At-risk-of-poverty threshold variance estimations using Gaussian kernel and smoothing splines in R package vardpoor

Juris Breidaks

August 23, 2018

Table of contents

Introduction

- The Central Statistical Bureau of Latvia (CSB) in 2012 developed R (R Core Team, 2018) package *vardpoor* (Breidaks, Liberts, Ivanova, 2017).
- The package *vardpoor* was developed with the objective to modernize the sample error estimation in sample surveys.

Before package vardpoor

Sampling errors were estimated using the chargeable software SUDAAN. Use of SUDAAN had several shortcomings:

- Only obsolete SUDAAN version was available at CSB, which had to be updated;
- Updating of SUDAAN version would require financial resources;
- It is difficult to integrate SUDAAN to work with other data processing programs (IBM SPSS Statistics or R).

Before package vardpoor

- With the help of SUDAAN it was possible to linearize only non-linear statistics, as the ratio of two totals, but in the EU-SILC survey there were several other nonlinear statistics, which had to be linearized separately;
- SUDAAN sampling error estimation did not include the effect of weight calibration.

Package vardpoor

- Given the above shortcomings, it was decided to develop the vardpoor and to designed it as an R extension (package).
- R is an open-source free statistical calculation environment;
- R is currently the most popular computing environment among statisticians;
- R environment is very convenient and suitable for development of such solutions.

Theoretical basis of vardpoor

- Guillaume Osier and Emilio Di Meglio (2012). The linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from the second wave onwards?
- Guillaume Osier (2009). Variance estimation for complex indicators of poverty and inequality. Journal of the European Survey Research Association, Vol.3, No.3, pp. 167-195, ISSN 1864-3361, URL http://ojs.ub.uni-konstanz.de/srm/article/view/369.

Sampling error estimation mechanism

- Calculation of the domain-specific study variables, if the sampling error is to be estimated for population domains;
- At-risk-of-poverty threshold linearization using smoothing splines (Asmuss, S., Breidaks, J., Budkina,, 2016) and Gaussian kernel (Osier, G., 2009);
- Calculation of regression residual if the weights are calibrated;
- Variance estimation with the ultimate cluster method (Hansen, Hurwitz and Madow, 1953);
- Variance estimation for the simple random sampling design.

Calculation of the domain-specific study variables

Often separate estimates for subpopulations are needed. Subpopulations are called domains. The domains concerned are denoted as $(U_1, ..., U_d, ..., U_D)$. It is assumed that y total value in each domain must be estimated. The aim is to estimate $(Y_1, ..., Y_d, ..., Y_D)$, where

$$Y_d = \sum_{k \in U_d} y_k, d = 1, ..., D$$

Calculation of the domain-specific study variables

The domain total can be expressed with a new variable y_{dk}, constructed from y specifically for domain U_d (Särndal, Swensson, 1992). The new variable is denoted with y_{dk} and its values for each element k are defined as

$$y_{dk} = \begin{cases} y_k, & \text{if } k \in U_d, \\ 0, & \text{if } k \notin U_d. \end{cases}$$

Then Y_d can be expressed as a total from the new variable y_{dk} for the whole population:

$$Y_d = \sum_{k \in U} y_{dk}$$

Linearization approach

The linearisation method (Särndal, Swensson and Wretman, 1992; Deville, 1999; Wolter, 2007; Osier, 2009) uses Taylor-like series approximation to reduce non-linear statistics to a linear form, justified by asymptotic properties of the estimator (Verma and Betti, 2005).
The method based on influence functions (Deville, 1999) is general enough to handle all the complex non-linear indicators of poverty and inequality based on EU-SILC such as the at-risk-of-poverty rate. The estimated variance of the estimator $\widehat{\theta}$ can be approximated by a linear function of the sample observations:

$$\widehat{Var}(\widehat{Y}) \cong \widehat{Var}\Big(\sum_{k \in s} w_k \cdot \widehat{u}_k\Big),$$

Linearization approach

where the value of the estimated linearized variable $\widehat{u}_k$ is determined by calculating the following functional derivative:

$$\widehat{u}_k = \lim_{t \to 0} \frac{T\left(\widehat{M} + t\delta_k\right) - T\left(\widehat{M}\right)}{t},$$

where the estimated population parameter $\widehat{\theta}$ is expressed T as a functional of the measure $\widehat{M}$, i.e.,

$$\widehat{\theta} = T\left(\widehat{M}\right),$$

and the measure $\widehat{M}$ allocates the sample weight w_k to each unit k in the sample s:

$$\widehat{M}(k) = \widehat{M}_k = w_k, k \in s,$$

Weighted quantile estimation in the domain

Let n_D be the number of observations in the domain D of the sample, let $x_D := (x_1, ..., x_{n_D}^{'})$, denote the equalized disposable income with $x_1 \leq ... \leq x_{n_D}$, and let $w_D := (w_1, ..., w_{n_D}^{'})$ be the corresponding personal sample weights. Weighted quantiles for the estimation of the population values in the domain D according are then given (Alfons, Templ, 2014) by

$$\widehat{Q}_{D;p} := \begin{cases} \frac{1}{2}(x_j + x_{j+1}), & \text{if } \sum_{i=1}^{j} w_i = p \sum_{i=1}^{n_D} w_i, \\ x_{j+1}, & \text{if } \sum_{i=1}^{j} w_i < p \sum_{i=1}^{n_D} w_i < \sum_{i=1}^{j+1} w_i. \end{cases}$$

