SAMPLE Seminar - Warsaw 24 March 2010

Comparison of two bootstrap methods of standard error estimation for some poverty measures

Jan Kordos
Warsaw School of Economics

Agnieszka Zięba
Warsaw School of Economics

Robert Wieczorkowski
Central Statistical Office

Agenda

- The Polish experience in standard error estimation from complex surveys
- Data basis for comparison: the Polish SILC for 2007
- Some indicators of poverty measures
- Experiments with bootstrap methods
  - The McCarthy-Snowden method
  - Bootstrap percentile confidence intervals
- Regression adjustment
- Comparison of the McCarthy-Snowden method and percentile method
- Concluding remarks
- Annexes

The Polish experience in precision estimation from complex household sample surveys

In Poland the following estimation methods for standard error estimators from complex sample surveys have been used:

- the interpenetrating sub-samples (Kordos, 1985, 2002; Popiński, 2006; Szarkowski & Witkowski, 1994),
- the Taylor series linearization (Szarkowski & Witkowski, 1994; Popiński, 2006)
- the jackknife replication techniques,
- the balanced repeated replication (BRR) (Särndal et al, 1992; Walter, 1985),
- the bootstrap methods. (Faucher, 2003, McCarthy & Snowden,1985)

The Polish EU-SILC

- We used data from the EU-SILC carried out by Central Statistical Office of Poland in 2007.
- The sample contained:
  - 34 888 interviewed persons (while 42 852 individuals were analyzed)
  - in 14 286 households
  - from 5120 primary sampling units (PSUs)
  - allocated in 211 strata.
- Formulas for variance estimation and other calculation were obtained using SAS 9.1 software.

Some indicators of poverty measures

We involved in our calculation five most common income poverty indicators:

- At-risk-of-poverty rate after social transfers (ARPR),
- Relative median poverty risk gap (RMPG),
- Gini coefficient (GINI),
- Income quintile share ratio (S80/S20),
- Mean equivalised income (MEAN_EQINC).

These indicators were calculated using **equivalised income distribution** based on EU-SILC data.

Experiments with bootstrap methods

We selected two bootstrap methods for our experiments:

- McCarthy, P. J. and Snowden, C. B. (1985) bootstrap method,
- Percentile bootstrap method for confidence intervals estimation.

In our study we focus on finding answers for such question as:

- What criteria should be accepted for comparison?
- Is there any dependence between number of bootstrap samples and standard error estimates values?
- Which method is more appropriate for calculating variance estimates for income poverty indicators?

The McCarthy and Snowden (1985) bootstrap method

This bootstrap procedure is an asymptotically valid method in assessing the variability of direct estimators for stratified multistage designs (Shao, 2003).

Variance of direct estimator is calculated in the following way:

$$\psi = V(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^{*})^2$$

where $\hat{\theta}$ is the direct estimator of poverty indicator, $\hat{\theta}^{*b}$ is the bootstrap estimator obtained from b^{th} bootstrap sample and $\hat{\theta}^{*}$ is computed as:

$$\hat{\theta}^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$$

Every b^{th} bootstrap sample is obtained drawing with replacement random sample of n_h-1 PSU's out the n_h , sampled in each stratum h $(h=1,2,\ldots,L)$. To obtain $\hat{\theta}^{*b}$ using b^{th} bootstrap sample the original weights are properly rescaled:

$$w_j(b) = w_j \frac{a_h}{a_h - 1} m_j(b)$$

where a_h is the number of PSUs in stratum h, $w_j(b)$ is the weight for person from j^{th} household in b^{th} bootstrap sample, w_j is the original weight for person from j^{th} household and $m_j(b)$ – number of how many times PSU from j^{th} household is included in b^{th} bootstrap sample ($b=1,2,\ldots,B$).

Some experiments with the McCarthy and Snowden (1985) bootstrap method

Study of a relation between coefficient of variation of standard error, i.e. CV_SE, was conducted for 5 poverty measures respectively and different B (=100, 200, 500, 700, 1000, 1500).

