Marcin Szymkowiak[1]

Some aspects of using calibration in polish surveys

1. Weighting - introduction

Weighting is a statistical technique commonly used and applied in practice to compensate for nonresponse and noncoverage. It is also used to make weighted sample estimates conform to known population external totals. In recent years a lot of theoretical work has been done in the area of weighting and there has been a rise in the use of these methods in many statistical surveys conducted by National Statistical Offices around the world. It is worth noting that there are many weighting methods which can be used in practice. One of the most popular method is calibration which will be wider discussed in the second part of this elaboration. Others include postratification, raking, GREG weighting, logistic regression weighting, mixture approach and logit weighting. A review of the weighting method with examples can be found in Kalton and Flores-Cervantes (2003).

2. Theoretical background of calibration

One of the most important weighting technique is calibration, whereby sampling weights are adjusted to reproduce known population totals. This method is successfully used by statistical offices of many countries in different kind of surveys including censuses, surveys based on sampling and surveys based on administrative registers. This technique is especially used in all statistical surveys because of existing nonresponse problem which is one of the major type of non-random error. Calibration estimation, whereby sampling weights are adjusted to reproduce known population totals, is commonly used in survey sampling. The milestone was article by Deville and Särndal (1992) in which calibration was described in details. A full definition of calibration approach was formulated by C-E Särndal (2007). According to Särndal, the calibration approach to estimation for finite populations consists in:

(a) the computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s),

(b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units,

(c) satisfying an objective of obtaining nearly design unbiased estimates given that nonresponse and other nonsampling errors are absent.

[1] Statisticall Office in Poznań, University of Economics in Poznań m.szymkowiak@stat.gov.pl, m.szymkowiak@ue.poznan.pl

Throughout the later part of this elaboration, we will assume that we are interested in computing the total value $Y = \sum_{i=1}^{N} y_i$ of variable Y. This total value can denote, for example, enterprise revenue, number of employees etc. Let us assume that the whole population $U = \{1,\ldots,N\}$ consists of N elements. From this population we draw, according to a certain sampling scheme, a sample $s \subseteq U$, which consists of n elements. Let π_i denote first order inclusion probability $\pi_i = P(i \in s)$ and $d_i = 1/\pi_i$ the design weight. The well known, classical estimator, of total value is the Horvitz-Thompson one which is given by the formula $\hat{Y}_{HT} = \sum_s d_i y_i = \sum_{i=1}^{n} d_i y_i$. If information for the variable y is not known for some units drawn to the sample then the Horvitz-Thompson estimator is biased and its variance is high. It is because of the nature of nonresponse, which is not random and errors are the consequence of differences between respondents and nonrespondents. Let $r \subseteq s$ denote a set of respondents for which the value of the variable y is known. Let us assume that this set consists of m elements, $m \leq n$. In a situation where in a survey the variable y is affected by the nonresponse Horvitz-Thompson estimator is given by $\hat{Y}_{HT} = \sum_r d_i y_i = \sum_{i=1}^{m} d_i y_i$. This weighted sum is usually underestimated compared to the real total value. According to the calibration paradigm, design weights should be changed to compensate for the loss of information as a result of nonresponse. In such situations we look for new weights (the so called calibration weights) for all units drawn to the sample for which we have information about the variable y. Let w_i denote calibration weight $i = 1,\ldots,m$. Our main goal is to look for new weights w_i which are as close as possible to the design weights d_i and which allow us to reduce bias. The process of constructing calibration weights depends on the properly chosen distance function. In our elaboration we assumed that the distance function is expressed by the formula:

$$D(\mathbf{w},\mathbf{d}) = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i}$$

Let $x_1,\ldots,x_k$ denote auxiliary variables which will be used in the process of finding calibration weights and let $\mathbf{X}_j$ denote the total value for the auxiliary variable x_j, $j = 1,\ldots,k$, i.e. $\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}$, where x_{ij} denotes the value of j-th auxiliary variable for the i-th unit. Moreover, let $\mathbf{X}$ denote the known vector of population totals for the vector of auxiliary variables: $\mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2},\ldots,\sum_{i=1}^{N} x_{ik}\right)^T$. The calibration estimator for total takes the form $\hat{Y}_{\mathbf{X}} = \sum_{i=1}^{m} w_i y_i$ where the vector of calibration weights $\mathbf{w} = (w_1,\ldots,w_m)^T$ is obtained as the

minimization problem $\mathbf{w} = \arg\min_{\mathbf{v}} D(\mathbf{v}, \mathbf{d})$ subject to the calibration constraints $\mathbf{X} = \tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik}\right)^T$. It can be proved that if the matrix $\sum_{i=1}^{m} d_i \mathbf{x}_i \mathbf{x}_i^T$ is nonsingular than the solution of the minimization problem is a vector of calibration weights $\mathbf{w} = (w_1, \ldots, w_m)^T$, whose elements are described by the formula $w_i = d_i + d_i (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum_{i=1}^{m} d_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \mathbf{x}_i$, where $\hat{\mathbf{X}} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik}\right)^T$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$ is the vector consisting of values of all auxiliary variables for the i-th respondent $i = 1, \ldots, m$.

