International Journal of Economics, Commerce and Management
United Kingdom ISSN 2348 0386 Vol. IX, Issue 1, Jan 2021

http://ijecm.co.uk/

VARIANCE ESTIMATION FOR THE INCOME AND LIVING CONDITIONS SURVEY

Liljana BOÇI
Statistical Institute of Albania
lboci@instat.gov.al

Abstract

The European Union Statistics on Income and Living Conditions (EU-SILC) aims at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. Albania has adapted the methodology of the SILC to its national conditions. Up to now poverty indicators were published at the national level and at the level of particularly interesting domains defined by age class, gender and employment status. In order to judge the accuracy of the figure for the main indicators of this survey, different approximations of the variance calculation were taken into account with the aim of estimating them accurately taking into account the whole complexity of being in the conditions of a complex survey of stratified and important non-linear indicators such as poverty, respectively the Poverty Risk Rate (ARPR) and the Poverty-or-Exclusion Risk (AROPE). For the ARPR the variance estimation with the naive estimator and the simple Bootstrap estimator at the prefecture level and at the level of the considered domains are very close.

Keywords: EU-SILC, at-risk-of-poverty-rate, at-risk-of-poverty-or-social-exclusion, Variance Estimation, Accuracy of Poverty Indicators

INTRODUCTION

Nowadays, users of statistical information are demanding high quality data on social, demographic, industrial, economic, financial, political, and cultural aspects of society with a great level of detail. National statistical institutes fulfill a central role in providing such high quality statistical information. Income and Living Conditions Survey (EU-SILC), European

Commission (2019), is source of comparative statistics on income and living conditions in Member States of the European Union.

The European Statistics on Income and Living Conditions (EU-SILC) project was launched for the first time in Albania in 2017 and has been carried out on annual basis. This survey is the main source of national statistics on income distribution, poverty and social exclusion. In this survey, data are collected at household level:

- Income- Gross income components at household level
- Housing- Dwelling type, tenure status and housing conditions, Housing costs
- Social Exclusion- Non-monetary household deprivation indicators, including problems in making ends meet, extent of debt and enforced lack of basic necessities, Physical and social environment

The data collections in survey are a potential source of errors. The main problem is the estimation of the variance taking into account the complex stratified and clustered sample design and the non-linear estimators of important poverty indicators, namely the at-risk-of-poverty-rate (ARPR) and the at-risk-of-poverty-or-social-exclusion (AROPE). The study is carried out using the statistical software environment R Core Team (2020) and as the main function for the design the R-package survey is used.

VARIANCE ESTIMATION

The Primary Selection Unit (PSU) carries the information about the population already: the number of the households in the stratum and the number of households per PSU. Theoretically it is possible that a PSU is selected twice or even up to four times (due to the rotation of the sample). This is neglected to not complicate the cross-sectional analysis. To calculate the inclusion probabilities we take into account the realized number of PSU in a stratum (nsh) which is matched to each person in the sample.

The number of households in a PSU is M_{hi}. This information must be present on the data file to be able to calculate the psu inclusion probability. In addition, the number of households in the stratum must be known. The psu inclusion probabilities can be calculated as: $\pi_{hi1} = n_{hi} M_{hi} / \sum_{i \in Uh} M_{hi}$.

Theoretically for the sample design the household splits must not be taken into account. However, for the calculation of indicators and for calibration hh_split must be taken into account. The assumption is that within a psu (identified by the sample psu number) the household number is unique (say no two hh have the same number). Thus the psu number in the sample can be used to identify a hh uniquely. The inclusion probabilities of the ssu (π_{hij2}) are calculated assuming simple random sampling of ssu within psu. The net sample sizes are taken into account, i.e. $\pi_{hij2} = m_{hij}/M_{hi}$.

Thus the inclusion probability of a household j in psu i in stratum h is $\pi_{hi1} \cdot \pi_{hij2}$. This is also the inclusion probability of any person in that household. The sample has been reweighted to adjust to the population margins in terms of age, gender etc. The resulting weights w_{hij} are derived from the inclusion probabilities by calibration. The svydesign function of package survey accepts both inclusion probabilities (mainly to calculate inclusion probabilities, but also for the calculation of variance estimators) and the weights.

SURVEY DESIGN IMPLEMENTATION

The survey package of R is used to estimate the means and variances, T. Lumley (2020). The impact of the calibration to known population totals on the variance estimation is neglected here, though the correct weights are used to ensure that the point estimators coincide with simple weighted means. However the splitting of households is not taken into account. The option for adjusting the variance estimator for domains where only a single PSU is in the stratum centers the stratum at the estimate for the population mean. The variance of proportions for deprivation, severe deprivation, and low work intensity (Deprivation, Severe Deprivation, and Lower Working Intensity) can be estimated with this set-up. Also the variance of domain estimates can be estimated.

The point estimates may slightly deviate from the results when household splits are taken into account. For the variance estimation this will have a minimal impact since only very few households actually split.

