QUANTITATIVE METHODS IN ECONOMICS
Volume XXII, No. 1, 2021, pp. 14 – 28

AN APPLICATION OF THE INTERVAL ESTIMATION FOR THE AT-RISK-OF-POVERTY RATE ASSESSMENT

Marcin Dudziński https://orcid.org/0000-0003-4242-8411
Joanna Kaleta https://orcid.org/0000-0001-6628-4251
Institute of Information Technology
Warsaw University of Life Sciences – SGGW, Poland
e-mails: marcin_dudzinski@sggw.edu.pl; joanna_kaleta@sggw.edu.pl

Abstract: In the document [Eurostat (Your Key to European Statistics) 2020], *At-Risk-of-Poverty Rate* (*ARPR* in short) is defined as the percentage of population with an income not exceeding 60% of the general population median income. Extensive and thorough research on the estimation of this measure has been conducted since its introduction. For example, in the paper of [Zieliński 2009a] a non-parametric, distribution-free confidence interval for *ARPR* has been constructed. An example of application of the confidence interval proposed by [Zieliński 2009a] has been given in [Zieliński 2009b]. Some other interesting approach regarding the interval estimation of *ARPR* has been proposed in [Luo and Qin 2017], where the authors introduced new concepts of the interval estimation for the so-called *Low-Income Proportion* (*LIP*) measure, which is a generalization of *ARPR*. The *LIP* measure and thus, the *ARPR* parameter in particular, are important indexes describing the inequality in an income distribution. Based on the construction of the point smoothed kernel estimate for *LIP*, [Luo and Qin 2017] established a smoothed jackknife empirical likelihood approach leading to the introduction of some new non-parametric confidence intervals for the *LIP* measure and consequently, for the *ARPR* index as well. In our work, we aim to apply the most interesting ideas of *LIP* and *ARPR* point and interval estimation for data consisting of 13057 observations concerning an equalised disposable income of households in Poland from 2003. We also discuss the accuracy and adequacy of the empirical results relating to the *ARPR* interval estimation, obtained by the implementation of the constructed confidence intervals.

Keywords: Low-Income Proportion (*LIP)*, At-Risk-of-Poverty Rate (*ARPR*), confidence intervals for *LIP* and *ARPR*, Nonparametric estimation, Kernel estimation

JEL classification: C51, C52

https://doi.org/10.22630/MIBE.2021.22.1.2

INTRODUCTION

At-Risk-of-Poverty Rate (or *ARPR* in short) is a measure that enables to determine the inequality in an income distribution. According to [Eurostat (Your Key to European Statistics) 2020], it is defined as the proportion of general population with an income not exceeding 60% of the median income in the whole population. Using mathematical terms, we may describe this measure in the following pattern. Namely, let $\mathrm{EQ\_INC}_j$ denote an equivalised disposable income of the j-th individual (person or household) and suppose that weight_i stands for the weight of individual i. Firstly, we shall determine the so-called *At-Risk-of-Poverty Threshold* (or *ARPT* in short). It is expressed as (see also [Zieliński 2009a-b])

$$ARPT = \textit{At-Risk-of-Poverty Threshold} = 60\%\mathrm{EQ\_INC}_{\mathrm{MEDIAN}},$$

where

$$\mathrm{EQ}_{\mathrm{INC}_{\mathrm{MEDIAN}}} = \begin{cases} \frac{1}{2}\left(\mathrm{EQ\_INC}_j + \mathrm{EQ\_INC}_{j+1}\right), & \text{if } \sum_{i=1}^{j} \mathrm{weight}_i = \frac{W}{2} \\ \mathrm{EQ\_INC}_{j+1}, & \text{if } \sum_{i=1}^{j} \mathrm{weight}_i < \frac{W}{2} < \sum_{i=1}^{j+1} \mathrm{weight}_i \end{cases},$$

where in turn,

$$W = \sum_{All\ persons} \mathrm{weight}_i.$$

Thus, we can directly come to stating the definition of *ARPR*. Since it is clear now that this measure denotes the percentage of individuals from the whole population with an equivalised disposable income not greater than *ARPT*, then the *ARPR* index is calculated as (see also [Zieliński 2009a-b])

$$ARPR = \frac{\sum_{\substack{All\ persons\ with \\ \mathrm{EQ\_INC} \le ARPT}} \mathrm{weight}_i}{W} \times 100.$$

We are now in a position to discuss the estimation methods for *ARPR*. We will start from the point estimation of this measure. Suppose that $X_1, X_2, \ldots, X_n$ is a sample of the equivalised disposable incomes of randomly drawn n individuals and let Med be the corresponding sample median. A straightforward point estimate for *ARPR* is given by (see, e.g., [Zieliński 2009a-b])

$$\widehat{ARPR} = \frac{1}{n}\#\{X_i\colon X_i \le 0.6 \cdot Med\},$$

with # standing for the cardinality of the considered set. It is obvious that in terms of probability distribution, the *ARPR* index is determined as

$$\theta = ARPR = F\big(0.6 \cdot F^{-1}(0.5)\big),$$

where: F denotes the cumulative distribution function (cdf) of an equivalised disposable income in the investigated general population, F^{-1} is the corresponding quantile function. *ARPR* is a special case of the so-called *Low-Income Proportion* (*LIP*) measure, which is an index defined for two parameters, usually denoted as α and β. Namely, if X denotes an income variable with a cdf F, then *LIP* is given by

$$LIP = \theta_{\alpha,\beta} = P\big(X \le \alpha \cdot \xi_\beta\big) = F\big(\alpha \cdot \xi_\beta\big) = F\big(\alpha \cdot F^{-1}(\beta)\big),$$

