R Laboratory
Estimating Poverty Indicators with R

Francesco Schirripa

francesco.schirripa@ec.unipi.it

October 24, 2019

R package laeken: Introduction

- The R package laeken is a very useful toolkit for the estimation of the most widely used social exclusion and poverty indicators.
- It allows to estimate these indicators from complex surveys (i.e. sampling units are drawn from finite populations, most often with unequal inclusion probabilities).
- A general calibrated bootstrap method to estimate the variance of indicators is included.
- The package contains simulated data based on the European Union Statistics on Income and Living Conditions (EU-SILC) and the Structure of Earnings Survey (SES).

Harmonized data sources

- In order to maximise the cross-country comparability of the common indicators, it is necessary to have:
  - common calculation algorithms
  - common harmonised data sources for their computation.

Harmonized data sources (cont.)

- **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- From 2003 the ECHP has been replaced by the **European Union Statistics on Income and Living Conditions (EU-SILC)**.

EU-SILC

- EU-SILC is a yearly data collection survey conducted by Eurostat.
- It provides annual data for 28 European Union countries, Iceland, Norway, Switzerland and Turkey.
- The main goal of the EU-SILC is to provide comparable data on income, poverty, social exclusion and living conditions for monitoring the poverty and social inclusion in the EU. Moreover, it covers other topic like house conditions, leisure and cultural activities ...
- Eurostat defines a set of target variables and defines a number of quality criteria in regards to data collection and the NSIs are responsible for the data collection efforts in their country (Output harmonization)

EU-SILC

- The EU-SILC includes both a longitudinal and cross- sectional component
  - Cross-sectional data: collected for one year
  - Longitudinal data pertaining to individual-level changes over time, observed periodically over a four year period.

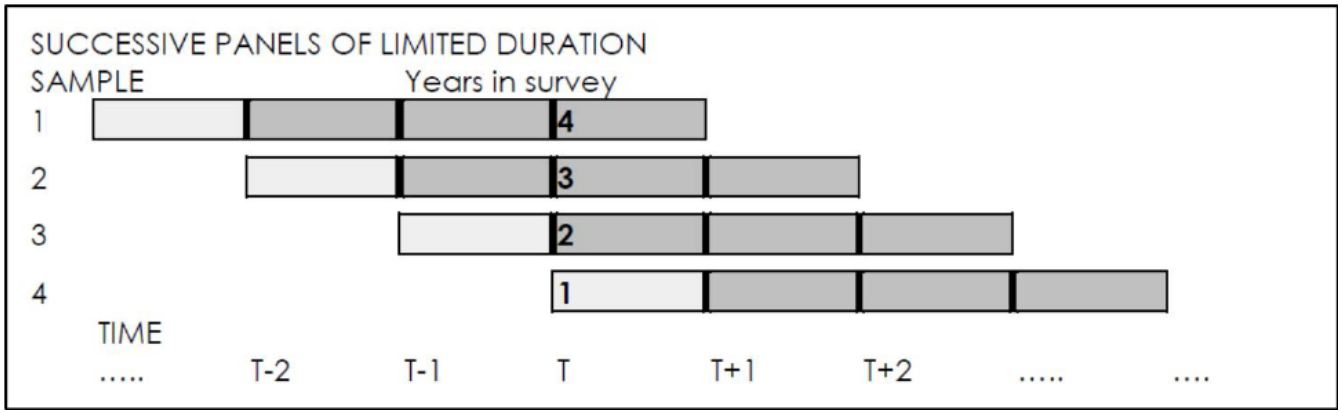

Figure: EU-SILC Rotating Panel

EU-SILC

- EU-SILC provides a harmonised lists of target primary (annual) and secondary (every four years or less frequently) variables to be transmitted to Eurostat. For example EU-SILC 2015 contains the module on social and cultural participation and material deprivation; EU-SILC 2017 contains the module on Health and children's health...
- 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. The most used sampling design is stratified multistage sampling: applying one or more stratification criteria, mainly a geographical stratification.

- 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.

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.

Laeken Indicators (cont.)

