STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 12, Number 2, October 2011

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Roman Popiński, Ph.D.: Secretary
r.popinski@stat.gov.pl; Phone number 00 48 22 — 608 33 66,
Waldemar Orlik: Technical Assistant

Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 237—238

FROM THE EDITOR

As one of the key objectives of our journal's mission is to provide a platform for cooperation and exchange of ideas and information among statisticians in the broadly conceived region, I am pleased to have published in this issue a set of papers based on presentations at ***The Third Baltic-Nordic Conference on Survey Statistics.*** The papers were collected and prepared for publication by Professor Daniel Thorburn, who kindly accepted our invitation to act as a guest editor of this part (which is the major part) of the issue.

Other articles are devoted to either estimation or dynamic analysis problems. The former are represented by two paper. One, by **V. Archana and K. A. Rao,** ***Improved Estimators of Coefficient of Variation in a Finite Population,*** is devoted to elaboration of several new versions of the coefficient of variation (V.V.) for the describing dispersion in the finite populations, based on the regression estimators. The regression estimator using the information on the Population C.V of the auxiliary variable emerges as the best estimator.

In another paper, ***Improved Separate Ratio Exponential Estimator for Population Mean Using Auxiliary Information,*** **R. Yadav, L. N. Upadhyaya, H. P. Singh** and **S.Chatterjee** present the improved separate ratio exponential estimator for population mean using the information based on auxiliary variable in stratified random sampling. The theoretical and numerical comparisons are carried out to show the efficiency of the suggested estimator over sample mean estimator, usual separate ratio and separate product estimator.

The issue of forecasting a number of new firms within a birth-and-death framework motivates **H. Gurgul and P. Zając** paper ***The Dynamic Model of Birth and Death of Enterprises***. The model aims to describe the dynamics of bankruptcy and foundation of new enterprises and to forecast the number of new firms founded using dynamic mathematical tools

In concluding section, two international **conferences** are reported briefly. The first ,The ISI Satellite Conference on Improving Statistical Systems Worldwide - Building Capacity was held in Krakow, Poland from 18 to 19 August 2011, as a satellite of the 58th ISI Congress held in Dublin, Ireland, from 21 to 26 August 2011. The conference was organized by the Polish Statistical Office in cooperation with the World Bank, African Development Bank, Paris 21 and the International Statistical Institute. It was financed by the World Bank,

Polish Statistical Office and ISI. Second conference, on Survey Sampling in Economic and Social Research, took place 18-20 September 2011 in Katowice and was organized by the Department of Statistics of the University of Economics in Katowice with the cooperation of the Department of Statistical Methods of the University of Łódź and Polish Statistical Association.

Włodzimierz OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 239

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition – new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series.*

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl., followed by a hard copy addressed to
Prof. Wlodzimierz Okrasa,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://www.stat.gov.pl/pts/15_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 241—242

FROM THE GUEST EDITOR

The Third Baltic-Nordic Conference on Survey Sampling (BaNoCoSS) was held in June 2011 in a small Swedish village called Norrfällsviken. The conference belongs to a long series of scientific conferences and workshops, which was initiated in 1997 by Professor Gunnar Kulldorff from the University of Umeå. The two previous BaNoCoSS conferences were held Ammarnäs, Sweden and Kuusamo, Finland. Workshops on survey sampling theory and methodology have been organized annually in different Baltic, Nordic and Ukrainian countries by the Baltic-Nordic Network in Survey Sampling. The network has during the last been extended to include also Ukraine. It consists of people from university departments, national statistics institutes and statistical societies of the respective countries. There were 60 participants at this conference from 16 different countries with 45 different contributions or invited speakers.

Statistics in Transition has always been so kind to publish a special issue with a selection of the talks. This journal is a good choice. With the impact of modern technology and the changing society, survey sampling is in rapid transition. When I was a young Ph D student in the seventies I read the book by Cochran. At that time one felt that everything in sampling theory was done and nothing remained. Later, when I got a job at Statistics Sweden in 1975 I had the same feeling. I also felt that traditional sampling was a side track on the tree of statistics and was going to be replaced by the same type of statistics as everywhere else where the observations were modelled as random variables.

On the first issue I was wrong. The computers have given rise to many new possibilities and opportunities. It is now feasible to use more complicated sampling schemes and estimators and also to use much more auxiliary variables. The BaNoCoss conferences are proofs of the vigorous development in modern sampling. At this conference e.g. the keynote speaker Jean Claude Deville talked about the development towards an extensive use of auxiliary variables. But on the second issue I may still be right. Models, random variables and processes are creeping into sampling. The two other keynote speakers were good examples of this. Giovanna Ranalli has taken steps towards integrating traditional methods like kernel estimation into sampling. Steven Thomson has broadened sampling towards spatial and network sampling.

We have selected ten papers in this special issue of *Statistics in Transition* from the 45 topics presented at the BaNoCoSS-conference. Together they show the diversity and the vitality characterising the field of Survey sampling today. When I started at Statistics Sweden I was also struck by the deep cleft between sampling theory and nonsampling errors. Institutes like Bureau of the Census and

people and professor Tore Dalenius were talking about "Total Survey Models" but almost all statisticians worked with one or the other at a time. One might say that the cleft is still there today but some parts are being filled in. In this issue there are e.g. three papers on the technical treatment of non-response. Edgar Bueno Castellanes discusses how to correct for non-response in a Colombian survey. Andre Neri and Giovanna Ranalli discuss the balance between non-response and measurement errors and try to correct for both in the Italian Survey on Household Income and Wealth. Nicklas Pettersson discusses multiple imputation and shows how techniques from kernel estimation can be used there to improve the precision of nonresponse correction.

Another area where models play an important role is estimation for smaller domains. Vilma Nekrasaite considers using panel type data for improving small area estimation by using also previous data and Kari Lumiste is interested in consistent estimation under cross classification. Ari Veijanen and Risto Lehtonen use a mixed model to estimate an income distribution. But to make the estimates to lie closer to survey data the derived distribution is adjusted to bring the percentiles closer to the observed values.

In survey sampling it may sometimes be important to protect the integrity of the respondents. Artem Shcherbina and Maiboroda Rostyslav look at mixtures of populations and try to estimate the components so that the respondents are protected by the mixtures. Dalius Pumputis and Andris Ciginas generalises linear calibration methods to handle also second order functions.

Julia Orlova has made a study to see how statisticians allocate their working time between different working tasks. Anna Larchenko finally advocates that certain statistics should get a higher priority in order to improve the situation for women, mothers and children in Belarus.

I am very proud of this issue, the conference and its mix of contributions, which really lie on the forefront of modern survey sampling. I hope that you will enjoy this issue and that you will be tempted to attend the next conference. It will probably be arranged in four years time. Finally, I want to thank all the authors and referees for their work, the participants of the conference for making it so interesting and finally *Statistics in Transition* for publishing this issue.

Daniel Thorburn
Guest Editor
University of Stockholm

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 243—252

THE MOTHERS' INDEX AS AN INDICATOR OF DEMOGRAPHIC SECURITY

Anna Larchenko[1]

ABSTRACT

The Mothers' Index was developed in 1998 and has been in use for ranking countries since 2000. The author advocates that this indicator should not only be used for cross-country analysis of maternal and child health but also as be part of the demographic indicators of national security in Belarus. A grouping of more developed countries in terms of the Mothers' Index is proposed. The Mothers', the Women's and the Children's Indices from 2007 to 2009 are calculated for Belarus, the post-Soviet countries and some EU countries.

Key words: demographic security, depopulation, the Mothers' Index, the Women's Index, the Children's Index.

1. Introduction

The Mothers' Index is annually calculated by the international humanitarian organization "Save the Children" (2003-2011) in order to compare the conditions of motherhood in different countries and to identify countries where mothers face the greatest problems. The quality of children's life depends on the health, safety and welfare of their mothers, as it is primarily the mother who cares for her child. The countries are divided into three groups after their degree of development. Belarus is included in the group of more developed countries, and the Mothers' Index is calculated for it since 2003.

In Belarus' socio-demographic situation, the issue of reproductive health has become most important. The situation is characterized by depopulation, aging population, low birth rates, high mortality rates and the degradation of the family institution. Maternal health, child health and social status of the mother are today considered only in general terms but the age-specific fertility and mortality should be analyzed in some detail.

[1] Belarus State Economic University, Minsk.E-mail:hanna-larchenko@mail.ru

In this paper the author discusses the Mothers' Index in more detail and also proposes to include the Mothers' Index into the National System of Demographic Indicators.

2. The methods for the Mothers' Index calculation

The Mothers' Index was developed in 1998. It has been used since 2000 with some changes in 2006. In turn, the Mothers' Index includes two sub-indices – the Women's Index and the Children's Index (Tables 1 and 2). Before 2006 the Women's Index was calculated in the same way for all countries but not exactly the same indicators.

For calculation of the Mothers' Index each indicator is standardized in the group. The standardized indicators with negative impact are multiplied by minus one. i.e. the lifetime risk of maternal mortality, child mortality rate, ratio of underweight children under 5. In the author's opinion the percent of women using modern contraception ought also to have negative impact. On the one hand, the use of contraceptives can be considered as a positive indicator (preventing the spread of sexually transmitted infections. On the other hand, the most common contraceptives in the more developed countries are hormonal contraceptives which in the future may lead to a violation of women's health.(see also Anokhin, 1995, Barabanov, 2004)

Table 1. The list of indicators in the Women's Index for 2007-2009 and country groups

Country group	The list of indicators								
	Health status				Educational status	Economic status			Political status
	Lifetime maternal mortality	Women using modern contraception %	Female life expectancy at birth	Births attended by health personnel %	Number of years of formal female schooling	Maternity leave length	Maternity leave benefit %	Ratio of female to male income	Participation of women (% seats)
More developed	+	+	+	-	+	+	+	+	+
less developed	+	+	+	+	+	-	-	+	+
Least developed	+	+	+	+	+	-	-	+	+

To avoid distortion of data by school systems where pupils do not start training on time or fail to achieve certain educational standards some adjustments are made. For example if the gross pre-primary enrollment ratio and gross primary enrollment ratio lies between 100 and 105 percent, they are discounted to 100 percent. Any value over 105 percent is decreased by 5%.

The standardized indicators are weighted together to determine the different indices. For the women's economic status, the weights are: ratio of estimated

female to male earned income – 75.0%; duration of maternity leave – 12.5%; size of the payment – 12.5%.

The Women's Index is calculated with the weights: indicator of women's health –30%; indicator of educational status –30%; indicator economic status – 30%; indicator of political status –10%. The Children's Index is calculated as simple arithmetic mean.

The Mothers' Index is finally calculated with the following weights: indicator of children 's status – 30%; indicator of women's health – 20%; indicator of educational status – 20%; indicator economic status – 20%; indicator of political status – 10%.

For a more thorough description see Save the Children (2003-2011).

Table 2. The list of indicators included into the Children's Index for various periods and country groups

Country group	The list of indicators						
	Under-5 mortality rate (per 1,000 live births)	Gross pre-primary enrollment ratio %	Gross primary enrollment ratio (% of total)	Gross secondary enrollment ratio (% of total)	children under 5 moderately or severely underweight for age %	Percent of population with access to safe water	Ratio girls/boys enrolled in primary school
More developed	+	+	-	+	-	-	-
less developed	+	-	+	+	+	+	-
Least developed	+	-	+	-	+	+	+

3. The Women's Index, the Children's Index and the Mothers' Index calculation for the Republic of Belarus

3.1. More developed countries according to Save the Children

In the period from 2007 to 2009 the list of more developed countries included 41 countries (Table 5). The means and standard deviations of the indicators are calculated in Table 3

Table 3. Mean and standard deviation calculated by the group of more developed countries, and standardized values of indicators for Belarus in 2007

Indicators included into the Index	Block	Mean	Standard deviation	z-value of Belarus
Lifetime risk of maternal mortality	Health status	$3.259 \cdot 10^{-4}$	$3.164 \cdot 10^{-4}$	-0.7258
Percent of women using modern contraception		53.27	18.56	0.2839
Female life expectancy at birth		79.98	3.25	-1.8388
Expected number of years of formal female schooling	Educational status	15.76	1.97	-0.3833

Table 3. Mean and standard deviation calculated by the group of more developed countries, and standardized values of indicators for Belarus in 2007

Indicators included into the Index	Block	Mean	Standard deviation	z-value of Belarus
Maternity leave benefit (length)	Economical status	148.51	77.73	-0.2896
Maternity leave benefit (% wages paid)		85.27	23.70	0.6216
Ratio of estimated female to male earned income		0.61	0.09	0.3809
Participation of women in national government	Political status	22.50	10.00	0.6501
Under-5 mortality rate	Children's Index	7.76	4.64	-0.9142
Gross pre-primary enrollment ratio		79.27	16.93	1.2243
Gross primary enrollment ratio		91.83	8.73	0.1342

Based on Table 3 and the weights given above, the indicator of women's economic status in Belarus in 2007 is 0.327 and the indicator of women's health state is:-0.6030. In the same way the Women's Index becomes -0.1327 and the Children's Index is 0.1481. Finally the Mothers' Index is equal to -0.0224. Similar calculations were made for 2008 and 2009. Interim data and the results of the calculations are presented in Table 4.

In Table 5 the 41 more developed countries are divided into three groups depending on their Mothers' Index (high over 0.2, middle -0.6-0.2 - and low below -0.6). The composition of each group varies between the years, albeit marginally. Belarus is in the middle level of the Mothers' Index group (without changing its position), while Latvia, Russia, Ukraine and some other countries change their positions.

3.2. Two special subgroups: former Soviet Union countries and some EU countries

The author computed the Mothers' Index for two separate subgroups: the post-Soviet countries in the European region, and a group of some EU countries. Seven countries in EU (the most and least developed) were selected for comparison with the seven post-Soviet countries in Europe. Using the formulas 1-3 the author has

Table 4. Mean and standard deviation, calculated by the group of more developed countries, and standardized values of indicators for Belarus in 2008-2009

Indicators included into the Index	2008			2009		
	Mean ($\bar{x}$)	Standard dev	*z-value*	Mean ($\bar{x}$)	Standard dev	z-value
Lifetime risk of maternal mortality	$1.676 \cdot 10^{-4}$	$3.073 \cdot 10^{-4}$	0.3254	$1.676 \cdot 10^{-4}$	$1{,}337 \cdot 10^{-4}$	-0,0735
Percent of women using modern contraception	55.27	17.56	0.7558	57.07	16.35	0.0656
Female life expectancy at birth	80.12	3.28	-1.8637	80.83	3.22	-1.4985
Expected number of years of formal female schooling	16.00	2.04	-0.4911	15.95	2.04	-0.4673
Maternity leave benefit (length)	134.63	63.22	-0.1366	170.46	103.73	-0.4286
Maternity leave benefit (% wages paid)	83.27	24.35	0.6870	81.85	24.37	0.7445
Ratio of estimated female to male earned income	0.61	0.09	0.2322	0.62	0.09	0.1483
Participation of women in national government	23.27	10.16	0.5639	24.34	10.24	0.7481
Under-5 mortality rate	7.61	5.14	-1.0494	6.78	3.74	-1.6608
Gross pre-primary enrollment ratio	81.17	16.72	1.1261	83.76	15.91	1.0210
Gross primary enrollment ratio	91.39	9.18	0.3932	91.85	8.74	0.3599
Women's Index (I_W)	-0.0963			-0.1309		
Children's Index (I_C)	0.1566			-0.0933		
Mothers' Index (I_M)	0.0016			-0.0903		

Table 5. Country groups with high, medium and low Mothers' Index among more developed countries 2007-2009

Countries with high Mothers' Index	
2007	Australia, Belgium, Denmark, Finland, Germany, Iceland, Spain, Lithuania, New Zealand, Norway, Portugal, Spain, Sweden
2008	Australia, Denmark, Estonia, Finland, Germany, Iceland, Italy, Latvia, Lithuania, Netherlands, New Zealand, Norway, Portugal, Slovenia, Sweden, Switzerland
2009	Australia, Denmark, Estonia, Finland, Germany, Iceland, Italy, Lithuania, Netherlands, New Zealand, Norway, Sweden
Countries with middle level of Mothers' Index	
2007	Austria, Belarus, Bulgaria, Canada, Croatia, Czech Republic, Estonia, France, Greece, Hungary, Ireland, Italy, Japan, Latvia, Luxembourg, Malta, Netherlands, Poland, Slovakia, Slovenia, Switzerland, United Kingdom, United States
2008	Austria, Belarus, Belgium, Bulgaria, Canada, Croatia, Czech Republic, France, Greece, Hungary, Ireland, Japan, Luxembourg, Malta, Poland, Romania, Russian Federation, Slovakia, Spain, United Kingdom, United States
2009	Austria, Belarus, Belgium, Bulgaria, Canada, Croatia, Czech Republic, France, Greece, Hungary, Ireland, Japan, Latvia, Luxembourg, Malta, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Switzerland, Ukraine, United Kingdom, United States
Countries with low Mothers' Index	
2007	Albania, Macedonia, Moldova, Romania, Russian Federation, Ukraine
2008	Albania, Macedonia, Moldova, Ukraine
2009	Albania, Macedonia, Moldova, Russian Federation

determined the countries ranking on three indices: Mothers', Women's and Children's for 2007-2009 (Tables 6 and 7). The leading positions for the three indices in the former Soviet Union group are held by the three Baltic countries: Latvia, Lithuania and Estonia, which in 2004 joined the European Union. Belarus ranked second among post-Soviet countries on the Children's Index during 2007-2008. But in 2009 it lost to fourth position. The situation deteriorated also for the Russian Federation.

In the EU-group the leading position was held by Finland and Sweden. In these countries favorable conditions for mothers and children have been created. In addition, these countries have the highest female life expectancy in Europe. The last position in the ranking occupies Poland, due to the relatively high lifetime risk of maternal mortality and under-5 mortality rate in the country, as well as low gross pre-primary enrollment ratio.

Table 6. Mothers' Index, Women's Index and the Children's Index in European countries for the period 2007-2009

Country	2007			2008			2009		
	Women's Index	Children's Index	Mothers' Index	Women's Index	Children's Index	Mothers' Index	Women's Index	Children's Index	Mothers' Index
Post-Soviet countries									
Belarus	0.169	0.487	0.308	0.119	0.648	0.324	-0.086	0.275	0.078
Estonia	0.144	1.475	0.539	0.130	1.498	0.544	0.281	0.879	0.462
Latvia	0.505	0.239	0.408	0.652	0.488	0.583	0.467	1.166	0.667
Lithuania	1.194	0.378	0.939	1.126	0.059	0.787	0.799	0.516	0.693
Moldova	-0.619	-1.642	-0.891	-0.609	-1.599	-0.873	-0.183	-1.853	-0.662
Russian Federation	-0.639	-0.476	-0.613	-0.605	-0.259	-0.511	-0.661	-0.782	-0.707
Ukraine	-0.754	-0.461	-0.690	-0.811	-0.835	-0.853	-0.617	-0.200	-0.531
EU countries									
Czech Republic	-0.883	0.455	-0.479	-0.907	0.508	-0.484	-0.607	0.765	-0.211
Germany	-0.141	0.619	0.108	-0.202	0.652	0.074	-0.166	0.856	0.162
Finland	0.851	1.753	1.126	0.564	0.884	0.682	0.854	1.282	0.991
France	-0.330	2.788	0.578	-0.027	2.448	0.690	-0.430	2.271	0.367
Italy	-0.236	0.886	0.084	0.096	0.678	0.238	0.020	0.599	0.172
Poland	-0.203	-1.350	-0.557	-0.253	-1.134	-0.530	-0.339	-1.072	-0.572
Sweden	0.942	1.225	1.053	0.731	1.034	0.852	0.668	1.194	0.859

4. Preconditions for inclusion the Mothers' Index into the Indicators of Demographic Security

In accordance with the Law of the Republic of Belarus "On the demographic security of Belarus" (2010) the indicators of demographic threats are: net reproduction rate; the coefficient of depopulation; the total fertility rate; death rate of working age population, including mortality rates for men and women of working age; life expectancy; rate of population aging; balances of migration exchange between rural and urban areas, including by gender and age; education level; number of illegal migrants; marriage rates and divorce rates.

Table 7. Ranking of European countries on the Mothers' Index, Women's Index and the Children's Index for 2007-2009

Country	2007			2008			2009		
	Women's Index	Children's Index	Mothers' Index	Women's Index	Children's Index	Mothers' Index	Women's Index	Children's Index	Mothers' Index
Post-Soviet countries									
Belarus	4	2	4	4	2	4	4	4	4
Estonia	3	1	2	3	1	3	3	2	3
Latvia	2	4	3	2	3	2	2	1	2
Lithuania	1	3	1	1	4	1	1	3	1
Moldova	5	7	7	6	7	7	5	7	6
Russian Federation	6	6	5	5	5	5	7	6	7
Ukraine	7	5	6	7	6	6	6	5	5
EU countries									
Czech Republic	7	6	6	7	6	6	7	5	6
Germany	3	5	4	5	5	5	4	4	5
Finland	2	2	1	2	3	3	1	2	1
France	6	1	3	4	1	2	6	1	3
Italy	5	4	5	3	4	4	3	6	4
Poland	4	7	7	6	7	7	5	7	7
Sweden	1	3	2	1	2	1	2	3	2

There is no Mothers' Index in this list. However, these indicators include demographic, social, and economic characteristics of the mothers' and children's states which, in the author's opinion, motivate their inclusion among the indicators characterizing demographic situation and, in particular the demographic security. We argue that the following two indicators of women's health Mothers' Index should also be included in Belarus with negative influence: The ratio of abortion for 1000 live birth and The level of women's infertility (the percentage of women between 15 and 49 who from medical, biological or social reasons do not get pregnant (Anokhin, 1995).

The level of the indicators is rather high for Belarus (in 2009 ratio of abortion for 1000 live birth was 281.94; the level of women's infertility in 2009 – 63.5 per 100 000 persons of relevant sex). Here is the official data obtained from reports of the state medical institutions. The Mothers' Index calculated for Belarus with the proposed indicators will reflect the mothers' real health status.

Most medical services in Belarus are free of charge. This fact significantly affects their quality. To improve the quality of health services a compulsory health insurance would be appropriate, as well as the training of medical personnel abroad.