3. Calibration in European and polish surveys

Calibration as a method of weighting is used by many statistical offices in many surveys. For instance see. Särndal and Lundström (2005), Cassel, Lundquist and Selén (2002), Éltetö and László (2002). It is also worth noting that in many surveys calibration as a method of weighting and adjusting initial weights in order to reconstruct the known totals of auxiliarry variables is recomended by Eurostat. This recommendation concerns primarily the European Union Survey on Income and Living Conditions (EU-SILC). For details see Eurostat (2004). In business statistics calibration is also used in practice. This method was used, for instance by ISTAT, in the survey of Structural Business Statistics for small-medium enterprises. For more details see Casciano, Giorgi, Oropallo and Siesto (2012). Calibration was also used as a weighting technique for the Structural Business Survey on enterprises at Statistics Belgium. For details see Vanderhoeft (2001).

In Poland the calibration approach is also used by the Central Statistical Office. For instance the surveys which make use of calibration to compensate for the high percentage of non-response are the European Survey on Income and Living Conditions (EU-SILC) and the National Census of Population and Housing 2011 (Central Statistical Office in Poland, 2011).

Up to now, the calibration approach has not been applied to business statistics in Poland. Anyway some simulation studies which aim was to check usefulness of calibration in the context of using administrative business registers, were conducted. For example, so called MEETS project, involved an attempt to assess the properties and feasibility of using the calibration estimator to estimate the average monthly and annual revenues of enterprises. The simulation study investigated a few variables. The monthly (January) and annual revenue was the response or output variable (Y). It's the basic characteristic of business activity and the choice of the two periods (monthly and annual) was motivated by the need to determine whether the calibration approach can be useful for short-term (monthly-based) and annual statistics of enterprises. The list of input or auxiliary variables included:

enterprise size (large and medium), selected PKD sections (construction, manufacturing, trade and transport) and VAT information. Data about the first two variables (enterprise size and PKD section) came from the DG-1 survey[2]. The VAT variable came from the VAT register.

To conduct the simulation study, a pseudo-population (further referred to as the MEETS real dataset) was created consisting of all enterprises included in the DG-1 survey for which information about the 3 auxiliary variables was available. Enterprises which reported zero revenue in the DG-1 survey, were excluded from the dataset. Taking advantage of a strong correlation between the pseudo-population and the VAT register, it was possible to match VAT information with records in the pseudo-population. The resulting dataset consisted of about 20,000 records containing complete information about the variables under analysis.

Average revenue was estimated on the basis of samples of different size drawn from the MEETS real data. Simulation-based estimates were computed and evaluated at the country level, regardless of enterprise size and PKD section. First of all, the average value of monthly and annual revenue was estimated; the obtained estimates were then divided by the sum of design weights to produce an estimate of average revenue. During the simulation study, 5%, 10% and 15% samples were drawn from the MEETS real dataset, using simple random sampling without replacement. After obtaining a sample, information about revenue (dependent variable Y) for some enterprises was replaced with missing data. As a result, a given sample contained complete information about enterprise size, PKD section and VAT for each sampled unit, but incomplete data about revenue.

Three different approaches were used to generate missing data. In the first one missing data were generated in a random fashion (option 1). In the second (option 2) and third (option 3), missing data were attributed to enterprises with the lowest and highest revenue respectively. In addition, in each sample the percentages of missing data could be either 5%, 10% or 15%. For each sample fraction (3 options), fraction of missing data (3 options) and method of their generation (3 options) 500 iterations were performed to estimate the expected value of revenue, the expected value of the bias of the estimators and their empirical variance as well as relative estimation errors. Separate simulation runs were performed for annual and monthly revenue. Below some chosen results in the form of tables are presented only for annual revenue[3].

2 DG-1 – polish monthly statistical survey of enterprises on economic activity

[3] The average revenue calculated on the basis of the MEETS real data set was at the level of 45 500 (in thousand PLN).

