Austrian Journal of Statistics
February 2020, Volume 49, 33–44.
http://www.ajs.or.at/
doi:10.17713/ajs.v49i1.882

Variance Estimation by Linearisation for the At Risk of Poverty or Social Exclusion (AROPE) Rate

Stefan Zins
GESIS - Leibniz-Institute for the Social Sciences

Abstract

The At Risk of Poverty or Social Exclusion (AROPE) Rate is the key indicator for monitoring the European Commissions 2020 Strategy poverty target. But the variance of the AROPE Rate is not straightforward to estimate. Re-sampling methods can be used, but they are difficult to adapt to complex sampling design, that are often used for the surveys that provide the data source for estimating the AROPER. The presented work fills a methodological gap by providing a linearisation of the AROPE Rate estimator that can be used with well known variance estimators and therefore facilitate the reporting of appropriate inference for this important indicator. The precision of the developed variance estimators based on linearisation is assessed via simulation studies and compared with a bootstrap variance estimator, as an alternative.

Keywords: variance estimation, linearisation, poverty measure, AROPE, Monte-Carlo study.

1. Introduction

In 2010 the European Commission proposed with Europe 2020 its strategy for the development of the economy of the European Union (EU) over the next 10 years. The target for social inclusion was to lift at least 20 million people in the EU out of the risk of being poor or socially excluded by 2020 (European Commission 2018). To operationalize this target an indicator was defined as: *At risk of poverty or social exclusion, abbreviated as AROPE, refers to the situation of people either at risk of poverty, or severely materially deprived or living in a household with a very low work intensity. The AROPE rate, the share of the total population which is at risk of poverty or social exclusion, is the headline indicator to monitor the EU 2020 Strategy poverty target.* (Eurostat 2018b)

By the definition above, AROPE is a combination of three different indicators, which are:

1. At risk of being poor, after social transfers: A person that has an equivalised disposable income, after social transfers, below the poverty threshold, which is set at 60% of the national median equivalised disposable income after social transfers (Eurostat 2018c).

2. Living under material deprivation: A person that is not able to afford, at least three of the following nine items:

- pay for rent, mortgage or utility bills
- keep their home adequately warm
- face unexpected expenses
- eat meat or protein regularly
- go on holiday
- a television set
- a washing machine
- a car
- a telephone

(Eurostat 2018e).

3. Living in a household with low work intensity: A persons living in a household where the members of working age worked less than 20% of their total potential during the previous 12 months (Eurostat 2018d).

If any of the above three indicators is positive for a person, AROPE is positive. The instruments used to estimate the AROPE rate (AROPER) for monitoring the Europe 2020 poverty target are the EU Statistics on Income and Living Conditions (EU-SLIC), sample surveys that collect data on households and persons in the 28 European Union countries, Iceland, Norway, Switzerland and Turkey (Eurostat 2018a). Hence statistical inference is needed to test whether observed changes in AROPER estimates are significantly different from zero or not. The purpose of this paper is to present an estimator for the sampling variance of the AROPER estimator that can be applied under complex sampling designs, as they are common for the EU-SILC surveys. Because the AROPER estimator is a non-linear function of the observed data, its variance cannot be displayed in closed form and instead approximations to the variance must be computed. The two common approaches to approximate the variance of a non-linear estimator are linearisation and re-sampling (Münnich 2008). While re-sampling techniques have to be adapted to the sampling design of the study, they are generic with respect to the estimator. The opposite can be said about linearisation, where a linearised version of the estimator has to be derived and then generic variance estimators can be used.

The proposed type of variance estimator in this work is based on linearisation using the influence function of the AROPER estimator. This method has been popularized by Deville (1999) and been used for statistical inference of various poverty and inequality measures (Osier 2009). Linearisation is also the method for variance estimation mainly featured in the *Handbook on standard error estimation and other related sampling issues in EU-SILC* (Berger, Goedemé, and Osier 2013). Standard errors have been reported for AROPER estimates that are based on variance estimates using linearisation (Eurostat 2013), but they assume the poverty threshold to be a constant, not an estimate from the sample data, which it is.

Section 2 presents a linearisation of the AROPER estimator, that does not assume the poverty threshold to be constant. Based on this linearisation a type of variance estimators is proposed that can also be adopted for complex sampling designs. In Section 3 the precision of two estimators based on linearisation is evaluated by a simulation study. For this the relative bias and the coverage rate of corresponding confidence intervals are computed by repeated sampling and estimation. As a comparison a non-parametric Bootstrap variance estimator is also included in the simulation study. Finally Section 4 closes with some concluding remarks and possible applications for the developed estimators.