Calculation of the at-risk-of-poverty threshold in domain and its linearization

The at-risk-of-poverty threshold (ARPT) in the domain D is defined as 60% of the median income in the domain D:

$$ARPT_D = 0.6 \cdot F_D^{-1}(0.5)$$

$$\widehat{ARPT}_D = 0.6 \cdot \widehat{Q}_{D;p}^{-1}(0.5)$$

Calculation of the at-risk-of-poverty threshold in domain and its linearization

The linearized variable of the ARPT in the domain D is defined by Osier (2009):

$$\widehat{u}_{D;k}^{ARPT} = I(ARPT_D)_k = 0.6 \cdot I(\widehat{Q}_{D;0.5})_k = \\ = \frac{-0.6}{f(\widehat{Q}_{D;0.5})} \cdot \frac{1_{[k \in D]}}{\widehat{N}_D} \left[1_{[y_k \leq \widehat{Q}_{D;0.5}]} - 0.5\right],$$

where y_k is k-th equalized disposable income, $\widehat{N}_D$ is estimated size of the population in the domain D.

Calculation of the at-risk-of-poverty threshold in domain linearization using Gaussian kernel

Deville (1999) and Osier (2009) suggest using Gaussian kernel estimation for the calculation of the density function. The density functions can be estimated on the basis of the Gaussian kernel function as follows (Preston, 1995)

$$f_D(x) = \frac{1}{\widehat{N}_D \widehat{h}_D} \sum_{i \in D} w_i K\left(\frac{x - y_i}{h_D}\right) \tag{1}$$

where

$$K(o) = \frac{1}{h_D \sqrt{2\pi}} e^{-\frac{o^2}{2}} \tag{2}$$

is the Gaussian kernel.

Calculation of the at-risk-of-poverty threshold in domain linearization using Gaussian kernel

$\widehat{N}_D = \sum_{i \in D} w_i$ is the Horvitz and Thompson (1952) estimator of the population size in domain D; h_D is the bandwidth parameter in the domain D. For normally distributed population densities, the following bandwidth parameter was recommended by Silverman (1986)

$$\widehat{h}_D = \widehat{\sigma}_D \widehat{N}_D^{-0.2} \tag{3}$$

$\widehat{\sigma}_D$ is the estimated standard deviation of the empirical income distribution:

$$\widehat{\sigma}_D = \frac{1}{\widehat{N}_D} \sqrt{\widehat{N}_D \sum_{i \in s_D} w_k y_k^2 - \Big(\sum_{i \in s_D} w_k y_k \Big)^2}. \tag{4}$$

Calculation of the at-risk-of-poverty threshold in domain linearization using smoothing splines function

The density functions can be estimated on the basis of the smoothing splines function as follows

$$f_D(x) = \frac{1}{\widehat{N}_D \widehat{h}_D} \sum_{i \in D} w_i s\left(\frac{x - y_i}{h_D}\right) \tag{5}$$

where $s(x)$ is the smoothing spline.

Calculation of the at-risk-of-poverty threshold in domain linearization using smoothing splines function

$\widehat{N}_D = \sum_{i \in D} w_i$ is the Horvitz and Thompson (1952) estimator of the population size in domain D; h_D is the bandwidth parameter in the domain D. For normally distributed population densities, the following bandwidth parameter was recommended by Silverman (1986)

$$\widehat{h}_D = \widehat{\sigma}_D \widehat{N}_D^{-0.2} \tag{6}$$

$\widehat{\sigma}_D$ is the estimated standard deviation of the empirical income distribution:

$$\widehat{\sigma}_D = \frac{1}{\widehat{N}_D} \sqrt{\widehat{N}_D \sum_{i \in s_D} w_k y_k^2 - \left(\sum_{i \in s_D} w_k y_k \right)^2}. \tag{7}$$

Calculation of the at-risk-of-poverty threshold in domain and its linearization using smoothing splines

Smoothing spline s is solution for the following problem of histopolation in the Sobolev space $W_2^q[a, b]$.

