QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 1, 2013, pp. 283 – 292

ESTIMATION OF POPULATION PARAMETERS USING INFORMATION FROM PREVIOUS PERIOD IN THE CASE OF OVERLAPPING SAMPLES – SIMULATION STUDY

Barbara Kowalczyk
Department of Econometrics, Warsaw School of Economics
e-mail: bkowal@sgh.waw.pl

Abstract: The paper concerns the problem of estimating population parameters for repeated rotating surveys. Coefficients required for theoretical BLUE estimator for rotating surveys are for actual real surveys usually not known. There are no theoretical papers relating to this problem. It is therefore necessary to conduct suitable simulation studies. Broad simulation analyses conducted in the paper are carried out on the basis of two populations: generated from a multivariate normal distribution and based on real data derived from agricultural censuses.

Keywords: survey methodology, rotating surveys, repeated surveys overlapping samples

INTRODUCTION

Theory connected with overlapping samples, also known as theory of rotating surveys or theory of rolling samples, started with the papers [Jessen 1942, Eckler 1955]. The theory was growing in 20th century [Patterson 1950, Rao and Graham 1964, Kordos 1967, Scott and Smith 1974, Jones 1980, Binder and Dick 1989, Fuller 1990, Szarkowski and Witkowski 1994] and it remains of utmost importance in 21 century [Feder 2001, Fuller and Rao 2001, Kowalczyk 2003a, Kowalczyk 2003b, Kowalczyk 2004, Steel 2004, Nedyalkova et al. 2009, Steel and McLaren 2009, Berger and Priam 2010, Wesołowski 2010, Ciepiela et al. 2012, Kordos 2012, Kowalczyk 2013].

The established role of rotating surveys theory is connected with the role of repeated rotating surveys in central statistical offices. Many of the most important surveys, both in Poland and other countries, are rotating surveys, e.g.

labour force surveys, household budget surveys. Repeated surveys are usually of multi-purpose nature. They aim to estimate population parameters on each current occasion, to estimate difference between two successive population means (i.e. net changes), ratio of two population means, various components of individual changes, combined population means from several periods etc. Additionally repeated surveys also aim to aggregate sample in time, which is of particular importance in measurement of rare events and rare populations. To take into account conflicting aims of repeated surveys they are often conducted in rotating manner, which means that they are based on overlapping samples. More precisely, a sample on each occasion consists of two parts: a part that has been also examined on previous occasion (matched part) and a part that is new in the sample, i.e. has not been examined on previous occasion (unmatched part). For more than two occasions rotating scheme becomes more complicated.

THE BASIS OF THE PROBLEM

Among many problems connected with rotating surveys the following one is of particular importance: no auxiliary information is available, we base only on a sample (overlapping) from all previous occasions and we want to increase precision of the population mean estimation on the current occasion by using all information from the sample, also from prior occasions. The problem for model approach for rotating scheme without holes was solved by Patterson 1950, in randomized approach it was given by [Kowalczyk 2002]. Model approach for rotating schemes with holes under different assumptions was considered by [Kowalski 2009, Kowalski and Wesołowski 2010, Wesołowski 2010, Ciepiela et al. 2012].

In the present paper we first give theoretical results for two periods in randomized approach to introduce general problem and divergence between theory and practice of rotating surveys. Kowalczyk [2002] has proved that for two periods the best linear unbiased estimator of the population mean on the second occasion for rotating surveys of a finite population is the estimator of the form:

$$e_2 = a\bar{y}_{1U} - a\bar{y}_{1M} + c\bar{y}_{2U} + (1-c)\bar{y}_{2M}, \tag{1}$$

where

$$a = \frac{\dfrac{n_M}{n_2}\dfrac{n_{1U}}{n_1}}{1-\rho^2(Y_2,Y_1)\dfrac{n_{2U}}{n_2}\dfrac{n_{1U}}{n_1}}\,\frac{C(Y_2,Y_1)}{S^2(Y_1)},$$

and

$$c = \frac{\frac{n_{2U}}{n_2}\left(1-\rho^2(Y_2,Y_1)\frac{n_{1U}}{n_1}\right)}{1-\rho^2(Y_2,Y_1)\frac{n_{2U}}{n_2}\frac{n_{1U}}{n_1}}.$$

Its variance is given by:

$$D^2(e_2) = \left(\frac{1}{n_2}\frac{1-\rho^2(Y_2,Y_1)\frac{n_{1U}}{n_1}}{1-\rho^2(Y_2,Y_1)\frac{n_{2U}}{n_2}\frac{n_{1U}}{n_1}} - \frac{1}{N}\right)S^2(Y_2)\ .$$

Notation used here is the following:

n_t – sample size on the t-th occasion, $t=1,2$,

n_M – matched sample size,

n_t – unmatched sample size on the t-th occasion, $t=1,2$,

N – population size.

