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

Juris Breidaks[1]

[1]Central Statistical Bureau of Latvia, e-mail: juris.breidaks@csb.gov.lv

Abstract

The Central Statistical Bureau of Latvia (CSB) in 2012 developed R (R Core Team (2018)) package vardpoor (Breidaks *et al.* (2018)) (a set of functions for statistical calculation in programme R). The package vardpoor was developed with the objective to modernise the sample error estimation in sample surveys. Before the package was developed, sampling errors were estimated using the chargeable programme SUDAAN (www.rti.org/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 into work with other data processing programmes (IBM SPSS Statistics or R);
- 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 non-linear statistics, which had to be linearized separately;
- SUDAAN sampling error estimation did not include the effect of weight calibration.

Given the above shortcomings, it was decided to develop the vardpoor package, which would be designed as R extension. First of all, R is an open-source free statistical calculation environment; secondly, R is currently the most popular computing environment among statisticians; and thirdly R environment is very convenient and suitable for development of such solutions. It should also be mentioned that, upon developing vardpoor package as R extension, it was easily integrated in the statistical production processes.

The theoretical basis of vardpoor was borrowed from G. Osier article The Linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from the second wave onwards? (Osier & Di Meglio (2012)), which was presented at the workshop devoted to the evaluation of the standard errors and other issues related to the EU-SILC survey in March 2012.
Keywords: BNU2018, vardpoor, risk of poverty threshold, smoothing splines

1 Introduction

The Central Statistical Bureau of Latvia (CSB) in 2012 developed R (R Core Team (2018)) package vardpoor (Breidaks *et al.* (2018)) (a set of functions for statistical calculation in

programme R). The package vardpoor was developed with the objective to modernise the sample error estimation in sample surveys. Before the package was developed, sampling errors were estimated using the chargeable programme SUDAAN (www.rti.org/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 into work with other data processing programmes (IBM SPSS Statistics or R);
- 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 non-linear statistics, which had to be linearized separately;
- SUDAAN sampling error estimation did not include the effect of weight calibration.

Given the above shortcomings, it was decided to develop the vardpoor package, which would be designed as R extension. First of all, R is an open-source free statistical calculation environment; secondly, R is currently the most popular computing environment among statisticians; and thirdly R environment is very convenient and suitable for development of such solutions. It should also be mentioned that, upon developing vardpoor package as R extension, it was easily integrated in the statistical production processes.

The theoretical basis of vardpoor was borrowed from G. Osier article "The Linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from the second wave onwards?" (Osier & Di Meglio (2012)), which was presented at the workshop devoted to the evaluation of the standard errors and other issues related to the EU-SILC survey in March 2012.

2 Sampling error estimation mechanism

Sampling error estimation mechanism consists of a sequence of procedures:

1. Calculation of the domain-specific study variables, if the sampling error is to be estimated for population domains;
2. At-risk-of-poverty threshold linearization using Gaussian kernel (Osier (2009) and smoothing splines (Asmuss *et al.* (2016));
3. Calculation of regression residual if the weights are calibrated;
4. Variance estimation with the ultimate cluster method (Hansen *et al.* (1953));
5. Variance estimation for the simple random sampling design.

2.1 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

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

The domain total can be expressed with a new variable y_{dk}, constructed from y specifically for domain U_d (Lundström & Särndal (2001)). The new variable is denoted with y_{dk} and its values for each element k are defined as

$$y_{dk} = \begin{cases} y_k, & \text{if } k \in U_d, \\ 0, & \text{if } k \notin U_d. \end{cases} \tag{2}$$

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

$$Y_d = \sum_{k \in U} y_{dk} \tag{3}$$

2.2 Linearization approach

The linearisation method (Särndal *et al.* (1992), Deville (1999), Osier (2009)) uses Taylor-like series approximation to reduce non-linear statistics to a linear form, justified by asymptotic properties of the estimator (Verma & 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 threshold. The estimated variance of the estimator $\widehat{\theta}$ can be approximated by a linear function of the sample observations:

$$\widehat{Var}(\widehat{Y}) \cong \widehat{Var}\Big(\sum_{k \in s} w_k \cdot \widehat{u}_k\Big), \tag{4}$$

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

$$\widehat{u}_k = \lim_{t \to 0} \frac{T\big(\widehat{M} + t\delta_k\big) - T\big(\widehat{M}\big)}{t}, \tag{5}$$

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

$$\widehat{\theta} = T\big(\widehat{M}\big), \tag{6}$$

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

$$\widehat{M}\big(k\big) = \widehat{M}_k = w_k, k \in s, \tag{7}$$

δ_k is the Dirac measure at k: for each unit k in the sample, $\delta_k(i) = 1$ if and only if $k = i$. The functional derivative (18) is called the influence function.