The Mothers' Index for Belarus should be computed in two ways: in relation to the other post-Soviet countries and also in relation to other more developed countries. It should also be included in the "Women and men in the Republic of Belarus: statistical book" (last issue, 2010):

5. Concluding remarks

The proposed introduction of the Women's Index, the Children's Index and the Mothers' Index would make it possible to determine the level of the mothers' and children's health in the Republic of Belarus. This would help the government in obtaining a good and fair health system. These approaches also allow Belarus to carry out cross-country assessments, having selected among the European countries the group of post-Soviet countries, as well as to assess the rank of Belarus on the indicators for children and mothers.

REFERENCES

ANOKHIN, L.V. (1995): Infertility in Marriage: medico-social aspects. *Ryazan State Medical University,* 130.

BARABANOV, L.G. (2004): Modern principles of organization of medical care for patients with sexually transmitted infections. *Medicine*, 3, 14-16.

Law of the Republic of Belarus "On the demographic security of Belarus". January 4, 2002, № 80.

Save the Children (2003-2011). State of the World's Mothers int. report. *NY:*

Women and men in the Republic of Belarus: statistical book (2010): *Minsk: National Statistical Committee of the Republic of Belarus 2010*, 205.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 253—264

CONSISTENT ESTIMATION OF CROSS-CLASSIFIED DOMAINS

Kaur Lumiste[1]

ABSTRACT

Domain estimation has become an important area in survey sampling, however a lot of problems associate with it. One out of many is the lack of consistency between different surveys. The results of one survey do not coincide with the results of another survey done earlier or simultaneously, although the same variable is under study.
We study two methods, AC-calibration (A – auxiliary, C – common) and repeated weighting (RW), for achieving consistency. A short overview of these two methods is given, and we develop formulas for a specified case, for the cross-classified domains.
We assume that there are two sources of information on the study variables (either surveys or registers). The problem is that one source has information on domains formed by certain categorical variable, not considered or not identified in the other source. Instead, this second source has information on domains formed by another categorical variable. We are however interested in domains cross-classified with these categorical variables. A survey is done regarding these new domains, but the domain estimates will probably be inconsistent with marginal information, from earlier surveys. We require that the new domain estimates be consistent with the marginals. To achieve this we apply AC-calibration or repeated weighting.
The formulas of AC-calibration and RW for the cross-classified domains are tested in a simulation study. Simulations were done on a population composed of real data from the Estonian Household Survey.

Key words: Consistent estimation; calibration; AC-calibration; repeated weighting.

1. Introduction

In National Statistics Agencies it is standard practice to provide estimates for certain smaller groups or subpopulations in the population, called domains. The

[1] University of Tartu. E-mail: klumiste@ut.ee

main research topics in domain estimation are how to increase accuracy and small area estimation (Estevao and Särndal, 2004).

Today's statistics consumers demand high quality statistics and several National Statistics Agencies have formulated their own definitions to the term "quality in (official) statistics" (Platek and Särndal, 2001). Elvers and Rosén (1999) give five principal dimensions: (1) Contents, (2) Accuracy, (3) Timeliness, (4) Coherence and Comparability, and (5) Availability and Clarity. Each has a number of sub-dimensions.

In modern practice it is quite usual that several surveys are carried out on the same population with some variables common in two or more surveys. Consistency between different surveys is one important indicator of quality, falling under Coherence and Comparability in Elvers and Rosén's quality dimensions. Here the term "consistent" refers not to "randomization consistent" but to "consistent with known aggregates".

In this study we assume that population totals, or totals of larger domains, for certain variables are estimated in one survey, that we call *the reference survey* (RFS). Estimates of the same variables are produced, but in greater detail (totals of domains) in another survey, that we call *the present survey* (PRS). The two surveys are carried out independently. Naturally consumers would assume that these two surveys produce numerically consistent estimates - e.g. domain total estimates in PRS sum up to the population total estimate in RFS. Sadly this does not always hold. Särndal and Traat (2011) treat the inconsistency problem and they propose a new method, the AC-calibration (A – auxiliary variables, C – common variables). Statistics Netherlands has also studied the problem and several articles have been published on repeated weighting (RW). The current study aims to use these two methods to achieve consistency in a more specific case.

We assume that there are two sources of information on the study variables (either surveys or registers). But the problem is that one source has information on domains formed by a certain categorical variable, not considered or not identified in the other source. Instead, this second source has information on domains formed by another categorical variable. We are however interested in domains formed by the cross-classification of the previous categorical variables. A survey is done regarding these new domains, but the domain estimates will probably be inconsistent with earlier information. So we take the earlier information, insert this as marginals to our 2-way table, and demand that the new found domain estimates are consistent with the marginals. To achieve this we apply AC-calibration or repeated weighting and test the new formulas in a simulation study.

First the necessary notations and terminology is given, then the concept of consistency is discussed, followed by an overview of the AC-calibration and the RW method. Formulas for our special case are derived in Section 4 and tested in a simulation study in Section 5.

2. Preliminaries

Let $U = (1, 2, \dots, N)$ denote a finite population of N units that is divided into D non-overlapping domains U_d, $d \in \mathfrak{D} = \{1, 2, \dots, D\}$. Let a random vector (design vector) $\mathbf{I} = (I_1, I_2, \dots, I_N)$ describe the sampling process on U and I_i is the sample inclusion indicator for unit $i \in U$. The probability sampling design generates for element i a known inclusion probability, $E(I_i) = \pi_i > 0$, and a corresponding sampling design weight $a_i = 1/\pi_i$. The column vector $\mathbf{y}_i : m \times 1$ of the study variables $\mathbf{y}$ is recorded for all $i \in s$ (complete response). Our objective is to estimate population totals $\mathbf{Y} = \sum_U \mathbf{y}_i$ and domain totals $\mathbf{Y}_d = \sum_{U_d} \mathbf{y}_i$.

The basic design unbiased estimator of $\mathbf{Y}$ is $\widehat{\mathbf{Y}} = \sum_s \mathrm{a_i}\mathbf{y_i}$, the Horwitz-Thopmson (HT) estimator. For domains it can be written $\widehat{\mathbf{Y}}_d = \sum_{s_d} a_i \mathbf{y}_i = \sum_s a_i \delta_i^d \mathbf{y}_i$, $d \in \mathfrak{D}$, where

$$\delta_i^d := \begin{cases} 1, & i \in U_d, \\ 0, \text{ otherwise.} \end{cases}$$

It is, however, inefficient when strong auxiliary information is available for use at the estimation stage.

Auxiliary information vector $\mathbf{x}_i : p \times 1$ can be found for every element $i \in s$ (or every $i \in U$ if it is compiled from comprehensive registers). It is assumed that the auxiliary variable totals $\mathbf{X} = \sum_U \mathbf{x}_i$ are known.

One way of using the auxiliary information is by calibrating. There are several approaches to calibration, but the two main methods are the distance minimization, used by Deville and Särndal (1992), and the instrument vector method considered in Estevao and Särndal (2000) and Kott (2006). A good overview of both methods is given by Särndal (2007), and he shows that the distance minimization method is a special case of the instrument vector method (also shown by Estevao and Särndal (2000) and Kott (2006)). We will use the instrument vector approach, since it is a more general method.

We find the new weights w_i, such that they satisfy the calibration equations

$$\sum\nolimits_U \mathbf{x}_i = \sum\nolimits_s w_i \mathbf{x}_i = \mathbf{X}. \tag{2.1}$$

The new weights are in the form

$$w_i = a_i(1 + \boldsymbol{\lambda}'\mathbf{z_i}) \tag{2.2}$$

where $\boldsymbol{\lambda}' = \left(\mathbf{X} - \widehat{\mathbf{X}}\right)' \mathbf{M}^{-1}$, $\mathbf{M} = \sum_s a_i \mathbf{z}_i \mathbf{x}_i'$, $\widehat{\mathbf{X}} = \sum_s a_i \mathbf{x}_i$ and vector $\mathbf{z}_i : p \times 1$ is called an instrument vector and is chosen freely.

3. Consistent estimation

Let us assume that the population totals **Y** are known, either from registers or RFS where the population was not divided into desired domains. It is natural to demand that the domain totals in the PRS sum up to the corresponding known totals.

For this, however, we have to make the following assumptions:

- data from registers and surveys is taken on the same time moment;
- all sources consider the same population;
- all common variables have the same definition.

If any of these assumptions is not met, then we are trying to achieve numerical consistency with variables that do not have to be consistent. Fulfilling these assumptions can become an obstacle. The first assumption is satisfied if the RFS data is taken from a register which is constantly updated in real-time, RFS and PRS are carried out simultaneously or the time between them is deemed acceptable. The satisfaction of the last two assumptions can be ensured in the planning phase of the PRS - using the same population and defining variables as they are in RFS.

The two methods discussed below can be used to achieve consistency with common variables.

AC-calibration

In this section we give a short overview of the AC-calibration method developed by Särndal and Traat (2011). The authors named it after the two sources of information used to calibrate new weights - auxiliary (A) and common (C) variables (further the terms A-variables, A-information and C-variables, C-information are used).

AC-calibration is basically standard calibration where the auxiliary variable vector is extended with C-variables. For each element $i \in s$, we can find vectors $\mathbf{x}_i$: $p \times 1$ and $\mathbf{y}_i$: $m \times 1$. We define new $(p + m)$ dimensional vectors:

$$\begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}, i \in s, \qquad \begin{pmatrix} \mathbf{X} \\ \mathbf{Y}_0 \end{pmatrix}, \qquad \begin{pmatrix} \widehat{\mathbf{X}} \\ \widehat{\mathbf{Y}} \end{pmatrix}, \tag{3.3}$$

where $\mathbf{X} = \sum_U \mathbf{x}_i$, $\mathbf{Y}_0$ is the vector of known C-information totals from RFS and $\widehat{\mathbf{X}} = \sum_s a_i \mathbf{x}_i$ and $\widehat{\mathbf{Y}} = \sum_s a_i \mathbf{y}_i$ are HT estimates found from PRS. The calibration weights are then

$$w_{ACi} = a_i(1 + \boldsymbol{\lambda}'_{AC}\mathbf{z}_{ACi}) \tag{3.4}$$

where $\mathbf{z}_{ACi}$ is the instrument vector and

$$\boldsymbol{\lambda}_{AC}^{'} = \begin{pmatrix} \mathbf{X} - \widehat{\mathbf{X}} \\ \mathbf{Y}_0 - \widehat{\mathbf{Y}} \end{pmatrix}^{'} \mathbf{M}^{-1}, \qquad \mathbf{M} = \sum_s a_i \mathbf{z}_{ACi} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}^{'}.$$

The new weights are consistent with A-information as well as C-information, meaning

$$\sum_s w_{ACi} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y}_0 \end{pmatrix}.$$

The inverse of the matrix $\mathbf{M}$ exists if $\mathbf{x}$ and $\mathbf{y}$ are linearly independent.

Repeated weighting

The repeated weighting method has been developed in Statistics Netherlands. The following is a brief overview of the method, more can be read from Kroese and Renssen (1999), Houbiers (2004), Knottnerus and Van Duin (2006).

The idea of calibration is that we change the initial weights a_i so that they satisfy the calibration equations (2.1). Repeated weighting is basically calibrating the already calibrated weights w_i once again, but in the second step the C-information $\mathbf{Y}_0$ is used in the calibration equations.

With RW we get weights

$$w_{RWi} = w_i(1 + \boldsymbol{\lambda}_{RW}^{'} \mathbf{z}_{RWi}) \tag{3.5}$$

where $\mathbf{z}_{RWi}$ is the instrument vector and

$$\boldsymbol{\lambda}_{RW}^{'} = \left(\mathbf{Y}_0 - \widehat{\mathbf{Y}}^{CAL}\right)^{'} \mathbf{M}^{-1}, \qquad \mathbf{M} = \sum_s w_i \mathbf{z}_{RWi} \mathbf{y}_i^{'}, \qquad \widehat{\mathbf{Y}}^{CAL} = \sum_s w_i \, \mathbf{y}_i.$$

The new weights are consistent only with C-information, meaning

$$\sum_s w_{RWi} \mathbf{y}_i = \mathbf{Y}_0.$$

4. Cross-classfied domain estimation

In the previous section we assumed that total of the C-variables were known from RFS and demanded consistency between them and domain estimates in the PRS. But what if we have access to more detailed information - domain level totals? Then by using them we can achieve more accurate results. Let us assume that there are two or more sources of common variable(s), but all had different categorical variables for domain identification. Let these variables be measured in the new, present survey. We are interested in the domains obtained by crossing these categorical variables. In order to use the known C-information we define cross-classified domains, insert the known information as marginals and demand that the row and column sums are consistent with them.

5. Cross-classified domains by two variables

From this point forward we assume that our population is divided into domains by two categorical variables B and D. Let variable B have r levels and variable D c levels, resulting in a $r \times c$ 2-way table. Naturally one could want domains by crossing three (or more) variables, then assuming that two of them come from the same source, we can redefine a $r \times c \times d$ table into a $(rc) \times d$ table. If all three (or more) categorical domain indication variables come from different sources, then 2-way tables can be applied as many times as needed. If a 3-way table still has to be used then a large sample size is also needed, since the number of cells in the table increases by $r \times c$ with each level of the third categorical variable.

We denote the cross-classified domains as U_{kl} $(k = \{1, 2, \dots, r\}; l = \{1, 2, \dots, c\})$ so that $\bigcup_{k=1}^{r} \bigcup_{l=1}^{c} U_{kl} = U$. Row marginals are denoted $U_{k\cdot} := \bigcup_{l=1}^{c} U_{kl}$ and column marginals $U_{\cdot l} := \bigcup_{k=1}^{r} U_{kl}$. In order to estimate the domains and marginals we have to define a row indicator:

$$\delta_i^k = \begin{cases} 1, & i \in U_{k\cdot}, \\ 0, & \text{otherwise} \end{cases} \tag{5.6}$$

and a column indicator

$$\gamma_i^l = \begin{cases} 1, & i \in U_{\cdot l}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.7}$$

By combining indicators into vectors $\boldsymbol{\delta}_i = (\delta_i^1, \delta_i^2, \dots, \delta_i^r)'$ and $\boldsymbol{\gamma}_i = (\gamma_i^1, \gamma_i^2, \dots, \gamma_i^c)'$ we can obtain an indicator for the 2-way table:

$$\mathbb{I}_i = \boldsymbol{\delta}_i \boldsymbol{\gamma}_i' = \begin{pmatrix} \delta_i^1 \gamma_i^1 & \cdots & \delta_i^1 \gamma_i^c \\ \vdots & \ddots & \vdots \\ \delta_i^r \gamma_i^1 & \cdots & \delta_i^r \gamma_i^c \end{pmatrix}.$$

The vectors $\boldsymbol{\delta}_i$ and $\boldsymbol{\gamma}_i$, and the matrix $\mathbb{I}_i$ contain only one „1“, all other elements are zeros. Using these indicators one can easily find the HT estimates for all $\mathbf{y}_i$. But instead of a normal multiplication, we have to use the Kronecker product($\otimes$)

$$\hat{\mathbf{Y}} = \sum_s a_i \mathbb{I}_i \otimes \mathbf{y}_i = \sum_s a_i (\boldsymbol{\delta}_i \boldsymbol{\gamma}_i') \otimes \mathbf{y}_i. \tag{5.8}$$

As a result we get a $(rm) \times c$ matrix, where we have (rc) m-dimensional vectors, each being a vector of m total estimates of one domain.

To get estimates for marginals we can use the earlier indicators (5.6) and (5.7):

$$\hat{\mathbf{Y}}_{k+} = \sum_U a_i \delta_i^k \mathbf{y}_i, \quad k = \{1, 2, \dots, r\}, \tag{5.9}$$

$$\hat{\mathbf{Y}}_{+l} = \sum_U a_i \gamma_i^l \mathbf{y}_i, \quad l = \{1, 2, \dots, c\}. \tag{5.10}$$

Already Deming and Stephan (1940) adjusted multi-dimensional tables with known marginals to achieve consistency, but they achieved it iteratively. With the formulas in the next section we can achieve consistency within one calculation step. They did however develop formulas for a 3-way table, not discussed here. Deville, Särndal and Sautory (1993) dealt with calibration on marginal sums, but they used marginals as A-information and with post-stratification in mind i.e. the known marginals were the sizes of margin domains. Knottnerus (2003, Chapter 12.9) touches on the issue at hand in the chapter on the General Restriction (GR) estimator, but he uses the covariance matrix of initial estimators to achieve consistency. Knottnerus and van Duin (2006) derive RW method formulas for a contingency table where marginal sums are known, but they use a GREG estimator and consider a one-dimensional case with domain sizes N_{kl}.

AC-calibration in caseS of cross-classified domains

In order to get AC-calibration estimators for our special case we first have to operate with the marginals by putting them into a single vector:

$$\mathbf{Y}_{0marg} := (\mathbf{Y}'_{1+}, \mathbf{Y}'_{2+}, \dots, \mathbf{Y}'_{r+}, \mathbf{Y}'_{+1}, \mathbf{Y}'_{+2}, \dots, \mathbf{Y}'_{+c})'.$$

The same has to be done with initial estimates:

$$\widehat{\mathbf{Y}}_{marg} := \left(\widehat{\mathbf{Y}}'_{1+}, \widehat{\mathbf{Y}}'_{2+}, \dots, \widehat{\mathbf{Y}}'_{r+}, \widehat{\mathbf{Y}}'_{+1}, \widehat{\mathbf{Y}}'_{+2}, \dots, \widehat{\mathbf{Y}}'_{+c}\right)'$$

Additionally we will need a combination of the indicator vectors $\boldsymbol{\delta}_i$ and $\boldsymbol{\gamma}_i$:

$$\boldsymbol{\Gamma}_i := (\boldsymbol{\delta}'_i \boldsymbol{\gamma}'_i)' = (\delta_i^1, \delta_i^2, \dots, \delta_i^r, \gamma_i^1, \gamma_i^2, \dots, \gamma_i^c)'.$$

From (3.3) we get

$$\begin{pmatrix} \mathbf{x}_i \\ \boldsymbol{\Gamma}_i \otimes \mathbf{y}_i \end{pmatrix}, i \in s, \qquad \begin{pmatrix} \mathbf{X} \\ \mathbf{Y}_{0marg} \end{pmatrix}, \qquad \begin{pmatrix} \widehat{\mathbf{X}} \\ \widehat{\mathbf{Y}}_{marg} \end{pmatrix}.$$

Similarly to (3.4) the new weights are then

$$w_{ACi} = a_i(1 + \boldsymbol{\lambda}'_{AC}\mathbf{z}_{ACi}) \tag{5.11}$$

where $\mathbf{z}_{ACi}$ is the instrument vector and

$$\boldsymbol{\lambda}'_{AC} = \begin{pmatrix} \mathbf{X} - \widehat{\mathbf{X}} \\ \mathbf{Y}_{0marg} - \widehat{\boldsymbol{Y}}_{marg} \end{pmatrix}' \mathbf{M}^{+}, \qquad \mathbf{M} = \sum_s a_i \mathbf{z}_{ACi} \begin{pmatrix} \mathbf{x}_i \\ \boldsymbol{\Gamma}_i \otimes \mathbf{y}_i \end{pmatrix}'.$$

Since $\mathbf{M}$ is singular because of the definition of $\boldsymbol{\Gamma}_i$, we have to use Moore-Penrose generalized inverse of the matrix $\mathbf{M}$ (denoted by $\mathbf{M}^{+}$). Deville, Särndal and Sautory (1993) suggest to leave one element out in the vector $\boldsymbol{\Gamma}_i$ when finding the matrix $\mathbf{M}$, e.g. γ_i^c (vectors $\mathbf{Y}_{0marg}$ and $\widehat{\mathbf{Y}}_{marg}$ also have to be shortened correspondingly). Then the linear dependency in matrix $\mathbf{M}$ is gone and $\mathbf{M}^{-1}$ can be found, the consistency remains. Since $\mathbf{M}^{+}$ is approximated in most of the cases the author suggests to use $\mathbf{M}^{-1}$. During simulations the use of $\mathbf{M}^{+}$ produced inconsistent results, meaning that the domain totals did not sum up to the

marginal totals, although, the differences were very near to zero. When applying $\mathbf{M}^{-1}$, this disturbing fact disappeared.

Finally from (5.8) and (5.11) we get

$$\hat{\mathbf{Y}}^{AC} = \sum_s w_{ACi} \mathbb{I}_i \otimes \mathbf{y}_i.$$

When finding new weights w_{ACi} we only use those study variables $\mathbf{y}_\mathrm{i}$ with which we wish to achieve consistency, but these weights can be used to calculate our 2-way table domain total estimates for all other variables of interest.

RW in caseS of cross-classified domains

With RW we first found weights that were calibrated on A-information. Since no assumptions have been made on auxiliary variables, then A-calibrated estimates weights w_i can be found using (2.2). Replacing weights a_i with w_i in (5.9) and (5.10) we get the A-calibrated estimates of the marginal sums $\hat{\mathbf{Y}}_{k+}^{CAL}$ and $\hat{\mathbf{Y}}_{+l}^{CAL}$, $k = \{1, 2, \dots, r\}; l = \{1, 2, \dots, c\}$. We define a new vector

$$\hat{\mathbf{Y}}_{marg}^{CAL} := \left(\hat{\mathbf{Y}}_{1+}^{CAL\prime}, \hat{\mathbf{Y}}_{2+}^{CAL\prime}, \dots, \hat{\mathbf{Y}}_{r+}^{CAL\prime}, \hat{\mathbf{Y}}_{+1}^{CAL\prime}, \hat{\mathbf{Y}}_{+2}^{CAL\prime}, \dots, \hat{\mathbf{Y}}_{+c}^{CAL\prime}\right)'.$$

The definitions of $\mathbf{Y}_{0marg}$ and $\mathbf{\Gamma}_\mathrm{i}$ remain the same. The auxiliary information vector takes the form $(\mathbf{\Gamma}_i \otimes \mathbf{y}_i), i \in s$ and by recalibrating, so that $\sum_s w_{RWi} \mathbf{\Gamma}_i \otimes \mathbf{y}_i = \mathbf{Y}_{0marg}$, from (3.5) we get new weights

$$w_{RWi} = w_i(1 + \boldsymbol{\lambda}_{RW}' \mathbf{z}_{RWi}) \tag{5.12}$$

where

$$\boldsymbol{\lambda}_{RW}' = \left(\mathbf{Y}_0 - \hat{\mathbf{Y}}_{marg}^{CAL}\right)' \mathbf{M}^+, \qquad \mathbf{M} = \sum_s w_i \mathbf{z}_{RWi} (\mathbf{\Gamma}_\mathrm{i} \otimes \mathbf{y}_i)'.$$

The argumentation around $\mathbf{M}^+$ is the same as in the last subsection.