We have:

$$n_t = n_M + n_{tU}\,,\ t=1,2\,.$$

As it can be seen, coefficients a and c in formula (1) include population parameters, namely correlation coefficient $\rho(Y_1,Y_2)$ and regression coefficient $C(Y_1,Y_2)/S^2(Y_1)$, which in real surveys are usually not known. That problem is common for rotating surveys theory in general, also for model approach. The same applies for analogous estimators considered by [Patterson 1950, Kowalski 2009, Kowalski and Wesołowski 2010, Wesołowski 2010 and Ciepiela at al. 2012].

So important question arises. What happens if we substitute in formula (1) unknown population correlation coefficient and unknown population regression coefficient by its estimates given on the basis of the sample? Does this procedure still increase precision of the estimation? No mathematical theory is given relating to this problem because of the complicacy of coefficients a and c.

Moreover, as most rotating surveys are of multi-purpose nature, what happens to other population parameters estimation? Kowalczyk [2013] gave the following theoretical results for net changes estimation:

- if $\rho(Y_1,Y_2)>0$, then for all n_M we have:

  $D^2(e_2-\bar{y}_1) \le D^2(\bar{y}_2-\bar{y}_1)$,

- if $\rho(Y_t,Y_{t+1})<0$, then we have:

$$D^2(e_2-\bar{y}_1) \le D^2(\bar{y}_2-\bar{y}_1) \Leftrightarrow \frac{n_{2U}}{n_1} \le \frac{-S^2(Y_2)}{2C(Y_2,Y_1)}$$

and for combined population means from two successive periods:

- if $\rho(Y_1,Y_2)>0$, then we have:

$$D^2\left[(e_2+\bar{y}_1)\right]\le D^2\left[(\bar{y}_2+\bar{y}_1)\right]\Leftrightarrow \frac{n_{2U}}{n_1}\le\frac{S^2(Y_2)}{2C(Y_2,Y_1)}, \tag{2}$$

- if $\rho(Y_1,Y_2)<0$, then for all n_M we have:

$$D^2\left[(e_2+\bar{y}_1)\right]\le D^2\left[(\bar{y}_2+\bar{y}_1)\right].$$

Still the question has to be answered. Are the results valid if we substitute in estimator e_2 given by formula (1) unknown population correlation and regression coefficients by their estimates based on a sample?

To answer all the questions broad simulation study will be presented in the next section.

SIMULATION STUDY

Description of the population

For simulation study two finite populations[1] are considered. Population 1 is finite population generated from a multivariate normal distribution. Generated finite population parameters look as follows:

$$\begin{bmatrix}\bar{Y}_1\\ \bar{Y}_2\end{bmatrix}=\begin{bmatrix}4{,}916\\ 9{,}833\end{bmatrix},\quad S=\begin{bmatrix}6{,}109\\ 7{,}031\end{bmatrix},\quad \rho=\begin{bmatrix}1 & 0{,}601\\ 0{,}601 & 1\end{bmatrix},\quad N=10000$$

Population 2 is based on real data taken from agricultural censuses of 2002 and 1996. The population consists of 1575 rural areas and variable under study are:

Y_1 – sawn area of spring wheat in 1996,

Y_2 – sawn area of spring wheat in 2002.

Finite population2 parameters look as follows:

$$\begin{bmatrix}\bar{Y}_1\\ \bar{Y}_2\end{bmatrix}=\begin{bmatrix}19743{,}24\\ 19515{,}67\end{bmatrix},\quad S=\begin{bmatrix}21387{,}08\\ 23511{,}31\end{bmatrix},\quad \rho=\begin{bmatrix}1 & 0{,}7964\\ 0{,}7964 & 1\end{bmatrix},\quad N=1575$$

Description of the sample

Two different sample sizes were considered in the simulation study:

$$n_1=n_2=n=100 \text{ and } n_1=n_2=n=50.$$

[1]Details of populations and of a sampling scheme are given in Kowalczyk B. (2013). In the book the same populations were discussed but the problems considered were of different nature.