2.3 Weighted quantile estimation in the domain

Quantiles are defined as $Q_{D,p}^{-1} = F_D^{-1}(p)$, where F_D is the income distribution function on the population in the domain D, i.e.,

$$F_{D,y}(x) = \frac{1}{N_D} \sum_{k \in U_D} 1_{[y_k \leq x]} \tag{8}$$

and $0 \leq x \leq 1$. The median is given by $p = 0.5$. For the following definitions, 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 (M. (2013)) by

$$\widehat{Q}_{D;p} = \widehat{Q}_{D;p}(x_D, w_D) := \begin{cases} \frac{1}{2}(x_j + x_{j+1}), & \text{if } \sum_{i=1}^{j} w_i = p \sum_{i=1}^{n_D} w_i, \\ x_{j+1}, & \text{if } \sum_{i=1}^{j} w_i < p \sum_{i=1}^{n_D} w_i < \sum_{i=1}^{j+1} w_i. \end{cases} \tag{9}$$

2.4 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:

$$ARPT_D = 0.6 \cdot F_D^{-1}(0.5) \tag{10}$$

$$ARPT_D = 0.6 \cdot \widehat{Q}_{D;p}^{-1}(0.5) \tag{11}$$

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

$$\widehat{u}_{D;k}^{ARPT} = I(ARPT_D)_k = 0.6 \cdot I(\widehat{Q}_{D;0.5})_k = \frac{-0.6}{f(\widehat{Q}_{D;0.5})} \cdot \frac{1_{[k \in D]}}{\widehat{N}_D} \big[1_{[y_k \leq \widehat{Q}_{D;0.5}]} - 0.5\big], \tag{12}$$

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

$f(.)$ is estimator of the density function which in the next subsections will be descripted using smoothing splines estimation and Gausian kernel estimation.

2.4.1 Calculation of the density function using Gaussian kernel estimation

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))

$$f_D(x) = \frac{1}{\widehat{N}_D \widehat{h}_D} \sum_{i \in D} w_i K\Big(\frac{x - y_i}{h_D}\Big) \tag{13}$$

where

$$K(o) = \frac{1}{h_D \sqrt{2\pi}} e^{-\frac{o^2}{2}} \tag{14}$$

is the Gaussian kernel. $\widehat{N}_D = \sum_{i \in D} w_i$ is the Horvitz and Thompson (Horvitz & 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 (Silverman (1986))

$$\widehat{h}_D = \widehat{\sigma}_D \widehat{N}_D^{-0.2} \tag{15}$$

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

$$\widehat{\sigma}_D = \frac{1}{\widehat{N}_D} \sqrt{\widehat{N}_D \sum_{i \in s_D} w_k y_k^2 - \Big(\sum_{i \in s_D} w_k y_k\Big)^2}. \tag{16}$$

2.4.2 Calculation of the density function using smoothing splines function estimation

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

$$f_D(x) = \frac{1}{\widehat{N}_D \widehat{h}_D} \sum_{i \in D} w_i s\Big(\frac{x - y_i}{h_{Di}}\Big) \tag{17}$$

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

$$\widehat{h}_D = \widehat{\sigma}_D \widehat{N}_D^{-0.2} \tag{18}$$

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

$$\widehat{\sigma}_D = \frac{1}{\widehat{N}_D} \sqrt{\widehat{N}_D \sum_{i \in s_D} w_k y_k^2 - \Big(\sum_{i \in s_D} w_k y_k \Big)^2}. \tag{19}$$

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

$$\int_a^b (g^{(q)}(t))^2 \, dt \longrightarrow \min_{\substack{g \in W_2^q[a,b], \\ \int_{t_{i-1}}^{t_i} g(t)dt = f_i h_i, \;\; i = 1, \ldots, n.}}$$

A solution of the spline s is in the form

$$s(t) = \sum_{j=0}^{r-1} \varrho_j t^j + \frac{(-1)^{r+1}}{(2r)!} \sum_{i=1}^{n} \alpha_i ((t - t_i)_+^{2r} - (t - t_{i-1})_+^{2r}) \tag{20}$$

with the following conditions on the coefficients:

$$\sum_{i=1}^{n} \frac{\alpha_i}{j+1} (t_i^{p+1} - t_{i-1}^{p+1}) = 0, \quad p = 0, 1, \ldots, r-1. \tag{21}$$