Finally we recieve RW estimates for all domain totals

$$\hat{\mathbf{Y}}^{RW} = \sum_s w_{RWi} \mathbb{I}_i \otimes \mathbf{y}_i.$$

6. Simulations

In order to test the formulas developed in the previous section and to compare the two methods discussed above, a simulation study was carried out in the SAS IML environment. A population was composed on real data from the Estonian Household Survey, with 17 540 households.

Simulation set-up

The population was divided into 12 cross-classified domains according to gender and economic activity of the head of the household (the variables were

combined to form one dimension), and education level of the head of the household (second dimension). The resulting domains and their sizes are presented in Table 1, immediately one can notice that sizes differ quite considerably. The largest being 23,3% of the population, and three domains contain only about 2% of households. Since we will use simple random sampling to produce samples, we have an opportunity to study the behavior of our estimators in cases of small areas.

Table 1. Domain sizes and percentage of the population total

Head of the household		Education level						
Activity	Gender	Primary		Secondary		Higher		TOTAL
Employed	Male	712	4,1%	4 094	23,3%	1 840	10,5%	6 646 37,9%
	Female	364	2,1%	2 393	13,6%	2 015	11,5%	4 772 27,2%
Unemployed	Male	923	2,0%	1 172	6,7%	370	2,1%	2 465 14,1%
	Female	1 559	8,9%	1 478	8,4%	620	3,5%	3 657 20,8%
TOTAL		3 558	20,3%	9 137	52,1%	4 845	27,6%	17 540 100%

(1) Household net income, (2) household expenditure and (3) domain membership were taken as study variables. Auxiliary variable vector consisted of (1) income from benefits, (2) number of household members and (3) number of children in the household under the age of 16.

In the simulation study we will explore two important cases:

a) The marginals $\mathbf{Y}_{0marg}$ are known;

b) The marginals $\mathbf{Y}_{0marg}$ are estimated in a previous survey (RFS).

To emphasize the flexibility of the calibration technique and previously developed formulas, we will assume that only marginals for net income and domain size are known (estimated for the second case) and later use the obtained weights to calculate domain estimates for household expenditure.

As said, the instrument vectors in formulas (2.2), (5.11) and (5.12) can be chosen freely (as long as the dimension coincides with the auxiliary vector), so they were fixed to $\mathbf{z}_i = \mathbf{x}_i,\ \mathbf{z}_{ACi} = (\mathbf{x}_i', (\mathbf{\Gamma}_i \otimes \mathbf{y}_i)')'$ and $\mathbf{z}_{RWi} = (\mathbf{\Gamma}_i \otimes \mathbf{y}_i)'$, which is a standard choice (Särndal, 2007). The auxiliary variable vector $\mathbf{x}_i$, study variable vector y_i and the matrix $\mathbf{\Gamma}_i$ are constructed using the variables mentioned above, with the exeption that household expenditure is not used in the calibration step.

Performance of the AC and RW estimators was assessed by relative root mean square error (RRMSE) and relative bias (RB).

$$RRMSE\left(\hat{Y}_d\right) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{Y}_d^{(m)} - Y_d\right)^2}}{Y_d}, \qquad RB\left(\hat{Y}_d\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\hat{Y}_d^{(m)} - Y_d}{Y_d}$$

where M is the number of sample repetitions, $\hat{Y}_d^{(m)}$ is the domain total estimate of the study variable in the m-th sample and Y_d is the actual domain total. To set a baseline for comparison and assessing performance HT estimates were also found, accompanied by RRMSE and RB.

Simulations and results

In case (a) 1000 independent samples were drawn with simple random sampling (SRS), each sample consisted of 1000 households (5,7% of the population) from which domain estimates with HT, AC and RW methods were found. For case (b) marginals were first estimated with HT estimator from a sample of 2000 households (SRS sample) and then the domain estimates were calculated using weights received from calibrating on the estimated marginals (and auxiliary variables), this was repeated 1000 times. Finally the performance indicators were calculated over the simulations for both cases. Resulting RRMSE and RB can be seen in Table and Table .

Consistency was achieved using AC-calibration and RW in both cases, when marginal sums were known, and when they were first estimated. From Table it can be seen that both the RW and the AC methods have lower RRMSE than the HT estimator. Even when the marginal sums were estimated, AC and RW outperform the HT estimator in all domains. The AC estimator has relatively the same RRMSE as RW in all domains and in all cases. When we compare the performance of the estimators in cases (a) and (b), then it is clear that using actual information gives better results.

In Table we see that using AC and RW typically increases bias compared to the HT estimator, but the bias is still close to 0. This is because calibration is approximately unbiased. HT showed worse results only where the variance within the domain was small. In the case of known marginals the bias is typically slightly lower than with estimated marginals, in some cases it is vice versa, but the differences are small.

Table 2. RRMSE of net income, domain size and expenditure over simulations

		Net income					Domain size					Expenditure				
			$\mathbf{Y}_{0marg}$		$\hat{\mathbf{Y}}_{0marg}$			$\mathbf{Y}_{0marg}$		$\hat{\mathbf{Y}}_{0marg}$			$\mathbf{Y}_{0marg}$		$\hat{\mathbf{Y}}_{0marg}$	
D	n_d	HT	AC	RW	AC	RW	HT	AC	RW	AC	RW	HT	AC	RW	AC	RW
1	712	,20	,12	,12	,17	,17	,15	,12	,12	,14	,14	,21	,15	,15	,19	,19
2	4 094	,08	,03	,03	,06	,06	,06	,03	,03	,04	,04	,08	,06	,06	,07	,07
3	1 840	,14	,06	,06	,11	,11	,09	,06	,06	,08	,08	,13	,09	,09	,12	,12
4	364	,28	,26	,26	,27	,27	,21	,19	,19	,20	,20	,32	,31	,30	,31	,31
5	2 393	,11	,08	,08	,10	,10	,08	,05	,05	,07	,07	,11	,09	,09	,10	,10
6	2 015	,13	,07	,07	,10	,10	,08	,05	,05	,07	,07	,13	,09	,09	,11	,11
7	923	,17	,14	,14	,16	,16	,14	,10	,10	,12	,12	,19	,16	,16	,18	,18
8	1 172	,17	,11	,11	,14	,14	,12	,08	,08	,10	,10	,21	,20	,20	,22	,21
9	370	,32	,26	,26	,29	,29	,22	,19	,19	,20	,20	,30	,28	,28	,30	,30
10	1 559	,13	,10	,10	,12	,12	,10	,06	,06	,08	,08	,14	,12	,12	,13	,13
11	1 478	,14	,10	,10	,12	,12	,10	,07	,07	,09	,09	,15	,12	,12	,13	,13
12	620	,24	,19	,19	,21	,21	,16	,13	,13	,15	,15	,27	,25	,25	,26	,26
U	17 540	,04	0	0	,02	,02	0	0	0	,01	,01	,04	,03	,03	,03	,03

Table 3. RB of net income, domain size and expenditure over simulations ($\times 10^{-2}$)

		Net income					Domain size					Expenditure				
			$\mathbf{Y}_{0marg}$		$\hat{\mathbf{Y}}_{0marg}$			$\mathbf{Y}_{0marg}$		$\hat{\mathbf{Y}}_{0marg}$			$\mathbf{Y}_{0marg}$		$\hat{\mathbf{Y}}_{0marg}$	
D	n_d	HT	AC	RW	AC	RW	HT	AC	RW	AC	RW	HT	AC	RW	AC	RW
1	712	-,17	-1,56	-1,63	-1,78	-1,86	-,32	-,31	-,32	-,27	-,29	-,27	-1,14	-1,14	-1,25	-1,28
2	4 094	,06	,30	,30	,31	,32	-,01	,09	,10	,10	,11	-,18	,12	,15	,12	,15
3	1 840	,02	-,10	-,10	,15	,15	-,13	-,08	-,09	,24	,23	-,21	,10	,16	,38	,46
4	364	-,19	-,10	,05	-,43	-,26	-,61	-,74	-,66	-,82	-,72	-,37	-,17	-,03	-,54	-,31
5	2 393	-,25	-,72	-,70	-,85	-,84	,06	-,13	-,13	-,14	-,14	-,21	-,32	-,29	-,44	-,41
6	2 015	-,08	,64	,61	,85	,82	,12	,29	,27	,55	,54	-,02	,82	,78	1,02	,99
7	923	-,71	1,81	1,86	1,65	1,70	-,42	,26	,27	,16	,17	-,84	1,59	1,64	1,38	1,42
8	1 172	-,50	-,62	-,74	-,88	-,98	-,13	-,18	-,21	-,44	-,46	,39	1,89	1,77	1,56	1,45
9	370	,50	-1,74	-1,51	-1,36	-1,19	,50	-,08	,02	,08	,13	,12	,88	,99	,94	1,02
10	1 559	-,44	,56	,54	,44	,42	-,24	,16	,14	,21	,20	-,52	,45	,46	,39	,41
11	1 478	1,37	,62	,64	,45	,46	,59	,10	,11	,05	,05	1,56	1,04	1,04	,87	,84
12	620	-,07	-2,41	-2,42	-2,65	-2,64	,34	-,64	-,61	-,42	-,40	,59	-,46	-,41	-,28	-,19
U	17 540	-,03	0	0	,02	,02	0	0	0	,05	,05	-,07	,28	,30	,30	,33

In terms of comparing AC *versus* RW they both give almost equal estimates. Neither estimator is superior to the other over all domains. The same conclusion was reached by Särndal and Traat (2011).

In the simulations AC and RW outperformed the HT estimator, but all of the mentioned estimators performed not so well in smaller domains, hence they are not appropriate for small area estimation.

Acknowledgments

The author would like to thank associate professor Imbi Traat, from the University of Tartu, Estonia for her many improving suggestions on the subject and Professor Emeritus Gunnar Kulldorff, from Umeå University, Sweden for his support and help. This work was supported by the Estonian Science Foundation grant 8789 and the European Social Fund DoRa program.

REFERENCES

Deming, W.E., Stephan, F.F., 1940. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected marginal Totals are Known. *The Annals of Mathematical Statistics*, 11, pp.427-444.

Deville, J-C., Särndal, C.-E., 1992. Calibration Esitmators in Survey Sampling. *Journal of the American Statistical Association*, 87, pp.376-382.

Deville, J.-C., Särndal, C.-E., Sautory, O., 1993. Generalized Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88, pp.1013-1020.

Elvers, E., Rosén, B., 1999. Quality Concept for Official Statistics. In: S. Kotz, C.B. Read, and D.L. Banks eds. *Encyclopedia of Statistical Sciences*, *Update Volume 3*. New York: Wiley, pp.621-629.

Estevao, V.M., Särndal, C.-E., 2000. A Functional Approach to Calibration. *Journal of Official Statistics*, 16, pp.379-399.

Estevao, V.M., Särndal, C.-E., 2004. Borrowing Strength Is Not the Best Technique Within a Wide Range of Design-Consistent Domain Estimators. *Journal of Official Statistics*, 20, pp.645-669.

Houbiers, M., 2004. Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. *Journal of Official Statistics*, 20, pp.55-75.

Knottnerus, P., 2003. *Sample Survey Theory. Some Pythagorean Perspectives*. New York: Springer

Knottnerus, P., van Duin, C., 2006. Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, pp.565-584.

Kott, P.S., 2006. Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, pp.133-142.

Kroese, A.H., Renssen, R.H., 1999. Weighting and Imputation at Statistics Netherlands. *Proceedings of IASS Satellite Conference on Small Area Estimation*, Riga:Latvia, pp.109-120.

Platek, R., Särndal, C.-E., 2001. Can a Statistician Deliver?. *Journal of Official Statistics*, 17, pp.1-20.

Särndal, C.-E., 2007. The Calibration Approach in Survey Theory and Practice. *Survey Methodology*, 33, pp.99-119.

Särndal, C.-E., Traat, I., 2011. Domain Estimators Calibrated on Information from Another Survey. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, vol. 15 (to appear).

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 265—280

SOME APPLICATIONS OF PANEL DATA MODELS IN SMALL AREA ESTIMATION

Vilma Nekrašaitė-Liegė[1]

ABSTRACT

This study uses a real population from Statistics Lithuania to investigate the performance of different types of estimation strategies. The estimation strategy is a combination of sampling design and estimation design. The sampling designs include equal probability design (SRS) and unequal probability designs (stratified SRS and model-based sampling designs). Design-based direct Horvitz-Thompson, indirect model-assisted GREG estimator and indirect model-based estimator are used to estimate the totals in small area estimation. The underlying panel-type models (linear fixed-effects type or linear random-effects type) are examined in both stages of estimation strategies: sample design and construction of estimators.

Key words: Small area estimation; panel-type data; model-based; model-assisted

1. Introduction

In this paper, the accuracy of several small area estimation strategies (a pair comprising a sample design and an estimator) is investigated. The focus on small area estimation (SAE) is made because SAE is an important objective of many surveys. Small areas almost always have small sample sizes, so standard survey estimation methods, which only use information from the small area samples, are unreliable for these areas. In this context SAE methods that borrow strength via statistical models (Rao 2003) are used to produce reliable estimates.

Nowadays, official statistics repeats the same surveys from year to year, so for most of the population elements it is possible to get information for the same variable in several time periods. It means that for many surveys individual data for some objects are known for at least one previous time point. We will call this panel-type data even when it is not part of the design. Also, in some cases it is possible to use information collected from the other sources (tax offices,

[1] Vilnius Gediminas Technical University, Vilnius. E-mail: nekrasaite.vilma@gmail.com

jobcentres, etc.). Such dataset of a large amount of auxiliary information might improve the quality of the estimation strategy as compared with a strategy based on the current sample alone.

The use of panel-type data in estimation strategy means that a prediction theory based on a superpopulation model is used. A superpopulation model can be used not only in estimation stage, but for sample selection as well. Such use of the superpopulation model is discussed for example by Royall (1970) and Nedyalkova and Tille (2008). For the estimation strategy they used linear regression model as a superpopulation model. In our research, a basic superpopulation model is an incomplete panel data model (Hsiao 2003).

The advantage of using a panel data model in model-based sample design for small areas has been noticed by Nekrašaitė-Liegė, Radavičius and Rudys (2011). In this research we place emphasis on application of high-dimensional multi-level fixed effect panel data model using comprehensive exploratory analysis and model selection technique. The results obtained using such model are compared with the results obtained using panel data model with random effect for domains and panel data model with few fixed effect that are the same for the whole population.

In this research not only different panel data models are compared, but also two types of estimators: model-assisted and model-based. Almost all papers devoted to model-based estimation in small areas (see, e.g., Rao 2005, Omrani, Gerber and Bousch 2009) deal with samples from a limited number of areas and the case where a set of auxiliary covariates are used to obtain estimates in the areas that are not actually sampled. In this research a sample from all areas is selected (number of selected elements in the areas is random) and a set of auxiliary information is known for all elements in the population from the previous surveys and other administrative sources. Such type of auxiliary information might be available in short term statistics where, at a current time, data are collected using small samples and in the future the data for the same period of time is expanded and updated using large samples or administrative sources. Thus the use of such type of auxiliary information might help us to find elaborate fixed effect panel data model appropriate for the effective model-based estimation strategy.

Hence, in this paper several estimation strategies are used to answer the following problems: what type of model (with fixed or random effects), sample design and estimator (design-based or model-based) should be used in small area estimation.

The paper consists of six sections. The main notation and definitions used in survey statistics are introduced in Section 2. In Section 3, different types of estimators and estimates are presented. Some panel data models applicable in survey sampling are described in Section 4. We end with simulation results (Section 5) and concluding remarks (Section 6).

2. Definitions and notations

Let us start with a common framework of finite population survey sampling. A finite population $U=\{u_1, u_2, ..., u_N\}$ of the size N is considered. For simplicity, in the sequel we identify a population element u_k as its index k. Hence $U=\{1, 2, ..., N\}$.

The elements k (k=1, 2, ..., N) of the population U has two components y and **x**.

The component y defines the value of a study variable (variable of interest), and the component $\mathbf{x}=\{x_1, x_2, ..., x_J\} \in \boldsymbol{R}^J$ defines the values of the J auxiliary variables.

In this study a panel-type data is considered. This means that y_k and $\mathbf{x}=\{x_{1,k}, x_{2,k}, ..., x_{J,k}\}$ are assumed to be time series,

$$y_k = \left(y_k(t), t = 1,2,...\right), \quad \mathbf{x}_k = \left(\mathbf{x}_k(t), t = 1,2,...\right). \tag{1}$$

The population is divided into D nonoverlapping domains (subpopulations) U(d) of size N(d), where d=1, 2, ..., D. Domain indicator variables define whether k ∈ U belongs to a given domain:

$$q_k^{(d)} = \begin{cases} 1, & \text{if } k \in U^{(d)} \\ 0, & \text{otherwise} \end{cases}, \forall k \in U, d = 1,...,D. \tag{2}$$

It is assumed, that an element can not change domain during the time period, thus $q_k^{(d)}$ do not depend on time.

The parameter of interest is a domain total in time t:

$$T^{(d)}(t) = \sum_{k \in U} q_k^{(d)} y_k(t) = \sum_{k \in U^{(d)}} y_k(t), \quad d = 1,...,D. \tag{3}$$

Actually in this study, the parameter of interest is a domain total in 4 different quarters for a given year. If time variable t denotes quarters and T+1 denotes the first quarter of interest, then the parameter of interest is

$$T^{(d)}(T+l) = \sum_{k \in U} q_k^{(d)} y_k(T+l) = \sum_{k \in U^{(d)}} y_k(T+l), \quad d = 1,...,D, \quad l = 1,...,4. \tag{4}$$

To estimate $T^{(d)}(T+l)$, we need information about unknown variable y in time $T+l$. This information is collected by sampling. The sampling vector $\underline{\mathbf{S}}(T+l) = (\underline{S}_1(T+l), \underline{S}_2(T+l), ..., \underline{S}_N(T+l))$ is a random vector whose elements $\underline{S}_k(T+l)$ indicate the number of selections for k element in time point $T+l$. In this research we are interested just in the sampling without replacement (WOR), thus the

largest number of selections for k element in $T+l$ is one: $\underline{S}_k(T+l)=1$ if element k is selected and $\underline{S}_k(T+l)=0$ if it is not selected. Also, in this paper, the sampling vector is the same for all 4 quarters of interest, thus it can be notated as $\underline{\mathbf{S}}(T+l)=\underline{\mathbf{S}}=(\underline{S}_1, \underline{S}_2, ..., \underline{S}_N)$, $l=1,\ldots,4$. It means that the sampling vector is determined for the first quarter of interest and it is repeated for other 3 quarters. The realization $\mathbf{S}(T+l) = \mathbf{S} = (S_1, S_2, ..., S_N)$ is called a sample. Let $\boldsymbol{S}$ be the set of all samples $\mathbf{S}$. The sampling vector $\underline{\mathbf{S}}$ (and its realization $\mathbf{S}$) define the sample set $\underline{s}$ (and the corresponding s) as

$$\underline{s}(T+l)=\underline{s}=\{k:k\in U,\underline{S}_k=1\},\quad s(T+l)=s=\{k:k\in U,S_k=1\}. \tag{5}$$

The difference between sample $\mathbf{S}$ and sample set s is that s is a subset of U whereas $\mathbf{S}$ is a N-dimensional vector of indicators.

The distribution of $\underline{\mathbf{S}}$, denoted by $p(\cdot)$, is called a sample design. The sampling design assigns a probability $\mathbf{P}(\underline{\mathbf{S}} = \mathbf{S}) = p(\mathbf{S})$ to every sample $\mathbf{S}$. First and second order inclusion probabilities π_k and π_{kl} for sampling without replacement (WOR) are defined as

$$\pi_k=\mathbf{P}(\underline{S}_k=1)=\sum_{\mathbf{S}:S_k=1}p(\mathbf{S})=\mathbf{E}(\underline{S}_k), \tag{6}$$

$$\pi_{kl}=\mathbf{P}(\underline{S}_k=1,\underline{S}_l=1)=\sum_{\mathbf{S}:S_kS_l=1}p(\mathbf{S})=\mathbf{E}(\underline{S}_k,\underline{S}_l), \tag{7}$$

where $\pi_{kk} = \pi_k$ and $\operatorname{cov}(\underline{S}_k,\underline{S}_l)=\Delta_{kl}=\pi_{kl}-\pi_k\pi_l$. Thus the samples for each quarter are the same as for the first quarter, the inclusion probabilities do not depend on time.

The sampling weights for WOR designs are defined as

$$w_k=\begin{cases}\pi_k^{-1}, & \text{if } k\in s\\ 0, & \text{otherwise}\end{cases}. \tag{8}$$

The sample size and the sample set in domain $U^{(d)}$ are

$$n^{(d)}=\sum_{k\in U^{(d)}}S_k,\quad s^{(d)}=s\cap U^{(d)}. \tag{9}$$

There are two types of domains:

1. Planned domains. (Singh, Gambino and Mantel 1994) For planned domains the sample size $n^{(d)}$ in domain sample is fixed in advance, so really these domains are strata with possible different allocations.

2. Unplanned domains. If the sample size $n^{(d)}$ in domain sample is random, domains are unplanned. The disadvantage of unplanned domains is that, there might be domains with zero elements in the sample **S**.

In this research domains are unplanned. It is assumed that the number of the elements in each domain $U^{(d)}$, d=1, 2, ..., D, is known, but the domains are not used in the sample design. This means that the sample part in each domain, $s^{(d)}$, has a random size.

3. Estimators and estimates

An estimator is a rule or algorithm that defines how to estimate the parameter of interest (in our case: domain total). It is a random variable, which value depends on the sample and the auxiliary information. An estimate is the realized value of an estimator. In general, an estimator and an estimate are denoted, respectively, as $\hat{\theta}(\underline{\mathbf{S}})$ and $\hat{\theta}(\mathbf{S})$, or briefly as $\underline{\hat{\theta}}$ and $\hat{\theta}$. For parameter $T^{(d)}(T+l)$, the estimator and estimate are $\underline{\hat{T}}^{(d)}(T+l)$ and $\hat{T}^{(d)}(T+l)$, l=1,…,4.