2. Estimating AROPER

Consider a finite population U of N persons with an associated set of indices $U = \{k\}_{k=1}^{N}$ from which a sample $s \subset U$ of size n is selected. We define a sampling design as a probability distribution function P on a collection of subsets of U, called support S (Tillé (2006), p. 14). That is, we have

$$\sum_{s \in S} P(s) = 1 \; . \tag{1}$$

Further let

$$I_k = \begin{cases} 1 & \text{if } k \in s \\ 0 & \text{else} \end{cases} \tag{2}$$

and π_k be the probability of including k into a sample s, i.e. $\pi_k = E(I_k) = \sum_{s \in S} I_k P(s)$. Also let $\pi_{k\,l} = E(I_k I_l) = \sum_{s \in S} I_k I_l P(s)$ be the probability of including k and l jointly into a sample s. Every element $k \in s$ has a survey weight w_k associated with it. The principal purpose of the survey weights is to estimate totals for population U. If $w_k = \pi_k^{-1}$, the survey weights are so called design weights and used in the Horvitz-Thompson Estimator for the unbiased estimation of a population total (Särndal, Swensson, and Wretman 1992, p. 42)

The statistic of interest AROPER, which we will denote with μ_{A}, is defined as the population mean of an indicator variable AROPE. As described in Section 1 the AROPE variable will assume a value of one for person $k \in U$ if any of the following indicator variables also assumes a value one:

$\mathrm{L}_k = 1$, if person k is living in a household with low work intensity, else $\mathrm{L}_k = 0$

$\mathrm{M}_k = 1$, if person k is living under material deprivation, else $\mathrm{M}_k = 0$

$\mathrm{P}_k = 1$, if person k is at risk of being poor, else $\mathrm{P}_k = 0$

The indicator variable AROPE for person $k \in U$ is then given by

$$\begin{aligned} \mathrm{A}_k &= \mathrm{P}_k \vee \mathrm{L}_k \vee \mathrm{M}_k \\ &= 1 - (1 - \mathrm{P}_k)(1 - \mathrm{L}_k)(1 - \mathrm{M}_k) \; . \end{aligned}$$

Variables L_k and M_k are directly observable from the survey data. However P_k can only be predicted, as the value of the poverty threshold is unknown.

Now we can write $\mu_{\mathrm{A}} = \tau_{\mathrm{A}}/N$, where

$$\tau_{\mathrm{A}} = \sum_{k \in U} g_k \mathrm{P}_k + (1 - g_k) \; ,$$

with

$$g_k = 1 - \mathrm{L}_k - \mathrm{M}_k + \mathrm{L}_k \mathrm{M}_k \; . \tag{3}$$

An estimator for μ_{A} is given by a weighted sample mean (Särndal *et al.* 1992, p. 182) or Hájek estimator (Hájek 1971):

$$\hat{\mu}_{\mathrm{A}} = \frac{\hat{\tau}_{\mathrm{A}}}{\hat{N}} \tag{4}$$

where

$$\begin{aligned} \hat{\tau}_{\mathrm{A}} &= \sum_{k \in s} w_k \left(g_k \hat{\mathrm{P}}_k + (1 - g_k) \right) \; , \\ \hat{N} &= \sum_{k \in s} w_k \; , \end{aligned} \tag{5}$$

and $\hat{\mathrm{P}}_k$ is determined by using the estimated poverty threshold based on the empirical equivalised disposable income distribution of the sample.

In the following we consider two variables of interest. y, the equivalised disposable income of a person after social transfers, and g, as defined for Equation 3 for all persons $k \in U$. The value of (y, g) for person k is $(y_k, g_k) \in \mathbb{R} \times [0, 1] \quad \forall\, k \in U$.

2.1. Linearisation of AROPER

We now linearise $\hat{\mu}_{\mathrm{A}}$ following the approach of Deville (1999), by deriving its influence function. In contrast to the influence function as introduced by Hampel (1974), Deville (1999) uses a finite discrete measure M for the size of the population and not a theoretical distribution function on which the functional is defined on. The statistic of interest is described by statistical functional T of M, i.e. as $T(M)$. The influence function of statistic T(M) at point $y \in \mathbb{R}$ is then defined as

$$IF\left(T, M, y\right) = \lim_{\epsilon \to 0} \frac{T\left(M + \epsilon\delta_k(y)\right) - T(M)}{\epsilon} \, , \tag{6}$$

where $\delta_k(y)$ is the Dirac measure at point $y \in \mathbb{R}$, with

$$\delta_k(y) = \begin{cases} 1 & \text{if } y = y_k \text{ and } k \in U \\ 0 & \text{else} \end{cases}$$

and $M = \sum_{k \in U} \delta_k(y_k)$.

To estimate $T(M)$ estimator $T(\hat{M})$ is used, a functional of stochastic measure $\hat{M}$, which is associated with the survey weights $\{w_k\}_{k=1}^{n}$, and close to the population size N (Goga, Deville, and Ruiz-Gazen 2009).

For example, an estimator for a population total $\tau = \sum_{k \in U} y_k$ for $y_k \in \mathbb{R}$ $\forall k \in U$ can be written as an estimator $\hat{M}$ of M. Because, $\tau = \sum_{k \in U} y_k = \int_U y dM = \sum_{k \in U} y_k \delta_k(y_k)$ is a functional of M. An obvious choice for an estimator $\hat{M}$ would be $\hat{M} = \sum_{k \in s} w_k \delta_k(y_k) = \hat{N}$, (Deville 1999).