where ξ_β denotes the β-th quantile of an income distribution. Thus, for the fixed α and β, *LIP* is the fraction of individuals with an equivalised disposable income not exceeding $\alpha \cdot \xi_\beta = \alpha \cdot F^{-1}(\beta)$ (in other words, it is the proportion of population with an income not greater than the given fraction α of the β-th quantile from an income distribution). It is clear that *LIP* equals *ARPR* for $\alpha = 0.6$ and $\beta = 0.5$. The *Low-Income Proportion*, and consequently the *At-Risk-of-Poverty Rate* as its special case, are the measures that have been extensively used by governing bodies and government experts, as well as by business managers and advisors or academics from different areas of interest, in order to gain a great deal of valuable information and conclusions. It is particularly convenient and useful in the assessment of potential inequalities regarding the socio-economic status. For example, the employees with earnings not exceeding 60% of the population median income are treated as the low-earners by the European Statistical Office 'Eurostat'. Since, as it has already been mentioned, *ARPR* is equivalent to *LIP* with $\alpha = 0.6$ and $\beta = 0.5$, the high values of *ARPR* indicate relatively large social inequalities in the wealth structure, as well as the social instability and uncertainty. All this together should serve as a warning signal for the state decision-makers. Except for the state authorities and business entrepreneurs, *LIP* and *ARPR* have attracted much attention of scholars from various fields of interest. In particular, a large number of inference methods related to both the point and the interval estimation of *LIP* have been proposed or developed. Among numerous research papers devoted to the subject of *LIP* and *ARPR* evaluation, or the risk measures assessment in general, the works of [Gong et al. 2010, Jing et al. 2009, Li et al. 2011, Luo and Qin 2017, Wei et al. 2009, Wei and Zhu 2010] and [Zieliński 2009a-b] - are especially worthwhile to mention. Roughly speaking, there exist two essential concepts concerning the estimation of *LIP*, and *ARPR* in particular. With reference to the issue of *LIP* estimation, these two primary approaches - commonly known as the empirical and kernel methods - may be illustrated as follows. Let $X_1, X_2, \dots, X_n$ denote a simple sample from the income distribution having a cdf F. Then, the empirical estimate for parameter $\theta_{\alpha,\beta} = LIP$ is defined by (see also [Luo and Qin 2017])

$$\hat{\theta}_{\alpha,\beta} = F_n\big(\alpha\hat{\xi}_\beta\big) = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \le \alpha\hat{\xi}_\beta) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,\alpha\hat{\xi}_\beta]}\,(X_i),$$

where: F_n stands for the empirical distribution function of $X_1, \ldots, X_n$, $\hat{\xi}_\beta = F_n^{-1}(\beta)$ is the β-th quantile of the empirical distribution function F_n, while I_A denotes the indicator function of a given set A. Sadly, an application of the empirical point estimate has a relatively serious drawback, which consists in the fact that $\hat{\theta}_{\alpha,\beta} = F_n(\alpha\hat{\xi}_\beta)$ is a non-smoothing estimator of $\theta_{\alpha,\beta}$, as it is a non-smoothing function of the sample quantile $\hat{\xi}_\beta$, whereas $LIP = \theta_{\alpha,\beta} = F(\alpha\xi_\beta)$ is a function related to the smoothing income distribution function F. Therefore, instead of employing a non-smoothing empirical estimator $\hat{\theta}_{\alpha,\beta}$, [Luo and Qin 2017] suggested using the kernel method in order to obtain a smoothed estimator for $\theta_{\alpha,\beta}$. A comprehensive study has shown an advantage of the kernel estimation over an assessment based on the implementation of the empirical estimator (see, e.g., [Falk 1983]-[Falk 1985] in this context). The kernel estimator of the *LIP* index $\theta_{\alpha,\beta}$ is given by the formula

$$\hat{T}_n(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\alpha\hat{\xi}_\beta - X_i}{h}\right),$$

where K is the so-called kernel function and h denotes the chosen bandwidth (h is also interchangeably known under the name of smoothing parameter). It turns out that the kernel estimator $\hat{T}_n(\alpha,\beta)$ has a slightly smaller *Mean Squared Error (MSE)* than the empirical estimator $\hat{\theta}_{\alpha,\beta}$ (see Table 1 in [Luo and Qin 2017]). Apart from the fact that the kernel estimator is more suitable for the *LIP* assessment, this kind of estimator is also used in the definition of a smoothed version of the jacknife empirical likelihood ratio statistic for *LIP*. This smoothed version may by later applied in the constructions of the corresponding confidence intervals. One of significant difficulties arising in calculation of the smoothed estimator $\hat{T}_n(\alpha,\beta)$ is the problem of an appropriate choice of bandwidth h for this kernel estimator. Many methods of bandwidth selection have been proposed so far (see, e.g., [Bowman et al. 1998], among others, for a comprehensive review regarding this matter). In particular, [Luo and Qin 2017] use the twofold cross-validation method for bandwidth selection in order to estimate *LIP* and, after conducting some simulation research, they recommend using bandwidth of the form $h = cn^{-1/3}$, where c is some constant not depending on n. On the other hand, it is worthwhile to mention that intensive simulation studies indicate that the selection of kernel *K* itself is not of so high importance, since the change of kernel does not affect the obtained estimation results too much. In the cited work of [Luo and Qin 2017], the authors use the triweight kernel density function $K(t) = \frac{35}{32}(1-t^2)^3 I(|t| \leq 1)$ in order to evaluate *LIP* for data relating to annual salaries of Professors, Associate Professors and Assistant Professors, employed in the Units of University System or in Military Colleges from a State of Georgia, U.S., during the 2012 fiscal year. Our primary objective is to use some chosen estimation procedures for both the point and the interval estimation of *LIP* and *ARPR*, for evaluation of *ARPR* on the basis of dataset

containing 13057 observations of an equivalised disposable income of Polish households in 2003 from [Statistical Publishing Establishment, Warsaw, 2004]. Above all, we apply and develop the methods proposed by [Luo and Qin 2017] and [Zieliński 2009a-b], as - in our view - these approaches include the most valuable and reliable ideas leading to the assessment of *ARPR* and to the evaluation of other similar poverty (or social inequality) measures. The remainder of our paper is structured as follows. In Section SELECTED CONCEPTS OF THE LOW-INCOME PROPORTION AND THE AT-RISK-OF-POVERTY RATE ESTIMATION, we introduce the essential concepts of *LIP* and *ARPR* assessment, which we later aim to implement in our empirical analyses. In Section EMPIRICAL STUDY, we present the details regarding computational techniques that allow for the corresponding point and interval estimation, as well as we conduct our empirical research concerning the point and interval evaluation of the mentioned risk measures for a dataset containing information on an equivalised disposable income of households in Poland from the year 2003. Finally, Section SUMMARY summarizes and concludes our study. All of our computations have been carried out using the software environment R.

