European Conference on Quality in Official Statistics (Q2014)
Vienna, 3-5 June 2014

Estimation of Poverty in Small Areas

Agne Bikauskaite
agne.bikauskaite@gmail.com

Abstract: The European Statistics System‘s (ESS) objectives is to produce and disseminate the highest quality of statistics. Data has to be precise and comparable between Member States (MS). A very important issue is development and implementation of a framework enabling the production of small area estimates for ESS social surveys (for instance poverty and social exclusion, unemployment rate, etc.).

One of the main aims of the Europe 2020 strategy is the reduction of poverty. The EU target is to lift at least 20 million people out of the risk of poverty and social exclusion by 2020 compared to the level in 2008. A qualitative estimation of poverty in MS is needed to better implement, monitor and determine national areas where support is most required.

The problem of small area estimation (SAE) is the production of reliable estimates in areas with small samples. The precision of estimates in strata deteriorate (i.e. the precision decreases when the standard deviation increases), if the sample size is smaller. In these cases traditional direct estimators may be not qualitative and therefore pointless. Currently there are many indirect methods for SAE. The purpose of this paper is to analyze several different types of techniques which produce small area estimates of poverty and to calculate indicators of poverty in order to compare the results.

1. Introduction

The focus of this analysis is persons and their income. Estimated parameters are the following: the average household income, the poverty indicators and their variances. Poverty indicators are presented in section 3. All parameters have been estimated using the Horvitz-Thompson, the Generalised Regression (GREG), and the Synthetic estimation methods. These methods are described in section 4. The Jack-Knife method has been used for the estimation of variances to indicate the precision of the estimates. See section 5. The Absolute Relative Bias (ARB) was applied to compare the performance of the different estimators for 1000 simulations. Methodology is given in section 6. The results shown in section 7 indicate the best method for poverty estimation in small areas. Section 8 summarize the results and offers some suggestions.

2. Analysed population. Sampling scheme. Estimated parameters

2.1 *Analysed population*

Canadian household survey data was used for the simulation. The analysed population U = (1, …, i, …, N) consisted of 3000 individuals with income values obtained $(y_1,\ldots,y_N)$. The gender and age of individuals have been used as auxiliary information. This population is actually a simple random sample but was treated as a population for simulation purposes.

2.2. Stratified sampling

A simple random sample drawn from the population can be homogeneous. In order to have more precise estimates of the population the data set has to be divided into H mutually exclusive strata $U_1,U_2,\ldots,U_H$.

For the analysis a stratified simple random sample s composed of seven strata with n_h elements in each has been drawn and y_h values observed. The size of the sample s is $n = n_1 + \ldots + n_h$.

Table 1: Strata size

Number of strata	The population size N_h	The sample size n_h
1	496	50
2	333	33
3	177	18
4	119	12
5	92	9
6	794	79
7	989	99
Total	3000	300

The sample design probability when element i belongs to strata h is $\pi_{ih} = \frac{n_h}{N_h}$; the sampling weight for selected person i from the h^{th} strata is $w_{ih} = \frac{1}{\pi_{ih}} = \frac{N_h}{n_h}$. The value of the observed variable y into h^{th} strata of the i^{th} element is y_{hi}, $i = 1,2,\ldots,N_h$, $h = 1,2,\ldots,H$. Then the sum of observed values y in h^{th} strata is $t_h = \sum_{i=1}^{N_h} y_{hi}$ and the mean $\mu_h = \frac{t_h}{N_h} = \frac{1}{N_h}\sum_{i=1}^{N_h} y_i$.

The sum and the mean of y values observed through the whole population are accordingly $t = \sum_{h=1}^{H} t_h$ and $\mu = \frac{t}{N}$.

2.3. *Estimated parameters*

The average incomes, the poverty threshold, the headcount and poverty gap indices and the variances of these indicators have been calculated. 1000 samples have been drawn to verify the best of three applied methods for small area estimation. The estimated indicators and variances have been compared with the real values. Parameters have been estimated for every strata separately and also for the total the population.

3. The poverty indicators

Persons or households with disposable income lower than poverty threshold are considered as living in poverty or social exclusion because there is no possibility of participating fully in society life. In countries with high quality of life conditions not all residents below the poverty threshold lack money. However, they have a significantly lower potential to meet their needs compared with the rest of community but they may live in good enough conditions.