Now we describe τ_{A} and μ_{A} as statistical functionals of the following form

$$\begin{aligned} \tau_{\mathrm{A}}(M) &= \int_U g \nVdash_{\{y \leq 0.6 \cdot Q_y(0.5, M)\}} dM + \int_U (1-g) dM \\ &= \sum_{k \in U} g_k 1_{y_k \leq 0.6 \cdot Q_y(0.5, M)} + \sum_{k \in U} (1 - g_k) \, , \\ \mu_{\mathrm{A}}(M) &= \frac{\tau_{\mathrm{A}}(M)}{\int_U dM} \\ &= \frac{\tau_{\mathrm{A}}(M)}{N} \, , \end{aligned}$$

with $Q_y(0.5, M)$ as the median of observations $\{y_k\}_{k=1}^{N}$. That is $Q_y(0.5, M) = F_y^{-1}(\alpha = 0.5, M)$ with $= F_y^{-1}(\alpha, M) = \inf\{q \in \mathbb{R} | F_y(q, M) \geq \alpha\}$, and

$$F_y(q, M) = \frac{\int_U \nVdash_{\{y \leq q\}} dM}{\int_U dM} \tag{7}$$

$$= \frac{\sum_{i \in U} \nVdash_{\{y_i \leq q\}}}{N} \, , \tag{8}$$

and a fixed $\alpha \in (0, 1)$.

We are now looking for the influence function of $\mu_{\mathrm{A}}(M)$. First we need to derive the influence function of the first term in $\tau_{\mathrm{A}}(M)$, the total of variable g for the part of the population for which $y \leq 0.6 \cdot Q_y(0.5, M)$. For this we will follow the approach of Langel and Tillé (2011) for deriving the influence function of the Quintile Share Ratio, an inequality measure. We can write the influence function of a partial sum

$$G_{y \leq \beta \cdot Q_y(\alpha, M)} = \sum_{k \in U} g_k \nVdash_{\{y_k \leq \beta \cdot Q_y(\alpha, M)\}} \tag{9}$$

with fixed $0 < \alpha, \beta, < 1$, at point y_k as

$$IF(G_{y \leq \beta \cdot Q_y(\alpha, M)}, y_k, M) = g_k \nVdash_{\{y_k \leq \beta \cdot Q_y(\alpha, M)\}} + \beta \tilde{G}'_{y \leq \beta \cdot Q_y(\alpha, M)} IF(Q_y(\alpha, M), y_k) \tag{10}$$

where $\tilde{G}'_{y\le\beta\cdot Q_y(\alpha,M)}$ is the derivative of a smoothed function of $G_{y\le\beta\cdot Q_y(\alpha,M)}$. The influence function for a quantile can be found in Osier (2009), with

$$IF(Q_y(\alpha,M),y_k) = -\frac{\nVdash_{\{y_k\le Q_y(\alpha,M)\}} - \alpha}{\tilde{F}'_y(Q_y(\alpha,M))N}\ , \tag{11}$$

where $\tilde{F}'_y(q)$ is the derivative of a smoothed function of $F_y(q,M)$. However to avoid having to derive smooth approximation to $G_{y\le\beta\cdot Q_y(\alpha,M)}$ and $F_y(q,M)$, in order of compute $IF(G_{y\le\beta\cdot Q_y(\alpha,M)},y_k,M)$ we can use the same algebra as Langel and Tillé (2011). For this we define

$$L(q) = \frac{\tilde{G}_{y\le q}}{\tau_g}\ , \tag{12}$$

with $\tau_g = \sum_{k\in U} g_k$. Then we can write

$$\begin{aligned} L(q) &= \frac{\int_0^q\int_0^1 v\,d\tilde{F}_g(v|y=u)\,d\tilde{F}_y(u)}{\int_0^\infty E(g|y=u)\,d\tilde{F}_y(u)} \\ &= \frac{\int_0^q E(g|y=u)\,d\tilde{F}_y(u)}{\int_0^\infty E(g|y=u)\,d\tilde{F}_y(u)} \\ &= \frac{N\int_0^q E(g|y=u)\,d\tilde{F}_y(u)}{\tau_g}\ , \end{aligned}$$

where $\tilde{F}_g(v|y)$ is the smooth cdf of variable g conditional on y. Accordingly $E(g|y=u)$ is the conditional expectation of variable g given that y has value u. Hence

$$L'(q) = \frac{N\,E(g|y=q)\tilde{F}'_y(q)}{\tau_g}\ , \tag{13}$$

given that $L'(q)\tau_g = \tilde{G}'_{y\le q}$ we get

$$\begin{aligned} IF(G_{y\le\beta\cdot Q_y(\alpha,M)},y_k,M) = g_k&\nVdash_{\{y_k\le\beta\cdot Q_y(\alpha,M)\}} \\ &- \beta\cdot\frac{\tau_g\,L'(\beta\cdot Q_y(\alpha,M))}{N\,\tilde{F}'_y(Q_y(\alpha,M))}\left(1_{\{y_k\le Q_y(\alpha,M)\}}-\alpha\right) \\ = g_k&\nVdash_{\{y_k\le\beta\cdot Q_y(\alpha,M)\}} \\ &- \beta\cdot E(g_k|y_k=\beta\cdot Q_y(\alpha,M))\left(1_{\{y_k\le Q_y(\alpha,M)\}}-\alpha\right)\ . \end{aligned} \tag{14}$$

