Task

- To estimate sampling error for Gini coefficient estimated from social sample surveys (EU-SILC)
- Estimation of sampling errors for totals and ratios was also analysed

Variance Estimation in EU-SILC Survey

Mārtiņš Liberts
Central Statistical Bureau of Latvia

8 November 2006 EU-SILC Methodological Workshop

Gini coefficient

- The Gini coefficient is a measure of inequality of a distribution, defined as the ratio of area between the Lorenz curve of the distribution and the curve of the uniform distribution, to the area under the uniform distribution
- It is often used to measure income inequality
- It is a number between 0 and 1, where 0 corresponds to perfect equality (e.g. everyone has the same income) and 1 corresponds to perfect inequality (e.g. one person has all the income, and everyone else has zero income)

Estimates of Gini Index

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>31</td>
<td>31</td>
<td>32</td>
<td>31</td>
<td>34</td>
<td>NA</td>
<td>34</td>
<td>35</td>
<td>36</td>
</tr>
</tbody>
</table>

8 November 2006 EU-SILC Methodological Workshop
EU-SILC

- Target population – population of Latvia living in private households
- Main variables of interest – income at household and individual level
- Yearly survey – organised once in a year
- Started in 2005 in Latvia

Sampling Design

- Two-stage sampling design
  - The first stage – stratified systematic pps sampling of census counting areas
  - The second stage – simple random sampling of households (addresses)
  - All individuals belonging to selected household are surveyed

First stage sampling

- The list of census counting areas was created for the last population census in 2000 (4279 areas)
- Census counting area is relatively small geographical area
- The size of area is defined by number of households in area
- Areas are stratified in four strata by urbanisation degree (Riga, 6 other cities, towns and rural areas)

Estimation of sampling errors

- Re-sampling methods are used
  - Dependent random group (DRG) method
  - Jackknife method
- Methods are used at the PSU level
- Both methods use the same resampling mechanisms – by dividing the whole sample in non-overlapping sub-samples
DRG and Jackknife
- From each sub-sample the parameter $\theta$ is estimated
- The estimate of variance $\nu_{\text{est}}(\theta) = \frac{1}{n(n-1)} \sum (\hat{\theta} - \bar{\theta})^2$

Resampling
- Resampling can be done in two ways:
  - Using the same sampling scheme
  - Using randomisation

Randomisation
- PSUs can be grouped in sub-samples in random order
- In this case variance estimate will differ each time the variance estimator is applied

Linearization
- In case of complex statistics the variance estimator becomes also more complex compared to variance estimator of total
- The linearization technique can be applied for complex statistics to get approximate variance estimate
- The goal of linearization is to find $z_i$ for each unit in the sample so that variance estimate of parameter $\hat{\theta}$ could be approximated by

$$
\psi(\hat{\theta}) = \sqrt{\sum_{i=1}^{n} \frac{z_i}{\pi_i}}
$$
Linearization

- In case of differentiable functions (for example, Ratio of two totals), the expansion of estimator in Taylor series can be applied to linearize the estimator.
- In case of non-differentiable functions, the expanded theory (Deville, 1999) can be used.

Linearization of Gini Index

- Gini Index can be linearized by

\[
IR(M) = \frac{2\sum x_i \left( \sum k \left( x_i - x_k \right) \right) + 2\sum x_i \left( \sum k_1 \left( x_i - x_{k_1} \right) \right) - \eta_1}{n \sum x_i}
\]

Program

- The program is written in SPSS mainly using the macro commands.
- It is possible to estimate sampling error for arbitrary two stage sampling design using DRG un Jackknife methods at PSU level.
- Sampling error is estimated for totals (SUM), ratios of two totals (RATIO) and Gini index (GINI).

Features of Program

1. The nonresponse correction at user defined groups
2. Poststratification of weights
3. Linearization of RATIO and GINI
4. Selection of number of subgroups
5. PSU ordering in random or user-defined order
6. PSU grouping
7. Estimation of parameters for sampling units or sublevel units
Procedures in program

- !linrat – linearization of RATIO
- !lingini – linearization of GINI
- !estim – estimator
- !weight – weighting
- !e_tion – estimation
- !proc – basic procedure
- !proc_u – main procedure

Parameters

- File – sample data file
- Strata – stratification variable
- Psu – PSU variables
- Diz_sv – variable for design weights
- Meth – method of estimation of sampling errors
- E_tor – estimators
- Lin – linearization
- Div – numbers of subgroups
- Other parameters

Example

!proc_u
dir="C:\DRG\SILC\files"
file="C:\DRG\SILC\SILC2005_data_var02.sav"
p_file="C:\DRG\SILC\dem_info.sav"
strata=prl / psu=atk iecirk / hh_id=bi030 /
per_sk=per_sk diz_sv=diz_sv
resp=resp resp_gr=atk iecirk / p_gr=prl /
p_var=per_sk p_tot=ieds_sk
method=jack / zorder=0 / repeat=1 psv_gr=sel_nr /
order=sel_nr /
div=2 4 8 12 16 / e_tor=sum ratio gini / lin=0 /
level=0 / eqscale per_sk / var=hh07n hs13n /
fast=1.

Variables

- SILC
  - Total housing cost
  - Lowest monthly income to make ends meet
Estimates of parameters

<table>
<thead>
<tr>
<th>SILC (estimates)</th>
<th>Housing</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>GINI</td>
<td>39.7</td>
<td>30.3</td>
</tr>
<tr>
<td>RATIO</td>
<td>0.117</td>
<td>8.566</td>
</tr>
</tbody>
</table>

| SUM              | 89 351 779 | 357 070 866 |

Results SILC (Gini)

<table>
<thead>
<tr>
<th>SILC (CV, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subgroups</td>
</tr>
<tr>
<td>Housing</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>32</td>
</tr>
</tbody>
</table>

Results SILC (Gini)

<table>
<thead>
<tr>
<th>SILC (mean CV, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subgroups</td>
</tr>
<tr>
<td>Housing</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>GINI</td>
</tr>
<tr>
<td>RATIO</td>
</tr>
<tr>
<td>TOTAL</td>
</tr>
</tbody>
</table>

Random ordering

- The stability (variance) of variance estimates was analysed
- Both methods works similarly and higher number of subgroups gives more stable results
### Linearization

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Variance1</th>
<th>Varianced</th>
<th>CV1</th>
<th>CV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATIO</td>
<td>DRG</td>
<td>1.160045</td>
<td>0.00019685</td>
<td>0.00018942</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>JACK</td>
<td>1.160046</td>
<td>0.00019486</td>
<td>0.00019628</td>
<td>1.20</td>
</tr>
<tr>
<td>GINI</td>
<td>DRG</td>
<td>35.90869</td>
<td>0.908</td>
<td>0.925</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td>JACK</td>
<td>35.90869</td>
<td>1.205</td>
<td>1.198</td>
<td>3.09</td>
</tr>
</tbody>
</table>

### Conclusions

- The DRG and JACK methods gives similar results
- The variance estimates are dependent on ordering of PSUs
- Higher number of subgroups gives more stable results
- Linearization can be used to simplify variance estimation