The headcount and the poverty gap indices concentrate attention on those individuals below the poverty threshold. The headcount index P_0 shows which part of society is below the poverty threshold. The poverty gap shows the average lack of finance and how much income has to increase so that the poverty threshold is reached.

3.1. *The poverty threshold*

The poverty threshold is defined as 60 per cent of the median equivalent disposable income $z = 60\%M$. This indicator depends on the income distribution in society and varies according to the changes of the general living conditions in the area.

3.1.1. *The poverty threshold estimation*

To estimate the poverty threshold, the median $\hat{M}$ of the income has to been estimated. Firstly units $y_1, \ldots, y_n$ of s^{th} sample have been sorted in ascending order $y_{1:s} \le y_{2:s} \le \ldots \le y_{n:s}$ and inclusion into s^{th} sample probabilities accordingly $\pi_{1:s}; \pi_{2:s}; \ldots; \pi_{n:s}$. Accumulative totals of sampling weights have been counted $B_1 = \frac{1}{\pi_{1:s}}$, $B_2 = \frac{1}{\pi_{1:s}} + \frac{1}{\pi_{2:s}}$, …, $B_l = \sum_{j=1}^{l} \frac{1}{\pi_{j:s}}$ while one of the l satisfied the following condition $B_{l-1} < 0{,}5\hat{N}$ *and* $B_l > 0{,}5\hat{N}$.

Then the estimated number of the population is $\hat{N} = B_n = \sum_{j=1}^{n} \frac{1}{\pi_{j:s}} = \sum_{s} \frac{1}{\pi_i}$ and the median estimate is $\hat{M} = \begin{cases} y_{l:s}, & if\ B_{l-1} < 0{,}5\hat{N} < B_l \\ \frac{1}{2}(y_{l:s} + y_{l+1:s}), & if\ B_l = 0{,}5\hat{N} \end{cases}$.

Then the estimate of the poverty threshold is defined by formula $\hat{z} = 60\%\hat{M}$.

3.2. *The headcount index*

The headcount index is defined as the number of persons below the poverty threshold divided by the population number $P_0 = \frac{1}{N} \sum_{i=1}^{N} I_{(y_i < z)} = \frac{N_q}{N}$, here $I_{(y_i < z)} = \begin{cases} 1, & y_i < z \\ 0 & y_i \geq z \end{cases}$. N_q defines the number of individuals whose income is below the poverty threshold $N_q = \sum_{i=1}^{N} I_{(y_i < z)}$.

3.2.1. *The headcount index estimation*

The headcount index estimator is $\hat{P}_0 = \frac{1}{\hat{N}} \sum_{i=1}^{n} w_i I_{(y_i < z)} = \frac{\hat{N}_q}{\hat{N}}$, here $\hat{N}$ is estimated number of the population elements; $\hat{N}_q$ is the estimated number of individuals in the population living in poverty or social exclusion.

3.3. *The poverty gap index*

The poverty gap G_n is defined as an amount of difference between poverty threshold and income value y_i of i^{th} person living in poverty or social exclusion $G_i = (z - y_i) I_{(y_i < z)}$. The poverty gap index is a proportion of the poverty gap and the poverty threshold $P_1 = \frac{1}{N} \sum_{i=1}^{q} \frac{G_i}{z} = \frac{1}{N} \sum_{i=1}^{N} \frac{z - y_i}{z} I_{(y_i < z)}$, here q is number of individuals in poverty or social exclusion.

3.3.1. The poverty gap index estimation

Then the direct estimate of the poverty gap index is defined by formula $\hat{P}_1 = \frac{1}{\hat{N}} \sum_{i=1}^{n} \frac{\hat{z} - y_i}{\hat{z}} w_i I_{(y_i < \hat{z})}$.

4. Small Area Estimation. Direct and indirect estimators

4.1. Small Area Estimation

An area is regarded as large if the sample drawn from that area is large enough to get direct estimates of adequate precision. An area is regarded as small if the sample is not large enough to get simple direct estimates of adequate precision. The variance of the estimate decreases through enlarging the size of the sample.

In order to have better quality estimates in areas, unbiased auxiliary variables have to be used from the same areas. This kind of estimation is defined as direct. For indirect estimation the auxiliary information has to be taken from adjacent areas.

4.2. The Horvitz-Thompson estimator

The Horvitz-Thompson estimator of the sum is $\hat{t}_\pi = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} w_i y_i$.

