R Laboratory

Poverty and Social Exclusion indicators based on EU-SILC using R

Francesco Schirripa

francesco.schirripa@ec.unipi.it

November 08, 2018

Harmonized data sources

- In order to maximise the cross-country comparability of the common indicators, it was necessary in addition to defining their calculation algorithms, common harmonised data sources are also required for their computation.
- **The European Community Household Panel (ECHP)** was designed to complement the two main social surveys co-ordinated at EU level (the Labour Force Survey and the Household Budget Survey). It is a panel survey of 15 European countries that ran from 1994 to 2001, covering a wide range of topics such as income, health, education, housing, demographics and employment characteristics.
- From 2003 the ECHP has been replaced by the **EU statistics on income and living conditions (EU-SILC)**

EU-SILC

- EU-SILC is the main source for the compilation of statistics on income, social inclusion and living conditions.
- It provides two types of annual data for 28 European Union countries, Iceland, Norway, Switzerland and Turkey:
  - Cross-sectional data pertaining to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions
  - Longitudinal data pertaining to individual-level changes over time, observed periodically over a four year period.
- EU-SILC provides a harmonised lists of target primary (annual) and secondary (every four years or less frequently) variables to be transmitted to Eurostat.
- The reference population in EU-SILC includes all private households and their current members residing in the territory of the countries at the time of data collection. All household members are surveyed, but only those aged 16 and more are interviewed.

Sampling design in EU-SILC

- Various sampling strategies are in place in different countries. A specific sampling design is chosen according to the structure of the country and the population and taking into account budgetary constraints. The most used sampling design is stratified multistage sampling: applying one or more stratification criteria, mainly a geographical stratification. Although one characteristic of EU-SILC is flexibility in terms of sampling design, Eurostat recommends a rotational design with four sub-samples or replications.
- It is important to note that EU-SILC data are in practice conducted through complex sampling designs with different inclusion probabilities for the observations in the population, which results in different weights for the observations in the sample. Furthermore, calibration is typically performed for non-response adjustment of these initial design weights. Therefore, the sample weights have to be considered for all estimates, otherwise biased results are obtained.

Income in EU-SILC

- Income is the core of the EU-SILC. The Commission regulation on definitions is mainly focussed on the detailed definition of income.
- Four main aggregates are computed from EU-SILC:
  1. total disposable household income
  2. total disposable household income before transfers with old-age and survivors' benefits
  3. total disposable household income before transfers without old-age and survivors' benefits
  4. total gross income

Calculation of the equivalized disposable income

- Equivalised disposable income: in order to reflect differences in a household's size and composition, the total (net) household income is divided by the number of 'equivalent adults', using a standard (equivalence) scale. Usually we use the modified OECD scale; this scale gives a weight to all members of the household (and then adds these up to arrive at the equivalised household size):
  - 1.0 to the first adult;
  - 0.5 to the second and each subsequent person aged 14 and over;
  - 0.3 to each child aged under 14.

  the equivalised disposable income and is attributed equally to each member of the household.
- The total disposable income of a household is calculated by adding together the personal income received by all of the household members plus the income received at the household level.
- In practice, the equivalized disposable income needs to be computed from the income components included in EU-SILC for the estimation of the indicators on social exclusion and poverty.

Laeken Indicators

The Laeken European Council in December 2001 endorsed a first set of 18 common statistical indicators for poverty and social inclusion, which will allow monitoring in a comparable way of Member States' progress towards the agreed EU objectives.

Primary indicators

The 11 most important indicators on social exclusion and poverty according to Eurostat (2004) are:

1. At-risk-of-poverty rate (or Head Count Rario - HCR) broken down by various charcateristics, such as: age and gender; most frequent activity status and gender; household type; accommodation tenure status; gender amon workers; work intensity.
2. Inequality of income distribution: S80/S20 income quintile share ratio
3. At-persistent-risk-of-poverty rate by age and gender (60% median)
4. Relative median at-risk-of-poverty gap (Poverty Gap - PG), by age and gender

Secondary indicators

5. Dispersion around the at-risk-of-poverty threshold
6. At-risk-of-poverty rate anchored at a moment in time
7. At-risk-of-poverty rate before social transfers by age and gender
8. Inequality of income distribution: Gini coefficient
9. At-persistent-risk-of-poverty rate, by age and gender (50% median)

   Other indicators
10. Mean equivalized disposable income
11. The gender pay gap

Gini coefficient

The Gini coefficient is a measure of inequality of a distribution. It is defined as a ratio with values between 0 and 1.
For our aim, the Gini coefficient is used to measure income inequality: 0 corresponds to perfect income equality (i.e. everyone has the same income) and 1 corresponds to perfect income inequality (i.e. one person has all the income, while everyone else has zero income).

Variance estimation: Naive bootstrap

Let θ denote a certain indicator of interest and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ denote a survey sample with n observations. The naive bootstrap algorithm for estimating the variance and confidence interval of an indicator can be summarized as follows:

1. Draw R independent bootstrap samples $\mathbf{X}_1^\star, ..., \mathbf{X}_R^\star$ from $\mathbf{X}$
2. Compute the bootstrap replicate estimates $\hat{\theta}_r^\star = \hat{\theta}(\mathbf{X}_r^\star)$ for each bootstrap sample $\mathbf{X}_r^\star$
3. Estimate the variance $V(\hat{\theta})$ by the variance of the R bootstrap replicate estimates:

$$V(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\theta}_r^\star - \frac{1}{R} \sum_{s=1}^{R} \hat{\theta}_s^\star \right)$$