SELECTED CONCEPTS OF THE LOW-INCOME PROPORTION AND THE AT-RISK-OF-POVERTY RATE ESTIMATION

Let $X_1, X_2, \dots, X_n$ be a simple random sample from the income distribution and $X_{1:n} \leq \cdots \leq X_{n:n}$ denote a sequence of the corresponding order statistcs. In view of the definition of the *Low-Income Proportion (LIP)* measure from the previous Section, the estimator of $\theta_{\alpha,\beta} = LIP$ may be defined as

$$\hat{\theta}_{\alpha,\beta} = \frac{1}{n}\#\{X_i\colon X_i \leq \alpha \cdot X_{M:n}\},$$

where $M = \lfloor \beta n \rfloor + 1$ (with $\lfloor x \rfloor$ denoting the largest integer not exceeding x) and # stands for the cardinality of a given set; obviously, $X_{M:n}$ is an estimator for the β-th quantile ξ_β of an income distribution. Therefore, an estimator of the *At-Risk-of-Poverty Rate (ARPR)* may be given by putting $\alpha = 0.6$ and $M = \lfloor 0.5n \rfloor + 1$ into the formula above, i.e. it can be determined as

$$\widehat{ARPR} = \hat{\theta}_{0.6,0.5} = \hat{\theta} = \frac{1}{n}\#\{X_i\colon X_i \leq 0.6 \cdot X_{(\lfloor 0.5n \rfloor+1):n}\}.$$

(Clearly, $X_{(\lfloor 0.5n \rfloor+1):n}$ is an estimator for a median of an income distribution)

As the first example of a confidence interval for $\theta_{\alpha,\beta} = LIP$ (and consequently for $\theta_{0.6,0.5} = ARPR$), we wish to examine an interval constructed in [Zieliński 2009b]. Namely, let ξ be the number of those among $X_1, X_2, \dots, X_n$, which are not greater than $\alpha \cdot X_{(\lfloor \beta n \rfloor+1):n}$, i.e.

$$\xi = \#\{X_i\colon X_i \leq \alpha \cdot X_{(\lfloor \beta n \rfloor+1):n}\}.$$

Then, assuming the fixed confidence level $\gamma \in (0,1)$, the following confidence interval for $\theta_{\alpha,\beta} = LIP$ has been introduced in [Zieliński 2009a]

$$(\beta \cdot B^{-1}(\xi, M-\xi+1; \frac{1-\gamma}{2});\ \beta \cdot B^{-1}(\xi+1, M-\xi; \frac{1+\gamma}{2})),$$

where: $M = \lfloor \beta n \rfloor + 1$, and $B^{-1}(a,b;q)$ denotes a quantile of order q for the beta distribution with parameters a, b. Thus, as a straightforward conclusion, we can write that an interval below is the corresponding confidence interval for $\theta_{0.6,0.5} = \theta = ARPR$

$$(l_0; u_0) =$$
$$\left(0.5 \cdot B^{-1}\left(\tilde{\xi}, \widetilde{M} - \tilde{\xi} + 1; \frac{1-\gamma}{2}\right);\ 0.5 \cdot B^{-1}\left(\tilde{\xi}+1, \widetilde{M} - \tilde{\xi}; \frac{1+\gamma}{2}\right)\right),$$

where:

$$\widetilde{M} = \lfloor 0.5n \rfloor + 1,\ \tilde{\xi} = \#\{X_i\colon X_i \le 0.6 \cdot X_{\widetilde{M}:n}\}.$$

As it has already been mentioned in our preliminary Section, the empirical estimate of $\theta_{\alpha,\beta} = LIP$ may be given by

$$\hat{\theta}_{\alpha,\beta} = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \le \alpha\hat{\xi}_\beta).$$

Furthermore, in view of [Preston 1995], this estimate satisfies the following property

$$\sqrt{n}\left(\hat{\theta}_{\alpha,\beta} - \theta_{\alpha,\beta}\right) \to N\left(0, \sigma^2_{\alpha,\beta}\right),$$

where

$$\sigma^2_{\alpha,\beta} = \theta_{\alpha,\beta}\left(1-\theta_{\alpha,\beta}\right) - 2\alpha(1-\beta)\theta_{\alpha,\beta}\frac{f\left(\alpha\xi_\beta\right)}{f\left(\xi_\beta\right)} + \alpha^2\beta(1-\beta)[\frac{f\left(\alpha\xi_\beta\right)}{f\left(\xi_\beta\right)}]^2,$$

with f standing for density of the corresponding income distribution.

Obviously, it means that $\theta_{\alpha,\beta}$ is asymptotically normal and hence, the following $(1-\delta)$-level normal approximation-based confidence interval for $\theta_{\alpha,\beta}$ may be established

$$(l_1; u_1) = (\hat{\theta}_{\alpha,\beta} - \frac{z_{1-\delta/2} \cdot \hat{\sigma}_{\alpha,\beta}}{\sqrt{n}};\ \hat{\theta}_{\alpha,\beta} + \frac{z_{1-\delta/2} \cdot \hat{\sigma}_{\alpha,\beta}}{\sqrt{n}}),$$

where $z_{1-\delta/2}$ stands for the $(1-\delta/2)$-th quantile of the standard normal distribution and $\hat{\sigma}_{\alpha,\beta}$ denotes a consistent estimator of the standard deviation $\sigma_{\alpha,\beta}$. However, since - as it has already been noted in our introductory part - $\hat{\theta}_{\alpha,\beta}$ is a non-smoothing estimate of $\theta_{\alpha,\beta}$, another approach, adopting the concept of the kernel estimation, has been proposed in order to obtain the smoothed estimator of *LIP*. Based on a simple random sample $X_1, X_2, \ldots, X_n$, the corresponding kernel estimate for $\theta_{\alpha,\beta}$ is determined as follows

$$\hat{T}_n(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\alpha\hat{\xi}_\beta - X_i}{h}\right),$$

where: K, h are the selected kernel and the stated bandwidth, respectively, and $\hat{\xi}_\beta$ is the β-th empirical quantile of the considered distribution.