For a stratified simple random sample the Horvitz-Thompson variance of the sum estimate is $D\hat{t}_\pi = \sum_{i,j \in s} (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j}$. The Horvitz-Thompson variance estimate of the sum estimate is $\hat{D}\hat{t}_\pi = \sum_{i,j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j}$, here $\pi_i = \frac{n_i}{N_i}, i \in U_h$; $\pi_{ij} = \frac{n_h}{N_h} \cdot \frac{n_h - 1}{N_h - 1}$, *when* $i, j \in U_h$ and $\pi_{ij} = \frac{n_h}{N_h} \cdot \frac{n_s}{N_s} = \pi_i \pi_j$, *when* $i \in U_h$, $j \in U_s$. π_{ij} is the inclusion into the sample probability of two elements *(i, j)* . If $i = j$ *then* $\pi_{ii} = \pi_i$.

4.3. The Generalised Regression Model (GREG)

y_i, is the values of the income and the value of the vector $\boldsymbol{x}$ is defined as the auxiliary information $\mathbf{x}_i=(x_{1i}, \ldots, x_{ji}, \ldots, x_{Ji})'$.

The sum of the dominant elements y is the GREG estimator of the sum t_y defined by the following formula $\hat{t}_{y,GREG} = \hat{t}_{y\pi} + \sum_{j=1}^{J} \hat{B}_j \left(t_{X_j} - \hat{t}_{X_j \pi} \right)$, where j is the number of several auxiliary pieces of information about the individual. The Horvitz-Thompson estimator of the sum t_{x_j} is $\hat{t}_{x_j\pi} = \sum_{i=1}^{n} \frac{\mathbf{x}_{ji}}{\pi_i}$. $\hat{B}_1, \hat{B}_2, ..., \hat{B}_J$ are the estimated components of the vector $\mathbf{x}$ $\hat{\mathbf{B}} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_k' q_k}{\pi_i} \right)^{-1} \sum_{i=1}^{n} \frac{\mathbf{x}_i y_i q_i}{\pi_i}$.

The GREG estimation method is appropriate to estimate parameters in non-responses. Then the GREG estimator of the sum is $\hat{t}_{wy} = \sum_r w_i y_i$, where r is the set of the respondents. The calibrated weights are $w_i = \left(1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x \right)' \left(\sum_r \frac{\mathbf{x}_i \mathbf{x}_i' q_i}{\pi_i \hat{\theta}_i} \right)^{-1} \mathbf{x}_i q_i \right) \times \frac{1}{\pi_i \hat{\theta}_i} = \frac{g_i}{\pi_i \hat{\theta}_i}$, where $\hat{\theta}_k$ is the estimator of element's i response to the survey probability.

The calibrated estimate of the sum $\hat{t}_{yw}$ is biased. When N is large but sampling rate $\frac{n}{N}$ small then the bias estimate is slight.

4.4. Simple Synthetic estimator

The stratified population U_h splits up into k mutually exclusive groups $G_1, \ldots, G_K$, $U_h = G_{h1} \cup \ldots \cup G_{hK}$ and $U = G_1 \cup \ldots \cup G_K$.

The mean of the elements from h^{th} strata and k^{th} group is $\mu_{yhk} = \frac{\sum_{i=0}^{n_{hk}} w_i y_i}{\sum_{i=0}^{n_{hk}} w_i}$, here $w_0 y_0 = 0$, when $n_{hk} = 0$, i.e. if the element from h^{th} strata and k^{th} group in the population does not exist then the sum is $t_{yhk} = 0$. The sum of population in strata h is $t_{yh} = \sum_{k=0}^{n_k} \mu_{yhk} N_{hk}$. The sum estimator of the sample is $\hat{t}_{yh}^{\sin t} = \sum_{k=0}^{K} \hat{\mu}_{yhk} N_{hk}$ and its variance is $D\hat{t}_{yh}^{\sin t} = D(\sum_{k=1}^{K} \hat{\mu}_{yhk} N_{hk})$.

The synthetic estimator is unbiased when $\mu_{yhk} = \mu_{yh}$, here $h = 1, \ldots, H$, $k = 1, \ldots K$. If this is the opposite, it is biased.

5. The Jack-Knife method for variance estimation

To estimate the precision of estimated parameters the Jack-Knife variance estimation method has been used.