For each B and each poverty measure coefficient of variation of standard error of standard errors for s – simulations was obtained:

$$CV\_SE(I_{ks}|B) = \frac{SE\_SE(I_{k,s}|B)}{MEAN\_SE(I_{k,s}|B)}$$

$I_{k,s}$ – estimate of poverty measure of k and subsample s (k = 1, 2, 3, 4 ,5; s = 10, 20, 30).

$MEAN\_SE(I_{k,s}|B)$ – the mean of standard errors for the poverty calculated according to the bootstrap algorithm for s simulations.

$SE\_SE(I_{k,s}|B)$ – the estimate of standard error of standard errors for the I_k poverty calculated according to the bootstrap algorithm for s simulations (s = 10, 20, 30)

Nr of replicates	$CV\_SE(I_{ks}\|B)$ (in %)				
B	**ARPR**	**RMPG**	**GINI**	**S80/S20**	**Mean_Inc**
100	7.35	4.30	6.30	5.66	7.54
200	4.58	4.66	4.75	4.13	4.10
500	3.40	2.20	2.91	3.07	3.07
700	1.89	2.56	2.71	2.77	2.80
1000	1.80	1.63	2.18	2.15	2.24
1500	1.76	1.91	1.33	1.43	1.52

Table 1. Comparison of CV for SE in case of different numbers of bootstrap replicates (B) for 30 sub-samples ($s = 30$).
Source: own calculation of the basis of the GUS: European Statistics on Income and Living Conditions 2007 (EU-SILC)

Regression adjustment

Power regression equation can be written as $y = ax^b$ where dependent variable y is CV_SE and independent variable x is number of bootstrap samples B, i.e.

$$CV\_SE(I_{k,s} \mid B) = aB^b$$

To estimate coefficients a and b the linear regression model was used.

$$\ln y = \ln a + b \ln x$$

	R^2	a	b
TOTAL*	**0.797**	**52.396**	**-0.469**
ARPR**	0.767	63.471	-0.503
RMPG	0.720	33.908	-0.408
GINI	0.876	72.184	-0.520
S80/S20	0.802	38.683	-0.422
Mean_Inc	0.890	65.712	-0.490

Table 2. Power regression analysis results (for dependence CV_SE on B)

Source: own calculation of the basis of the GUS: European Statistics on Income and Living Conditions 2007 (EU-SILC)

*) results for 5 indicators, 5 number of bootstrap samples (B) and 3 random subgroups altogether