The estimator is accurate if its bias and variance are small. The bias is the difference between the parameter expectation and the true value: $BIAS(\underline{\hat{\theta}}) = \mathbf{E}(\underline{\hat{\theta}}) - \theta$. If $BIAS(\underline{\hat{\theta}}) = 0$, the estimator is unbiased. The bias might come with respect to the design or to the model. The symbols **E**, *var* denote, respectively, the expected value and the variance under the sample design. They are defined as

$$\mathbf{E}(\underline{\hat{\theta}}) = \sum_{\mathbf{S} \in S} p(\mathbf{S})\,\hat{\theta} \tag{10}$$

$$var(\underline{\hat{\theta}}) = \sum_{\mathbf{S} \in S} p(\mathbf{S})\left[\hat{\theta} - \mathbf{E}(\underline{\hat{\theta}})\right]^2 \tag{11}$$

In this research two types of estimators of the domain total are used:

1. Design-based estimators. The design-based estimators can be divided in two groups (Särndal, Swensson and Wretman 1992, Lehtonen and Veijanen 2009): design-based direct estimators, which are design unbiased by definition and design-based model-assisted indirect estimators, which are nearly design unbiased irrespective of the model choice.
2. Model-based estimators. A model-based estimator usually has smaller variance then a design-based estimator, and it is possible to use them even when there is no selected unit in the domain. Still model-based estimator is design- biased and in some cases it might have a large bias.

Two types of estimators can be used for estimation in domains:

1. Direct estimators. A direct estimator uses values of the variable of interest only from the time period of interest and only from units in the domain of interest (U.S. office of management and budget 1993).
2. Indirect estimators. An indirect domain estimator uses values of the variable of interest from a domain and/or time period other than the domain and time period of interest (U.S. office of management and budget 1993).

A convenient direct estimator is Horvitz - Thompson (HT) estimator (Narain, 1951, and Horvitz and Thompson, 1952) for the domain $\underline{\hat{T}}_{HT}^{(d)}(T+l)=\sum_{k\in\underline{s}^{(d)}} w_k y_k(T+l)$ and it's estimate is $\hat{T}_{HT}^{(d)}(T+l)=\sum_{k\in s^{(d)}} w_k y_k(T+l)$, l=1,…,4.

Another estimator is the generalized regression (GREG) estimator (Särndal, Swensson and Wretman 1992). The estimator and estimate for the domain total are

$$\underline{\hat{T}}_{GREG}^{(d)}(T+l)=\sum_{k\in\underline{U}^{(d)}}\hat{y}_k(T+l)+\sum_{k\in\underline{s}^{(d)}} w_k(y_k(T+l)-\hat{y}_k(T+l)) \text{ and}$$
$$\hat{T}_{GREG}^{(d)}(T+l)=\sum_{k\in U^{(d)}}\hat{y}_k(T+l)+\sum_{k\in s^{(d)}} w_k(y_k(T+l)-\hat{y}_k(T+l)) \qquad (12)$$

The last estimator is Model-based (MB) estimator defined by

$$\underline{\hat{T}}_{MB}^{(d)}(T+l)=\sum_{k\in\underline{U}^{(d)}\setminus\underline{s}^{(d)}}\hat{y}_k(T+l)+\sum_{k\in\underline{s}^{(d)}} y_k(T+l) \text{ and}$$
$$\hat{T}_{MB}^{(d)}(T+l)=\sum_{k\in U^{(d)}\setminus s^{(d)}}\hat{y}_k(T+l)+\sum_{k\in s^{(d)}} y_k(T+l). \qquad (13)$$

For the both (GREG and MB) estimators, $\hat{y}_k(T+l), k\in U^{(d)}$, are predicted values of study variable y for each element in $U^{(d)}$ in time $T+l$, l=1,…,4. The prediction algorithm is described in Section 4. Due to the prediction algorithm GREG and MB estimators are indirect. Thus in this paper direct design-based (HT), indirect design-based (GREG) and indirect model-based (MB) estimators are compared.

4. Panel data models in survey sampling

The use of model-assisted or model-based estimators is impossible unless some model is considered. In the most papers a linear regression model is exploited (see, e.g., Royall 1970, Nedyalkova, Tille 2008, and references therein), however in some cases a generalized linear mixed model (Saei and Chambers

2003, Lehtonen, Särndal and Veijanen 2003, 2005, Lehtonen and Veijanen 2009) is applied.

In this study a panel-type data is considered. The problem is to find an effective strategy for estimating the totals of $y_k(T+l)$, $k \in U^{(d)}$, l=1,…,4, given the ("historical", i.e. prior to the sample selection) auxiliary information

$$AI := \left(\mathbf{x}_k(t), y_k(t), t \in \mathrm{T}_k \subset \{1,2,...,T\}, k \in U\right). \tag{14}$$

Let $y_k(t)$ and $\mathbf{x}_k(t)$, k = 1, ..., N, be the realizations of random variables $\underline{y}_k(t)$ and $\underline{\mathbf{x}}(t)=\{\underline{x}_{1,k}(t), \underline{x}_{2,k}(t), ..., \underline{x}_{J,k}(t)\}$ of the superpopulation model $\boldsymbol{M}$:

$$y_k(t) = \beta_{0,g(k)}(t) + r_{0,k}(t) + \sum_{j=1}^{J} \beta_{j,g(k)}(t)\, \underline{x}_{j,k}(t) + \sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t) + \varepsilon_k(t), \quad k \in U. \tag{15}$$

Here $\underline{x}_{j,k}(t)$, j=1, 2, ..., J, are fixed-effects variables, $\beta_{0,g(k)}(t), \beta_{1,g(k)}(t), ..., \beta_{J,g(k)}(t)$ are the unknown fixed-effects model coefficients, which are the same in group $g(k)$. The groups $g(k)$ divides population U into G nonoverlaping groups which in some special cases can be the same as domains. The unknown random-effects models coefficient is denoted as $r_{0,k}(t)$ $\left(r_{0,k}(t) \sim IID\left(0, \lambda^2_{0,g(k)}(t)\right), g(k) = 1,..., G(k)\right)$.

The model error is denoted as $\varepsilon_k(t)$ $\mathbf{E}_{\mathbf{M}}\left(\varepsilon_k(t)\right) = 0, var_{\mathbf{M}}\left(\varepsilon_k(t)\right) = \upsilon_k^2 \sigma^2, \forall k \in U$ and $\operatorname{cov}\left(\varepsilon_k(t), \varepsilon_l(v)\right) = 0$ when $(k,t) \neq (l,v)$). It should be noticed that model error $\varepsilon_k(t)$ and the random-effects model coefficient $r_{0,k}(t)$ are conditionally independent if values of $\underline{x}_{j,k}(t)$, j=1, 2, ..., J, are given. The component $\sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t)$ represents a time trend. The structure of this component depends on "historical" auxiliary information (14) and is specified using exploratory analysis.

Some special cases of general panel data model (15) are the following:

1. Fixed effect panel data model:

$$\underline{y}_k(t) = \beta_{0,g(k)} + \sum_{j=1}^{J} \beta_{j,g(k)}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t) + \varepsilon_k(t), \quad k \in U. \tag{16}$$

Here models coefficients $\beta_{0,g(k)}, \beta_{1,g(k)}, ..., \beta_{J,g(k)}$ do not depend on time which means they are the same for the all periods of time. Such model is very useful in practice since it enables one to find model coefficients just using data from the past. The current data might be use just for prediction.

2. Random effect panel data model:

$$\underline{y}_k(t) = \beta_{0,g(k)} + r_{0,k} + \sum_{j=1}^{J} \beta_{j,g(k)}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t) + \varepsilon_k(t), \quad k \in U. \tag{17}$$

Here the random effect $r_{0,k}$ is included into the previous model. Random effect also does not depend on time and hence it is also possible to find all model coefficients from the past data.

Thus the use of models which coefficients are known before the sample is selected might improve not only the estimators but the sample design as well. The application of the same model in both stages (sample selection and estimation) might be very useful.

5. Simulation

5.1. Population

For the simulation experiment, a real population from Statistics Lithuania is used. Enterprisers which are responsible for adult and other education and have less than 50 employers are taken as the finite population. Information about these enterprisers is taken 16 times – each quarter from 2005 till 2008. The average number of enterprises in each quarter is 650 (Number in population).

The study variable y_k is the income of an enterprise k and the auxiliary variables are the number of employers $x_{1,k}$, tax of value added (VAT) $x_{2,k}$ and various indicators (specification of enterprise (5 indicators), size of enterprise (2 indicators)) $x_{j,k}$, $j = 3, \ldots, 9$.

The study parameter is the total income $T^{(d)}$, in the domain d. The domain is chosen as counties (there are 10 counties in Lithuania). The number of enterprises in each domain varies from 6 to 323 (see table 1.).

Table 1. Domain size in population

Domain size	Number of enterprisers in domain	Number of domains in one quarter	Total number of domains of interest
Small	6 – 25	5	20
Medium	25 – 50	2	8
Large	>50	3	12

The total income in a domain in each quarter in 2008 is chosen as the parameter of interest ($T + l$, T=12, l=1,...,4). So, in this research the study variables are elements of a time series with 4 elements and the total number of domains of interest is 40 (see table 1.).

The overall available auxiliary information is divided into two sets: the "historical" data *AI* (formula (14)) available before the sample selection, i.e. in the sample design stage, and new auxiliary information with the true observations

$$AI(l) := \begin{pmatrix} \mathbf{x}_u(T+l), u \in U \\ y_k(T+l), k \in s \end{pmatrix}, \tag{18}$$

which is available at estimation stage for each quarter l under consideration (l=1,...,4).

5.2. Estimation strategy

Before selecting a sample, the three different panel-type data models were analyzed using *AI*. A detailed exploratory data analysis has been performed in order to construct an appropriate model for the data. For instance, model (FI) has been selected from a quite large set of alternative models using model selection technique. In particular, panel models with the enterprise-specific slopes and/or the seasonal components have been tried out. These models are as follows:

1. Linear fixed effect panel data model (FC):

$$\underline{y}_k(t) = \beta_{0,h} + \sum_{j=1}^{3} \beta_{j,h}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{3} \alpha_{i,h} s_i(t) + \varepsilon_k(t), \quad k \in U. \tag{19}$$

Here, the index $h \in \{1,2\}$ denotes the size of an enterprise (small or medium), auxiliary variables are the number of employers $\underline{x}_{1,k}(t)$, tax of value added $\underline{x}_{2,k}(t)$ and $\underline{x}_{3,k}(t)$, which indicates whether the enterprise engages in a specific activity (learning to drive) or not. The variable s_i, $i \in \{1,2,3\}$, is the indicator of the i-th quarter.

2. Linear mixed panel data model with domain-specific random effects (RD):

$$\underline{y}_k(t) = \beta_{0,h} + r_o^{(d)} + \sum_{j=1}^{3} \beta_{j,h}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{3} \alpha_{i,h} s_i(t) + \varepsilon_k(t), \quad k \in U. \tag{20}$$

The difference between RD and FC model is the additionally included random effect $r_0^{(d)}$, $\left(r_0^{(d)} \sim IID\left(0, \lambda_0^{(d)2}\right)\right)$ for domains.

3. Fixed effect panel data model with different intercepts for enterprisers (FI):

$$\underline{y}_k(t) = \beta_{0,k} + \sum_{j=1}^{2} \beta_{j,h}\, \underline{x}_{j,k}(t) + \gamma_{1,h} s_1(t) \underline{x}_{2,k}(t) + \sum_{j=2}^{8} \gamma_{2,1}^{(d)} s_1(t)\, \underline{x}_{j,k}(t) + \sum_{i=1}^{3} a_{i,h} s_i(t) + \varepsilon_k(t), \quad k \in U. \tag{21}$$

Here the intercept $\beta_{0,k}$ is different for each enterprise, the component $\gamma_{1,h} s_1(t) \underline{x}_{2,k}(t)$ indicates the difference of $\underline{x}_{2,k}(t)$ in the first quarter and the component $\sum_{j=3}^{8} \gamma_{2,1}^{(d)} s_1(t) \underline{x}_{j,k}(t)$ represents the difference between small enterprise specifications in the first quarter. This effect was revealed in the explorative analysis of "historical" data.

Using these three models a model-based sample design is applied (Nekrašaitė-Liegė, Radavičius, Rudys 2011). It consists of three steps.

1. In the first step (FC) model is fitted to the available auxiliary information *AI*.
2. In the second step the prediction errors (residuals) $\varepsilon_k(t) = \hat{y}_k(t) - y_k(t), t \in \mathrm{T}_k$ are calculated and the variance of prediction error $var_{\mathbf{M}}(\varepsilon_k(t)), \forall k \in U$ for each enterpriser is estimated. This is possible to do, because $var_{\mathbf{M}}(\varepsilon_k(t))$ does not depend on time (see formula (15)).
3. Finally, in the third step the (approximately) optimal sample design $p(\mathbf{S})$ based on the estimated variances is constructed. In this case, the stratified probability proportional to size variable sample design is used where the size variable is the variance of prediction error for each enterpriser. Thus the less model-based prediction accuracy for the enterprise the greater its probability to be selected into the sample.

The same is done and using (RD) and (FI) models. Hence three different model-based (MB-FC, MB-RD, MB-FI) sample designs are used for selecting sample. A sample of n=230 enterprisers is selected for the whole 2008 year, thus the selected enterprisers are the same for all 4 quarters. Since the performance of estimation is investigated for one year, the rotation has no effect and is not considered in the paper.

For the comparison, two more sample designs are constructed: Simple random sampling with n=230 enterprisers and Stratified simple random sampling with the same number of enterprisers. For stratification the size of enterprisers is used to define strata. There are two strata: small (160 enterprisers are selected from 545) and medium (70 enterprisers are selected from 105. All sample designs are without replacement, i.e. each enterprise cannot be selected more than one time in one sample.

For each sample design, three types of estimators are considered: Horvitz - Thompson (HT), Generalized regression (GREG, see equation (12)) and Model-

based (MB, see equation (13)) estimators. For the last two estimators, the predicted values are calculated in three ways using (FC), (RD) and (FI) models, respectively. The model coefficients are estimated using the auxiliary information *AI*. Thus, model coefficients are the same for all quarters, the auxiliary information with true observations *AI*(*l*) is used just for estimation of the predicted values.

5.3. Simulation results

To compare the performance of the different estimators (the estimation strategies) a design-based relative root mean squared error (*RRMSE*) for M = 1000 simulations is evaluated:

$$RRMSE(t) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{T}_m^{(d)}(t) - T^{(d)}(t)\right)^2}}{T^{(d)}(t)}. \tag{22}$$

Here $\hat{T}_m^{(d)}(t)$ is the estimate of the total for m-th simulation in the domain d and $T^{(d)}(t)$ refers to the true population total in the same domain. There are 40 domains of interest, so for the better comparison these regions are grouped into three domain sample size classes by the average number of elements in the domain sample (small 0 – 9, medium 10 – 39 and large >40). A mean of relative root means square error (*MRRMSE*) in each class is calculated (see Tables 2–4).

Table 2. HT estimator results

Estimator	Sample design	*MRRMSE* in domains		
		Small domains	Medium domains	Large domains
HT	MB-FI	36,6	21,7	12,6
	MB-FD	36,8	21,8	12,7
	MB-RD	36,7	21,9	12,6
	SRSS	37,9	23,8	15,9
	SRS	40,4	29,1	20,7

Table 3. GREG estimator results

Estimator, model	Sample design	*MRRMSE* in domains		
		Small domains	Medium domains	Large domains
GREG, FI model	MB-FI	14,8	10,3	6,7
	SRSS	15,0	11,8	7,7
	SRS	16,0	12,5	9,5
GREG, FC model	MB-FC	16,8	13,6	8,3
	SRSS	15,7	12,5	8,8
	SRS	15,9	14,4	11,4
GREG, RD model	MB-RD	18,7	13,5	6,5
	SRSS	15,4	12,4	8,7
	SRS	15,8	14,1	11,2

Table 4. MB estimator results

Estimator, model	Sample design	*MRRMSE* in domains		
		Small domains	Medium domains	Large domains
MB, FI model	MB-FI	11,0	11,4	15,1
	SRSS	11,4	9,9	14,9
	SRS	12,2	10,3	14,4
MB, FC model	MB-FC	21,7	25,6	24,5
	SRSS	21,3	23,4	22,6
	SRS	21,4	23,9	21,7
MB, RD model	MB-RD	21,4	26,3	24,4
	SRSS	21,3	24,9	22,7
	SRS	21,0	24,8	21,6

The results for the HT estimators show that the best sample design strategy is model-based strategy. Nevertheless, the accuracy of the HT estimator is twice less than the GREG estimator. For the other two estimators, it is difficult to indicate the best strategy using *RRMSE*. It seems, however, that overall performance of GREG estimator is better whereas MB estimator is better only for the FI model in case of the small domains.

The performance of the hypothesis testing of equality of two variances is taken as an additional criterion for the comparison. Sample designs under the different models and models under the different sample designs are compared (see tables 5–6).

Table 5. Sample designs comparison

Sample design	GREG estimator			MB estimator		
	domains where *var* is significant smaller, %			domains where *var* is significant smaller, %		
	Small domains	Medium domains	Large domains	Small domains	Medium domains	Large domains
FI model						
MB-FI vs SRSS	35,0	50,0	50,0	80,0	100,0	100,0
MB-FI vs SRS	40,0	62,5	75,0	80,0	100,0	100,0
SRSS vs SRS	35,0	50,0	100,0	80,0	50,0	100,0
FC model						
MB-FC vs SRSS	40,0	12,5	58,3	40,0	50,0	100,0
MB-FC vs SRS	40,0	37,5	66,7	40,0	50,0	100,0
SRSS vs SRS	35,0	50,0	100,0	40,0	50,0	100,0
RD model						
MB-RD vs SRSS	20,0	25,0	75,0	40,0	50,0	100,0
MB-RD vs SRS	15,0	50,0	91,7	40,0	50,0	100,0
SRSS vs SRS	40,0	62,5	100,0	60,0	50,0	100,0

Table 6. Models comparison

Models	GREG estimator			MB estimator		
	domains where *var* is significant smaller, %			domains where *var* is significant smaller, %		
	Small domains	Medium domains	Large domains	Small domains	Medium domains	Large domains
MB sample design						
FI vs FC	55,0	37,5	25,0	80,0	50,0	66,7
FI vs RD	40,0	37,5	50,0	60,0	50,0	66,7
RD vs FC	60,0	25,0	8,3	100,0	50,0	33,3
SRSS sample design						
FI vs FC	50,0	37,5	8,3	100,0	100,0	66,7
FI vs RD	50,0	37,5	8,3	100,0	100,0	66,7
FC vs RD	30,0	12,5	0,0	40,0	0,0	0,0
SRS sample design						
FI vs FC	45,0	50,0	16,7	100,0	100,0	66,7
FI vs RD	55,0	37,5	16,7	100,0	100,0	33,3
FC vs RD	30,0	12,5	0,0	40,0	0,0	0,0

Tables 5 and 6 show a percentage of domains, for each model and each sample design, respectively, with significantly smaller variance of the estimators in the first case (respectively, the sample design or the model) as compared to the second. For the rest of domains, the hypothesis about equality of the two variances is not rejected.

The comparison of the sample designs demonstrates that for the GREG estimator and FI model, the use of MB-FI sample design reduces the variance of the estimator for 40% of the small domains and 50% of the medium or the large domains as compared with SRSS or SRS designs. The variance of the MB estimator is smaller for more domains than the GREG estimator for the same model and sample design. For the large domains, using SRSS instead of SRS always reduces the variance for both (GREG and MB) estimators, however, for the small domains the efficiency of SRSS design is much smaller (in some cases it is just 40%).

The comparison of the models shows that the impact of the model is larger for the small domains (especially for the MB estimator) than for the large domains.

The results in table 6 demonstrate that the best prediction model is FI for both estimators.

The analysis of tables 5–6 reveals that the best strategy is to use MB-FI sample design with FI model for both (GREG and MB) estimators and the analysis of tables 3–4 shows that for the small domains the *RRMSE* of MB estimator is quite small, however for the large domains GREG estimator is the best.

6. Concluding remarks

In this paper three different models have been used. The fixed effect panel data model for the whole population (FC) has been taken as the basic panel data model. Then the model has been extended by adding a random effect for domains (RD). This extension has affected the variance of the small domains – it is smaller for more than 30% of the domains (Table 6.). For the large domains, the impact is negligible. A detailed exploratory analysis has been performed in order to improve the underlying panel data model. An attempt to identify enterprise-dependent and significant fixed effects has been made resulting in a smaller variance for more than 50% of the domains (especially when the model-based estimator is used). Nevertheless, for some domains the differences between the models do not significantly affect the variance of the estimators. A reasonable explanation of this observation is that there were structural changes in some enterprisers in 2008 and these changes are not captured by the fitted model. Thus constructing a design-based (nonparametric) test for structural changes in the population is a challenging problem for the further research.

Another aspect investigated in this research, is choice of sample design. The results (Table 5) show that for more than half domains (especially for large ones) the model-based sample design decrease the variance of all estimators (including HT (Table 2)) in particular when FI model is fitted.

This investigation confirms, for the case of panel data model, an empirical observation (Lehtonen, Särndal and Veijanen 2003, 2005) that the design-based estimators (especially model-assisted) show in practice a better design-based performance than model-based estimators. The model-assisted approach enables one to avoid a large bias of an estimator even when there are only few selected elements in the small area. When there are no selected elements in a small area – a model-based estimator is the only choice.

In summary, the comparison of different estimation strategies for the real Lithuanian data has shown that, in the case where a large amount of panel type data is available, the estimation strategy with the FI model based design and the model-assisted estimator (GREG) might be a reasonable choice in small area estimation.

REFERENCES

HORVITZ, D. G. AND THOMPSON, D. J., (1952). A generalization of sampling without eplacement from a finite universe, *Journal of the American Statistical Association*, 47:663–685.

HSIAO, C. (2003). Analysis of Panel Data, Economic Society monographs no.

34, 2nd edition, New York: Cambridge University Press.

LEHTONEN, R. AND VEIJANEN, A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in C.R. Rao and D. Pfeffermann (Eds.) *Handbook of Statistics*, Vol 29B, *Sample Surveys: Inference and Analysis.* New York: Elsevier.

LEHTONEN, R., SÄRNDAL, C.-E. AND VEIJANEN, A. (2003). The effect of model choice in estimation for domains, including small domains, *Survey Methodology* 29:33-44.

LEHTONEN, R., SÄRNDAL, C.-E. AND VEIJANEN, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains, *Statistics in Transition* 7:649-673.

NARAIN, R. D., (1951). On sampling without replacement with varying probabilities, *Journal of the Indian Society of AgriculturalStatistics*, 3:169–174.

NEDYALKOVA, D., TILLE, Y., (2008). Optimal sampling and estimation strategies under the linear model, *Biometrika* 95[3]:521-537.