For a given sample size different matched fraction were taken into account:

$$p = \frac{n_M}{n_2} = \frac{n_M}{n_1} = 0{,}1; 0{,}2; 0{,}4; 0{,}6; 0{,}8; 0{,}9 \,.$$

For instance, $n = 100$ and $p = 0{,}1$ means that on both occasions 100 element samples were examined, out of which only 10 elements were examined on the first and second occasions together. Analogously, $n = 100$ and $p = 0{,}9$ means that out from 100 elements examined on the first occasion, 90 were also examined on the second occasion and 10 were additionally resampled. Unknown correlation and regression coefficients used in estimator e_2 given by formula (1) are estimate on the basis of 10 elements only in the first example and on the basis of 90 elements in the second example, although in both examples total sample sizes are the same.

For every sample size and every matched fraction sampling was repeated 1000 times.

Simulation results

In Tables 1-3 average absolute differences in percentage are juxtaposed. Average absolute difference for correlation coefficient ρ, regression coefficient β and estimator e_2 is defined respectively as:

$$\tfrac{|\hat{\rho}-\rho|}{\rho} \cdot 100\% \,,\ \tfrac{|\hat{\beta}-\beta|}{\beta} \cdot 100\% \,,\ \tfrac{|\hat{e}_2 - e_2|}{e_2} \cdot 100\% \,,$$

where ρ, β are real population values, $\hat{\rho}, \hat{\beta}$ are values assessed on the basis of a sample, e_2 is theoretical estimator given by formula (1), $\hat{e}_2$ is available in practice estimator constructed in such a way that the population correlation and regression coefficients that appear in formula (1) are substituted by their estimates on the basis of the sample.

Table 1. Average absolute difference in % for population 1, $n = 100$

p	0,1	0,2	0,4	0,6	0,8	0,9
Correlation coeff.	29,5	20,6	13	11,2	9,7	8,8
Regression coeff.	38,6	26,2	16,8	14,4	12,5	11,2
e_2	1	0,7	0,4	0,3	0,2	0,1

Source: own calculations

Table 2. Average absolute difference in % for population 2, $n = 100$

p	0,1	0,2	0,4	0,6	0,8	0,9
Correlation coeff.	17,6	13,5	10,6	8,9	8,1	7,9
Regression coeff.	34,4	26,9	20,7	17,8	16,3	15,3
e_2	2,7	1,9	1,2	0,7	0,5	0,3

Source: own calculations

Table 3. Average absolute difference in % for population 2, n=50

p	0,1	0,2	0,4	0,6	0,8	0,9
Correlation coeff.	24,8	17	13,5	11,2	10,3	9,8
Regression coeff.	48,6	33,9	26,3	22,4	20,6	20,5
e_2	6,1	3,4	2,1	1,4	0,9	0,6

Source: own calculations

Although correlation and regression coefficients assessed on the basis of the sample can differ substantially from real population values (from 7,9% up to 48,6%), substituting that real values by their assessments based on the sample in formula (1) does not influence estimator e_2 substantially (it changes the value of the estimator from 0,1% up to 6,1%).

In tables 4-6 efficiency of the estimation of estimatorse_2 and $\hat{e}_2$ compared to common sample mean is presented for different populations, sample sizes and matched fractions of the sample. Efficiency of the estimators e_2 and $\hat{e}_2$, i.e. theoretical estimator and estimator available in practice are defined respectively as:

$$eff(e_2) = \frac{MSE(\bar{y}_2)}{MSE(e_2)}, \ eff(\hat{e}_2) = \frac{MSE(\bar{y}_2)}{MSE(\hat{e}_2)}.$$

Table 4. Efficiency of mean estimation on the second occasion for population 1, n = 100

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff(e_2)$	1,060	1,089	1,094	1,083	1,049	1,032
$eff(\hat{e}_2)$	0,977	1,078	1,094	1,081	1,047	1,028

Source: own calculations

Table 5. Efficiency of mean estimation on the second occasion for population 2, n = 100

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff(e_2)$	1,166	1,227	1,250	1,240	1,130	1,079
$eff(\hat{e}_2)$	1,089	1,172	1,241	1,230	1,132	1,084

Source: own calculations

Table 6. Efficiency of mean estimation on the second occasion for population 2, n = 50

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff(e_2)$	1,169	1,165	1,234	1,200	1,152	1,085
$eff(\hat{e}_2)$	0,888	0,990	1,206	1,162	1,147	1,084

Source: own calculations

Substituting unknown correlation coefficient and regression coefficient in estimator e_2 by their assessments based on the sample in most cases increased efficiency of the population mean estimation on the second occasion compared to common sample mean. Efficiency of the estimation decreased only in the case of very low number of elements examined on both occasions, namely not greater than 10.