**) results for indicator ARPR, 5 number of bootstrap samples (B) and 3 random subgroups altogether

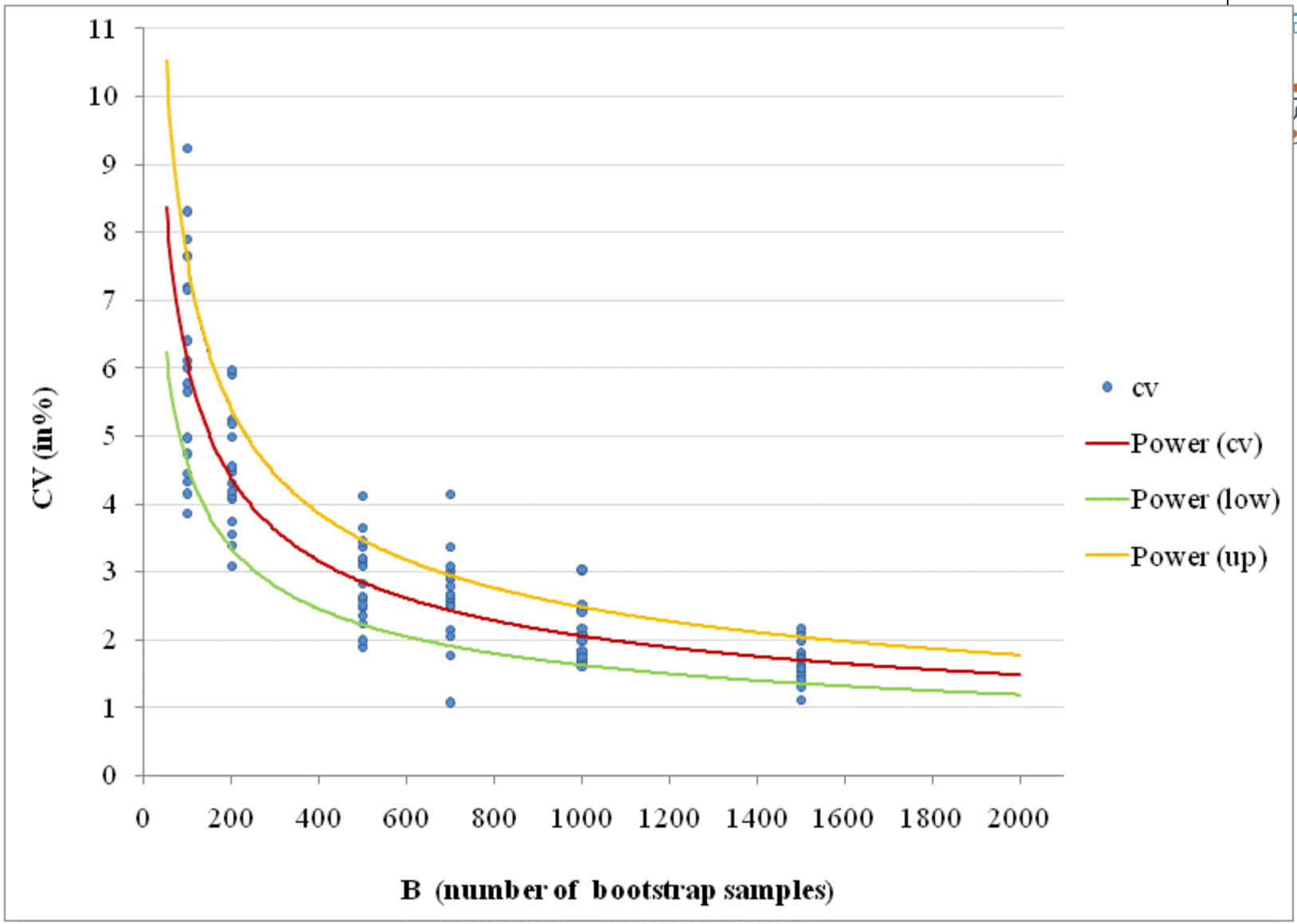


Chart 1: Power regression analysis results.
Source: own calculation of the basis of the GUS: European Statistics on Income and Living Conditions 2007 (EU-SILC)

Bootstrap percentile confidence intervals method

- Using the same algorithm as for bootstrap estimates in previous section, each bootstrapped estimates $I_{k,B}$, (where B = 100, 200, 500, 700, 1000, 1500), were sorted from the smallest to the largest value .
- It means that to estimate confidence intervals, we simply generate a large number of bootstrapped statistics and sort them in ascending order.
- For example, the 95% confidence interval then can be estimated simply by selecting the bootstrapped statistics at the 2.5-th and 97.5-th percentiles, i. e. L_ci and U_ci respectively.
- Assuming normality of estimated parameters, and using estimated confidence intervals, $SE_p(I_k)$ is calculated using formula:

$$SE(I_k) = \frac{U\_ci - L\_ci}{2 \cdot perc(1-\frac{\alpha}{2})}$$

where: U_ci and L_ci stands for upper and lower limits of confidence interval respectively.

Some experiments with the bootstrap percentile confidence intervals metod

Using *s* simulations for each B and estimated parameter, we obtain:

$$CV\_SE_p(I_{k,s}|B) = \frac{SE_p\_SE(I_{k,s}|B)}{MEAN_p\_SE(I_{k,s}|B)}$$

where:

$I_{k,s}$ – estimate of poverty measure of k and subsample s (k = 1, 2, 3, 4 ,5 ; s= 10, 20, 30).