$$\int_a^b (g^{(q)}(t))^2 \, dt \longrightarrow \min_{g \in W_2^q[a,b],}$$

$$\int_{t_{i-1}}^{t_i} g(t)dt = f_i h_i, \quad i = 1, \ldots, n.$$

Calculation of the at-risk-of-poverty threshold in domain and its linearization

$$s(t) = \sum_{j=0}^{r-1} \varrho_j t^j + \frac{(-1)^{r+1}}{(2r)!} \sum_{i=1}^{n} \alpha_i ((t - t_i)_+^{2r} - (t - t_{i-1})_+^{2r}) \quad (8)$$

with the following conditions on the coefficients:

$$\sum_{i=1}^{n} \frac{\alpha_i}{j+1} (t_i^{p+1} - t_{i-1}^{p+1}) = 0, \quad p = 0, 1, \ldots, r-1. \quad (9)$$

Regression residual calculation

If the weights are calibrated, then calibration residual estimates $\widehat{e}_k$ are calculated (Lundstrom, Sarndal, 2001) by formula

$$\widehat{e}_k = y_k - x_k^{'}\widehat{B},$$

where

$$\widehat{B} = \Big(\sum_{k \in s} d_k q_k x_k x_k^{'}\Big)^{-1} \Big(\sum_{k \in s} d_k q_k x_k y_k\Big)$$

Variance estimation with the ultimate cluster method

If we assume that $n_h \geq 2$ for all h, that is, two or several primary sampling units (PSUs) are sampled from each stratum, then variance of $\widehat{\theta}$ can be estimated from the variation among the estimated PSU totals of y (Hansen, Hurwitz, Madow, 1953; Osier, Meglio, 2012; Berger, Goedeme, Osier, 2013):

$$\widehat{V}(\widehat{\theta}) = \sum_{h=1}^{H}(1 - f_h)\frac{n_h}{n_h - 1}\sum_{k=1}^{n_h}(y_{hk*} - \bar{y}_{h**})^2$$

Variance estimation with the ultimate cluster method

where

- $y_{hk*} = \sum_{j=1}^{m_{hk}} w_{hkj} y_{hkj}$
- $y_{h**} = \frac{\sum_{k=1}^{n_h} y_{hk*}}{n_h}$
- f_h is a sampling fraction of PSUs for stratum h,
- h is the stratum number, with a total of H strata,
- k is the number of PSU within the sample of stratum h, with a total of n_h PSUs,
- j is the household number within PSU k of stratum h, with a total of m_{hi} households,
- w_{hkj} is the sampling weight for household j in PSU k of stratum h,
- y_{hkj} denotes the observed value of study variable y for household j in PSU k of stratum h.

Function varpoord description

Function varpoord is used to estimate sampling errors for indicators on social exclusion and poverty. Data is given at the person level, but information for the calibration is given at the household level. At the beginning of the function execution a range of tests is performed in order to test if there are any mistakes in data. Function varpoord consist argument type, if it is chosen *linarpT*, then calculate the at-risk-of-poverty threshold (ARPT) in the domain

- linearized values in the domain D using Gaussian kernel (Osier (2009))
- linearized values in the domain D using smoothing splines (Asmuss, Breidaks (2016))

Function varpoord description

If calibration matrix X and g weights are used at household level, function calculates the residuals at the household level. Function varpoord outputs several results:

- point estimates for statistics,
- variance estimates,
- relative standard error,
- absolute margin of error,
- relative margin of error,
- lower and upper bound of the confidence interval,
- variance of HT estimator under current design,
- variance of calibrated estimator under SRS,
- the sample design effect, the estimated design effect of estimator,
- the overall design effect of sample design and estimator,
- the effective sample size.

varpoord function testing results

Function was tested on simulated Austria data of EU-SILC. In this function will test ARPT quality indicator using smoothing splines, the function varpoord() is used:

```
smooth_cal <- varpoord(inc = "INC_ekv20",
w_final = "db090", income_thres = "INC_ekv20",
wght_thres = "db090", ID_household = "db030n",
H = "db050", PSU = "db060", sort = NULL,
dataset = dataset2, type = c("linarpt"),
method = "smooth_splines", r = 2, ro = 0.01)
```

varpoord function testing results

In this function will test ARPT quality indicator using Gaussian kernel (Osier (2009), the function varpoord() is use

```
gausian_cal <- varpoord(inc = "INC_ekv20",
w_final = "db090", income_thres = "INC_ekv20",
wght_thres = "db090", ID_household = "db030n",
H = "db050", PSU = "db060", sort = NULL,
dataset = dataset2, type = c("linarpt"),
method = "Gaussian")
```

varpoord function testing results

Method	Estimation	Standart error	CV
Gaussian kernal	1876.67	50.59	2.69
Smoothing spline r=2 $\rho = 0.01$	1876.67	70.18	3.74

Thank you!