Due to Theorem 2.1 in [Luo and Qin 2017], we directly get that $\sqrt{n}\big(\hat{T}_n(\alpha,\beta) - \theta_{\alpha,\beta}\big) \to N\big(0,\sigma^2_{\alpha,\beta}\big)$, where $\sigma^2_{\alpha,\beta}$ is the same as earlier. Thus, the following $(1-\delta)$-level normal approximation-based confidence interval for $\theta_{\alpha,\beta}$ may be obtained

$$(l_2;u_2) = (\hat{T}_n(\alpha,\beta) - \frac{z_{1-\delta/2}\cdot\hat{\sigma}_{\alpha,\beta}}{\sqrt{n}}; \hat{T}_n(\alpha,\beta) + \frac{z_{1-\delta/2}\cdot\hat{\sigma}_{\alpha,\beta}}{\sqrt{n}}).$$

The definition of smoothed estimator $\hat{T}_n(\alpha,\beta)$ is applied in establishing the so-called smoothed log jackknife empirical likelihood ratio statistic for *LIP* and later, for creating the corresponding confidence interval. In order to introduce the estimation concepts leading to both of the mentioned constructions, it is needed to define the so-called jackknife pseudo-values for *LIP*. By [Tukey 1958], the jackknife pseudo-values for *LIP* are defined as follows

$$\hat{V}_k(\alpha,\beta) = n\hat{T}_n(\alpha,\beta) - (n-1)\hat{T}_{n-1,k}(\alpha,\beta), \quad k = 1,2,\dots,n,$$

where: $\hat{T}_{n-1,k}(\alpha,\beta) = \frac{1}{n-1}\sum_{j\neq k}^{n} K\left(\frac{\alpha\hat{\xi}_{\beta,-k} - X_j}{h}\right)$ refers to the determined smoothed estimator $\hat{T}_n(\alpha,\beta)$, but it is computed on $n-1$ observations $X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_n$, and $\hat{\xi}_{\beta,-k} = F^{-1}_{n,-k}(\beta)$ is the β-th quantile from an empirical distribution $F_{n,-k}(x) = \frac{1}{n-1}\sum_{j\neq k}^{n} I\left(X_j \le x\right)$, based on $n-1$ observations (i.e., on all observations except for the *k*-th one).

Using the jackknife pseudo-values $\hat{V}_k(\alpha,\beta)$, $k = 1, \dots n$, we may define the log jackknife empirical likelihood ratio statistic in the form as below

$$l_n\big(\theta_{\alpha,\beta}\big) = -2\log L_n\left(\theta_{\alpha,\beta}\right) = 2\sum_{k=1}^{n} \log\{1 + \lambda\big(\hat{V}_k(\alpha,\beta) - \theta_{\alpha,\beta}\big)\},$$

where $L_n\big(\theta_{\alpha,\beta}\big)$ denotes the jackknife empirical ratio statistic for $\theta_{\alpha,\beta}$ and $\lambda = \lambda\big(\alpha,\beta,\theta_{\alpha,\beta}\big)$ is the solution to

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\hat{V}_k(\alpha,\beta) - \theta_{\alpha,\beta}}{1 + \lambda\big(\hat{V}_k(\alpha,\beta) - \theta_{\alpha,\beta}\big)} = 0.$$

It is known that, under certain conditions, $l_n\big(\theta_{\alpha,\beta}\big) \to \chi^2(1)$, where $\chi^2(1)$ stands for the chi-squared distribution with one degree of freedom. Thus, the $(1-\delta)$-level confidence interval for $\theta_{\alpha,\beta} = LIP$ may be given in the form

$$(l_3;u_3) = \{\theta: l_n(\theta) \le \chi^2_{1,1-\delta}\},$$

where $\chi^2_{1,1-\delta}$ denotes the $(1-\delta)$-th quantile of the $\chi^2(1)$ distribution. It is worthwhile to mention here that the variance $var(\sqrt{n}\hat{T}_n(\alpha,\beta))$ can be estimated by the sample variance of jackknife pseudo-values $\{\hat{V}_1(\alpha,\beta),\dots\hat{V}_n(\alpha,\beta)\}$ and that the jackknife variance estimator of $T_n(\alpha,\beta)$ is determined as

$$v_{JACK}(\alpha,\beta) = var\left(\sqrt{n}\hat{T}_n(\alpha,\beta)\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta) - \frac{1}{n}\sum_{j=1}^{n}\hat{V}_j(\alpha,\beta))^2.$$

In view of Theorem 3.1 in [Luo and Qin 2017], we get $v_{JACK}(\alpha,\beta) \to \sigma^2_{\alpha,\beta}$, where $\sigma^2_{\alpha,\beta}$ is the same as in the earlier considerations. Thus, the following $(1-\delta)$-level normal approximation-based confidence interval for $\theta_{\alpha,\beta}$ may be introduced

$$(l_4;u_4) = (\hat{T}_n(\alpha,\beta) - \frac{z_{1-\delta/2}\cdot\sqrt{v_{JACK}(\alpha,\beta)}}{\sqrt{n}}; \hat{T}_n(\alpha,\beta) + \frac{z_{1-\delta/2}\cdot\sqrt{v_{JACK}(\alpha,\beta)}}{\sqrt{n}}).$$