Variance estimators

Means and proportions can be calculated with closed form variance estimators. To avoid double inclusion probabilities for the IPPS design the Brewer approximation is used in the survey design. While a variance for the median equalized income can be derived from the above confidence interval it is more involved to get variance estimation of the at-risk-of-poverty-rate (ARPR) and of the at-risk-of-poverty-or-exclusion (AROPE). Due to the correlation between the at-risk-of-poverty-threshold (ARPT) and the proportion below it, as well as with the other variables involved in AROPE (Severe Deprivation and Lower Working Intensity) both ARPR and AROPE need special attention. Once the indicator for ARPR or AROPE is calculated, a naive estimator would just use the above method for linear estimator. However to check the impact of the ARPT a comparison with two Bootstrap methods was investigated.

Check with Bootstrap

The direct variance calculation can be compared with Bootstrap variance estimates. This is mainly a test for the Brewer approximation to avoid the necessity of double inclusion probabilities. However, the Bootstrap versions available in the survey package do not cover multistage designs with unequal PSU inclusion probabilities. Preston's multi-stage bootstrap is used, Preston, J. (2009). The bootstrap by Canty, Davison, Hinkley and Ventura (2002), seems not to work. The sampling is assumed simple random sampling of PSU. With 500 replicates the variability of the Boostrap-Variance estimate seems to be sufficiently small.

Explicit calculation of replicates for non-linear estimators

The main problem with the variance calculation of the at-risk-of-poverty rate (ARPR) and of the at-risk-of-poverty-and-social-exclusion rate (AROPE) is the at-risk-of-poverty threshold (ARPT) has its own variability and is correlated with ARPR and AROPE. The effect of the correlation is clearly visible in smaller populations like prefectures or domains of study according to gender, age or employment.

We call the variance estimator that assumes the threshold is fix, i.e. that neglects the variability of ARPT and, hence, the correlation with the proportion below the threshold, naive estimator.

To investigate the effect of the variance of the median on the ARPR the bootstrap weights must be made explicit. The boostratp weights are extracted from the replicate design and the loop over the replicates is programmed fully. The bootstrap weights add up (approximately) to the number of observations in the sample. To get extrapolation weights the bootstrap weights are multiplied with the sum of the original calibrated extrapolation weight. Then a new svydesign object is created with those new weights and the median, the threshold and the indicator are calculated under this bootstrap replicate design.

The variance of the ARPR can be calculated at the level of a prefecture in several ways, depending on the definition of the threshold:

- Naive variance estimator with national threshold fixed,
- Bootstrap variance estimator with national threshold fixed,
- Naive variance estimator with prefecture threshold fixed,
- Bootstrap variance estimator with prefecture threshold fixed (uses svyrepdesign),
- Bootstrap variance estimator with prefecture threshold recalculated for each replicate (uses explicit loop over replicates).

The last version of the Bootstrap with recalculated threshold should have least bias since it comes closest to the real situation. This complex Bootstrap constructs the survey design object for each replicate anew and takes several hours to calculate on the full data set of Albania. The bootstrap standard error is lower than for the naive variance estimate which assumes no variance for the estimation of the median.

This is in line with Zins (2020, Table 1) where the relative bias of the naive variance estimator is much larger than the one of the boostrap variance estimator (both have positive bias). It seems that the fixed threshold makes the percentage under the threshold more variable than with the adapted threshold.

A check on the variability of the inclusion probabilities of the psu in Berat yields a ratio of maximum to minimum of 2.45 in Stratum 1 and 5.2 in Stratum 2 and a standard deviation of 0.020 and 0.013 (Table 1). This indicates that the variability of the inclusion probabilities is moderate at most. Therefore, neglecting the IPPS feature of PSU by the Bootstrap variance estimators are sufficiently reliable.

Table 1 Inclusion probabilities of psu

Stratum	pi1	sd(pi1)	cv(pi1)	min(pi1)	max(pi1)	max(pi1)/min(pi1)
11	0.10322	0.01971	0.19093	0.0566	0.1385	2.45
12	0.05649	0.0131	0.23192	0.0174	0.08702	5

RESULTS

National results for the **At risk of Poverty (**ARPR) with naive and with simple Bootstrap, as well as with the complex Bootstrap that recalculates the poverty threshold for each replicate (from silc-varest8_last_version.html) are shown in
Table 2

Table 2 ARPR at level of Albania with different variance estimators

Variance estimator	ARPR	SE(ARPR)	Norse
naive estimator	0.23385	0.0091	0.01076239
simple BS	0.23385	0.0087	0.01024924
complex BS	0.23296	0.0074	0.00876946

The naive estimator yields the highest standard error. The simple Bootstrap estimator is a bit lower, while the complex Bootstrap with recalculated threshold is considerably lower than the naive estimator. Thus the naive estimator is the most conservative estimator of variance.

Prefectures

Results with naive estimator and with simple bootstrap estimator are shown in
Table 3.

