell'_m(\hat{\theta}_m) \]

- **Variance of the derived indicators**

By linearization with respect to the GB2 parameters.
The simulation set-up

Data source: AMELIA synthetic universe constructed from EU-SILC (Trier University),

Designs: the most frequently used designs in EU-SILC,

Sample Size: 6,000 households / 14,000 individuals,

Number of samples: 1,000 per design.
Main survey designs

1. Simple random cluster sampling,
2. Stratified simple random cluster sampling,
3. Stratified cluster sampling with probabilities proportional to the household size,
4. Stratified two-stage simple random cluster sampling.
# Performance, design 2, 1000 AMELIA samples

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>adj</th>
<th>varT</th>
<th>cvsT</th>
<th>cvdT</th>
<th>relbiasT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARPR</td>
<td>no</td>
<td>6.86e-02</td>
<td>9.12e-03</td>
<td>9.65e-03</td>
<td>8.76e-02</td>
<td></td>
</tr>
<tr>
<td>ARPR</td>
<td>yes</td>
<td>7.63e-02</td>
<td>9.99e-03</td>
<td>9.20e-03</td>
<td>4.60e-02</td>
<td></td>
</tr>
<tr>
<td>eARPR</td>
<td>no</td>
<td>2.75e-01</td>
<td>1.99e-02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMPG</td>
<td>no</td>
<td>3.71e-01</td>
<td>1.31e-02</td>
<td>1.37e-02</td>
<td>6.85e-02</td>
<td></td>
</tr>
<tr>
<td>RMPG</td>
<td>yes</td>
<td>3.97e-01</td>
<td>1.43e-02</td>
<td>1.28e-02</td>
<td>8.68e-03</td>
<td></td>
</tr>
<tr>
<td>eRMPG</td>
<td>no</td>
<td>1.99e+00</td>
<td>3.23e-02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QSR</td>
<td>no</td>
<td>9.30e-02</td>
<td>2.92e-02</td>
<td>3.06e-02</td>
<td>8.94e-02</td>
<td></td>
</tr>
<tr>
<td>QSR</td>
<td>yes</td>
<td>7.06e-02</td>
<td>2.83e-02</td>
<td>2.65e-02</td>
<td>-2.14e-02</td>
<td></td>
</tr>
<tr>
<td>eQSR</td>
<td>no</td>
<td>8.26e-02</td>
<td>2.99e-02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gini</td>
<td>no</td>
<td>1e-05</td>
<td>1.00e-02</td>
<td>1.03e-02</td>
<td>2.27e-02</td>
<td></td>
</tr>
<tr>
<td>Gini</td>
<td>yes</td>
<td>1e-05</td>
<td>1.02e-02</td>
<td>1.02e-02</td>
<td>-1.97e-03</td>
<td></td>
</tr>
<tr>
<td>eGini</td>
<td>no</td>
<td>2e-05</td>
<td>1.04e-02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Uses of parametric estimation: Some results*
Discussion

- We have concentrated on the comparison between the direct indicator estimates and their parametric counterparts in the EU-SILC context.
- The variance of the parametrically estimated indicators is smaller than the direct estimates.
- Bias can occur, especially for RMPG.
- It won’t be difficult to add other indicators as well.
- Basic methods are at disposal in the R package GB2 (Graf and Nedyalkova (2010)).
- More information in Deliverable 2.1 and 7.