In Tables 7-9 efficiency of the estimation of net changes is presented. Efficiency of the estimation for estimators $e_2 - \bar{y}_1$ and $\hat{e}_2 - \bar{y}_1$ compared to difference of two usual sample means is defined respectively:

$$eff\,(e_2 - \bar{y}_1) = \frac{MSE(\bar{y}_2 - \bar{y}_1)}{MSE(e_2 - \bar{y}_1)},\ eff\,(\hat{e}_2 - \bar{y}_1) = \frac{MSE(\bar{y}_2 - \bar{y}_2)}{MSE(\hat{e}_2 - \bar{y}_1)}.$$

Table 7. Efficiency of net changes estimation for population 1, n = 100

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff\,(e_2 - \bar{y}_1)$	1,072	1,096	1,131	1,113	1,081	1,035
$eff\,(\hat{e}_2 - \bar{y}_1)$	1,054	1,088	1,139	1,114	1,081	1,032

Source: own calculations

Table 8. Efficiency of net changes estimation for population 2, n = 100

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff\,(e_2 - \bar{y}_1)$	1,024	1,359	1,453	1,364	1,248	1,138
$eff\,(\hat{e}_2 - \bar{y}_1)$	1,219	1,417	1,472	1,368	1,252	1,155

Source: own calculations

Table 9. Efficiency of net changes estimation for population 2, n = 50

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff\,(e_2 - \bar{y}_1)$	1,072	1,096	1,131	1,113	1,081	1,035
$eff\,(\hat{e}_2 - \bar{y}_1)$	1,054	1,088	1,139	1,114	1,081	1,032

Source: own calculations

In the case of net changes estimation for multi-purpose surveys, substituting unknown correlation coefficient and regression coefficient in estimator e_2 given by (1) by their estimates in all cases increased efficiency of net changes estimation, even for low number of elements examined on both occasions.

In Tables 10-12 efficiency of the estimation of combined sample means from two periods is presented. Efficiency of the estimation for estimators $e_2 + \bar{y}_1$ and $\hat{e}_2 + \bar{y}_1$ compared to summing usual sample means is defined respectively:

$$eff\,(e_2 + \bar{y}_1) = \frac{MSE(\bar{y}_2 + \bar{y}_1)}{MSE(e_2 + \bar{y}_1)},\ eff\,(\hat{e}_2 + \bar{y}_1) = \frac{MSE(\bar{y}_2 + \bar{y}_2)}{MSE(\hat{e}_2 + \bar{y}_1)}.$$

Table 10. Efficiency of the net changes estimation for population 1, n = 100

P	0,1	0,2	0,4	0,6	0,8	0,9
$eff\,(e_2 + \bar{y}_1)$	1	1,013	1,006	1,015	1,09	1,014
$eff\,(\hat{e}_2 + \bar{y}_1)$	0,934	1,010	1,001	1,013	1,008	1,012

Source: own calculations

Table 11. Efficiency of the net changes estimation for population 2, n = 100

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff\,(e_2 + \bar{y}_1)$	0,990	0,958	1,008	1,043	1,049	1,034
$eff\,(\hat{e}_2 + \bar{y}_1)$	0,806	0,872	0,990	1,021	1,044	1,034

Source: own calculations

Table 12. Efficiency of the net changes estimation for population 2, n = 50

p	0,1	0,2	0,4	0,6	0,8	0,9
$eff\,(e_2 + \bar{y}_1)$	0,995	0,965	0,999	1,055	1,031	1,026
$eff\,(\hat{e}_2 + \bar{y}_1)$	0,926	0,907	0,989	1,048	1,032	1,027

Source: own calculations

According to formula (2) applying estimator e_2 does not always increase precision of the combined population means estimation. The main problem considered in the paper is the influence of substituting unknown population coefficients by its estimates based on the sample. So we focus only on cases in which the effect of using estimator e_2 is different from that of using $\hat{e}_2$. This is the case of population 1, n = 100, p = 0,1 and population 2, n = 100, p = 0,4 only.

CONCLUSIONS

Substituting unknown population correlation and regression coefficients by its estimates on the basis of the sample and applying estimator that uses information from previous period caused decrease of the population mean estimation in three extreme cases only, namely for $np \le 10$. In all other cases efficiency of the mean estimation on the second occasion increased compared to applying usual sample mean. In the case of multi-purpose surveys using previous information and estimator $\hat{e}_2$ increased efficiency of net changes estimation in all considered cases. Estimation of combined population mean from two successive periods posed more of a problem. But this population parameter is rarely used in practice. Population mean on each current occasion and net changes are of utmost importance in real surveys.