NEKRAŠAITĖ-LIEGĖ, V., RADAVIČIUS, M. and RUDYS, T. (2011). Model-based design in small area estimation. *Lithuanian Mathematical Journal.* 51[3]:417-424.

OMRANI, H., GERBER, P. AND BOUSCH. P (2009). Model-Based Small Area Estimation with application to unemployment estimates, *World Academy of Science, Engineering and Technology* 49, 793-800

RAO, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

Rao, J.N.K. (2005). Inferential issues in small area estimation: some new developments. *Statistics in Transition*, **7**, 513-526.

ROYALL, R. M., (1970). On finite population sampling theory under certain linear regression models *Biometrika* 57[2]:377-387.

SAEI, A. and CHAMBERS, R. (2003). Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. Methodology Working Paper No. M03/15. University of Southampton, UK.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*, Springer - Verlag, New York.

SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 314.

U.S. OFFICE OF MANAGEMENT AND BUDGET (1993). Indirect Estimators in Federal Programs, *Statistical Policy Worlang Paper 21,* NATIONAL Technical Information Service, Springfield, Virginia.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 281—300

TO MISREPORT OR NOT TO REPORT? THE CASE OF THE ITALIAN SURVEY ON HOUSEHOLD INCOME AND WEALTH

Andrea Neri[1], **M. Giovanna Ranalli**[2]

ABSTRACT

The objective of the paper is to adjust for the bias due to unit nonresponse and measurement error in survey estimates of total household financial wealth. Sample surveys are a useful source of information on household wealth. Yet, survey estimates are affected by nonsampling errors. In particular, when it comes to household wealth, unit nonresponse and measurement error can severely bias the estimates. Using the Italian Survey on Household Income and Wealth, we exploit the available auxiliary information in order to assess the magnitude of such a bias. We find evidence that for this kind of surveys, nonsampling errors are a major issue to deal with, possibly more serious than sampling errors. Moreover, in the case of SHIW the potential bias due to measurement error seems to outweigh by far that induced by nonresponse.

Key words: Unit Nonresponse; Measurement Error; Auxiliary Information; Subsampling; Imputation.

1. Introduction

Information on household financial wealth plays an important role in policy analysis. The information available from National Financial Accounts (NFAs) does not usually fulfill policy makers needs since it does not allow analysts to take into account household heterogeneity. Sample surveys are generally used to fill such a gap, since they make it possible to evaluate the impact of shocks, policies and institutional changes on various groups of individuals (European Central Bank, 2009). Yet, the measurement of household financial wealth through sample surveys is a difficult task.

1 Economic and Financial Statistics Department, Banca d'Italia, Rome, Italy. E-mail: andrea.neri@bancaditalia.it
2 Department of Economics, Finance and Statistics, University of Perugia, Italy. E-mail: giovanna@stat.unipg.it

The data we use in this work are from the Survey on Household Income and Wealth (SHIW) conducted by the *Banca d'Italia* (the Italian central bank) every two years. The main objective of the SHIW is to study economic behaviors of Italian households. The survey is used both for research and for the evaluation of economic policies. Previous studies show that survey estimates usually underestimate the corresponding aggregate figures. Even if national accounts can hardly be considered flawless, the comparison is useful to highlight some quality issues in the microdata. In general, the main sources of error for this kind of surveys are the low propensity of wealthy households to participate in the survey (D'Alessio and Faiella, 2002) and the measurement error that likely arises when collecting survey data of this type (Biancotti et al., 2008). These issues are particularly relevant when it comes to financial wealth. First, financial assets and liabilities are highly concentrated in the hands of wealthy households. Second, the increasing complexity of household financial portfolios increases the respondents' difficulty in retrieving a correct information.

From a data producer point of view, it is crucial to study all the potential survey error components in order to allocate the limited financial resources where most needed (Biemer, 2010). The objective of the paper is to quantify the magnitude of the two main sources of error (nonresponse and measurement error) on the estimator of total household wealth components using the auxiliary information available for the SHIW survey.

The analysis is based on two steps. We first deal with unit nonresponse. Nonresponse is considered as a second phase of sampling with unknown probabilities (see e.g. Särndal, Swensson and Wretman, 1992, Chapt. 9). To this end, we use individual response propensities estimated using data coming from a survey conducted on a subsample of unwilling-to-participate households and from past surveys for panel households (see Little, 1986; Ekholm and Laaksonen, 1991; Kim and Kim, 2007, when estimation is conducted using logistic models). Secondly, we deal with measurement error using a validation sample made of a survey of customers of a major Italian commercial bank, with survey data matched to the bank's administrative records. Measurement error is considered as a third source of uncertainty modeled using propensities to misreport on the validation sample. These propensities are then used to develop a simulation-based adjustment process for SHIW assets data.

The paper is organized as follows. In Section 2 a description of the sampling design employed for SHIW is provided. Then, Section 3 describes the proposed methodology to tackle nonresponse. Different models are developed and considered to estimate response probabilities for panel and non-panel households according to the available auxiliary information. It will be shown that nonresponse is driven by different factors for the two types of households. Section 4 provides details on the models used to estimate misreporting propensities and to obtain imputed values for the variables of interest. Finally, in Section 5 a comparison of the alternative estimators obtained using the aforementioned techniques is provided, together with an appraisal of the role of the auxiliary information employed for nonresponse and measurement error adjustments on the estimates for the survey at hand. Some concluding remarks are also provided to envision further and more general methodological developments suggested by the present application.

2. The sampling design used for the SHIW

The SHIW is a two stage survey, with municipalities and households as primary and secondary sampling units, respectively. PSUs are stratified by administrative region (NUTS 1 level) and population size (less than 20,000 thousand inhabitants; between 20,000 and 40,000; 40,000 or more). Within each stratum, PSUs are selected to include all those with a population of 40,000 inhabitants or more and those with panel households (self-representing municipalities), while smaller municipalities are selected using probability proportional to size sampling (without replacement). Individual households are then randomly selected from administrative registers.

Up to 1987 the survey was conducted with time-independent samples (cross sections) of households. In order to make it possible to analyze the change in the phenomena under investigation, since 1989 part of the sample has included households interviewed in previous surveys (panel households). The overall sample size for the 2008 edition is 7,977 households, with 4,345 panel households (54.5% of the sample). The rotation scheme for the panel component is as follows: households that had participated for at least two waves are all included in the sample, while the remaining panel households are selected randomly from among those interviewed only in the previous survey. As a result, the longitudinal component of the sample consists of a quite heterogeneous group of households as of the year of the first interview and the number of waves. For example, of the 4,345 panel households in 2008, 28 have participated since 1987, 146 since 1989, 347 since 1991 and 1,143 come from the previous 2006 edition.

The questionnaire used in the survey has a modular structure. It is made of a general part addressing aspects concerning all households and a series of additional sections containing questions that are relevant to specific subsets of households. Data collection is entrusted to a specialized company using about 200 professional interviewers. Substitutions are allowed under a strict protocol. In particular, interviewers have no influence on when a household can be dropped and which household to use as a substitute. Information is collected using the Computer-Assisted Personal Interviewing (CAPI) technique. Interviews last an average of 55 minutes. In addition, interviews are considered valid if they have no missing items on the questions regarding income and wealth. As a result, item nonresponse is negligible while, as it will become clearer soon, unit nonresponse is a major issue.

3. Unit nonresponse

In 2008, 14,209 households have been contacted and 7,977 have been interviewed (56.1%), while 32.4% has refused to cooperate and the remaining 11.5% is given by non-contacts (see Table 1). Nonresponse affects particularly non-panel households, in fact 41.1% refuses to participate in the survey, while

this percentage decreases for panel households to 18.5%. To study the factors that drive nonresponse and try to adjust for nonresponse bias, we use a two-phase approach: the selected sample is considered as the first phase sample, while the set of respondents is considered as a second phase sample. Each unit in the population has attached a probability of inclusion for such second phase sample, that is a response probability and, therefore, an unknown characteristic.

Table 1. Households contacted in 2008 and reasons for non-participation.

	Panel		Non-panel		Total	
	Number	%	Number	%	Number	%
Respondents	4,345	79.3	3,632	41.6	7,977	56.1
Refusals	1,012	18.5	3,589	41.1	4,601	32.4
Not at home	120	2.2	1,511	17.3	1,631	11.5
Total	5,477	100.0	8,732	100.0	14,209	100.0
Ineligible *	150	2.7	629	6.7	779	5.2

* Households not found at their address (wrong address, death, change of address).

More formally, given a finite population of N elements $U = \{1,\ldots,k,\ldots,N\}$, the aim is to estimate the vector of totals $\mathbf{t}_y = \sum_U \mathbf{y}_k$, where $\mathbf{y}_k$ is the value of the p-dimensional vector of variables of interest $\boldsymbol{y}$ for the k-th unit. We will use in general the shorthand $\sum_A$ for $\sum_{k\in A}$, with $A \subseteq U$ an arbitrary set. In our application $p = 6$: for each of two types of aggregated financial assets and for financial liabilities we have the number of households possessing the asset (or the liability) and the amount possessed. The two types of aggregated assets are: bonds (government + private bonds) and risky assets (shares + mutual funds + managed savings).

A sample s of size n is drawn from U according to the sampling design $p(s)$ that induces first order inclusion probabilities $\pi_k = P(k \in s)$. Since nonresponse occurs, the response set r of size n_r is obtained from the response mechanism given by the distribution $q(r \mid s)$, with $r \subseteq s$ and $n_r \le n$. Let $\delta_k = 1$ if unit k responds and zero otherwise. Then, $\theta_k = P(k \in r \mid k \in s) = P(\delta_k = 1)$ is the probability that unit k responds given that it was included in the sample. Since θ_k is considered as an individual characteristic defined for all units in the populations, $\theta_k = P(k \in r \mid k \in s) = P(k \in r)$. If these probabilities were known, the two-phase estimator

$$\hat{\mathbf{t}}_{y,2} = \sum_r \frac{\mathbf{y}_k}{\pi_k \theta_k}$$

would be unbiased for $\mathbf{t}_y$.

When auxiliary information is available for all units in s, these probabilities can be estimated using response propensties. One of the most common and simple technique to handle nonresponse is given by constructing response homogeneity groups: the population (or the sample s) is partitioned into groups such that units belonging to the same group are assumed to have the same response propensity. In the SHIW such propensities are currently estimated for a PSU l by the ratio between the effective number of components in the respondents set m_{lr} and the number of components in the original sample m_{ls}. Therefore, the estimated response propensity for household k is given by $\hat{\theta}_k^S = m_{l(k)r}/m_{l(k)s}$, with $l(k)$ denoting the PSU to which household k belongs to. Then, the estimator of the total is computed as

$$\hat{\mathbf{t}}_{y,\,SHIW} = \sum_r \frac{\mathbf{y}_k}{\pi_k \hat{\theta}_k^S}. \tag{1}$$

Another common but more flexible approach is to use a logistic model for the response indicator δ_k under the assumption of the classical binomial response model that δ_k is independent of δ_j for $k \neq j$, i.e. $\theta_{kj} = P(k \,\&\, j \in r) = \theta_k \theta_j$ (Little, 1986; Ekholm and Laaksonen, 1991). More in general, the response probability can be assumed to be the inverse of a known *link* function of an unknown (but estimable) linear combination of model variables (Folsom, 1991; Fuller et al., 1994; Kott, 2006). Asymptotic properties in the case of a logistic link are explored in (Kim and Kim, 2007). This is a reasonable approach here too, because the design foresees the sampling of full households. Note that response homogeneity groups and logistic models provide the same response propensities when the auxiliary variables used in the logistic model are the response group indicator variables.

In this application, two different models and data sources have been employed for panel and for non-panel households. In particular, we can partition the original sample s (and the respondents set r) into two sub-samples given by s_p and s_{np} (and by r_p and r_{np}) corresponding to panel and non-panel households, respectively, so that $s_p \cup s_{np} = s$ (and $r_p \cup r_{np} = r$). Once models are selected, estimates of θ_k for $k \in r_p$ and for $k \in r_{np}$ are obtained and denoted by $\hat{\theta}_k^M$. The estimator of the total is then computed as

$$\hat{\mathbf{t}}_{y,\,NR} = \sum_r \frac{\mathbf{y}_k}{\pi_k \hat{\theta}_k^M}. \tag{2}$$

3.1. Response model for panel households

To estimate response probabilities for panel households we exploit the information from the previous interview(s) and use additive logistic regression (Ruppert, Wand and Carrol, 2003). In particular, the effect of the age of the

interviewer has been modeled using nonparametric regression via p-splines given that there was evidence of a more complex relationship than a linear one. Figure 1 shows the shape of the effect of the age of the interviewer on the linear predictor scale. In general, younger interviewers tend to obtain lower response rates. Table 2 shows the coefficients for the other variables found significant through model selection from all those available.

Figure 1. The estimated effect of the age of the interviewer (and 95% confidence bounds) on the linear predictor scale from the additive logistic response probability model for panel households.

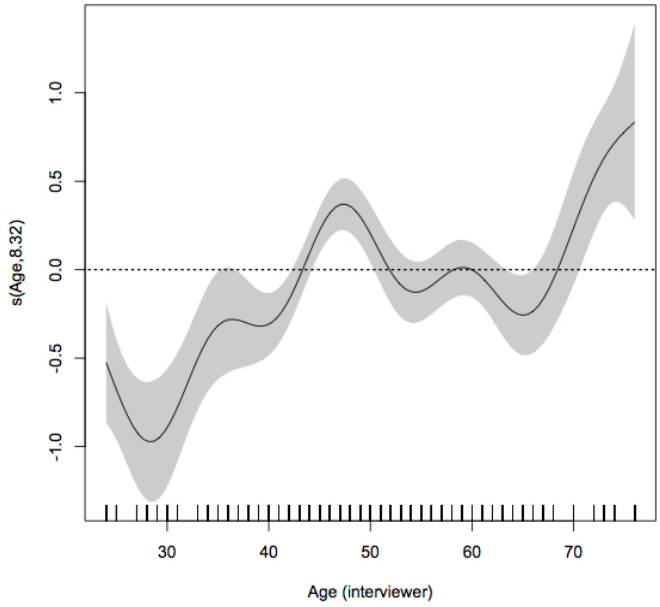


Table 2. Logistic response probability model for panel households - estimated coefficients, standard errors and *p*-values.

Variables	Coeff	Std. Err.	p-value
Intercept	-1.48	0.21	< .001
Municipalities with more than 500,000 inhabitants	-0.58	0.12	< .001
Household living downtown	0.34	0.10	< .001
Number of waves (household)	0.18	0.02	< .001
Number of members of household	0.11	0.03	< .001
High level of education (interviewer)	0.34	0.10	< .001
Number of waves (interviewer)	0.03	0.01	< .001
Good climate at previous interview	0.20	0.02	< .001
Workload of interviewer 21 - 100	-0.06	0.09	0.481
Workload of interviewer 101 - 300	-0.43	0.13	< .001
Workload of interviewer > 300	0.50	0.15	< .001

Pseudo $R^2 = 0.085$; 5,625 obs.

For panel households, responding appears to be mainly a matter of trust. On the contrary, household economic conditions, although included in the model as available auxiliary information, do not show to have an effect on the response propensity. The only household's attributes that have an effect on the response rate, are the number of members and the place where they live: numerous households and those living downtown are more likely to continue to participate, while those living in larger municipalities show higher attrition. On the other side, a major determinant of response propensity is the number of waves the household has already been interviewed successfully. Old panel households are more willing to continue to participate. For instance, households who have entered the panel in 2006 have an estimated response probability of about 0.68. This figure jumps to about 0.90 for those households who have been in the panel for more than 5 waves. One reason is the building of a relationship of trust between respondents and the survey and, in particular, with the interviewer. Households become progressively aware that there is no risk of a break in confidentiality. At the same time, their identification with survey aims increases as time passes. In order to preserve such a link with the respondents, panel households are usually assigned to the same interviewer.

Moreover, a climate judged as "good" by the interviewer at the previous interview provides higher household cooperation. Other important variables affecting response are connected to the characteristics of the interviewer. Interviewers with a relatively higher degree of education, who take larger workloads and have participated in a larger number of editions of the survey have better results. The estimated function of age and coefficients from Table 2 are used to predict response probabilities $\hat{\theta}_k^M$ for all $k \in r_p$, i.e. for the 4,345 interviewed panel households to be used in estimator (2).

3.2. Response model for non-panel households

In 2008, 8,732 non-panel households have been contacted and 3,632 (41.6%) have been interviewed. About 70% of the 5,100 non participating households has explicitly refused to cooperate, while the remaining 30% is not found at the address. The response propensity modeling for non-panel households is based on an *ad hoc* source of auxiliary information. Starting from 2006, the survey agency has to carry out a survey among non-panel households that refuse to participate. It is a telephone survey (CATI technique) run on a sample of non-respondents. This survey is conducted during the fieldwork, while trying to convert refusals. If the attempt is not successful, the interviewers ask whether the household is at least willing to reply to a five minutes telephone questionnaire. The survey agency has

to contact all the non-participating households. Among non-participating households, only 316 have agreed to the telephone interview, about 6% of those households which are selected but do not participate.

For non-panel households, auxiliary information is not known for each unit in the original sample s_{np}, but only for the respondents r_{np} and a subsample of units of $s_{np} \setminus r_{np}$. Nonetheless, we propose to estimate response probabilities using weighted logistic regression on a dataset made of the subsample of nonrespondents and the sample of respondents. In general, (*i*) nonrespondents should be given a weight equal to the inverse of the inclusion probability coming from the sub-sampling design, while (*ii*) respondents a weight equal to 1. For the case at hand, given that the sub-sample of nonrespondents is not a probabilistic sample, a sort of post-stratification is employed in which (*i*) nonrespondents are given a weight that sums up to the total number of nonrespondents by geographical area and size of the municipality resulting from the sample register file, while (*ii*) respondents are given a weight equal to 1 (Laaksonen and Chambers, 2006, use a similar approach when the variable of interest is observed on a sub-sample of non-respondents – follow-up sample). This approach assumes that sub-sampling is at random and that nonrespondents in the sub-sample can be considered similar to the others in the same post-stratum. We will discuss more in detail later whether such assumptions can be considered valid in the situation at hand.

Response probabilities are then estimated as a function of a set of variables that are available for both samples using additive logistic regression as for panel households. In particular, in this case the effect of the age of the head of the household has been modeled using nonparametric regression via p-splines given that there was evidence of a more complex relationship than a linear one. Figure 2 shows the estimated function of age on the linear predictor scale, while Table 3 shows the estimated coefficients for the other variables found significant.

The propensity to respond decreases steadily with age until the age of 30 where it stabilizes and then decreases again. A slight increase is then detected between 65 and 75. The horizontal dotted line shows that households with heads who are 50 or younger are more willing to participate than those with heads who are older than 50. Table 3 shows that response probabilities decrease for households whose head is self-employed, home owner, graduated, or retired. In addition, households living in the North/Centre of Italy and those with a larger number of members are less willing to participate. On the contrary, response propensity increases for households who in smaller municipalities. Finally, households with two (three or more) wage earners are less (more) likely to respond than those with only one.

Figure 2. The estimated effect of age (and 95% confidence bounds) on the linear predictor scale from the logistic response probability model for non-panel households.

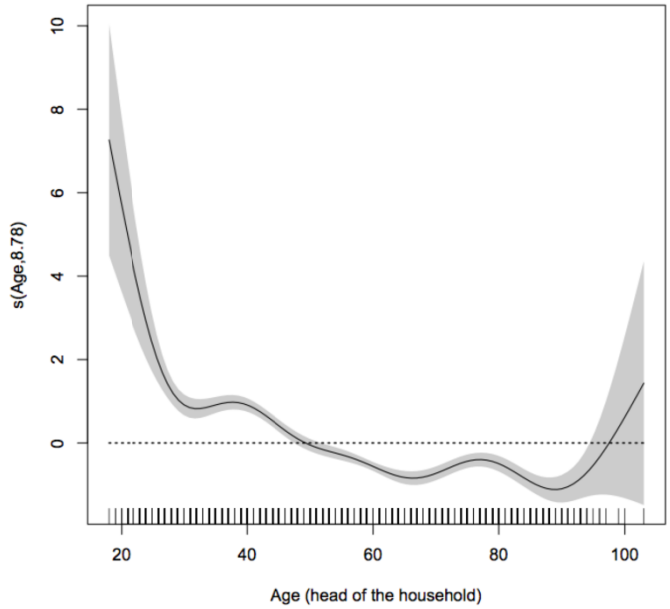


Table 3. Logistic response probability model for non-panel households - estimated coefficients, standard errors and p-values.

Variables *	Coeff	Std. Err.	p-value
Intercept	1.599	0.117	< .0001
Living in the North/Centre of Italy	-0.744	0.058	< .0001
Municipality with less than 500,000 inhabitants	0.362	0.059	< .0001
Household originally selected (vs substitute)	-0.280	0.057	< .0001
Workload of interviewer 21 -- 100	0.361	0.058	< .0001
Workload of interviewer 101 -- 300	0.880	0.077	< .0001
Workload of interviewer > 300	1.912	0.110	< .0001
Self-employed	-0.482	0.084	< .0001
Graduated	-0.265	0.078	0.0007
Retired	-0.291	0.109	0.0073
Home owner	-0.658	0.058	< .0001
Number of members of the household	-0.368	0.028	< .0001
Number of income earners = 2	-0.103	0.055	0.0639
Number of income earners ≥ 3	0.525	0.099	< .0001

* Demographic characteristics refer to the head of the household; Pseudo $R^2 = 0.357$; 3,948 obs.

In the response model for panel households, information about the interviewer was found to be significant. For non-panel households, such information is not available. Indeed the interviewers for the CATI survey are different from those running the CAPI survey. The only useful information is that of the workload of the original interviewer measured by the number of households to be interviewed. Those with a larger workload tend to have larger response rates than the others. A likely explanation for this result is that the survey agencies usually allocate a larger number of households to their best interviewers.

Note that no explicit income related items are surveyed on the sub-sample of nonrespondents given their refusal to participate to the SHIW. Therefore, there is no information available on this to be incorporated in the response model for non-panel household. Nevertheless, some of the variables found significant that are related to the head of the household are usually good predictors of wealthier households (being graduated, self-employed, home owner). The estimated function of age and coefficients are then used to compute estimated response probabilities $\hat{\theta}_k^M$ for all $k \in r_{np}$, i.e. for all 3,632 non-panel respondents to be used in estimator (2).