Now we can write the influence function of $\mu_{\mathrm{A}}(M)$ as

$$IF\left(\mu_{\mathrm{A}},y,M\right) = IF\left(\tau_{\mathrm{A}},y,M\right)/N \tag{15}$$

where

$$\begin{aligned} IF\left(\tau_{\mathrm{A}},y=y_k,M\right) = g_k&\mathrm{P}_k \\ &- 0.6\cdot E(g|y=\beta\cdot Q_y(0.5,M))\left(\mathrm{P}_k-0.5\right) \\ &+ \left(1-g_k\right)\ . \end{aligned} \tag{16}$$

Note that $\mathrm{P}_k = \nVdash_{\{y_k\le 0.6\cdot Q_y(0.5,M)\}}$.
We can express Estimator $\hat{\mu}_{\mathrm{A}}$ in Equation 4 as functional

$$\mu_{\mathrm{A}}(\hat{M}) = \frac{\tau_{\mathrm{A}}(\hat{M})}{\int_s d\hat{M}}\ , \tag{17}$$

with $\tau_{\mathrm{A}}(\hat{M}) = \hat{\tau}_{\mathrm{A}}$ and $\int_s d\hat{M} = \hat{N}$, as described in Equations 5. As such we can write its influence function as

$$IF\left(\hat{\mu}_{\mathrm{A}},y,M\right) = \left(IF\left(\tau_{\mathrm{A}},y,M\right)-\mu_{\mathrm{A}}\right)/N\ . \tag{18}$$

2.2. Variance estimation

If z_k is the value of the influence function of estimator in Equation 18 at point $y = y_k$, then under asymptotic conditions Deville (1999) shows that the variance of $\hat{\mu}_{\mathrm{A}}$ can be approximated by

$$V\left(\sum_{k\in s} w_k z_k\right) . \tag{19}$$

Values z_k are not observable, as the influence function in Equation 18 contains unknown quantities. Thus the function in Equation 18 has to be estimated, by replacing these unknown quantities with suitable estimates. As an estimator for the values of influence function in Equation 16 we use

$$\begin{aligned}\hat{v}_k = {} & g_k \hat{\mathrm{P}}_k \\ & - 0.6 \cdot \hat{E}(g_k | y_k = 0.6 \cdot \hat{Q}_y(0.5, \hat{M})) \left(\hat{\mathrm{P}}_k - 0.5\right) \\ & + (1 - g_k)\end{aligned} \tag{20}$$

where $\hat{\mathrm{P}}_k = \mathbb{1}_{\{y_k \leq 0.6 \cdot \hat{Q}_y(0.5, \hat{M})\}}$. Further, $\hat{Q}_y(0.5, \hat{M}) = \hat{F}_y^{-1}(0.5, \hat{M})$ with $\hat{F}_y^{-1}(0.5, \hat{M}) = \inf\{y \in \mathbb{R} | \hat{F}_y(q, \hat{M}) \geq 0.5\}$ and $\hat{F}_y(q, \hat{M}) = \sum_{k\in s} w_k \mathbb{1}_{\{y_k \leq q\}} / \hat{N}$. $\hat{E}(g|y = 0.6 \cdot Q_y(0.5, M))$ is the estimated conditional probability for $g_k = 1$, given that y_k is equal to the estimated poverty threshold. This probability can be estimated using, for example, a generalized linear model or a generalized additive model for binomial response data. Both of these models to estimate $E(g|y = 0.6 \cdot \hat{Q}_y(0.5, \hat{M}))$ are evaluated in the simulation study in Section 3.

Now we can construct the following estimator for the values of influence function of $\hat{\mu}_{\mathrm{A}}$

$$\begin{aligned}\hat{z}_k = {} & \widehat{IF}\left(\hat{\mu}_{\mathrm{A}}, y = y_k, \hat{M}\right) \\ = {} & \left(\widehat{IF}\left(\hat{\tau}_{\mathrm{A}}, y = y_k, \hat{M}\right) - \hat{\mu}_{\mathrm{A}}\right) / \hat{N} \\ = {} & \left(\hat{v}_k - \hat{\mu}_{\mathrm{A}}\right) / \hat{N} .\end{aligned} \tag{21}$$

In the case where $w_k = \pi_k^{-1} \ \forall \ k \in U$, we can estimate $V\left(\sum_{k\in s} w_k \hat{z}_k\right)$ by using the variance estimator for the Horvitz-Thompson estimator of a total (Särndal *et al.* 1992, Section 2.3), which is given by

$$\hat{V}\left(\sum_{k\in s} w_k \hat{z}_k\right) = \sum_{k\in s}\sum_{l\in s} \frac{(\pi_{k\,l} - \pi_k \pi_l)}{\pi_{k\,l}} \frac{\hat{z}_k}{\pi_k} \frac{\hat{z}_l}{\pi_l} \tag{22}$$

For a Simple Random Sample the estimator in Equation 22 can be written as

$$N^2 \left(1 - \frac{n}{N}\right) \frac{s_{\hat{z}}^2}{n} \tag{23}$$

with $s_{\hat{z}}^2 = \sum_{k\in s} (\hat{z}_k - \overline{\hat{z}})^2 / (n-1)$ and $\overline{\hat{z}} = \sum_{k\in s} \hat{z}_k / n$.