REFERENCES

Berger Y.G., Priam R. (2010) Estimation of Correlations between Cross-Sectional Estimates from Repeated Surveys – an Application to the Variance of Change, Proceedings of the 2010 Statistic Canada Symposium.

Binder D.A., Dick J.P. (1989) Modelling and estimation for repeated surveys, Survey Methodology, vol. 15, no. 1, pp. 29–45.

Ciepiela P., Gniado K., Wesołowski J., Wojty M. (2012) Dynamic K-Composite Estimator for an arbitrary rotation scheme, Statistics in Transition - New Series, vol. 13 no. 1, pp. 7–20.

Eckler A.R. (1955) Rotation Sampling, Annals of Mathematical Statistics, vol. 26, pp. 664–685.

Feder M. (2001) Time series analysis of repeated surveys: The state-space approach. Statistica Neerlandica, vol. 55, pp. 182–199.

Fuller W.A. (1990) Analysis of Repeated Surveys, Survey Methodology, vol. 16, no. 2, pp. 167–180.

Fuller W.A., Rao J.N.K. (2001) A Regression Composite Estimator with Application to the Canadian Labour Force Survey, Survey Methodology, vol. 27, no. 1, pp. 45–51.

Jessen R. (1942) Statistical investigation of a farm survey for obtaining farm facts, Iowa Agricultural Station Research Bulletin, vol. 304, pp. 54−59.

Jones R. (1980) Best linear unbiased estimators for repeated surveys, Journal of the Royal Statistical Society, series B, vol. 42, pp. 221–226.

Kordos J. (1967) Metoda rotacyjna w badaniach reprezentacyjnych, Przegląd Statystyczny, No. 4, pp. 373–394.

Kordos J. (2012) Review of application of rotation methods in sample surveys in Poland, Statistics in Transition – New Series, vol. 13, no. 2, pp. 47–64.

Kowalczyk B. (2002) Badania reprezentacyjne powtarzalne w czasie, PhD Thesis, Kolegium Analiz Ekonomicznych SGH, Warszawa.

Kowalczyk B. (2003a) Badania reprezentacyjne powtarzalne przy założeniu populacji stałej w składzie, in: Metoda reprezentacyjna w badaniach ekonomiczno-społecznych, eds. J. Wywiał, Akademia Ekonomiczna w Katowicach, Katowice, pp. 109–124.

Kowalczyk B. (2003b) Estimation of the Population Total on the Current Occasion under Second Stage Unit Rotation Pattern, Statistics in Transition, vol. 6, no. 4, pp. 503–513.

Kowalczyk B. (2004) Wykorzystanie estymatorów ilorazowych do estymacji indeksu dynamiki zmian wartości średniej w populacji, Roczniki Kolegium Analiz Ekonomicznych SGH, No. 13, pp. 47–58.

Kowalczyk B. (2013) Zagadnienia estymacji złożonej w badaniach reprezentacyjnych opartych na próbach rotacyjnych, Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie

Kowalski J. (2009) Optimal estimation in rotation patterns, Journal of Statistical Planning and Inference, vol. 139, no. 4, s. 2429–2436.

Kowalski J., Wesołowski J. (2010) Recurrence optimal estimators for rotation cascade patterns with holes (unpublished manuscript).

Nedyalkova D., Qualite L., Tille Y. (2009) General Framework for the Rotation of Units on repeated Survey sampling, Statistica Neerlandica, vol. 63, no. 3, pp. 269–293.

Patterson H.D. (1950) Sampling on successive occasions with partial replacement of units, Journal of the Royal Statistical Society, Series B 12, pp. 241−255.

Rao J., Graham J. (1964) Rotation designs for sampling on repeated occasions, Journal of the American Statistical Association, vol. 50, pp. 492−509.

Scott A., Smith T. (1974) Analysis of repeated surveys using time series methods, Journal of the American Statistical Association, vol. 69, pp. 674−678.

Steel D.G. (2004) Sampling in Time, in: Encyclopedia of Social Measurement, eds. K. Kempf–Leonard, Elsevier Academic Press, Amsterdam.

Steel D., McLaren C. (2009) Design and Analysis of Surveys Repeated over Time, in: Sample surveys: design, methods and applications, eds. D. Pfeffermann, C.R. Rao, Handbook of Statistics, vol. 29B, Elsevier, Amsterdam.

Szarkowski A., Witkowski J. (1994) The Polish Labour Force Survey, Statistics in Transition, vol. 1, no. 4, pp. 467–483.

Wesołowski J. (2010) Recursive optimal estimation in Szarkowski rotation scheme, Statistics in Transition – New Series, vol. 11, no. 2, pp. 267−285.