The Jack-Knife method's idea is to divide stratified sample s_h into K_h mutually exclusive subgroups. If $\hat{\theta}_h$ is the estimate of the parameter θ_h of the primary stratified sample s_h, then $\hat{\theta}_{(hk)}$ is parameter's θ estimator obtained by estimating the sample composed of h^{th} strata elements apart units from k^{th} $(k = 1, \ldots, K_h)$ group. The modified sampling weights were used to estimate $\hat{\theta}_{(hk)}$:

$$w_{i(hk)} = \begin{cases} w_i, & \text{when } i^{\text{th}} \text{ element does not belong to } h^{\text{th}} \text{ stratum,} \\ 0, & \text{when } i^{\text{th}} \text{ element belongs to } h^{\text{th}} \text{ stratum and } k^{th} \text{ subgroup,} \\ \dfrac{n_i}{n_i - 1} w_i, & \text{when } i^{\text{th}} \text{ element belongs to } h^{\text{th}} \text{ stratum.} \end{cases}$$

Then the Jack-Knife variance estimator of θ estimate is equal to $\hat{D}_{JACK}\hat{\theta}_{(hk)} = \sum_{h=1}^{H} \left[\frac{K_h - 1}{K_h} \right] \sum_{k=1}^{K_h} \left(\hat{\theta}_{(hk)} - \hat{\hat{\theta}}_{(hk)} \right)^2$, here $\hat{\hat{\theta}}_{(hk)} = \frac{1}{K_h} \sum_{k=1}^{K_h} \hat{\theta}_{(hk)}$.

6. The Absolute Relative Bias

The Absolute Relative Bias (ARB) assessed the accuracy of the estimates $ARB = \left| \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{\theta}_h - \theta_h}{\theta_h} \right|$, where K is the number of drawn samples; $\hat{\theta}_h$ is the estimate of the parameter in the strata h; θ_h is real value of parameter in the strata h.

7. Results

7.1. Estimates of parameters

The best ARB, estimating the average income and poverty gap index for the whole population, was through using the Horvitz-Thompson method. The headcount index estimates obtained the least ARB applying the GREG method.

The purpose of the paper was to find out the method which is the most accurate for the estimation in small areas. The results show that in the smallest, third, fourth and fifth strata which consist accordingly of 9, 12 and 18 elements in the sample, the Synthetic estimates of the average income are closest to the real values (see Table 2).

Table 2: The ARB of the average income estimates

Strata	Horvitz-Thompson estimate's ARB (%)	Generalised Regression estimate's ARB (%)	Synthetic estimate's ARB (%)
Population	**-0.06447544**	0.098310539	-0.08398375
1	-0.3211974	**-0.31518106**	-0.34310121
2	-0.02643092	**-0.014056**	-0.06902109
3	0.465571393	0.551799055	**0.403882282**
4	-0.81562095	-0.88208503	**-0.65375062**
5	0.485715332	0.510841272	**0.492216146**
6	-0.1417938	**-0.13401672**	-0.14913289
7	0.079252793	**0.090945055**	0.188597999

The Synthetic headcount index estimate's ARB in the smallest fifth strata is least. (see Table 3).

Table 3: The ARB of the headcount index estimates

Strata	Horvitz-Thompson estimate's ARB (%)	Generalised regression estimate's ARB (%)	Synthetic estimate's ARB (%)
Population	0.36396329	**0.147665664**	0.152247869
1	**-3.51959494**	-3.7958481	-3.8266288
2	1.468493151	1.192029888	**1.003491015**
3	**4.644761905**	4.757144543	5.058255185
4	2.859782609	**2.601086957**	2.877924901
5	-2.80634921	-2.90370419	**-1.68042706**
6	**-0.63675717**	-0.78252971	-0.8622043
7	1.344097079	**1.068357786**	1.298860988

In the same fifth strata the Synthetic poverty gap index estimate has the smallest ARB (0.02 per cent) compared with the Horvitz-Thompson and the GREG estimation methods.