For complex sampling designs their exist numerous approximations to the variance of a Horvitz-Thompson estimator, that can be estimated without having to know the second-order inclusion probabilities $\pi_{k\,l}$ (Matei and Tillé 2005). This gives this variances estimation strategy based on linearisation much flexibility towards its application under different complex sampling designs. If the w_k's are calibration weights, as described by (Deville and Särndal 1992), which is common in many surveys, e.g. to adjust for Unit-Nonresponse, the $\hat{z}_k$'s in Equation 22 can be replaced by the corresponding residuals of the estimated regression of $\hat{z}_k$ on the auxiliary variables used in the calibration. Although to count for the randomness of the response process a different kind of variance estimator should be used (Särndal and Lundström 2005, Chapter 11).

3. Monte-Carlo simulation study

We conduct a design-based simulation study to evaluate the precision of two variance estimators for $\hat{\mu}_{\mathrm{A}}$ that are based on linearisation as described in Equation 21. One where $E(g|y = 0.6 \cdot Q_y(0.5, M))$ is estimated by generalized linear model and one where a generalized additive model is used. The simulation study is implement using the R language (R Core Team 2018). To estimate $E(g|y = 0.6 \cdot Q_y(0.5, M))$ with the generalized linear model the svyglm function from the survey package was used (Lumley 2016). For the estimation with the generalized additive model the gam function from the mgcv package was used (Wood 2018). For both models the only predictor variable is the income variable y, where for the generalized additive model a regression spline smoother was used on the income. As a comparison we also include in the study a naive variance estimator, which ignores the variability of the estimated poverty threshold, and a non-parametric bootstrap. The bootstrap variance estimator was implement using the boot function from the boot package (Canty and Ripley 2017).

3.1. Synthetic EU-SLIC data

To generate variable y for a finite population we use the equivalised household income variable in the data set eusilc from the R-package laeken (Alfons and Templ 2013). The eusilc data set is synthetic and was generated based on the Austrian EU-SILC survey of 2006. It has 14827 observations. To generate a population larger than this we sample from the data by a Simple Random Sample with replacement with a sample size equal to the desired population size. A log-norm distributed error term is added to each selected equivalised household income to reduce the heaping in the generated empirical income distribution. To generate g a logistic model is used, where the probability for $g_k = 1$ depends on y_k, that is

$$E(g_k) = \frac{e^{8.4 \cdot 10^{-1} + 6 \cdot 10^{-5} \cdot y_k}}{(1 + e^{8.4 \cdot 10^{-1} + 6 \cdot 10^{-5} y_k})} \, . \tag{24}$$

The value of g_k is then generated as a realisation from a Bernoulli distribution. The positive relation between y_k and $E(g_k)$ in Equation 24 is motivate be the definition of g_k in Equation 3. Using the Fréchet inequalities we know that

$$\max(E(\mathrm{L}_k) + E(\mathrm{M}_k) - 1, 0) \leq E(\mathrm{L}_k \mathrm{M}_k) \leq \min(E(\mathrm{L}_k), E(\mathrm{M}_k))$$

from which follows that $E(\mathrm{L}_k \mathrm{M}_k) - E(\mathrm{L}_k) - E(\mathrm{M}_k) \leq 0$. Further, it is plausible to let $E(\mathrm{L}_k)$ and $E(\mathrm{M}_k)$ decrease with an increasing income and vice versa, thus introducing a positive dependency between $E(g_k)$ and y_k. To evaluate the asymptotic properties of the estimators we generate populations of different sizes but hold the sampling fraction constant at 0.1%, i.e. the sample size will change proportionally to the population size. From each of the different populations we take repeated samples, using Simple Random Sampling. Two measures are used to evaluate the precision of the variance estimates, the relative bias and the coverage rate of the confidence intervals. We define the relative bias for variance estimator $\hat{V}(\hat{\mu}_{\mathrm{A}})$ as

$$rb(\hat{V}(\hat{\mu}_{\mathrm{A}}) = \frac{E(\hat{V}(\hat{\mu}_{\mathrm{A}})) - V(\hat{\mu}_{\mathrm{A}})}{V(\hat{\mu}_{\mathrm{A}})} \, .$$

We use a Monte-Carlo approximations to calculate the relative bias of our variance estimators, where $E(\hat{V}(\hat{\mu}_{\mathrm{A}}))$ is approximated by the mean value of variance estimates from 1000 samples and $V(\hat{\mu}_{\mathrm{A}})$ by the variance of point estates $\hat{\mu}_{\mathrm{A}}$ from 10000 samples. A higher number of samples to approximate $V(\hat{\mu}_{\mathrm{A}})$ is chosen because the convergence for $V(\hat{\mu}_{\mathrm{A}})$ is thought to be slower than for $E(\hat{V}(\hat{\mu}_{\mathrm{A}}))$. The coverage rate is defined as

$$cr(\hat{V}(\hat{\mu}_{\mathrm{A}}), \hat{\mu}_{\mathrm{A}}) = Pr\left(\mu_{\mathrm{A}} \in \left[\hat{\mu}_{\mathrm{A}} \pm t_{0.975}\sqrt{\hat{V}(\hat{\mu}_{\mathrm{A}})}\right]\right)$$