The confident intervals $(l_1;u_1)$, $(l_2;u_2)$ and $(l_4;u_4)$ are established on the basis of normal approximation theorems. Such the approximation-based confidence intervals may perform poorly for the estimation of income-related ratios, since the income datasets tend to be skewed or have outliers. In order to overcome this drawback, another technique that enables to create the confidence intervals for the rates like *LIP*, and *ARPR* in particular, has been proposed in the case when asymptotic variance of the corresponding point estimator is unknown. This idea - called the bootstrap method - is due to [Efron 1979] and has become a celebrated estimation approach in recent decades. With reference to Efron's design, [Luo and Qin 2017] combined the bootstrap approach with the kernel estimation in order to obtain appropriate confidence intervals for $\theta_{\alpha,\beta}$. The concept introduced in the cited work of [Luo and Qin 2017] may be depicted as follows. Namely, assume that $(X_1^*, X_2^*, \dots, X_n^*)$ is a bootstrap sample from the original sequence $(X_1, X_2, \dots, X_n)$, i.e. $(X_1^*, X_2^*, \dots, X_n^*)$ is repeatedly drawn, with replacement, from $(X_1, X_2, \dots, X_n)$. Then, the bootstrap equivalent of the kernel estimate $\hat{T}_n(\alpha,\beta)$ is given by

$$\hat{T}_n^*(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\alpha\hat{\xi}_\beta^* - X_i^*}{h}\right).$$

After repeating the bootstrap procedure $B \geq 500$ times, i.e. after drawing $B \geq 500$ bootstrap samples $(X_{1b}^*, X_{2b}^*, \dots, X_{nb}^*)$, where $b = 1, \dots, B$, B bootstrap copies $\{\hat{T}_{nb}^*(\alpha,\beta)\}_{b=1,\dots,B} = \{\hat{T}_b^*\}_{1,\dots,B}$, of the estimate $\hat{T}_n(\alpha,\beta)$, are computed. Finally, based on the obtained bootstrap replicates $\{\hat{T}_1^*, \hat{T}_2^* \dots, \hat{T}_B^*\}$, the following $(1-\delta)$-level bootstrap kernel-based confidence intervals for $\theta_{\alpha,\beta}$ are established:

$$(l_5;u_5) = (\hat{T}_n(\alpha,\beta) - z_{1-\delta/2}\cdot\sqrt{V_T^*}; \hat{T}_n(\alpha,\beta) + z_{1-\delta/2}\cdot\sqrt{V_T^*}),$$

$$(l_6;u_6) = (\bar{T}^* - z_{1-\delta/2}\cdot\sqrt{V_T^*}; \bar{T}^* + z_{1-\delta/2}\cdot\sqrt{V_T^*}),$$

where: $\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{T}_b^*$, $V_T^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{T}_b^* - \bar{T}^*\right)^2$.

In the subsequent Section, we aim to use the above discussed point and interval estimation procedures in order to evaluate *ARPR* for data containing the equivalised disposable incomes of households from Poland, gained for the year 2003. We wish to pay our special attention to the issue of *ARPR* estimation methods which apply the non-smoothing kernel-based designs, in particular to those methods which combine the kernel estimation with the jackknife resampling technique.

EMPIRICAL STUDY

In our computations, we consider a sample of data $(x_1, x_2, \dots, x_n)$, comprising of n = 13507 observations referring to an equivalised disposable income of the Polish households in the year 2003, collected from [Statistical Publishing Establishment, Warsaw, 2004]. Before we compute the realizations of confidence intervals for *ARPR*, we need to check whether the given observations come from a random simple sample. For this reason, we apply the so-called *Runs Test*. Based on our dataset and assuming the most common confidence level 0.95, we obtain the value of test statistic $U = 0.4025805$ and the critical (rejection) region $(-\infty; -1.96 > \cup < 1.96; \infty)$. Consequently, we do not reject the null hypothesis that our observations come from a random simple sample. Thus, we may proceed to computation of the empirical confidence intervals for *ARPR* (i.e., to calculation of these confidence intervals realizations). As we have already mentioned, the point estimate of *ARPR* may be expressed in the form $\widehat{ARPR} = F_n\left(0.6 \cdot \hat{\xi}_{0.5}\right) = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \leq 0.6 \cdot \hat{\xi}_{0.5})$, or by $\widehat{ARPR} = \frac{1}{n}\#\{X_i\colon X_i \leq 0.6 \cdot X_{(\lfloor 0.5n \rfloor + 1):n}\}$.

For our dataset, we have:

$$\widetilde{M} = (\lfloor 0.5n \rfloor + 1) = 6529,\ X_{\widetilde{M}:n} = \hat{\xi}_{0.5} = F_n^{-1}(0.5) = 5570.073,$$
$$\tilde{\xi} = \#\{X_i\colon X_i \leq 0.6 \cdot X_{\widetilde{M}:n}\} = 2083,$$

and hence, $\widehat{ARPR} = \hat{\theta}_{0.6,0.5} = 0.1595$.

Due to the earlier given formula for the confidence interval $(l_0; u_0)$, introduced by [Zieliński 2009a], we immediately obtain the following confidence interval for *ARPR*, at the confidence level $\gamma = 0.95$,

$$(l_0; u_0) = (0.5 \cdot B^{-1}(\tilde{\xi}, \widetilde{M} - \tilde{\xi} + 1; \frac{1-\gamma}{2});\ 0.5 \cdot B^{-1}(\tilde{\xi} + 1, \widetilde{M} - \tilde{\xi}; \frac{1+\gamma}{2})) = (0.1539; 0.1652).$$