Table 3 ARPR with naive and simple Boostrap (estimate1, estimate1b)

	naive	naive	naive	naive	simpleBS	simpleBS	simpleBS
pref_survey	arpr	se	DEff.arpti	norse	arpr	se	norse
BERAT	0.28377	0.04588	13.11820	0.05089	0.28377	0.04707	0.05221
DIBER	0.30756	0.04976	15.41857	0.05391	0.30756	0.04978	0.05394
DURRES	0.07459	0.01445	7.42650	0.02750	0.07459	0.01414	0.02691
ELBASAN	0.36875	0.03124	9.64835	0.03237	0.36875	0.03341	0.03462
FIER	0.22900	0.02533	10.62910	0.03014	0.22900	0.02524	0.03004
GJIROKASTER	0.29079	0.06605	21.97437	0.07272	0.29079	0.06830	0.07520
KORCE	0.19087	0.02883	10.91963	0.03669	0.19087	0.02843	0.03617
KUKES	0.56754	0.05141	9.68301	0.05188	0.56754	0.05181	0.05229
LEZHE	0.32810	0.03811	7.43612	0.04058	0.32810	0.03705	0.03946
SHKODER	0.29913	0.03625	10.64123	0.03958	0.29913	0.03571	0.03900
TIRANE	0.19387	0.01804	8.21604	0.02281	0.19387	0.01789	0.02262
VLORE	0.15111	0.02819	8.13521	0.03936	0.15111	0.02977	0.04155

The point estimates are the same for the naive estimator and the simple Bootstrap estimator. The naive variance estimates and the simple Bootstrap variance estimates are very close.

What is remarkable is the high design effect. It varies from 7.42 to 21.97. Note that this highest design effect stems from Gjirokaster. It would need further investigation to determine which part of the design effect is due to the cluster effect of the primary sampling units and which part is due to the cluster effect of the household. The latter could be large because the equivalised personal income is the same for all household members.

Comparison with complex Bootstrap variance estimator

Table shows the ARPR with naive estimator and with complex Bootstrap to take the variability of the median income into account. The comparison of the standard errors shows that the naive estimator is very close to the complex Bootstrap estimator. This is the confirmation that also at the level of the prefectures the naive estimator is a good indicator of accuracy.

Table 4 ARPR with complex and naive estimator at the level of prefectures

	complex bootstrap			naive estimator		
prefecture	arprcomplex	secomplex	norsecomplex	arprnaive	senaive	norsenaive
BERAT	0.2749	0.0468	0.0525	0.2838	0.0459	0.0509
DIBER	0.3047	0.0488	0.0530	0.3076	0.0498	0.0539
DURRES	0.0754	0.0147	0.0279	0.0746	0.0145	0.0275
ELBASAN	0.3687	0.0314	0.0325	0.3688	0.0312	0.0324
FIER	0.2289	0.0250	0.0297	0.2290	0.0253	0.0301
GJIROKASTER	0.2845	0.0647	0.0717	0.2908	0.0661	0.0727
KORCE	0.1941	0.0257	0.0325	0.1909	0.0288	0.0367
KUKES	0.5621	0.0525	0.0529	0.5675	0.0514	0.0519
LEZHE	0.3222	0.0369	0.0395	0.3281	0.0381	0.0406
SHKODER	0.3033	0.0351	0.0381	0.2991	0.0363	0.0396
TIRANE	0.1920	0.0168	0.0213	0.1939	0.0180	0.0228
VLORE	0.1510	0.0281	0.0392	0.1511	0.0282	0.0394

CONCLUSION

The naive estimator has a tendency to overestimate the variance of the ARPR. In principle, this is only the case for the national level, because at the prefecture level and for the domains considered always the national poverty threshold is used. This is in line with the European recommendation. When using the national threshold for the ARPR the variance estimation with the naive estimator and the simple Bootstrap estimator at the prefecture level and at the level of the considered domains are very close.

So, using the naïve estimation for the variance calculation we are in the safety side and more conservative for the publication and as a conclusion the naive variance estimator that neglects the variability of the poverty threshold is sufficient to inform the users about the accuracy of the poverty indicators at the prefecture and domain level. It has the advantage that its calculation is rather simple and fast. Neglect the Bootstrap variance estimations even for the ARPR- at risk of poverty rate indicator which needs quite a long time to be produces and have no distinguish difference with naïve estimation method.

REFERENCES

Canty, A. J., Davison, A. C., Hinkley, D. V. & Ventura, V. (2006), ‘Bootstrap diagnostics and remedies’, The Canadian Journal of Statistics/La revue canadienne de statistique 34(1), 5–27

European Commission (2019), Methodological Guidelines and description of EU-SILC target variables, EU-SILC doc 65/01. Luxembourg: Eurostat.

Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. Survey Methodology, 35(2), 227-234.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0

Zins, S. (2020). Variance Estimation by Linearisation for the At Risk of Poverty or Social Exclusion (AROPE) Rate. Austrian Journal of Statistics, 49(1), 33-44.