4. Measurement error

Financial assets collected in the SHIW are also likely to be affected by misreporting of the financial tools and amounts by households. Such misreporting may well be due to a malicious behavior, with underreporting being the most likely outcome. However, it can also be done in *bona fides*, given the respondents' difficulty in retrieving a correct information due to the increased complexity of household financial portfolios. For these reasons, the value for the variables of interest reported by unit k, which we will denote by $\tilde{\mathbf{y}}_k$, may differ from the true value $\mathbf{y}_k$.

Bias caused by measurement error could be adjusted for by selecting a subsample m of the respondents where a more accurate measurement of the study variable(s) is taken (e.g. Lessler and Kalsbeek, 1992). When the subsample is selected using a probabilistic sampling design, the framework is another example of two phase sampling. When nonresponse is present, as is the case of the SHIW, then a three phase framework arises: $m \subset r \subseteq s$ of dimension $n_m < n_r \leq n$ is selected using the design $p_m(m \mid r, s)$ with conditional inclusion probabilities $\tau_k = P(k \in m \mid k \in r)$. Then, the three phase estimator

$$\hat{\mathbf{t}}_{y,3} = \sum_m \frac{\mathbf{y}_k}{\pi_k \theta_k \tau_k}$$

would be unbiased for $\mathbf{t}_y$. Of course the efficiency of $\hat{\mathbf{t}}_{y,3}$ depends on the dimension of m: a compromise choice should be made according on how expensive it is to retrieve the correct information on units. The unbiased estimator $\hat{\mathbf{t}}_{y,3}$ is constructed using the subsample m alone. Other estimators that make a better use of the information on the respondents set r (given by the correlated surrogate variable $\tilde{\mathbf{y}}_k$ and some auxiliary information) can be proposed in a model assisted framework to improve efficiency, using GREG type or model calibration type estimators (e.g. see the hint in Wu and Luan, 2003, Section 6, in a two phase framework). These extensions go however beyond the scope of this paper.

Now, for this survey we have no such data available on a subsample of r and the three phase approach cannot be used as described earlier. However, we have data available from an independent experiment survey carried out by the *Banca d'Italia* and a major Italian bank group on a sample of customers of the latter. The experiment was carried out in 2003 on a sample of 1,681 households where at least one member was a customer of the bank group. In order to get data comparable with that coming from the SHIW, the questionnaire and the survey design were as close as possible to those used in the most recent edition of the SHIW (2002). Interviews were made by the same survey agency using the same interviewers and CAPI technique.

Survey data had then been matched with the bank customers database containing the amount of the assets actually held by the individuals selected in the sample. Since these amounts and those declared in the interview refer to the same period (year 2002), they were fully comparable. The two sets of data are merged through an operation of exact record linkage. The resulting dataset will be referred to as our "validation sample".

Although temporally misaligned, the validation sample gives us the possibility of studying misreporting behaviors and of trying to extrapolate it to the SHIW sample. This is accomplished in a two-step fashion. In fact, household wealth reporting in surveys is generally a two-stage process involving first the reporting of ownership of assets and liabilities and then the reporting of the amounts owned (Moore et al., 2000). Errors can occur at either of the two stages. An entire financial instrument can be either omitted or reported even if it is not actually owned. Alternatively, the ownership may be reported correctly but the amount can be misreported. Even if the respondent has fully understood the question, he/she may fail to retrieve the correct information. Lack of knowledge is the first cause. Even if in the SHIW the respondent is selected as the more knowledgeable person in the household, he or she may not know the true situation of all the other components.

In the final stage, after recalling the requested information, the respondent adopts a response strategy. Deliberate underreporting is probably the major cause of response error at this stage. Nonetheless, besides deliberate prevarication, there are further possible sources of error, like those coming from the interaction between the interviewer and the respondent. For instance, if the respondent belongs to a very rich household he/she may decide to underreport wealth because of a need of "social conformability" with the interviewer. This could be

considered as a special case of the so called "social desirability bias" (Bagozzi, 1994), namely, the tendency for an individual to present himself in a way that makes the person look positive to cultural norms or standards. On the opposite side, over-reporting may arise from a respondent willing to impress the interviewer.

Recall that we consider two types of aggregate financial assets – bonds and risky assets – and also financial liabilities. For each of these three quantities we have, therefore, two related variables: possession and amount possessed. For the former, we have in particular two variables defined as follows:

$$y_{pk} = \begin{cases} 1 & \text{if unit } k \textit{ possesses} \text{ financial instrument } p = 1,2,3 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\tilde{y}_{pk} = \begin{cases} 1 & \text{if unit } k \textit{ declares to possesses} \text{ financial instrument } p = 1,2,3 \\ 0 & \text{otherwise} \end{cases}.$$

Table 4. Logistic response probability model for ownership of bonds – estimated coefficients, standard errors and p-values.

Variables *	Coeff	Std. Err.	p-value
Intercept	-4.67	0.98	< .0001
Self-reported ownership of bonds	2.50	0.17	< .0001
Self-reported ownership of risky assets	0.45	0.12	0.0003
Employee	0.21	0.19	0.2589
Self-employed	0.23	0.20	0.2502
Secondary school diploma	0.39	0.15	0.0086
University degree	0.55	0.19	0.0045
Age	0.06	0.03	0.0477
Age^2	-0.01	0.03	0.6520
Living in municipalities with more than 30,000 inhabitants	-0.26	0.13	0.0432
Living in the North/Centre of Italy	0.86	0.20	< .0001
Living in a rural area	-0.51	0.24	0.0315
Living in the town outskirts	-0.36	0.23	0.1133
Number of income earners	0.02	0.05	0.6610
First quartile of household income	-0.05	0.16	0.7792
Fourth quartile of household income	-0.16	0.16	0.3139
First quartile of household real wealth	0.35	0.15	0.0203
Fourth quartile of household real wealth	0.06	0.16	0.7051
Client of more than one bank	0.06	0.13	0.6171
Respondent's level of understanding of the questions	0.01	0.07	0.8899
Respondent's easiness to answer the questions	0.00	0.07	0.9935
Reliability of the information provided by the respondent	-0.03	0.04	0.4168

* Demographic characteristics refer to the head of the household; Pseudo $R^2 = 0.364$; 1,681 obs.

The validation sample then allows to identify, in the first phase, households among those declaring not to own a given financial asset, that very likely own it and provided incorrect data and, symmetrically, to identify households declaring to own a financial asset, but unlikely to possess it. This is accomplished by estimating a logistic model for $P(y_{pk}=1)$ using a vector of socio-economic characteristics both at the household and at the head of household level as covariates, together with the declared value $\tilde{y}_{pk}$.

Tables 4, 5 and 6 report the results from the models for the variables bonds, risky assets and financial liabilities, respectively. The tables also show the p-values for the covariates considered and an overall measure of goodness of fit. Note that these models are fit with the aim of imputation. Therefore, model selection is based on the performance for out-of-sample predictions rather than on the amount of variability explained by the covariates. For this reason, also non-significant covariates can be found in the aforementioned Tables. A nonparametric term for the effect of the age of the head of the household has been tested, but provided no better performance than a quadratic function of age and was dropped in all models.

Table 5: Logistic response probability model for ownership of risky assets - estimated coefficients, standard errors and p-values.

Variables *	Coeff	Std. Err.	p-value
Intercept	-2.27	0.89	0.0105
Self-reported ownership of bonds	0.18	0.16	0.2640
Self-reported ownership of risky assets	2.65	0.15	< .0001
Employee	0.56	0.19	0.0038
Self-employed	0.07	0.21	0.7455
Secondary school diploma	0.31	0.15	0.0347
University degree	0.13	0.20	0.5184
Age	0.04	0.03	0.2298
Age^2	-0.01	0.03	0.6452
Municipalities with more than 30,000 inhabitants	0.09	0.13	0.4904
Living in the North/Centre of Italy	0.19	0.17	0.2667
Living in a rural area	0.41	0.25	0.0906
Living in the town outskirts	0.50	0.23	0.0332
Number of income earners	-0.12	0.06	0.0322
First quartile of household income	-0.20	0.16	0.2149
Fourth quartile of household income	0.23	0.17	0.1795
First quartile of household real wealth	-0.01	0.15	0.9243
Fourth quartile of household real wealth	0.45	0.17	0.0092
Client of more than one bank	0.04	0.13	0.7346
Respondent's level of understanding of the questions	-0.07	0.07	0.3640
Respondent's easiness to answer the questions	0.03	0.08	0.6768
Reliability of the information provided by the respondent	-0.04	0.04	0.2953

* Demographic characteristics refer to the head of the household; Pseudo $R^2 = 0.393$; 1,681 obs.

From Table 4, the probability of holding bonds increases for relatively less wealthy households living in the North/Center of Italy and in smaller municipalities. In addition it increases with the age of the head of the household and with his/her level of educational attainment. From Table 5, on the other hand, the probability of owning risky assets increases for relatively wealthier households with few components, which live in residential or rural areas. Finally, Table 6 shows that the probability of owing financial liabilities decreases for relatively less wealthy households living in the North/Centre of Italy, in smaller municipalities and in rural or residential areas.

Table 6. Logistic response probability model for ownership of financial liabilities - estimated coefficients, standard errors and p-values.

Variables *	Coeff	Std. Err.	p-value
Intercept	-5.93	1.94	0.0022
Self-reported ownership of liabilities	4.43	0.30	<.0001
Employee	-0.87	0.39	0.0255
Self-employed	0.40	0.40	0.312
Secondary school diploma	-0.18	0.30	0.5413
University degree	0.05	0.42	0.9063
Age	0.25	0.07	0.0009
Age^2	0.00	0.00	0.0003
Municipalities with more than 30,000 inhabitants	0.10	0.27	0.7211
Living in the North/Centre of Italy	0.00	0.34	0.9894
Living in a rural area	-1.43	0.46	0.0021
Living in the town outskirts	-0.94	0.41	0.0227
Number of income earners	0.12	0.11	0.2999
First quartile of household income	0.15	0.34	0.6562
Fourth quartile of household income	0.41	0.33	0.2146
First quartile of household real wealth	-0.18	0.31	0.5698
Fourth quartile of household real wealth	-0.04	0.35	0.8968
First quartile of household financial assets	0.44	0.29	0.1328
Fourth quartile of household financial assets	-0.73	0.36	0.046
Respondent's level of understanding of the questions	-0.33	0.17	0.0494
Respondent's easiness to answer the questions	0.31	0.17	0.0714
Reliability of the information provided by the respondent	-0.08	0.09	0.3653

* Demographic characteristics refer to the head of the household; Pseudo R^2= 0.403; 949 obs.

In a second phase misreporting on the amount held is estimated through a separate model for each of the three financial tools. In particular, for the three last variables of interest y_p, $p = 4,5,6$, let $r_{pk} = y_{pk} / \tilde{y}_{pk}$ be the ratio of the actual and the declared amount held by household k. Then $\log r_{pk}$ is modeled for each unit of the validation sample on a set of household characteristics, including household income and wealth classes, a synthetic judgmental variable on the reliability of the information provided in the interview expressed by the

interviewer (a discrete score ranging from 1 to 10), and the declared amount. Tables 7 and 8 report the results from the models for bonds and for risky assets, respectively. For financial liabilities the number of available observations is too small for proper modeling of r_{pk}. Therefore, a common mean model for the ratio is estimated, that takes value 1.064 for all units in the sample.

These two sets of models can be used to adjust measurement error in the SHIW as follows. If we assume that the misreporting behavior of the households in the bank experiment is the same as that of those in the SHIW, then parameter estimates from these two sets of models can be used to stochastically impute micro data for households in the SHIW (imputation for measurement error correction for distribution function estimation is explored in Durrant and Skinner, 2006). In particular, profiles of households given by unique combinations of covariate values are constructed from the SHIW, then predictions $\hat{\mathbf{y}}_k$ are obtained using parameter estimates from the aforementioned models that substitute the surveyed values $\tilde{\mathbf{y}}_k$. A random error term is then added to preserve variability. In particular, in the models for asset ownership a Bernoulli experiment is conducted to assign the imputed possession of a given asset class. As of the models related to the amount possessed, a random draw from a zero-mean normal distribution is added to the imputed value; the variance of the normal distribution is given by that of the residuals of the model fitted in the validation sample.

Table 7. Regression model for log of ratio between actual and declared amount of *bonds* - estimated coefficients, standard errors and p-values.

Variables *	Coeff	Std. Err.	p-value
Intercept	0.85	0.69	0.2197
Second quartile of household financial wealth in risky assets	-0.63	0.12	<.0001
Third quartile of household financial wealth in risky assets	-0.64	0.12	<.0001
Fourth quartile of household financial wealth in risky assets	-1.09	0.12	<.0001
Employee	0.38	0.13	0.0038
Self-employed	0.25	0.13	0.062
Secondary school diploma	-0.02	0.10	0.8718
University degree	-0.06	0.12	0.6204
Age	-0.02	0.02	0.5034
Age^2	0.00	0.00	0.2413
Living in the North/Centre of Italy	0.23	0.17	0.1666
First quartile of household income	-0.02	0.11	0.8534
Fourth quartile of household income	0.11	0.10	0.2782
First quartile of household real wealth	-0.08	0.11	0.4769
Fourth quartile of household real wealth	0.07	0.10	0.5081
Client of more than one bank	0.03	0.08	0.7526
Reliability of the information provided by the respondent	0.01	0.02	0.703

* Demographic characteristics refer to the head of the household; Pseudo R^2 = 0.453; 482 obs.

Table 8. Regression model for the log of ratio between actual and declared amount of *risky assets*- estimated coefficients, standard errors and p-values

Variables *	Coeff	Std. Err.	p-value
Intercept	0.43	0.31	0.1634
Second quartile of household financial wealth in risky assets	-0.04	0.07	0.5864
Third quartile of household financial wealth in risky assets	-0.34	0.06	<.0001
Fourth quartile of household financial wealth in risky assets	-0.32	0.07	<.0001
Employee	-0.08	0.08	0.3025
Self-employed	-0.04	0.08	0.6245
Secondary school diploma	0.14	0.06	0.0174
University degree	0.16	0.07	0.0318
Age	0.00	0.01	0.9866
Age squared	0.00	0.00	0.802
Living in the North/Centre of Italy	0.00	0.08	0.9887
First quartile of household income	0.13	0.07	0.0578
Fourth quartile of household income	0.12	0.06	0.0416
First quartile of household real wealth	-0.07	0.06	0.2726
Fourth quartile of household real wealth	-0.04	0.06	0.5137
Client of more than one bank	0.00	0.05	0.9368
Reliability of the information provided by the respondent	-0.04	0.01	0.0009

* Demographic characteristics refer to the head of the household; Pseudo $R^2 = 0.126$; 876 obs.

The final estimator of the total of the variables of interest adjusted for measurement error is essentially a two-phase estimator, and takes the two following forms according to whether nonresponse is adjusted for using the logistic models or not:

$$\hat{\mathbf{t}}_{\hat{y},ME} = \sum_r \frac{\hat{\mathbf{y}}_k}{\pi_k \hat{\theta}_k^S}, \tag{3}$$

$$\hat{\mathbf{t}}_{\hat{y},NRME} = \sum_r \frac{\hat{\mathbf{y}}_k}{\pi_k \hat{\theta}_k^M}. \tag{4}$$

5. Results and concluding remarks

In this section we report the final estimates of the six variables of interest (holding and amount possessed for bonds, risky assets and financial liabilities) obtained using the alternative estimators discussed in the previous sections. Table 9 reports the estimates of the total of the first three variables of interest

($p = 1,2,3$), i.e. the number of households holding the financial instrument, plus the estimate for the number of households holding either bonds or risky assets, or holding both (*total financial assets*). Note that the first two estimators are computed as in equations (1) and (2), respectively, in which $\mathbf{y}_k$ is replaced by the observed surrogate value $\tilde{\mathbf{y}}_k$. Table 10 reports, on the other hand, the estimates of the total of the last three variables ($p = 4,5,6$), i.e. the amount held (in billions of Euros) by households for the three financial instruments, plus the estimate of the total financial assets own. As a measure of the coverage of each estimate, the ratio of its value with respect to the corresponding estimate coming from the National Financial Accounts (NFAs) is also computed.

Our main findings may be summarized as follows. Underreporting and unit nonresponse emerge as particularly serious issues with regard to financial assets. The estimator $\hat{\mathbf{t}}_{\hat{y},\,NRME}$ of the total adjusted for both nonresponse and measurement error is about 3-4 times higher that the unadjusted $\hat{\mathbf{t}}_{\tilde{y},\,SHIW}$ when financial assets are considered (Table 9). When it comes to the estimation of the amounts held (Table 10), the bias increases: the SHIW estimates for the variables bonds and risky assets should be inflated by factors from 5 to 8 in order to get the figures obtained with $\hat{\mathbf{t}}_{\hat{y},\,NRME}$. The latter are those closest to the estimates coming from NFAs (last two columns of Table 10).

Correction for nonresponse and measurement error for financial liabilities seems to be less effective. As far as the amount is concerned, this may be due to the very simple measurement error model employed for this variable. In addition, the information on liabilities is generally easier to recall and less sensitive than the information on the assets. This usually results in general in a lower measurement error.

Table 9. Number of households (millions) holding a financial instrument using different estimators

Instrument	$\hat{\mathbf{t}}_{\tilde{y},\,SHIW}$	$\hat{\mathbf{t}}_{\tilde{y},\,NR}$	$\hat{\mathbf{t}}_{\hat{y},\,ME}$	$\hat{\mathbf{t}}_{\hat{y},\,NRME}$	$\frac{\hat{\mathbf{t}}_{\hat{y},\,NRME}}{\hat{\mathbf{t}}_{\tilde{y},\,SHIW}}$
Bonds	2.605	2.877	10.710	10.709	4.11
Risky assets	3.332	3.674	13.184	13.112	3.94
Total financial assets	4.832	5.310	16.021	16.068	3.33
Financial liabilities	5.646	5.881	6.518	6.800	1.20

Table 10. Total amount held (billions of euros) using different estimators, estimate from NFAs, and ratio between survey estimates and NFAs (percentages).

Instrument	$\hat{\mathbf{t}}_{\tilde{y},\,SHIW}$	% *of NFAs*	$\hat{\mathbf{t}}_{\tilde{y},\,NR}$	% *of NFAs*	$\hat{\mathbf{t}}_{\hat{y},\,ME}$	% *of NFAs*	$\hat{\mathbf{t}}_{\hat{y},\,NRME}$	% *of NFAs*	NFAs *
Bonds	102.3	13.5	127.3	16.8	486.2	64.1	501.8	66.1	758.8
Risky assets	126.8	11.5	155.2	14.1	994.3	90.3	1,005.2	91.3	1,100.6
Total financial assets	229.1	12.3	282.5	15.2	1,480.5	79.6	1,507.0	81.0	1,859.4
Financial liabilities	257.6	41.2	298.7	47.8	304.8	48.8	350.8	56.2	624.8

* The figures exclude: cash, bank and postal deposits, and technical insurances.

In order to give a sense of the magnitude of the impact of nonresponse and measurement error, it is useful to compare it with the magnitude of sampling errors: the relative standard error for the total financial assets and for financial liabilities is about 5 and 6 percent, respectively. These figures are negligible compared to the ones shown in the aforementioned tables. The main implication is that surveys on households' wealth require data producers to pay more attention to nonsampling errors rather than to the sampling error alone. Our results are likely driven by the specific features of the SHIW. Yet, when it comes to households' wealth, sampling errors are likely to play a negligible role compared to nonsampling errors. Data producers should therefore allocate their limited financial resources accordingly.

Moreover, in the case of the SHIW, the bias due to measurement error outweighs by far the bias due to unit nonresponse. This result is in part due to how the survey on nonrespondents is designed. The response rate for this survey is very low and there is likely a severe issue of self-selection underneath the sampling of nonrespondents. Therefore, the response model for non-panel households seems to be more a model for the propensity to response over that of providing at least a short interview, rather than over a refusal. The survey on nonrespondents seems to fail to provide information on the real refusing households. For this reason the adjustment for unit nonresponse should be considered as a lower bound.

Finally, measurement error has been dealt with by imputation using model estimates coming from an external validation sample. It would certainly be of interest to investigate the properties of an estimator based on a subsample of households on which an accurate measurement of the variables of interest can be taken. More study is required to determine the sub-sampling design and dimension. Finally, like with imputation for item nonresponse, a fully weighting or an imputation approach can be used to determine the final estimates. Both approaches have pros and cons. The former requires the dissemination of a

different set of weights for each variable of interest for which accurate measurements are taken. The latter, on the other hand, allows the computation of a single set of weights, but requires the dissemination of imputed values for units not in the subsample.

Acknowledgement

The work of Ranalli is supported by the Research of National Interest n. 2007RHFBB3PRIN awarded by the Italian government to the University of Perugia. The views expressed in the paper are solely those of the authors and do not necessarily correspond to those of the *Banca d'Italia*.

REFERENCES

Bagozzi, R. (1994). Measuring in market research: basic principal of questionnaire design. *Principles of Marketing Research in: Blackwell Business.*

Biancotti, C., G. D'Alessio, and A. Neri (2008). Measurement error in the Bank of Italy's survey of household income and wealth. *Review of Income and Wealth,* 54-3.

Biemer, P. (2010). Total survey error: design, implementation, and evaluation. *Public Opinion Quarterly* 74, 817-848.

D'Alessio, G. and I. Faiella (2002). Nonresponse behaviour in the Bank of Italy's survey of household income and wealth. *Temi di Discussione del Servizio Studi*, Banca d'Italia 462.

Durrant, G. B. and C. Skinner (2006). Using missing data methods to correct for measurement error in a distribution function. *Survey Methodology,* 32-1, 25-36.

Ekholm, A. and S. Laaksonen (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics* 7, 325-337.

European Central Bank (2009). Survey data on household finance and consumption. ECB, *Occasional paper series* (100).

Folsom, R. E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *In ASA Proceedings of the Social Statistics Section*, pp. 197-202.

Fuller, W. A., M. M. Loughin, and H. D. Baker (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology* 20, 75-85.

Kim, J. K. and J. J. Kim (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics* 35 (4), 501-514.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* 32 (2), 133-142.

Laaksonen S. and R. Chambers (2006) Survey estimation under informative nonresponse with follow-up, *Journal of Official Statistics* 22, 81-95

Lessler, J. T. and W. D. Kalsbeek (1992). *Nonsampling Error in Surveys*. John Wiley & Sons.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* 54, 139-157.

Moore, J., L. Stinson, and E. Welniak (2000). Income measurement error in surveys: A review. *Journal of Official Statistics* 16, 331-361.

Ruppert, D., M. P.Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, New York.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer, Berlin, New York.