Furthermore, it is easy to compute the realization of 95% (0.95-level) normal approximation-based confidence interval for $ARPR = \theta_{0.6,0.5}$. Namely, for $\alpha = 0.6$ and $\beta = 0.5$, we obtain that

$$(l_1; u_1) = (\hat{\theta}_{0.6,0.5} - \frac{z_{1-0.05/2} \cdot \hat{\sigma}_{0.6,0.5}}{\sqrt{13057}};\ \hat{\theta}_{0.6,0.5} + \frac{z_{1-0.05/2} \cdot \hat{\sigma}_{0.6,0.5}}{\sqrt{13057}}) = (0.1537; 0.1654).$$

Among the realizations of the kernel-based confidence intervals $(l_2; u_2)$-$(l_5; u_5)$, we shall consider the realization of $(l_3; u_3)$ first, as it is recommended by [Luo and Qin 2017], since the empirical study conducted there shows that, among the presented intervals, $(l_3; u_3)$ displays the best statistical performance in terms of coverage probabilities. Out of all the introduced confidence intervals for *ARPR*, $(l_3; u_3)$ seems to be relatively the most difficult one to evaluate, as a technique leading to the construction of $(l_3; u_3)$ combines the kernel estimation with the jackknife resampling concept. In particular, computing the realization of $(l_3; u_3)$ requires the selection of an appropriate kernel function K together with its bandwidth h. Although, it has been checked that the choice of K does not affect the accuracy of a calculated estimate too much, various research studies exhibit that it is not so in case of the bandwidth selection, since it has been shown that the change of h may have a significant impact on the estimator value. Thus, following the suggestions from [Luo and Qin 2017], we apply the triweight kernel $K(t) = \frac{35}{32}(1 - t^2)^3 I(|t| \leq 1)$ and implement a bandwidth $c \cdot n^{-1/3}$, where a constant c is selected on the grounds of the two-fold cross-validation method. The procedure leading to the calculation of c involves employing several steps, which may be described as follows.

Step 1°. We randomly split the given sample into two parts of possibly equal size, the first of which is the training sample, while the second is treated as the test sample;

Step 2°. Based on the training sample, we compute the kernel estimate $\hat{T}_{n,c}^{(1)}(\alpha = 0.6, \beta = 0.5) = \hat{T}_{n,c}^{(1)}(0.6,0.5)$ for $ARPR$ and based on the test sample, we compute its empirical estimate $\hat{\theta}^{(2)}_{\alpha=0.6,\beta=0.5} = \hat{\theta}^{(2)}_{0.6,0.5}$;

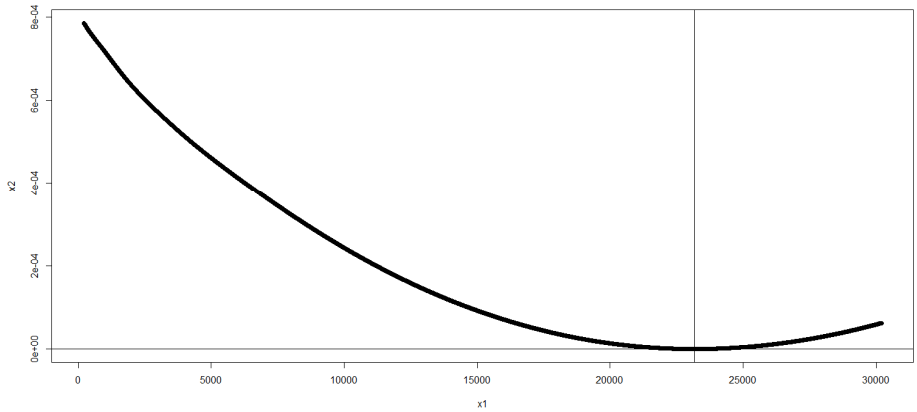
Step 3°. We repeat the random split and the computation, described in the previous steps, $L \geq 30$ times and obtain a set of L pairs consisting of the kernel estimates $\hat{T}_{n,c}^{(1,l)}(0.6,0.5)$ and the empirical estimates $\hat{\theta}^{(2,l)}_{0.6,0.5}$, where $l = 1, \dots, L$ (i.e., we get a set $\{\left(\hat{T}_{n,c}^{(1,l)}(0.6,0.5), \hat{\theta}^{(2,l)}_{0.6,0.5}\right) : l = 1, \dots, L\}$);

Step 4°. We choose a constant c by minimizing the following cross-validation estimate of *MSE*

$$CV_c = \frac{1}{n}\sum_{l=1}^{L}\left[\hat{T}_{n,c}^{(1,l)}(0.6,0.5) - \hat{\theta}^{(2,l)}_{0.6,0.5}\right]^2.$$

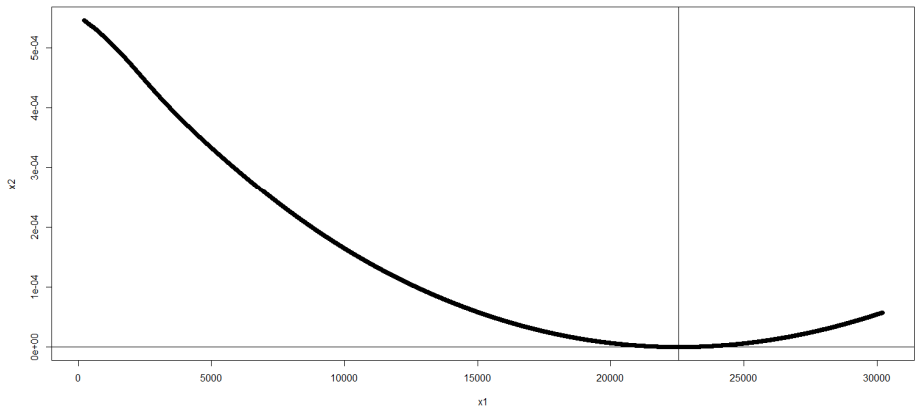
We conducted the steps 1°– 4° above for our data - composed of the equivalised disposable incomes of 13057 Polish households from the year 2003 - for the cases when: $L = 30,40,50,60$. As a result, CV_c reached its minimum for: $c = 23174.6$ - if $L = 30$, $c = 22553.6$ - if $L = 40$, $c = 23240.6$ - if $L = 50$, $c = 24634.1$ - if $L = 60$. That can be illustrated in the figures below.