2.5 Regression residual calculation

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

$$\widehat{e}_k = y_k - x_k^{'} \widehat{B}, \tag{22}$$

where

$$\widehat{B} = \Big(\sum_{k \in s} d_k q_k x_k x_k^{'} \Big)^{-1} \Big(\sum_{k \in s} d_k q_k x_k y_k \Big) \tag{23}$$

2.6 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 *et al.* (1953); Osier & Di Meglio (2012); Di Meglio *et al.* (2013)):

$$\widehat{V}(\widehat{\theta}) = \sum_{h=1}^{H}(1-f_h)\frac{n_h}{n_h-1}\sum_{k=1}^{n_h}\left(y_{hk*}-\bar{y}_{h**}\right)^2 \tag{24}$$

where

- $y_{hk*} = \sum_{j=1}^{m_{hk}} w_{hkj}y_{hkj}$
- $y_{h**} = \dfrac{\sum_{k=1}^{n_h} y_{hk*}}{n_h}$
- 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.

2.7 The design effect estimation and effective sample size

The design effect of sampling is estimated by

$$\widehat{Deff}_{sam}(\widehat{\theta}) = \frac{\widehat{Var}_{CUR,HT}(\widehat{\theta})}{\widehat{Var}_{SRS,HT}(\widehat{\theta})} \tag{25}$$

where $\widehat{Var}_{SRS,HT}(\widehat{\theta})$ is the variance of HT estimator under SRS, $\widehat{Var}_{SRS,HT}(\widehat{\theta})$ is the variance of HT estimator under current sampling design.

The design effect of estimator is estimated by

$$\widehat{eff}_{est}(\widehat{\theta}) = \frac{\widehat{Var}_{CUR,CAL}(\widehat{\theta})}{\widehat{Var}_{CUR,HT}(\widehat{\theta})} \tag{26}$$

where $\widehat{Var}_{CUR,CAL}(\widehat{\theta})$ is the variance of calibrated estimator under current sampling design.

The overall design effect of sampling and estimator is estimated by

$$\widehat{Deff}(\widehat{\theta}) = \widehat{Deff}_{sam}(\widehat{\theta}) \cdot \widehat{eff}_{est}(\widehat{\theta}) \tag{27}$$

The effective sample size is estimated by

$$\widehat{n}_{eff}(\widehat{\theta}) = \frac{n}{\widehat{Deff}_{sam}(\widehat{\theta})}, \tag{28}$$

where n is the sample size or the number of respondents (in case of non-response).

3 R package vardpoor

3.1 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 and linearized values in the domain D using Gaussian kernel (Osier (2009) and smoothing splines (Asmuss *et al.* (2016))

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.

3.2 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 (Asmuss *et al.* (2016)), 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)
```

Table 1: ARPT quality in 2012

method	estim	se	cv
Gaussian kernal	1876.67	50.59	2.69
Smoothing spline r=2 $\rho = 0.01$	1876.67	70.18	3.74

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

```
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")
```

In table was shown has calculated standard errors, coefficient of the variance.

References

Asmuss, S., Breidaks, J. & Budkina, N. (2016). On approximation of density function by shape preserving smoothing histospline. *Proceedings of the 15th Conference on Applied Mathematics (APLIMAT 2016)* , 30 – 43.

Breidaks, J., Liberts, M. & Ivanova, S. (2018). vardpoor: Variance estimation for sample surveys by the ultimate cluster , 1 – 27.

Deville, J. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* **25(2)**, 193 – 203.

Di Meglio, E., Osier, G., Goedemé, T., Berger, Y. G. & Di Falco, E. (2013). Standard error estimation in EU-SILC – first results of the Net-SILC2 project .

Hansen, M., H., W. N., Hurwitz & Madow, W. G. (1953). *Sample survey methods and theory*, vol. Volume I Methods and applications. Wiley.

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite verse , 663 – 685.

Lundstöm, S. & Särndal, C. E. (2001). *Estimation in the presence of nonresponse and frame imperfections*.

M., A. A. . T. (2013). Estimation of social exclusion indicators from complex surveys: The r package laeken. **54(15)**.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods* **3 (3)**, 167 – 195.

Osier, G. & Di Meglio, E. (2012). The linearisation approach implemented by eurostat for the first wave of EU-SILC: what could be done from second wave onwards? Tech. rep., Institut National de la Statistique et des Etudes Economiques (STATEC Luxembourg), Luxembourg.

Preston, I. (1995). Sampling distributions of relative poverty statistics. *Applied Statistics* **44**, 91 – 99.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Särndal, C., Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London:Chapman and Hall.

Verma, V. & Betti, G. (2005). *Sampling errors and design effects for poverty measures and other complex statistics*. Working Paper 53, Siena: Dipartimento di Metodi Quantitativi, Universita degli Studi.