Wu, C. and Y. Luan (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics* 19, 119-131.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 301—308

QUANTIFICATION OF THE FACTORS OF STATISTICAL WORK LABOR INPUT USING THE METHODS OF SAMPLE SURVEYS

Julia Orlova[1]

ABSTRACT

This research work aims at developing a methodological framework for evaluation of labor input of statistical work. Such an evaluation is necessary to determine the cost of services provided by state statistics on, as well as to determine the level of budget financing. During the study the method of a sample photo of the working day had been used. The aim of the study was to calculate the ratio of the working time (the cost of work which is not directly associated with an assignment to the cost of the operative time) in a general population of workers of the Minsk regional statistics department. Such ratio is necessary for further evaluation of complexity of statistical work on the methodology described in this paper.

Key words: labor input of statistical works, costs of statistical works, expenses on the state statistical bodies.

1. Introduction

The main objectives and principles of state statistics are stated in the Law of Belarus "On State Statistics" [1]. The main tasks are:

- Development of science-based methodology and its improvement in accordance with national and international standards in statistics;
- Collection, processing, compilation, storage and protection of statistical data (information) on the basis of statistical methodology;
- Providing statistical data (information) to the President of Belarus, the National Assembly of Belarus, Council of Ministers, Presidential Administration, the State Control Committee of Belarus, republican government authorities and other state organizations subordinate to the Council of Ministers, regional and Minsk City executive committees;
- Dissemination of the summary of statistical data (information).

[1] Belarus State Economics University, Minsk. E-mail: orlova-julia-gen@mail.ru

The main principles of state statistics are: objectivity and reliability, stability and comparability of statistical data, its accessibility and openness within the boundaries established by the legislation of the Republic of Belarus, the integration of statistics into accounting as the main source of economic information.

Science-based solution of the formulated problems, thread management, statistical information, the systematic organization of employees of the state statistics in accordance with its scientific principles, the development of statistical work plans and monitoring their performance are inextricably linked with the assessment of complexity of statistical work. Publications on this subject in the Republic of Belarus are almost absent but seem relevant delineated on the background theme of the present paper.

Defining the methodological framework for evaluation of labor input of statistical work described in this paper is currently one of the primary tasks of the state statistics bodies in Belarus in connection with the transition to a unified system of electronic documents and justification of the effectiveness of the system implementation.

2. Categories of Time Costs

The evaluation of labor input of statistical work needs identification and quantitative measurement of productive time employees of the statistics.

The proposed assessment suggests that the total unit of statistical work is divided on operation of the input and output information. Working with the input information includes collecting, recording, organizing statistical data, etc. Working with the output information is represented by processing, accumulation, storage and presentation of summary statistics on economic, demographic, social and environmental situation in Belarus.

The complexity of statistical work depends on the time-consuming structure of state statistical bodies. Our study has the dual purpose of evaluatingf the value of the operational costs per work unit and the subsequent adjustment to a number of factors that take into account the special properties and components of the work.

On the basis of a survey conducted in government statistics and presented in this study, the author proposes that the following correction factors are developed:

Input information:

1. complexity (*Kc*);
2. value (*Kv*);
3. occupancy (*Ko*);
4. amount of information (*Kai*).

Output information:

1. amount of information (*Kao*).

In assessing the complexity one should distinguish the following categories of the working time: operative time, set-up time, while observing the operation of the equipment. In this set-up time and time monitoring of equipment operation the categories of input and output information should be considered. At the same time the operational statistical work must be allocated to time, which is directly aimed at the implementation of the set tasks.

We define the operative time with the input includes time required for registration and reporting questionnaires and other statistical microdata, transferring primary input records to electronic media, merging of sets of information on state statistical reports received from districts (at the National Committee - from the statistical bodies), editing and checks of micro and macrodata and forming output tables.

Operative work with the output is represented by time-consuming tasks of the executive bodies, analytical writing and briefing notes, writing and editing of publications as well as the work with the statistical registers..

Set-up time is the time for preparation before the implementation of a given work but also the time associated with their ending. The set-up time is divided between input and output information.. Set-up time for the input includes consultations in the area (businesses) on issues arising in the process of monitoring, the workers' press reports and organization of materials for controlling details of statistical reporting.. Set-up while working with the output information includes decisions on what materials to write, on analyzing methods, and other materials for the statistical publications, documentation and methods for archival storage and systematization. We included also the time to perform other operations such as monitoring the operation of the equipment when working with input and output information.

The labor input of statistical work is proposed to be modeled by the following formula:

$$Ti = Top_i \times Kc \times Kv \times Ko \times Kai, \quad (1)$$

where Top_i – is the average operative time for work related to the processing of input data in hours.

The output information is modeled as:

$$To = Top_o \times Kao, \quad (2)$$

where Top_o – is the average operative time for work related to the processing of output information, hour.

The total operative time is defined as their sum: $Top = Ti + To$.

Finally the time required to set-up work and work on the supervision of equipment is modeled as follows:

$$Ts + Tm = To \times K1, \quad (3)$$

where Ts is the value of cost set-up time; Tm the value of time spent on monitoring the operation of the equipment; To the total operative time work and $K1$ the ratio of working time (proportion of work that is not directly associated with an operative time), determined on the basis of photo of working day.

The overall complexity of the statistical work is determined by summing the operating times for input and output information, set-up time costs and also time costs to perform work for the supervision of equipment

3. Quantification of the Labor Input of Statistical Work

Collection of baseline data to determine the above staff times is performed in the "Photo of Working Day Professional Bodies of State Statistics Survey". It is recommended that this evaluation should be done every three years. It should be noted that self photo with a list of activities (components of labor costs) is more convenient to process data, but does not fully reflect the use of time since it allows for subjectivity. However, for large scale monitoring and long-term use of it is considered to be feasible.

When carrying out this study it is important to properly define the scope of work (photo of working day of employees of state statistics including managers and specialists). The author has found that to get results with sufficient precision it is necessary to conduct observations on a random sample of 10-15 per cent of the specialists for this population. Because of the great complexity the method should be used with care as a tool to specify or check the time spent on certain types of work.

To test the sampling methodology for this study the Minsk Regional Statistics Department has been chosen. It is a regional agency of the State Statistics of the Republic of Belarus which is responsible for collecting and consolidating statistical information for Minsk. It is subordinate to the National Statistics Committee of Belarus, along with six other regional departments.

Information about the structure of the working time selective data on the time spent on different tasks was obtained as a result of photo of working day, (including data on time spent on tasks related to processing the statistical forms) .The data was extended to the general population which amounted to 127 managers and specialists of the Minsk regional department at the time of photography. Photo of working day was carried out of the Minsk regional statistics department. A stratified random sample of 15 people was selected in three divisions of the department.

Photo of working day was carried out during one day by the staff of the Research Institute of Statistics by means of questionnaires tailored to the subdivision procedures for all types of statistical work in the category of working time.

The grouping of the working time is presented in Table 1. Thus, e.g. a manager/specialist in the sample spent on average 5023/15 = 334.9 minutes. (5.58 hours) to perform the specified work during the studied day, 118.6 minutes (1.98 hours) to prepare for the implementation of a given work and activities related to its end, and 15.5 minutes (0.26 hours) to monitor the operation of the equipment.

Table 1. The grouping of working time of the sample of workers of the Minsk regional statistics department

Category of working time	Recur-rence of the time	The total time for 15 pers. minutes	Overlapping time, minutes	Per cent
Set-closing time (*Ts*)	131.00	1779.00	63.50	25.23
Operative time (*To*)	120.00	5023.00	36.00	71.25
Supervision of equipment time (*Tm*)	0.34	248.00	0.00	3.52
Total		7050.00		100.00

The time is divided into different categories of work in Table 2. For example the managers,/professionals in the sample spent 54.61 per cent of the total time on the processing of statistical reporting forms (the checks, writing analytical, explanatory notes), 43.34 per cent of the total cost set-up time - to prepare for the processing of forms (receipt, the union body of information, service print reports for follow-up, getting jobs, study guides, ordering information prior to the array of control) and actions related to its completion (for file transfer to the archive), 87.5 per cent of the total time spent observing the operation of equipment - to monitor the operation of the equipment (official seal records for transfer to the archive).

As follows from Tables 1-2, the differences in the structure of time-consuming to work with the forms of state statistical reporting and in the total working time categories are insignificant: if the overall proportion of staff time to set-up time was 25.23 per cent, in investment of time to work with the forms of statistical reporting, the same category of working time is 20.66 per cent, the share of operating time in the value of total working time is 71.25 per cent, the largest amount of time on form processing of statistical reports - 73.52 per cent, the share of time monitoring of performance of equipment in the magnitude of the total cost of working time is 3.52 per cent, and in the value of time spent on work with the forms of statistical reports 5.82 per cent.

The precision of these figures are presented in Table 3 based on the unrealistic assumption that the sample was a simple random sample. For example the

average operative time for a specialist in the sample was 334.87 minutes (5.58 hr.). The variation coefficient of cost of operative time was 31.64 per cent. The standard deviation of cost of operative time is 105.95 minutes and he average sampling error for the operative time costs of managers, specialists is 25.69 minutes. From this table it is also seen that the sampling error does not

Table 2. Grouping of the staff working time according to the photo survey of workers of the Minsk regional statistics department, minutes

Three divisions of the Minsk regional statistics department	Category of working time			Total
	Set-closing time (*Ts/f*)	Operative time (*To/f*)	Supervision of equipment time (*Tm/f*)	
Total time spent on work with statistical forms	771	2743	217	3731
The structure of time spent on work with statistical forms,%	20,66	73,52	5,82	100
Total working time	1779	5023	248	7050
The structure of working time,%	25,23	71,25	3,52	100
Percent of time to work with statistical forms	43,34	54,61	87,50	52,92
Percent of time spent on work not related directly to the processing of statistical forms	56,66	45,39	12,50	47,08

Table 3. The variation of working time of a simple random sample of workers of the Minsk regional statistics department

Category of working time	The total cost for a manager, a specialist in a sample, minutes	Coefficient of variation, per cent	The average sampling error, minutes	The limit sampling error, minutes	Standard deviation, minutes
Set-closing time (Ts)	334.87	31.64	25.69	54.72	105.95
Operative time (To)	118.60	67.98	19.55	41.64	80.62
Supervision of equipment time (Tm)	16.53	227.05	9.10	19.38	37.53

exceed 54.7197 minutes (25.69 × 2.13 = 54.72 minutes, where t=2.13) at a confidence level of 95 per cent [2]. The boundaries for the operative time in the general population are from 280 to 389 minutes. The other figures in the table are interpreted accordingly.

For the total sample the proportion of work not directly associated with an assignment is:

$$K1 = (Ts + Tm) / To = (118.6 + 16.53) / 334.87 = 0.404. \quad (4)$$

For the total population it is in the range of 0.2747 to 0.5035. Thus, for every 10 minutes of operative time cost a specialist of the Minsk statistics department spends from 2.80 to 4.98 minutes on set-up work and monitoring of the equipment.

Table 4. Results of the general population on the value of staff time

Index	Category of working time	The average for a sample (15 pers.), minutes	The boundaries of the average for the general population (127 pers.), minutes		The average for a sample (15 pers.), hour	The boundaries of the average for the general population (127 pers.), hour	
			Bottom	Top		Bottom	Top
A	B	1	2	3	4	5	6
Total working time	Operative time (To)	334.87	280.1503	389.5897	5.58	4.67	6.49
	Set-closing time (Ts)	118.6	76.9585	160.2415	1.98	1.28	2.67
	Supervision of equipment time (Tm)	16.53	0	35.913	0.28	0.00	0.60
	Total	**470**	**357.1088**	**585.7442**	**7.83**	**5.95**	**9.76**
Time spent on statistical forms processing	Operative time (To_f)	190.07	104.9126	275.2274	3.17	1.75	4.59
	Set-closing time (Ts_f)	65.13	28.0467	102.2133	1.09	0.47	1.70
	Supervision of equipment time (Tm_f)	14.47	0	33.9808	0.24	0.00	0.57
	Total	**255.2**	**132.9593**	**411.4215**	**4.25**	**2.22**	**6.86**

The same proportion for work connected with the input is $K2 = (Ts_f + Tm_f) / To_f = (65.13 + 14.47) / 190.07 = 0.42$. For the total population of managers(specialists the proportion lies in the range of 0.2673 to 0.4948. (Still assuming Simple random sampling)

For every 10 minutes of operative time costs connected with the state statistical reporting staff of the Minsk regional statistics department spends from

2.7 to 4.9 minutes on set-up work with forms and monitoring of the equipment in the processing of statistical reporting forms.

Thus, as the result of the survey the time-consuming structure of the general population of specialists of the Minsk regional statistical office has been determined using the methods of sampling; the resulting ratio of working time may be used for subsequent calculation of labor costs and labor regulation of the state statistics.

4. Conclusions

As a result of a simple random sample of workers of the Minsk regional statistics department the following conclusions can be obtained:

1. For every 10 minutes of operative time cost a specialist of the Minsk statistics department spends from 2.80 to 4.98 minutes on set-up work and monitoring of the equipment.
2. For every 10 minutes of operative time costs connected with the state statistical reporting staff of the Minsk regional statistics department spends from 2.7 to 4.9 minutes on set-up work with forms and monitoring of the equipment in the processing of statistical reporting forms.
3. The resulting ratio is necessary for the further isolation operative working time category from the total statistical working time for the subsequent adjustment for the coefficients of the complexity, value, occupancy and the amount of information and an objective assessment of the distribution of statistical work on time costs directly aimed at the implementation of the set tasks.

REFERENCES

1. Law of the Republic of Belarus "On State Statistics" dated 28.11.2004 / / NRPA Republic of Belarus. 2004. № 192.

2. Sharon L. Lohr. Sampling: Design and Analysis. 2010.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 309—330

ESTIMATION OF QUADRATIC FINITE POPULATION FUNCTIONS USING CALIBRATION

Dalius Pumputis[1], **Andrius Čiginas**[2]

ABSTRACT

Since the quadratic finite population functions can be expressed as totals over a synthetic population consisting of some ordered pairs of elements of the initial population, the traditional and penalized calibration technique is used to derive some calibrated estimators of the quadratic finite population functions. A linear combination of estimators discussed is considered as well. A comparison of approximate variances of the calibrated estimators is also presented. A simulation study is performed to analyze the empirical properties of the calibrated estimators of the finite population variance and covariance which appear as special cases of the quadratic functions. It is shown also how the calibrated estimators of the population covariance (variance) can be applied in regression estimation of the finite population total.

Key words: calibrated estimator; penalized calibration; auxiliary variables; approximate variance.

1. Introduction

In many statistical offices and official statistics, auxiliary information becomes more and more important at the estimation stage seeking to increase the accuracy of estimators of finite population parameters. To this end, the calibration approach is often used. The idea of the calibration technique for estimating the finite population totals is presented by Deville and Särndal (1992).

Since the population totals or means are the most popular parameters in survey practice, there exists a lot of scientific literature which deals with the estimation of these parameters using calibration methods. Some of them are discussed in the paper of Särndal (2007), where an overview of the calibration theory and its application in survey sampling are given.

[1] Lithuanian University of Educational Sciences, Vilnius, Lithuania.
E-mail: dalius.pumputis@vpu.lt
[2] Vilnius University, Vilnius, Lithuania. E-mail: andrius.ciginas@mif.vu.lt

The topic on the estimation of some quadratic finite population functions, such as the finite population variance, covariance or variance of the Horvitz-Thompson estimator (see e.g. Särndal, Swensson and Wretman, 1992, p. 43), is not often met in the literature of survey statistics. Plikusas and Pumputis (2007) introduced the calibrated estimators of the population covariance (variance), which use one weighting system defined by various calibration equations and distance measures. In the paper (Plikusas and Pumputis, 2010), the estimation of population covariance (variance) is considered using several systems of calibrated weights. The estimators, derived here, are applied to improve the regression (GREG) estimators of the finite population total. A more detailed description about that application is given in the Subsection 2.4.

Singh, Horn, Chowdhury and Yu (1999) proposed calibrated estimators of the variance of the Horvitz-Thompson estimator. Sitter and Wu (2002) extended the model calibration and pseudoempirical likelihood methods to obtain efficient estimators of quadratic finite population functions. Using a general expression of the new estimators, they also derived the corresponding model calibrated estimators of the population variance, the covariance and variance of the Horvitz-Thompson estimator, and analyzed their properties.

The structure of this paper is as follows. In the next section we derive some calibrated estimators of the quadratic population functions by employing Sitter and Wu's (2002) idea to express the quadratic population functions as the population totals and by applying Deville and Särndal's (1992) calibration method as well as the penalized calibration approach (Farrell and Singh, 2002). Subsection 2.3 provides a slightly different calibration which leads to a linear combination of the Horvitz-Thompson type estimator and calibrated estimators mentioned above. In Section 3 we first derive the approximate variances of the calibrated estimators and then we present a comparison of them. Some numerical results are presented in Section 4. Here we compare by simulation the calibrated estimators of the finite population variance and covariance which are both special cases of the quadratic functions. Section 5 is devoted to concluding remarks.

2. Estimators of quadratic functions

2.1. Deville and Särndal's calibration

Consider a finite population $\mathbf{U} = \{u_1, u_2, \ldots, u_N\}$ of N elements. Without loss of generality, we can assume $\mathbf{U} = \{1,2,\ldots,N\}$. Let $y^{(k)} : y_1^{(k)}, y_2^{(k)}, ..., y_N^{(k)}$, $k = 1,2,...,J$, be J study variables defined on the population $\mathbf{U}$ and taking fixed real values. The values of all variables are known only for sampled population elements. Denote $\mathbf{y}_i = (y_i^{(1)}, y_i^{(2)}, ..., y_i^{(J)})$.

We are interested in the estimation of the quadratic finite population function

$$T = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(\mathbf{y}_i, \mathbf{y}_j), \tag{1}$$

under a probability sampling design (of fixed size) with strictly positive second and fourth order inclusion probabilities. Here $\phi(\bullet,\bullet)$ is a symmetric function (a kernel of degree 2 for a *U*-statistic).

Well known finite population parameters, such as the finite population variance of $y^{(k)}$,

$$S^2_{y^{(k)}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(y_i^{(k)} - y_j^{(k)} \right)^2 ;$$

the finite population covariance between two study variables $y^{(k)}$ and $y^{(l)}$,

$$C\left(y^{(k)}, y^{(l)}\right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(y_i^{(k)} - y_j^{(k)} \right)\left(y_i^{(l)} - y_j^{(l)} \right);$$

and the variance,

$$V\left(\hat{t}_{HT,y^{(k)}}\right) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\pi_i \pi_j - \pi_{ij} \right)\left(y_i^{(k)} / \pi_i - y_j^{(k)} / \pi_j \right)^2 ,$$

of the Horvitz-Thompson estimator, $\hat{t}_{HT,y^{(k)}} = \sum_{i \in \mathbf{s}} d_i y_i^{(k)}$, of the population total $t_{y^{(k)}} = \sum_{i=1}^{N} y_i^{(k)}$, are as special cases of function *T*. Here π_i and π_{ij} are the first and second order inclusion probabilities, respectively; $\mathbf{s}$, $\mathbf{s} \subset \mathbf{U}$, denotes the probability sample set drawn from the population $\mathbf{U}$; $d_i = 1/\pi_i$ is the sample design weight of the element i, $i = 1,2,...,N$.

The presented alternative expressions of the finite population variance and covariance are useful in the context of our investigation. The variance $V\left(\hat{t}_{HT,y^{(k)}}\right)$ of the Horvitz-Thompson estimator $\hat{t}_{HT,y^{(k)}}$ is given in the Yates and Grundy (1953) form.

Let us arrange all the pairs $(ij), i < j$, of indexes of population elements in a sequence and number the elements of the sequence using $m = 1,2,...,N^*$, where $N^* = N(N-1)/2$ (For more details on the procedure see Sitter and Wu (2002)).

Then function *T* can be expressed in the following way:

$$T = \sum_{m=1}^{N^*} \phi_{ym} ,$$

where $\phi_{ym} = \phi(\mathbf{y}_i, \mathbf{y}_j)$ for a pair of indices $m = (ij)$. Now function T is viewed as a population total of the variable $\phi_y : \phi_{y1}, \phi_{y2}, ..., \phi_{yN^*}$, defined on a synthetic finite population $\mathbf{U}^* = \{1,2,\ldots, N^*\}$ of size N^*.

Thus, some calibration methods can be easily employed to derive the estimators of function T. But first, some elements of the sampling design in the population $\mathbf{U}^*$ should be defined. The sampling design in the population $\mathbf{U}^*$ is defined so that the corresponding sample of pairs is $\mathbf{s}^* = \{m = (ij) \,|\, i < j, i, j \in \mathbf{s}\}$ and it is treated as if it were drawn from population $\mathbf{U}^*$; the first order inclusion probabilities over the synthetic population $\mathbf{U}^*$ are coincident with the second order inclusion probabilities over the population $\mathbf{U}$: $\pi^*_m = \pi_{ij}$ for $m = (ij)$, where π_{ij} are assumed to be strictly positive. Then the sample design weights over the population $\mathbf{U}^*$ are equal to the inverse of second order inclusion probabilities: $d^*_m = 1/\pi^*_m = 1/\pi_{ij}$ for $m = (ij)$. Denote $d_{ij} = d^*_m$.

When sample design weights are defined and there is no auxiliary information, the quadratic finite population function (1) can be estimated using the Horvitz-Thompson type estimator:

$$\hat{T}_{HT} = \sum_{m \in \mathbf{s}^*} d^*_m \phi_{ym} = \sum_{i \in \mathbf{s}} \sum_{j > i} d_{ij} \phi(\mathbf{y}_i, \mathbf{y}_j). \tag{2}$$

As it is known, estimator (2) is unbiased but its variance is often relatively large.

The weights d_{ij} of the estimator $\hat{T}_{HT}$ can be modified using auxiliary variables and calibration methods to obtain estimators with a smaller variance. Let $x^{(k)}$ serve as an auxiliary variable for the study variable $y^{(k)}$, $k = 1,2,...,J$. Denote $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(J)})$. Assume also that the values of all auxiliary variables are known only for sampled population elements and that the total $T_{\phi_x} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(\mathbf{x}_i, \mathbf{x}_j)$ is known.