Table 1: Simulation results: synthetic EU-SLIC data

N	rb				cr			
	$\hat{V}_{glm}$	$\hat{V}_{gam}$	$\hat{V}_{naive}$	$\hat{V}_{boot}$	$\hat{V}_{glm}$	$\hat{V}_{gam}$	$\hat{V}_{naive}$	$\hat{V}_{boot}$
10^5	0.04944	0.05211	0.13058	0.07459	0.960	0.959	0.957	0.954
10^6	0.04559	0.04568	0.14123	0.01891	0.960	0.960	0.971	0.950
$5 \cdot 10^6$	0.04901	0.04903	0.14194	0.01905	0.956	0.956	0.965	0.950
10^7	0.03886	0.03887	0.13169	0.02117	0.957	0.957	0.961	0.949

where $t_{0.975}$ is the 97.5% quantile of the standard normal distribution. The coverage rate is then approximated by the proportion of the 1000 sample for which the above described confidence interval include the true value μ_{A}.

The first 4 columns of Table 1 contain the relative biases and the last 4 the coverage rates that were obtained from the simulation study. Column labels $\hat{V}_{glm}$, $\hat{V}_{gam}$, $\hat{V}_{naive}$, and $\hat{V}_{boot}$ correspond to the different variance estimate. $\hat{V}_{glm}$ and $\hat{V}_{gam}$ are variances estimators as described in Equation 23 with $E(g|y = 0.6 \cdot \hat{Q}_y(0.5, \hat{M}))$ estimated by a generalized linear model and generalized additive model, respectively. The naive variance estimator $\hat{V}_{naive}$ has the same form as the estimator in Equation 23, but instead of $\hat{z}_k$ the estimated AROPE indicator variable $\hat{\mathrm{A}}_k = g_k\hat{\mathrm{P}}_k + (1 - g_k)$ is used. Estimator $\hat{V}_{boot}$ is a non-parametric bootstrap variance estimator, using a Simple Random Sample with replacement of same size as the sample size and with 99 replications. The rows of Table 1 correspond to population sizes, 10^5, 10^6, $5 \cdot 10^6$, and 10^7 respectively.

If we look at the relative bias of the variance estimates we see that none of the them are negative and that the naive variance estimator has by far the largest bias. For the smallest sample size of 100, corresponding to the population of size 10^5, estimators $\hat{V}_{glm}$, $\hat{V}_{gam}$, and $\hat{V}_{boot}$ have very similar relative biases with the estimators based on linearisation showing a slightly lower bias. With an increasing population and sample size the bias of $\hat{V}_{boot}$ reduces, but for $\hat{V}_{glm}$, $\hat{V}_{gam}$, it stays stable and only reduce for the largest population. For populations larger than 10^5 the relative bias of the two estimators based on linearisation is larger than for the bootstrap estimator. Both appear not to converge to the true variance, or only very slowly. However the relative bias of $\hat{V}_{glm}$ and $\hat{V}_{gam}$ appears to be bounded for larger sample sizes (e.g. $n > 1000$), at or below 5%, which can be regarded as low and enables reasonable inference for μ_{A}, as is shown by the corresponding coverage rates. The coverage rate for $\hat{V}_{boot}$ for all populations is almost right on the target value 0.95. For estimators $\hat{V}_{glm}$ and $\hat{V}_{gam}$ the coverage rates are slightly higher, but still close to 0.95. Estimator $\hat{V}_{glm}$ and $\hat{V}_{gam}$ preform almost identical, which is to be expected given the relationship between g_k and y_k, as described in Equation 24.

In Section 3.2 a more complex data structure is examined, where the dependency between y_k and g_k is not model directly. Instead L_k is modelled based of the working month for a household, on which y_k depends as well. Material deprivation M_k will directly model after y_k.

3.2. Full synthetic data

To generate the three indicator variables that the AROPE indicator is composed of, a population of persons is generated with households specific variables. At first a household size variable is generated by drawing from a multinomial distribution. Then working months are associated with the household size. Possible values for working month per household member are 0, 3, 6, 9, 12. If the combined working months of the household of person k are below 20% of its maximal working months, 12 times the household's size, then $\mathrm{L}_k = 1$, else $\mathrm{L}_k = 0$. The income variable y_k is modelled after the working months of the household. Average monthly earnings are generated from an exponential distribution with mean 1.700. The average monthly earnings are then multiplied by the working months of a household to generate its yearly income. Then the household income is equivalised in relation to the household size to generate y_k. If the household is of size 1 it remains unchanged, for size 2 the household income is divided by 1.5, and for each additional households member the scaling factor increases by 0.5. If the household of person k has an equivalised income below 60% of the median equivalised income for all households $\mathrm{P}_k = 1$, else $\mathrm{P}_k = 0$. The indicator variable for living under material deprivation is directly dependent on y_k, with $E(\mathrm{M}_k) = 1/(y_k+1)^{0.25}$. The value of M_k is generated as a realisation of a Bernoulli distribution.