Figure 1. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 30; c_{min} = 23174.6



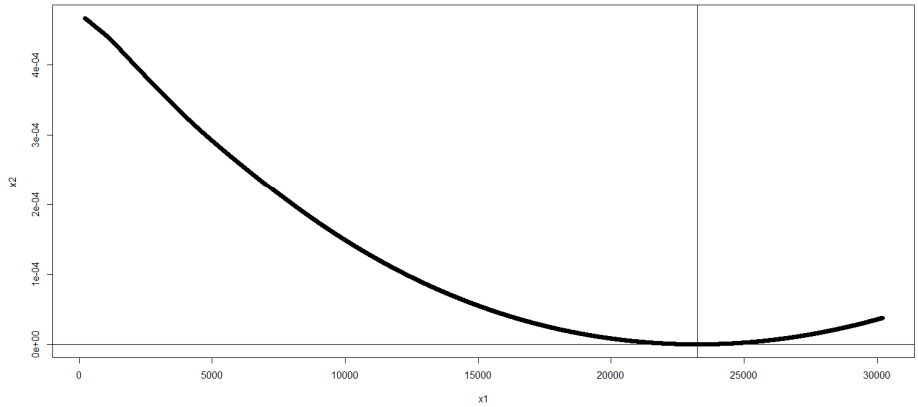
Source: own elaboration

Figure 2. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 40; c_{min} = 22553.6



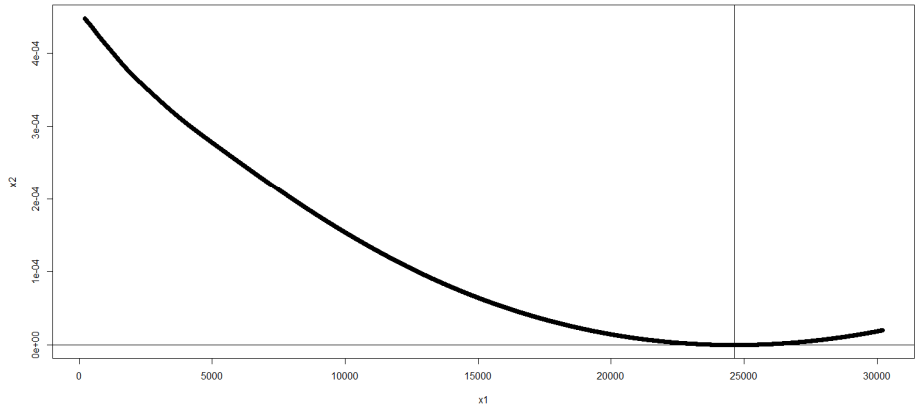
Source: own elaboration

Figure 3. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 50; c_{min} = 23240.6



Source: own elaboration

Figure 4. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 60; c_{min} = 24634.1



Source: own elaboration

Consequently, since the values for c_{min} were computed based on the subsamples of size $\lceil n/2 \rceil = 6529$, then - according to the recommendation in [Luo and Qin 2017] - we applied the following values for bandwidths: $h = 23174.6 \cdot (6529)^{-1/3} = 1239.929$, or $h = 22553.6 \cdot (6529)^{-1/3} = 1206.703$, or $h = 23240.6 \cdot (6529)^{-1/3} = 1243.46$, or $h = 24634.1 \cdot (6529)^{-1/3} = 1318.017$, for: $L = 30,40,50,60$, respectively. In Table 1 below, we collected the 95% (0.95-level) realizations of the kernel-based confidence intervals $(l_2; u_2)$-$(l_6; u_6)$ for $\theta_{\alpha,\beta}$. We

limited ourselves to the case when $\alpha = 0.6$ and $\beta = 0.5$, i.e. to evaluation of the realizations of confidence intervalas for *ARPR*. The values of B from this table denote the sizes of bootstrap copies used in computations of the bootstrap kernel-based confidence intervals $(l_5; u_5)$-$(l_6; u_6)$.

Table 1. The 95% realizations of the selected kernel-based confidence intervals for ARPR (the bootstrap kernel-based realizations $(l_5; u_5)$-$(l_6; u_6)$ were obtained for: B = 500 - if: L = 30, 40, 50, or B = 1000 - if L = 60)

I \ *L*	30	40	50	60
$(l_2; u_2)$	0.1481-0.1599	0.1443-0.1560	0.1486-0.1603	0.1572-0.1689
$(l_3; u_3)$	≤ 0.1548	≤ 0.1520	≤ 0.1544	≤ 0.1604
$(l_4; u_4)$	0.1540-0.1541	0.1501-0.1502	0.1544-0.1545	0.1630-0.1631
$(l_5; u_5)$	0.1433-0.1647	0.1398-0.1605	0.1439-0.1650	0.1540-0.1721
$(l_6; u_6)$	0.1428-0.1642	0.1393-0.1601	0.1435-0.1646	0.1539-0.1720