Table 4: ARB of the poverty gap index estimate

Strata	Horvitz-Thompson estimate's ARB (%)	Generalised regression estimate's ARB (%)	Synthetic estimate's ARB (%)
Population	**-0.1594528**	-0.35543944	-0.41065525
1	**-1.37126072**	-1.59705543	-1.57592157
2	**-0.9619282**	-1.23793553	-1.64985282
3	**-0.49766038**	-0.6453229	-0.69178013
4	**-1.21749012**	-1.2989069	-1.45047831
5	-0.73358855	-0.95628486	**0.02299011**
6	0.702989553	0.477610719	**0.357964671**
7	0.19625962	**0.02379632**	0.278143276

7.2. Estimated variances of parameters estimates

The largest over-estimations of the variance coefficients of averaged income estimates are in the smallest strata. Significantly better variance coefficients are obtained through the Horvitz-Thompson estimation (see Table 5). While the GREG and the Synthetic estimates are equally worse.

Table 5: Estimated variance coefficients of averaged income estimates

Strata	Sample size	Variance coefficient of the population	Horvitz-Thompson estimate's variance coefficient	GREG estimate's variance coefficient	Synthetic estimate's variance coefficient
Total	300	0.035	**0.039**	0.040	0.040
1	50	0.094	0.102	0.102	**0.101**
2	33	0.095	**0.104**	0.112	0.111
3	18	0.141	**0.135**	0.156	0.156
4	12	0.163	**0.181**	0.208	0.211
5	9	0.239	**0.252**	0.307	0.307
6	79	0.064	**0.068**	0.069	0.069
7	99	0.067	**0.072**	0.073	0.074

Concerning the variance coefficients of the headcount and poverty gap indices estimates, in most strata Horvitz-Thompson also produced the smallest overestimation (see Tables 6 and 7).

Table 6: Estimated variance coefficients of the headcount index estimates

Strata	Sample size	Real variation coefficient	Horvitz-Thompson variation coefficient's estimate	GREG variation coefficient's estimate	Synthetic variation coefficient's estimate
Total	300	0.104	**0.110**	0.115	0.117
1	50	0.422	**0.415**	0.410	0.440
2	33	0.408	0.477	0.477	**0.475**
3	18	0.544	0.483	0.484	**0.532**
4	12	0.698	0.602	**0.624**	0.818
5	9	0.172	**0.156**	0.206	0.186
6	79	0.232	**0.228**	0.266	0.284
7	99	0.171	0.217	0.217	**0.203**

Table 7: Estimated variance coefficients of the poverty gap index estimates

Strata	Sample size	Real variation coefficient	Horvitz-Thompson variance coefficient's estimate	GREG variance coefficient's estimate	Synthetic variance coefficient's estimate
Total	300	0.141	**0.151**	0.162	0.166
1	50	0.420	**0.458**	0.461	0.472
2	33	0.421	**0.451**	0.462	0.467
3	18	0.645	**0.638**	0.637	0.613
4	12	0.666	**0.697**	0.747	0.733
5	9	0.792	**0.873**	0.982	1.051
6	79	0.332	0.362	**0.361**	0.362
7	99	0.226	**0.246**	0.247	0.256

8. Conclusions

Consequently we can see that to get good quality data would be better to apply different estimation methods for large and for small areas.

It is therefore suggested that if poverty estimation in small areas is to be carried out and if auxiliary information from the adjacent areas can be taken into account, the Synthetic method should be used. If, however, that auxiliary information is not available, then given the simulation results in general, the most appropriate estimation method would be Horvitz-Thompson.

When comparing estimated variances of parameters estimates with real variances, large ARBs have been obtained. The best results of estimation in small and in large areas are given by the Horvitz-Thompson method.

Estimating the Jack-Knife variances calculation takes more time but the precision of the estimates increases when the group size is extremely small.

9. References

[1] Andersson C., Nordberg L., *A User's Guide to CLAN 97 – a SAS – Program for Computation of Point and Standart Error Estimats in Sample Survey*, Statistics Sweden, 1998.

[2] Molina I., Rao J. N. K., *Estimation of Poverty Measures in Small Areas*, http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/S3P3_ESTIMATION_OF_POVERTY_MEASURES_MOLINA_RAO.pdf

[3] Rao J. N. K., *Small Area Estimation*, Wiley, 2010.

[4] Krapavickaitė D., Plikusas A., *Imčių teorijos pagrindai*, Vilnius, Leidykla „Technika“, 2005.

[5] Wolter K. M., *Introduction to Variance Estimation*, New-Yourk, Springer, 1985.

Särndal S.-E., Swensson B., Wretman J., *Model Assisted Survey Sampling*, New York, Springer, 1992.