- At-risk-of-poverty threshold
- At-risk-of-poverty rate (or Head Count Rario - HCR)
- Dispersion around the at-risk-of-poverty threshold
- Inequality of income distribution: S80/S20 income quintile share ratio
- Relative median at-risk-of-poverty gap (Poverty Gap - PG)
- Inequality of income distribution: Gini coefficient

These indicators may be discriminated by various criteria (gender, age group, region, etc.).

Equivalised income

The indicators described in this lesson are estimated from EU-SILC data based on household income rather than personal income.

Equivalised disposable income: total household disposable income divided by equivalized household size.

The equivalized household size is defined according to the modified OECD scale, which gives a weight of 1.0 to the first adult, 0.5 to other household members aged 14 or over, and 0.3 to household members aged less than 14.

Each person in the same household receives the same equivalized disposable income.

At-risk-of-poverty threshold

At-risk-of-poverty threshold (ARPT): Establishing a poverty line, i.e. sets a threshold (a minimum acceptable standard of consumption or income) below which an individual or a household is considered poor.

The (National) ARPT is usually set at 60% of the national median equivalized disposable income.

Sometimes could be useful to use relative poverty lines (i.e regional poverty lines)

At-risk-of-poverty rate (or Head Count Rario - HCR)

- **Head Count Ratio (HCR)**
  The At-risk-of-poverty rate (Head Count Rario - HCR) is the share of people with an equivalised disposable income below the at-risk-of-poverty line, which is set at 60% of the national median equivalised disposable income.

$$\text{HCR} = \frac{1}{N}\sum_{i=1}^{N} I(y_i < z),$$

where y is the income; z is the poverty line; and $I(.)$ is an indicator function that is 1 if its argument is true, 0 otherwise (it is equal 1 when the income of the ith individual below the poverty line)

$$\text{HCR} = \frac{\sum_{i \in I(y_i < z)} w_i}{\sum_{i=1}^{n} w_i},$$

Dispersion around the at-risk-of-poverty threshold

This indicator is defined as the percentage of persons with an equivalised disposable income below respectively 40%, 50%, 60% and 70% of the national median equivalised disposable income.

Poverty Gap

- **Poverty Gap (PG)**
  The PG (also known as 'relative median at-risk-of-poverty gap') is calculated as the difference between the median equivalised total income of persons below the at-risk-of-poverty threshold and the at-risk-of-poverty threshold itself, expressed as a percentage of the at-risk-of-poverty threshold.

$$\mathrm{PG} = \frac{P.L. - \hat{q}_{0.5}(y_i, w_i)_{i \in I(y_i < z)}}{P.L.},$$

  It measures the extent to which individuals fall below the poverty line (the poverty gaps) as a proportion of the poverty line (in other words it measures the poverty intensity, since it takes into account the distance from the P.L.)

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.

- 0 corresponds to perfect income equality (i.e. everyone has the same income)
- 1 corresponds to perfect income inequality (i.e. one person has all the income, while everyone else has zero income)

It is defined as the relationship of cumulative shares of the population arranged according to the level of equivalized disposable income, to the cumulative share of the equivalized total disposable income received by them.

$$\text{Gini} = \frac{2\sum_{i=1}^{n}\left(w_i x_i \sum_{j=1}^{n} w_j\right) - \sum_{i=1}^{n} w_i^2 x_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i}$$

Quintile share ratio

The income quintile share ratio (QSR) is defined as the ratio of the sum of the equivalized disposable income received by the 20% of the population with the highest equivalized disposable income to that received by the 20% of the population with the lowest equivalized disposable income.

$$QSR = \frac{\sum_{i \in I > \hat{q}_{0.80}} w_i x_i}{\sum_{i \in I \leq \hat{q}_{0.20}} w_i x_i}$$

where $\hat{q}_{0.20}$ and $\hat{q}_{0.80}$ denote the weighted 20% and 80% quantiles, respectively.

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:

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

Variance estimation: Calibrated bootstrap

The calibrated version of the bootstrap thus results in more precise variance and confidence interval estimation, but comes with higher computational costs than the naive approach. In any case, the calibrated bootstrap algorithm is obtained by adding the following step between Steps 1 and 2 of the naive bootstrap algorithm

1bis. Calibrate the sample weights for each bootstrap sample $\mathbf{X}_1^\star, ..., \mathbf{X}_R^\star$. Generalized raking procedures are thereby used for calibration: either a multiplicative method known as raking, an additive method or a logit method.