Source: own elaboration

SUMMARY

The main goal of our study was to apply some selected estimation procedures in evaluation of the *Low-Income Proportion* (*LIP*) and *At-Risk-of-Poverty Rate* (*ARPR*) measures. We primarily focused on interval estimation of the *ARPR* index and computed the realizations of the corresponding confidence intervals for dataset consisting of 13057 observations referring to an equivalised disposable income of households in Poland from the year 2003, gained from [Statistical Publishing Establishment, Warsaw, 2004]. As recommended in [Luo and Qin 2017], we mainly considered the kernel-based confidence intervals. It may be easily seen that, depending on the number L (which is the number of random splits into the training and test samples), the ranges of lower/upper limits of the selected confidence intervals are (with an exception of the realizations of $(l_3; u_3)$) as follows: (i) if $L = 30$, then the lower limits of the obtained realizations range from 0.1428 to 0.1540 and its upper limits range from 0.1541 to 0.1647, (ii) if $L = 40$, then the lower limits of the obtained realizations range from 0.1393 to 0.1501 and its upper limits range from 0.1502 to 0.1605, (iii) if $L = 50$, then the lower limits of the obtained realizations range from 0.1435 to 0.1544 and its upper limits range from 0.1545 to 0.1650, (iv) if $L = 60$, then the lower limits of the obtained realizations range from 0.1539 to 0.1630 and its upper limits from 0.1631 to 0.1721. Furthermore, comparing the obtained realizations of the kernel-based confidence intervals $(l_2; u_2)$ and $(l_4; u_4)$-$(l_6; u_6)$ for *ARPR* with its empirical estimate $\widehat{ARPR} = \hat{\theta}_{0.6,0.5} = 0.1595$, we observe that: three out of four computed realizations of $(l_2; u_2)$ contain $\widehat{ARPR}$, none of four computed realizations of $(l_4; u_4)$ contains $\widehat{ARPR}$, all of four computed realizations of $(l_5; u_5)$ contain $\widehat{ARPR}$, and also all of four computed realizations of $(l_6; u_6)$ contain $\widehat{ARPR}$. Thus, it seems reasonable to limit our

attention to interpretation of the obtained realizations of confidence intervals $(l_2; u_2)$ and $(l_5; u_5)$-$(l_6; u_6)$. If we do so, then - taking into account all of the considered numbers of iterations L - we observe that the lower limits of these realizations range from 0.1393 to 0.1572, whereas the upper ones are between 0.1560 and 0.1721. Obviously, if we average the minimum and maximum of the considered lower limits range, we get an average 0.148 and by averaging the minimum and maximum of the considered upper limits range, we have an average 0.164. Thus, roughly speaking, we may state that, based on the obtained 95% realizations of selected confidence intervals for $ARPR$, the $ARPR$ measure ranges, on average, between 0.148 and 0.164. In other words, we may claim that with high probability, the percentage of Polish households with the equivalised disposable incomes not exceeding 60% of the whole population median amounted between 15% and 16% in the year 2003. That was the last year before Poland's entry to the European Union and a natural question arises, whether $ARPR$ has changed throughout the years since Poland has become a member of the EU. An answer to this question has been - at least partially - delivered in the report from [Eurostat Statistics Explained 2020] (which is an electronic publishing platform containing Eurostat's statistical information). It shows that in 2019, the $ARPR$ measure in Poland was between 15% and 16% - approximately the same in range that we obtained for the year 2003 using the chosen procedures of interval estimation. Thus, we may conclude that the percentages of low earners in Poland in the years 2003 and 2019 were roughly similar. It may seem slightly unusual that a reliable poverty measure was the same both in the year directly preceding Poland's accession to the EU and 15 years after that, especially since the results presented in [Eurostat Statistics Explained 2020] were computed for data including social transfers. It would be vital to study this issue in our further research. Also, it would be worthwhile to estimate $ARPR$ for dataset covering a period when various Coronavirus lockdown rules have been introduced. Directly before this period, the estimated value of this index for Poland, amounting to between 15% and 16%, has ranked Poland, in the group of EU countries with the $ARPR$ measure below the EU average of 21.1% and we think it would be desirable to check whether it is also the case after almost two turbulent years of SARS-CoV-2 era.

REFERENCES

Bowman A. W., Hall P., Prvan T. (1998) Cross-Validation for the Smoothing of Distribution Functions. Biometrika, 85, 799-808.

Efron B. (1979) Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7, 1-26.

Eurostat Statistics Explained (2020) File: At-Risk-of-Poverty Rate and At-Risk-of-Poverty Threshold 2019 LCIE20.png, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:At-risk-of-poverty_rate_and_at-risk-of-poverty_threshold,_2019_LCIE20.png.

Eurostat (Your Key to European Statistics) (2020) https://ec.europa.eu/eurostat/web/products-datasets/-/tespm010.

Falk M. (1983) Relative Efficiency and Deficiency of Kernel Type Estimators of Smooth Distribution Functions. Statistica Neerlandica, 37, 73-83.

Falk M. (1985) Asymptotic Normality of the Kernel Quantile Estimator. The Annals of Statistics, 13, 428-433.

Gong Y., Peng Y., Qi Y. C. (2010) Smoothed Jackknife Empirical Likelihood Method for ROC Curve. Journal of Multivariate Analysis, 101, 1520-1531.

Jing B., Yuan J., Zhou W. (2009) Jackknife Empirical Likelihood. Journal of American Statistical Association, 104, 1224-1232.

Li Z., Gong Y., Peng L. (2011) Empirical Likelihood Intervals for Conditional Value-at-Risk in Heteroscedastic Regression Models. Scandinavian Journal of Statistics, 38, 781-787.

Luo S., Qin G. (2017) New Non-Parametric Inferences for Low-Income Proportions. Annals of the Institute of Statistical Mathematics, 69(3), 599-626.

Preston I. (1995) Sampling Distributions of Relative Poverty Statistics. Applied Statistics, 44, 91-99 [Correction, 45, 399 (1996)].

Statistical Publishing Establishment, Warsaw (2004) Household Budget Surveys in 2003.

Tukey J. W. (1958) Bias and Confidence is Not-Quite Large Sample. The Annals of Mathematical Statistics, 29, 614.

Wei Z., Wen S., Zhu L. (2009) Empirical Likelihood-Based Evaluations of Value at Risk Models. Science in China Series A, Mathematics, 52, 1995-2006.

Wei Z., Zhu L. (2010) Evaluation of Value at Risk: An Empirical Likelihood Approach. Statistica Sinica, 20, 455-468.

Zieliński W. (2009a) A Nonparametric Confidence Interval for At-Risk-of-Poverty-Rate. Statistics in Transition, new series, 10 (3), 437-444.

Zieliński W. (2009b) A Nonparametric Confidence Interval for At-Risk-of-Poverty-Rate: Example of Application. Polish Journal of Environmental Studies, 18 (5B), 217-219.