$MEAN_p\_SE(I_{k,s}|B)$ – the mean of standard errors for the poverty calculated according to the bootstrap algorithm for *s* simulations.

$SE_p\_SE(I_{k,s}|B)$ – the estimate of standard error of standard errors for the I_k poverty calculated according to the bootstrap algorithm for s simulations (s=10,20, 30)

Indicator	B	CV_SE_p (in%)	CV_SE (in%)	CV_SE_p / CV_SE
ARPR	100	9.51	7.35	1.29
	200	5.60	4.58	1.22
	500	5.06	3.40	1.49
	700	2.40	1.89	1.27
	1000	2.74	1.80	1.52
	1500	2.19	1.76	1.25
RMPG	100	7.11	4.30	1.65
	200	6.31	4.66	1.35
	500	3.08	2.20	1.40
	700	3.72	2.56	1.46
	1000	2.32	1.63	1.42
	1500	2.63	1.91	1.38
GINI	100	10.83	6.30	1.72
	200	6.61	4.75	1.39
	500	4.02	2.91	1.38
	700	3.53	2.71	1.30
	1000	2.46	2.18	1.13
	1500	2.17	1.33	1.62
S80/S20	100	9.21	5.66	1.63
	200	4.92	4.13	1.19
	500	4.37	3.07	1.42
	700	3.84	2.77	1.38
	1000	3.10	2.15	1.44
	1500	2.10	1.43	1.47
Mean_Inc	100	7.53	7.54	1.00
	200	4.46	4.10	1.09
	500	4.15	3.07	1.35
	700	4.23	2.80	1.51
	1000	2.91	2.24	1.30
	1500	2.18	1.52	1.43

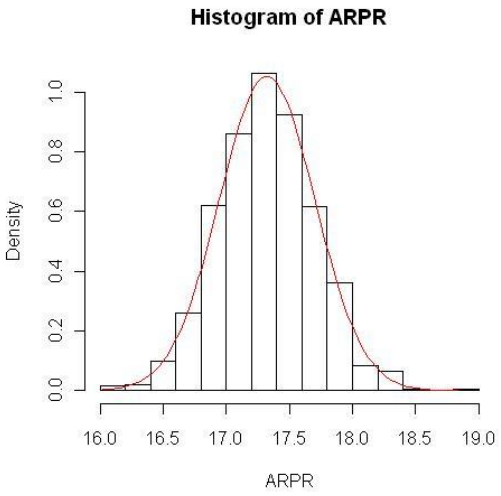
Table 3. Comparisons of CV standard errors obtained according to the McCarthy-Snowden bootstrap method CV_SE (in%) with bootstrap percentile method CV_SE_p (in%)

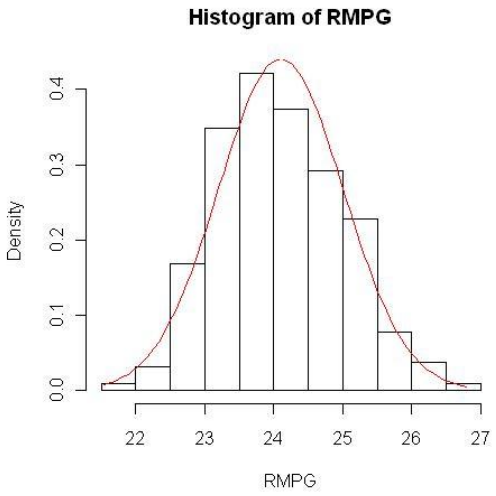
Source: own calculation of the basis of the GUS: European Statistics on Income and Living Conditions 2007 (EU-SILC)