Table 2: Simulation results: full synthetic data

N	rb				cr			
	$\hat{V}_{glm}$	$\hat{V}_{gam}$	$\hat{V}_{naive}$	$\hat{V}_{boot}$	$\hat{V}_{glm}$	$\hat{V}_{gam}$	$\hat{V}_{naive}$	$\hat{V}_{boot}$
10^5	0.11697	0.08280	0.91234	0.24923	0.956	0.955	0.987	0.969
10^6	0.11489	0.06337	0.92957	0.10707	0.949	0.941	0.994	0.947
$5 \cdot 10^6$	0.10029	0.04881	0.90661	0.05046	0.961	0.953	0.995	0.956
10^7	0.09083	0.03954	0.89039	0.02585	0.961	0.957	0.991	0.950

Table 2 shows the results of the simulation study for the fully synthetic population. The study was implemented the same way as the simulation with the synthetic EU-SILC data in Section 3.1. The naive variance estimator $\hat{V}_{naive}$ also over estimates, however to a larger extent than in the previously described study in Section 3.1. The consequence of this over estimation is shown in the corresponding coverage rates, which are close to one for all populations, but the smallest. If we look a the variances estimators based on linearisation, $\hat{V}_{glm}$ and $\hat{V}_{gam}$, we see that the relative bias for $\hat{V}_{glm}$ is higher for all populations. The generalized additive model seems more apt for modelling the relationship between g_k and y_k, with $\hat{V}_{gam}$ performing well even for low samples sizes. If we compare $\hat{V}_{gam}$ with $\hat{V}_{boot}$ we see that for two smallest population the linearisation method is more precise than the re-sampling method. For population size $5 \cdot 10^6$, the relative bias of both estimators is very similar and only for the largest population size $\hat{V}_{boot}$ outperforms $\hat{V}_{gam}$. The coverage rates for $\hat{V}_{boot}$ are also similar to $\hat{V}_{gam}$, apart from the smallest population, where the much higher bias of the re-sampling method leads to a coverage rate of 0.969 compared to 0.955 for $\hat{V}_{gam}$. For the smaller population sizes it can be observed that coverage rates are still close to their target value despite a high bias of some variance estimators. This can be explained by the fact that the distribution of point estimator $\hat{\mu}_{\mathrm{A}}$ is not close enough to normality, e.g. the $t_{0.975}$ quantile might not be appropriate for the confidence intervals. The two error sources, biased variance estimate and inappropriate confidence intervals seem to cancel each other out. For larger sample and population sizes the normal approximation of the point estimator works better and more precise variance estimators also lead to better coverage of the confidence intervals.

4. Conclusions

Deville (1999) linearisation technique, using influence functions, has been used to derive a class of variance estimators for the AROPER estimator. Two variance estimators based on linearisation have been proposed that can be applied to sample surveys with complex sampling designs. Both estimators haven been tested in two simulation studies and compared against alternative variance estimators. A naive approach, which ignores the complexity of the measure that AROPER is based on, and a bootstrap re-sampling method. The estimator based on the naive approach has been identified as inadequate for estimating the variance of the AROPER estimator. The results form the simulation studies for the estimators based on linearisation validate the method, showing that they can be used to making adequate inference for the AROPER indicator. Compared to the bootstrap method linearisation preforms better for small sample sizes (e.g. n=100). However the re-sampling method converges faster, while the linearisation method shows a positive bias that appears to bound around 4% in both simulation studies.

Both simulation studies in Section 3 use simple random sampling and not a complex sampling design. This was done to have a straightforward comparison between the different variance estimators. A complex sampling design will affect the variance of estimators in their own way, making it harder to attribute the asymptotic behaviour of variance estimators to the variance estimation technique (e.g. linearization or re-sampling) alone (Helga and Eckmair 2016). However, simulation studies with complex sampling designs would be a welcome addition to the presented simulation studies. In particular it would be of interest to evaluate how the estimation of $E(g|y = 0.6 \cdot Q_y(0.5, M))$ is affected by complex sampling designs.

Of the two estimates based on linearisation that have been considered, the estimator that uses a generalized additive model seem to be more adequate for modelling the relationship between variable g_k, as defined in Equation 3, and the income variable y_k. Thus the additional computational effort of using a generalized additive model, compared to a generalized linear model, seems to be justified. For surveys that have unequal weights there are estimators that allow for the usage of survey weights when estimating a generalized linear or generalized additive model. For example, the R functions mentioned at the beginning of Section 3 to estimate both model types allow for the usage of weighted data.