Table 1. The expected value of estimators of the average annual revenue for enterprises (in thousands of PLN)

		Horvitz-Thompson estimator			Calibration estimator		
sample size	% of missing data	option 1	option 2	option 3	option 1	option 2	option 3
5%	5%	46839	47388	16197	45555	44012	18718
	10%	45093	49955	11647	45411	42492	13542
	15%	45900	53392	9137	45758	40942	10684
10%	5%	46175	47290	16140	45801	44118	18264
	10%	45606	50843	11608	46079	42353	13218
	15%	45750	53303	9137	45458	40603	10502
15%	5%	45701	47862	16114	46113	44293	18078
	10%	45683	50761	11592	45802	42476	13085
	15%	45668	53254	9111	45920	40733	10404

Source: own tabulation based on the results of the simulation study

Table 2. The expected value of the bias of estimators of the average annual revenue for enterprises (in thousands of PLN)

		Horvitz-Thompson estimator			Calibration estimator		
sample size	% of missing data	option 1	option 2	option 3	option 1	option 2	option 3
5%	5%	9516	8734	29353	4574	4488	26832
	10%	9222	9145	33903	4247	5225	32007
	15%	9414	10786	36413	4931	6111	34866
10%	5%	7093	6389	29410	3157	3353	27286
	10%	6442	7716	33942	3471	4302	32332
	15%	7435	8961	36413	3614	5396	35048
15%	5%	5391	5272	29436	2697	2664	27471
	10%	5860	6600	33958	2941	3592	32465
	15%	5627	8373	36439	2878	4994	35146

Source: own tabulation based on the results of the simulation study

Table 3. The variance of estimators of the annual enterprise revenue

		Horvitz-Thompson estimator			Calibration estimator		
sample size	% of missing data	option 1	option 2	option 3	option 1	option 2	option 3
5%	5%	169 237 917	153 691 052	1 381 969	31 574 140	29 583 361	4 908 696
	10%	148 074 684	149 561 103	642 338	28 757 830	29 831 443	1 969 977
	15%	166 978 078	202 724 663	300 194	38 128 377	31 562 957	829 532
10%	5%	83 068 297	73 475 310	725 256	15 309 356	15 432 178	1 715 743
	10%	67 141 163	99 118 824	284 353	18 571 129	15 540 268	619 126
	15%	90 304 017	94 383 040	154 945	20 017 539	15 136 572	326 275
15%	5%	46 080 813	41 510 368	410 765	11 339 844	9 854 627	908 690
	10%	52 965 196	49 116 293	166 175	13 223 748	9 133 600	327 854
	15%	51 336 166	60 585 179	94 462	13 064 110	9 666 712	179 280

Source: own tabulation based on the results of the simulation study

Table 4. The relative estimation error of estimators of the annual enterprise revenue (in percent)

		Horvitz-Thompson estimator			Calibration estimator		
sample size	% of missing data	option 1	option 2	option 3	option 1	option 2	option 3
5%	5%	27.77	26.16	7.26	12.33	12.36	11.84
	10%	26.99	24.48	6.88	11.81	12.85	10.36
	15%	28.15	26.67	6.00	13.49	13.72	8.53
10%	5%	19.74	18.13	5.28	8.54	8.90	7.17
	10%	17.97	19.58	4.59	9.35	9.31	5.95
	15%	20.77	18.23	4.31	9.84	9.58	5.44
15%	5%	14.85	13.46	3.98	7.30	7.09	5.27
	10%	15.93	13.81	3.52	7.94	7.12	4.38
	15%	15.69	14.62	3.37	7.87	7.63	4.07

Source: own tabulation based on the results of the simulation study

4. Conclusion

In this elaboration only selected aspects of calibration were presented from point of view of polish statistical surveys. As it was shown, calibration can improve the quality of estimation not only in social surveys (like EU-SILC or LFS) but also in surveys devoted to business statistics, especially the quality of short-term and annual business statistics of medium-sized and large enterprises. The wide use of variables coming from administrative registers in Poland can in significantly way minimize negative effect and influence of nonresponse which can harm all statistical surveys, including business ones.

LITERATURE

1. Casciano M.C., Giorgi V., Oropallo F., Siesto G. (2012), *''Estimation of Structural Business Statistics for Small Firms by Using Administrative Data''*, Rivista Di Statistica Ufficiale, N. 2-3.

2. Cassel C., Lundquist P., Selén J. (2002), ‘*'Model-based calibration for survey estimation, with an example from expenditure analysis*'', R&D Report, Research-Methods-Development, Statistics Sweden.

3. Central Statistical Office in Poland (2011), ''*Incomes and Living Conditions of the Population in Poland (report from the EU-SILC survey of 2009)*'', Statistical Information and Elaborations, Warsaw 2011.

4. Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376-382.

5. Éltetö Ö., László M. (2002), ''*Household Surveys in Hungary*'', Statistics in Transition, Vol. 5, No. 4, 521-540.

6. Eurostat (2004), *Description of target variables: Cross-sectional and Longitudinal*, EU-SILC 065/04.

7. Kalton G., Flores-Cervantes I. (2003), *Weighting methods*, Journal of Official Statistics, vol. 19, No. 2, pp. 81-97.

8. Särndal C-E., Lundström S. (2005), ''*Estimation in Surveys with Nonresponse''*, John Wiley & Sons, Ltd.

9. Särndal C-E. (2007), „*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99-119.

10. Vanderhoeft C. (2001), ''*Generalised Calibration at Statistics Belgium. SPSS Module g-CALIB-S and Current Practises*'', Statistics Belgium.