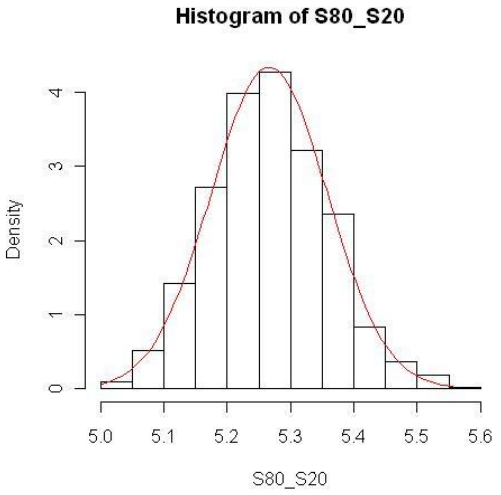
Concluding remarks

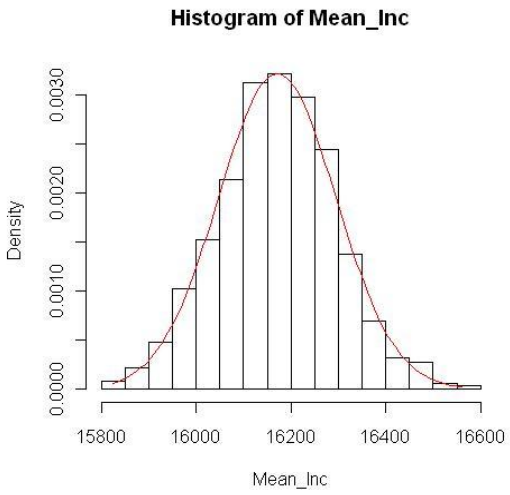
- Regression addjustment showed that in each case b coefficient is close to value -0.5 so it means that if B increases then CV_SE goes down at the rate (that was noticed by R. Tibshirani, 1985).
- From power regression chart (Chart 1) can be observed that above *B=1000* resamples the difference in CV_SE seems to be neglected.
- In all cases McCarthy and Snowden method perform smaller coefficient of variation (CV_SE) than coefficient of variation calculated based on bootstrap percentile confidence intervals (CV_CE_p). The value of CV_SE_p/CV_SE is always bigger than 1 for all simulations.
- After this preliminary research we suggest to use the McCarthy-Snowden bootstrap method.

Annex 1









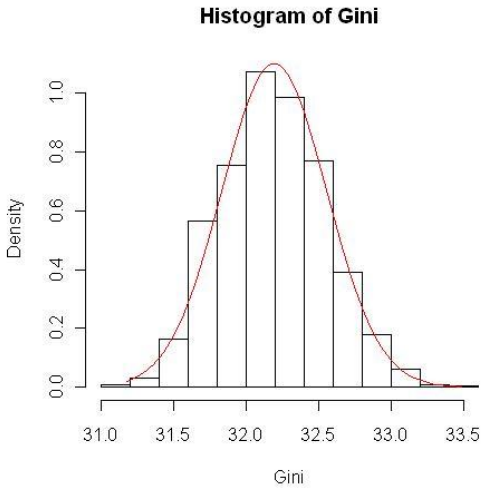
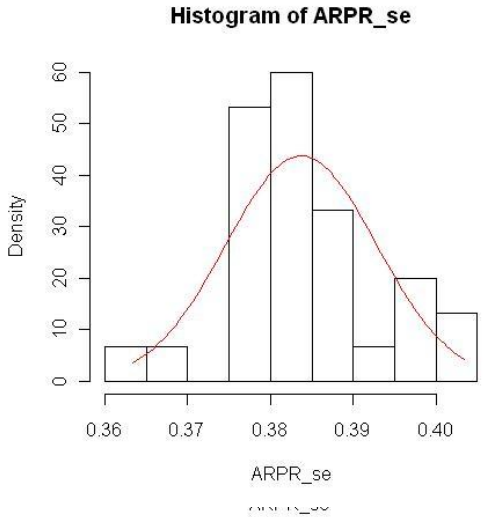
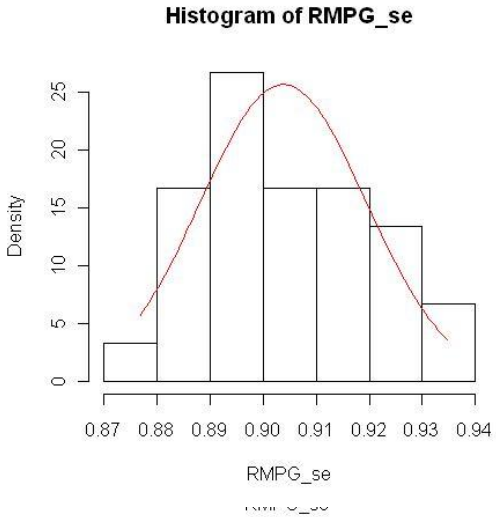
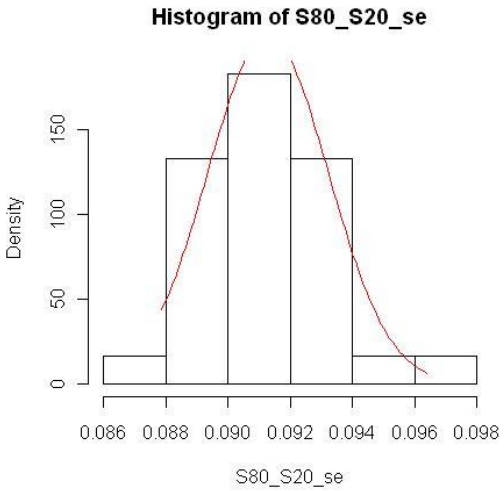