Variance estimators based on linearisation are not only flexible with regard to the sampling design but it is also straightforward to use for domain estimates. For surveys with complex sampling designs variance estimation can be made possible by providing re-sampling weights. However these re-sampling weights are usually constructed to do inference on the population, not on a subset of it. Their usability for domain estimation has to be built in when they are constructed. For linearisation methods this limitation does not apply and the method developed in this article is applicable as long as the design weights of a survey are known. Also for the estimation of change the developed linearisation of the AROPER estimator can easily be used in the variance of change estimators presented by Berger (2004) and Berger and Priam (2016), to assess whether an observed change in AROPER estimates is significant or not. Thus the method presented in this paper is considered a valuable addition to the tool set of analysts who want to report for the AROPER indicator confidence intervals, significant change over time, or inference for certain population domains.

References

Alfons A, Templ M (2013). "Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken." *Journal of Statistical Software*, **54**(15), 1–25. URL http://www.jstatsoft.org/v54/i15/.

Berger YG (2004). "Variance Estimation for Measures of Change in Probability Sampling." *Canadian Journal of Statistics*, **32**(4), 451–467. ISSN 03195724, 1708945X. doi:10.2307/3316027.

Berger YG, Goedemé T, Osier G (2013). *Handbook on Standard Error Estimation and Other Related Sampling Issues in EU-SILC Second Network for the Analysis of EU-SILC, Euro-Stat*.

Berger YG, Priam R (2016). "A Simple Variance Estimator of Change for Rotating Repeated Surveys: An Application to the European Union Statistics on Income and Living Conditions Household Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **179**(1), 251–272. ISSN 09641998. doi:10.1111/rssa.12116.

Canty A, Ripley BD (2017). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-20.

Deville JC (1999). "Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques." *Survey methodology*, **25**(2), 193–204.

Deville JC, Särndal CE (1992). “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association*, **87**(418), 376–382.

European Commission (2018). “Europe 2020 Strategy.” https://ec.europa.eu/info/business-economy-euro/economic-and-fiscal-policy-coordination/eu-economic-governance-monitoring-prevention-correction/european-semester/framework/europe-2020-strategy_en.

Eurostat (2013). “Standard Error Estimation for the EU-SILC Indicators of Poverty and Social Exclusion.” *Technical Report 2013 edition*, Eurostat.

Eurostat (2018a). “EU Statistics on Income and Living Conditions (EU-SILC) Methodology - Childcare Arrangements - Statistics Explained.” https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology_-_childcare_arrangements.

Eurostat (2018b). “Glossary: At Risk of Poverty or Social Exclusion (AROPE) - Statistics Explained.” https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:At_risk_of_poverty_or_social_exclusion_(AROPE).

Eurostat (2018c). “Glossary: At-Risk-of-Poverty Rate - Statistics Explained.” https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:At-risk-of-poverty_rate.

Eurostat (2018d). “Glossary: Persons Living in Households with Low Work Intensity - Statistics Explained.” https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Persons_living_in_households_with_low_work_intensity.

Eurostat (2018e). “Glossary:Material Deprivation - Statistics Explained.” https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Severe_material_deprivation_rate.

Goga C, Deville JC, Ruiz-Gazen A (2009). “Use of Functionals in Linearization and Composite Estimation with Application to Two-Sample Survey Data.” *Biometrika*, **96**(3), 691–709.

Hájek J (1971). “Comment on a paper by D. Basu: An Essay on the Logical Foundations of Survey Sampling, Part One*.” In VP Godambe, DA Sprott (eds.), *Selected Works of Debabrata Basu*, pp. 167–206. Springer New York, New York, NY. ISBN 978-1-4419-5824-2 978-1-4419-5825-9. doi:10.1007/978-1-4419-5825-9_24.

Hampel FR (1974). “The Influence Curve and Its Role in Robust Estimation.” *Journal of the American Statistical Association*, **69**, 383–393.

Helga W, Eckmair D (2016). “Simulation Studies for Complex Sampling Designs.” *Austrian Journal of Statistics*, **35**(4), 419–435. ISSN 1026597X. doi:10.17713/ajs.v35i4.352.

Langel M, Tillé Y (2011). “Statistical Inference for the Quintile Share Ratio.” *Journal of Statistical Planning and Inference*, **141**(8), 2976–2985.

Lumley T (2016). *survey: Analysis of Complex Survey Samples*. R package version 3.32.

Matei A, Tillé Y (2005). “Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size.” *Journal of Official Statistics*, **21**(4), 543–570.

Münnich R (2008). “Varianzschätzung in komplexen Erhebungen.” *Austrian Journal of Statistics*, **37**(3), 319–334.

Osier G (2009). “Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques.” *Survey Research Methods*, **3**(3), 167–195.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Särndal CE, Lundström S (2005). *Estimation in Surveys with Nonresponse.* John Wiley & Sons.

Särndal CE, Swensson B, Wretman J (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Tillé Y (2006). *Sampling Algorithms.* Springer Series in Statistics. Springer, New York.

Wood S (2018). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.* R package version 1.8-24.

Affiliation:

Stefan Zins
GESIS Leibniz-Institute for the Social Sciences
B2 1
D-68159 Mannheim, Germany
E-mail: stefan.zins@gesis.org
URL: https://www.gesis.org/en/home/

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 49 *Submitted:* 2018-09-30
February 2020 *Accepted:* 2018-12-06