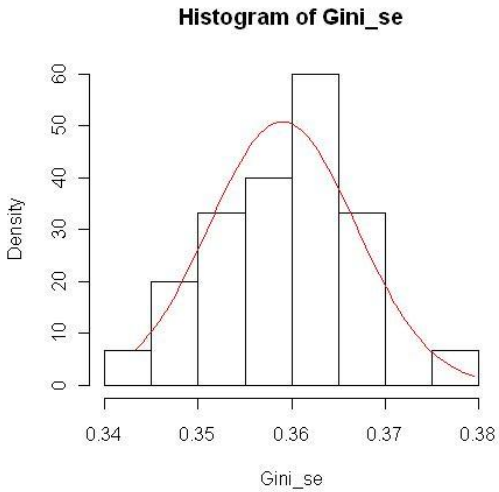
Figure1. Histograms of bootstrap estimators for *B=1000.* All of the histograms seem to be close to normal curve.

Annex 2









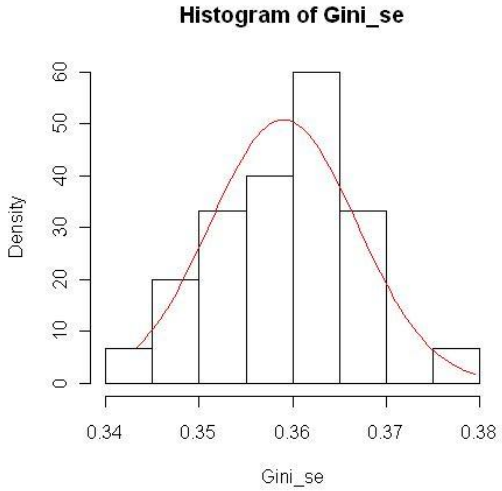


Figure 2. Histograms of standard errors calculated for every poverty indicator with bootstrap replicates *B=1000.*

Acknowledgment

Presented work was done under the S.A.M.P.L.E. project (Small Area Methods for Poverty and Living Condition Estimates). This research programme is funded by European Commission under the Seventh Framework (FP7) Programme of the European Union. (http://www.sample-project.eu/).

The authors would like to thank the Central Statistical Office of Poland for the production and provision of the survey data used, to Dr. Robert Wieczorkowski for computation of linearization and variance formula in SAS Software and to Mr. Bronisław Lednicki for sampling consultation.