Remark 1. The simple summary statistics of the auxiliary variables (e.g. a total of $x^{(k)}$) are independent of the survey and may be taken from an outside source, such as national statistical institutes. The second-order summary statistics T_{ϕ_x} are much more complicated and they are not often considered in real surveys. Thus, the direct access to such a type of auxiliary information is not very realistic. A situation when all the values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ are known (referred to as complete

auxiliary information) is more realistic and useful in practice. The auxiliary variables may be taken from the previous complete surveys of the same population, various administrative registers and databases. Knowing these variables, one can easily calculate T_{ϕ_x} and use it for the construction of the calibrated estimators.

We consider here the calibrated estimators of the quadratic finite population functions of the following shape

$$\hat{T}_{cal} = \sum_{i\in\mathbf{s}} \sum_{j>i} \omega_{ij}^{(cal)} \phi(\mathbf{y}_i, \mathbf{y}_j), \tag{3}$$

where new (calibrated) weights $\omega_{ij}^{(cal)}$ are defined under the following conditions:

The weights $\omega_{ij}^{(cal)}$ satisfy some calibration equation;

The distance between the weights d_{ij} and $\omega_{ij}^{(cal)}$ is minimal according to some distance measures.

First, by applying Deville and Särndal's (1992) calibration technique, we define the calibrated estimator of quadratic function *T*:

$$\hat{T}_{DS} = \sum_{i\in\mathbf{s}} \sum_{j>i} \omega_{ij}^{(DS)} \phi(\mathbf{y}_i, \mathbf{y}_j), \tag{4}$$

where the weights $\omega_{ij}^{(DS)}$ minimize the distance measure

$$L(\omega, d) = \sum_{i\in\mathbf{s}} \sum_{j>i} \frac{\left(\omega_{ij}^{(DS)} - d_{ij}\right)^2}{d_{ij} q_{ij}} \tag{5}$$

and satisfy the calibration equation

$$\sum_{i\in\mathbf{s}} \sum_{j>i} \omega_{ij}^{(DS)} \phi(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(\mathbf{x}_i, \mathbf{x}_j) = T_{\phi_x}. \tag{6}$$

Here $q_{ij}, i, j \in \mathbf{s}, i < j$, are free additional weights. The estimators can be modified by choosing q_{ij}.

Calibration equation (6) shows that the known quadratic function T_{ϕ_x} is estimated by $\hat{T}_{DS,\phi_x} = \sum_{i\in\mathbf{s}} \sum_{j>i} \omega_{ij}^{(DS)} \phi(\mathbf{x}_i, \mathbf{x}_j)$ without error. In the case of a quite high correlation between the variables ϕ_y and ϕ_x (where ϕ_x is defined similarly

as ϕ_y), it is natural to expect that the estimates of the function T are more accurate when the new weights $\omega_{ij}^{(DS)}$ are applied in (4).

The weights $\omega_{ij}^{(DS)}$ of estimator (4) are given by the following proposition that is actually a corollary which follows from the derivation of weights of a calibrated estimator of the finite population total (see Deville and Särndal, 1992).

Proposition 1. The weights $\omega_{ij}^{(DS)}$, $i, j \in \mathbf{s}$, $i < j$, which minimize the distance measure (5) and satisfy equation (6), are defined by the equations:

$$\omega_{ij}^{(DS)} = d_{ij}\left(1 + \frac{q_{ij}\phi(\mathbf{x}_i, \mathbf{x}_j)}{\sum\limits_{u\in\mathbf{s}}\sum\limits_{v>u} d_{uv}q_{uv}\phi^2(\mathbf{x}_u, \mathbf{x}_v)}\left(T_{\phi_x} - \sum_{u\in\mathbf{s}}\sum_{v>u} d_{uv}\phi(\mathbf{x}_u, \mathbf{x}_v)\right)\right).$$

A number of other calibrated estimators may be derived using different distance measures and calibration equations. In the following part of this paper, we will analyze some cases.

Remark 2. By replacing the values $\phi(\mathbf{y}_i, \mathbf{y}_j)$ and $\phi(\mathbf{x}_i, \mathbf{x}_j)$ in the expression of the estimator $\hat{T}_{DS}$ with $(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i^{(k)}}{\pi_i} - \frac{y_j^{(k)}}{\pi_j}\right)^2$ and $(\pi_i\pi_j - \pi_{ij})\left(\frac{x_i^{(k)}}{\pi_i} - \frac{x_j^{(k)}}{\pi_j}\right)^2$, we obtain an estimator $\hat{V}_{DS}\left(\hat{t}_{HT,y^{(k)}}\right)$ of the variance of the Horvitz-Thompson estimator $\hat{t}_{HT,y^{(k)}} = \sum\limits_{i\in\mathbf{s}} d_i y_i^{(k)}$. Assume $\pi_i\pi_j - \pi_{ij} > 0$ and let $q_{ij} = Q_{ij}/\left(\pi_i\pi_j - \pi_{ij}\right)$, where Q_{ij} are free additional constants. Then the estimator $\hat{V}_{DS}\left(\hat{t}_{HT,y^{(k)}}\right)$ reduces to that considered by Singh, Horn, Chowdhury and Yu (1999).

2.2. Penalized calibration estimators

Let us consider the estimator of quadratic finite population function T of the same form (3) and define the weights $\omega_{ij}^{(FS)}$, $i, j \in \mathbf{s}$, $i < j$, of it, using the same calibration equation (6), but a different distance measure

$$L_P(w,d) = \sum_{i\in\mathbf{s}}\sum_{j>i}\frac{\left(\omega_{ij}^{(FS)} - d_{ij}\right)^2}{d_{ij}q_{ij}} + \varphi^2\sum_{i\in\mathbf{s}}\sum_{j>i}\frac{\left(\omega_{ij}^{(FS)}\right)^2}{d_{ij}q_{ij}} \qquad (7)$$

the analog of which is proposed in the papers of Farrell and Singh (2002) and Singh (2003), and called a penalized one. Minimization of this distance measure subject to calibration equation (6) leads to the estimator with interesting features that can be described by the words of Farrell and Singh (2002, p. 965): "… φ is a positive quantity that reflects a penalty to be decided by the investigator based on prior knowledge, or the desire for certain levels of efficiency and bias…increasing φ results in a decrease in the mean square error of the estimator; unfortunately has the side effect of increasing the bias".

Denote by $\hat{T}_{FS}$ the new, just defined estimator. Since the function L_p is coincident with L as $\varphi = 0$, the new group of estimators

$$\hat{T}_{FS} = \sum_{i \in \mathbf{s}} \sum_{j>i} \omega_{ij}^{(FS)} \phi(\mathbf{y}_i, \mathbf{y}_j) \tag{8}$$

includes the calibrated estimators $\hat{T}_{DS}$.

Proposition 2. The weights $\omega_{ij}^{(FS)}, i, j \in \mathbf{s}, i < j,$ which minimize the distance measure (7) and satisfy the calibration equation (6), are defined by the equations:

$$\omega_{ij}^{(FS)} = \frac{d_{ij}}{1+\varphi^2}\left(1 + \frac{q_{ij}\phi(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{u \in \mathbf{s}} \sum_{v>u} d_{uv} q_{uv} \phi^2(\mathbf{x}_u, \mathbf{x}_v)}\left(\left(1+\varphi^2\right)T_{\phi_x} - \sum_{u \in \mathbf{s}} \sum_{v>u} d_{uv}\phi(\mathbf{x}_u, \mathbf{x}_v)\right)\right)$$

.

Proof. Let us take the distance measure (7) and calibration equation (6), and define the Lagrange function

$$\Lambda = \sum_{i \in \mathbf{s}} \sum_{j>i} \frac{\left(\omega_{ij}^{(FS)} - d_{ij}\right)^2}{d_{ij} q_{ij}} + \varphi^2 \sum_{i \in \mathbf{s}} \sum_{j>i} \frac{\left(\omega_{ij}^{(FS)}\right)^2}{d_{ij} q_{ij}} - \lambda\left(\sum_{i \in \mathbf{s}} \sum_{j>i} \omega_{ij}^{(DS)} \phi(\mathbf{x}_i, \mathbf{x}_j) - T_{\phi_x}\right).$$

By solving the equations

$$\frac{\partial \Lambda}{\partial \omega_{ij}^{(FS)}} = 0, \; i, j \in \mathbf{s}, \; i < j,$$

we get

$$\omega_{ij}^{(FS)} = \frac{1}{1+\varphi^2} d_{ij}\left(1 + \frac{1}{2}\lambda q_{ij}\phi(\mathbf{x}_i, \mathbf{x}_j)\right). \tag{9}$$

Then, multiplying (9) by $\phi(\mathbf{x}_i, \mathbf{x}_j)$, summing over the sample $\mathbf{s}^*$ elements and taking into account calibration equation (6), we get an expression for λ. Substituting this expression into (9), we get an equation for $\omega_{ij}^{(FS)}$.

One can note that penalized calibration is usually used to penalize the magnitude of the calibrated weights when a lot of calibration constraints are used and the sample is particularly unbalanced so that negative or very large weights occur after the calibration procedure (see Guggemos and Tillé, 2010). In this paper, we consider penalized calibration in the case of only one calibration equation, because we are seeking only to find out if the penalized distance measure may be more advantageous than function (5) when the resulting estimators are derived using the same calibration equation.

As it is shown below (see Subsections 3.2 and 4.2), the penalized estimator $\hat{T}_{FS}$ has the lower approximate and empirical variances as compared to that of the calibrated estimator $\hat{T}_{DS}$, but according to the results of Farrell and Singh (2002), the bias of $\hat{T}_{FS}$ becomes relatively large when the parameter φ is increasing. This is not a desirable property that could inspire for the development of an improved penalized estimator.

2.3. Linear combination of estimators

We consider here a slightly different calibration when the weights of estimator (3) are derived by calibrating the original design weights d_{ij}, multiplied by some correction factor. The estimator under consideration is

$$\hat{T}_{lin} = \sum_{i\in\mathbf{s}} \sum_{j>i} \omega_{ij}^{(lin)} \phi(\mathbf{y}_i, \mathbf{y}_j), \tag{10}$$

where the weights $\omega_{ij}^{(lin)}$ minimize the distance measure

$$\begin{aligned} L_{new}(\omega, \tilde{d}) &= \sum_{i\in\mathbf{s}} \sum_{j>i} \frac{\left(\omega_{ij}^{(lin)} - \tilde{d}_{ij}\right)^2}{d_{ij} q_{ij}}, \\ \tilde{d}_{ij} &= c d_{ij},\ c = \alpha + \beta + \gamma / (1+\varphi^2),\ \alpha + \beta + \gamma = 1, \end{aligned} \tag{11}$$

and satisfy the new calibration equation

$$\sum_{i\in\mathbf{s}} \sum_{j>i} \omega_{ij}^{(lin)} \phi(\mathbf{x}_i, \mathbf{x}_j) = (1-\alpha) T_{\phi_x} + \alpha \sum_{i\in\mathbf{s}} \sum_{j>i} d_{ij} \phi(\mathbf{x}_i, \mathbf{x}_j). \tag{12}$$

Note that the right side of the calibration equation (12) consists of two terms: the first one is the true value of T_{ϕ_x} multiplied by $1-\alpha$, and the second one – the estimate of T_{ϕ_x} multiplied by the coefficient α.

Proposition 3. Minimization of the distance measure (11) subject to the calibration equation (12) leads to the calibrated weights given by

$$\omega_{ij}^{(lin)} = d_{ij}\left(c + \frac{q_{ij}\phi(\mathbf{x}_i,\mathbf{x}_j)\left((1-\alpha)T_{\phi_x} + (\alpha - c)\sum_{u\in\mathbf{s}}\sum_{v>u} d_{uv}\phi(\mathbf{x}_u,\mathbf{x}_v)\right)}{\sum_{u\in\mathbf{s}}\sum_{v>u} d_{uv}q_{uv}\phi^2(\mathbf{x}_u,\mathbf{x}_v)} \right). \quad (13)$$

The proof is similar to that of Proposition 2.

By inserting the weights (13) into (10), we get the estimator

$$\hat{T}_{lin} = \alpha\hat{T}_{HT} + \beta\hat{T}_{DS} + \gamma\hat{T}_{FS},\ \alpha+\beta+\gamma = 1, \quad (14)$$

which is a linear combination of $\hat{T}_{HT}$, $\hat{T}_{DS}$ and $\hat{T}_{FS}$.

In expression (14) one can see that α can be interpreted as a weight of the Horvitz-Thompson estimate which is included into the expression of $\hat{T}_{lin}$. Therefore, the absolute value of α reflects the rate of an influence of the Horvitz-Thompson estimator on the accuracy of the estimator $\hat{T}_{lin}$. A similar discussion can be provided about the coefficients β and γ. As an example of a high influence of the Horvitz-Thompson estimator can be obtained by choosing a value close to one for α and the values close to zero for the coefficients β and γ ($\alpha+\beta+\gamma=1$). Then the estimator $\hat{T}_{lin}$ is almost unbiased with a variance similar to that of the Horvitz-Thompson estimator. The variance of $\hat{T}_{lin}$ can be reduced by choosing a value of α close to zero, but then the estimator $\hat{T}_{lin}$ may be more biased.

Thus, the statistical properties of $\hat{T}_{lin}$ can be controlled through the values of coefficients α, β and γ. Consequently, the (optimal) values of α, β and γ, which minimize the mean square error of the estimator $\hat{T}_{lin}$ subject to an unbiasedness constraint, are more preferable than any set of α, β and γ.

2.4. Some aspects from a practical perspective

The main purpose of this subsection is to present some possibilities for the practical applications of the calibrated estimators of some quadratic functions, such as the finite population variance and covariance.

Note that according to the formulation of our problem, there is only one auxiliary variable available when the estimated parameter is a finite population variance and two auxiliaries are used in the case of estimation of the finite population covariance. Further, for simplicity, we denote the study and auxiliary variables, corresponding to the cases of estimation of variance and covariance, by y and a, and by y, z and a, b.

By replacing the values $\phi(\mathbf{y}_i, \mathbf{y}_j)$ and $\phi(\mathbf{x}_i, \mathbf{x}_j)$ in the expressions of the estimators $\hat{T}_{HT}, \hat{T}_{DS}, \hat{T}_{FS}$ and $\hat{T}_{lin}$ with $\frac{1}{N(N-1)}(y_i - y_j)^2$ and $\frac{1}{N(N-1)}(a_i - a_j)^2$, we get four estimators of the finite population variance S_y^2. We denote them by $\hat{S}_{HT}^2$, $\hat{S}_{DS}^2$, $\hat{S}_{FS}^2$, and $\hat{S}_{lin}^2$, respectively. Analogously, the substitution $\frac{1}{N(N-1)}(y_i - y_j)(z_i - z_j)$ and $\frac{1}{N(N-1)}(a_i - a_j)(b_i - b_j)$ leads to the four estimators of the finite population covariance $C(y,z)$: $\hat{C}_{HT}, \hat{C}_{DS}, \hat{C}_{FS}$ and $\hat{C}_{lin}$.

The estimators derived can be useful in the following common situation. Let us say, we want to estimate a population total

$$t_y = \sum_{k=1}^{N} y_k .$$

In the case of only one known auxiliary variable, say a, one can take the simple regression estimator (see e.g. Särndal, Swensson and Wretman, 1992, p. 272)

$$\hat{t}_{yr} = \frac{N}{\hat{N}} \sum_{k \in \mathbf{s}} d_k y_k + \hat{B}\left(\sum_{k=1}^{N} a_k - \frac{N}{\hat{N}} \sum_{k \in \mathbf{s}} d_k a_k \right), \tag{15}$$

where $\hat{N} = \sum_{k \in \mathbf{s}} d_k$, $\hat{B} = \dfrac{\sum_{k \in \mathbf{s}} d_k \left(a_k - \sum_{l \in \mathbf{s}} d_l a_l \Big/ \hat{N} \right) \left(y_k - \sum_{l \in \mathbf{s}} d_l y_l \Big/ \hat{N} \right)}{\sum_{k \in \mathbf{s}} d_k \left(a_k - \sum_{l \in \mathbf{s}} d_l a_l \Big/ \hat{N} \right)^2}$.

If the variables y and a are well correlated, then estimator (15) is much more accurate as compared to the Horvitz-Thompson estimator. For the sample designs for which $\hat{N} = N$, regression estimator (15) reduces to

$$\hat{t}_{yr} = \sum_{k \in \mathbf{s}} d_k y_k + \frac{\hat{C}(y,a)}{\hat{S}_a^2}\left(\sum_{k=1}^{N} a_k - \sum_{k \in \mathbf{s}} d_k a_k \right), \tag{16}$$

where $\hat{C}(y,a) = \frac{1}{N-1}\sum_{k \in \mathbf{s}} d_k \left(a_k - \sum_{l \in \mathbf{s}} d_l a_l \Big/ N \right)\left(y_k - \sum_{l \in \mathbf{s}} d_l y_l \Big/ N \right)$ and $\hat{S}_a^2 = \frac{1}{N-1}\sum_{k \in \mathbf{s}} d_k \left(a_k - \sum_{l \in \mathbf{s}} d_l a_l \Big/ N \right)^2$ are standard only design based estimators of the population covariance $C(y,a)$ and variance S_a^2, respectively. As it is shown in (Plikusas and Pumputis, 2010), regression estimator (16) can be improved by replacing the standard estimators $\hat{C}(y,a)$ and $\hat{S}_a^2$ with more accurate ones. Thus the calibrated estimators $\hat{C}_{DS}$, $\hat{C}_{FS}$, $\hat{C}_{lin}$ and $\hat{S}_{DS}^2$, $\hat{S}_{FS}^2$, $\hat{S}_{lin}^2$ may be suitable for this purpose assuming that there are available two additional known variables x_y and x_a which serve as the auxiliaries for the variables y and a, respectively.

Beside that application of the calibrated estimators of the finite population variance and covariance, they can be used also to improve the estimates of other finite population parameters, such as a finite population correlation coefficient

$$\rho(y,z) = C(y,z) \Big/ \left(S_y \cdot S_z \right)$$

which is a ratio of the covariance $C(y,z)$ and product of the standard deviations $S_y = \sqrt{S_y^2}$ and $S_z = \sqrt{S_z^2}$. The simplest way to estimate the correlation coefficient $\rho(y,z)$ is to use the Horvitz–Thompson type estimators $\hat{C}_{HT}$, $\hat{S}_{HT,y}^2$ and $\hat{S}_{HT,z}^2$ for estimating covariance $C(y,z)$ and variances S_y^2, S_z^2, respectively, and to take the ratio

$$\hat{\rho}(y,z) = \hat{C}_{HT} \Big/ \sqrt{\hat{S}_{HT,y}^2 \cdot \hat{S}_{HT,z}^2} \tag{17}$$

as the estimator of the correlation coefficient. More accurate estimates may be obtained using in (17) the calibrated estimators instead of the corresponding Horvitz-Thompson estimators of the covariance $C(y,z)$ and variances S_y^2 and S_z^2.

3. Comparison of estimators

3.1. Approximate variances

For practical and theoretical purposes, it is good to have expressions of the exact or approximate variances of estimators, or even more, to know which estimator has the lowest variance. Since the estimators $\hat{T}_{DS}, \hat{T}_{FS}$ and $\hat{T}_{lin}$ are nonlinear functions of the Horvitz-Thompson estimators

$$\hat{T}_{HT} = \sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}\phi(\mathbf{y}_i,\mathbf{y}_j),\ \hat{T}_{HT,\phi_x} = \sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}\phi(\mathbf{x}_i,\mathbf{x}_j),$$

$$\hat{T}_{HT,q\phi_x^2} = \sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}q_{ij}\phi^2(\mathbf{x}_i,\mathbf{x}_j),\ \hat{T}_{HT,q\phi_x\phi_y} = \sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}q_{ij}\phi(\mathbf{x}_i,\mathbf{x}_j)\phi(\mathbf{y}_i,\mathbf{y}_j),$$

of the population totals

$$T,\ T_{\phi_x},\ T_{q\phi_x^2} = \sum_{i=1}^{N}\sum_{j=i+1}^{N} q_{ij}\phi^2(\mathbf{x}_i,\mathbf{x}_j),\ T_{q\phi_x\phi_y} = \sum_{i=1}^{N}\sum_{j=i+1}^{N} q_{ij}\phi(\mathbf{x}_i,\mathbf{x}_j)\phi(\mathbf{y}_i,\mathbf{y}_j),$$

respectively, we will use the Taylor linearization technique to derive expressions of the approximate variances.

According to the Result 6.6.1 of (Särndal, Swensson and Wretman, 1992, p. 235), the approximate variance of $\hat{T}_{DS}$ can be written as

$$AV\left(\hat{T}_{DS}\right) = V\left(\hat{T}_{HT} - B\hat{T}_{HT,\phi_x}\right), \tag{18}$$

where $B = T_{q\phi_x\phi_y} \big/ T_{q\phi_x^2}$.

Proposition 4. The approximate variances of calibrated estimators $\hat{T}_{FS}$ and $\hat{T}_{lin}$ can be expressed as follows

$$AV\left(\hat{T}_{FS}\right) = \frac{1}{\left(1+\varphi^2\right)^2} AV\left(\hat{T}_{DS}\right), \tag{19}$$

$$AV\left(\hat{T}_{lin}\right) = \left(\frac{1-\alpha+\beta\varphi^2}{1+\varphi^2}\right)^2 AV\left(\hat{T}_{DS}\right) + \alpha^2 V\left(\hat{T}_{HT}\right) + 2\alpha\frac{1-\alpha+\beta\varphi^2}{1+\varphi^2} C\left(\hat{T}_{HT} - B\hat{T}_{HT,\phi_x}, \hat{T}_{HT}\right).$$

Proof. By substituting the weights $\omega_{ij}^{(FS)}$ into (8), we obtain

$$\hat{T}_{FS} = \hat{T}_{HT} \big/ \left(1+\varphi^2\right) + \left(T_{\phi_x} - \hat{T}_{HT,\phi_x} \big/ \left(1+\varphi^2\right)\right) \cdot \hat{T}^{-1}_{HT,q\phi_x^2} \cdot \hat{T}_{HT,q\phi_x\phi_y}$$
$$= f\left(\hat{T}_{HT}, \hat{T}_{HT,\phi_x}, \hat{T}_{HT,q\phi_x^2}, \hat{T}_{HT,q\phi_x\phi_y}\right).$$

Thus, the estimator $\hat{T}_{FS}$ can be viewed as a nonlinear function depending on the Horvitz-Thompson estimators $\hat{T}_{HT}, \hat{T}_{HT,\phi_x}, \hat{T}_{HT,q\phi_x^2}$ and $\hat{T}_{HT,q\phi_x\phi_y}$ that are unbiased, i.e.

$$E\hat{T}_{HT} = T, \;\; E\hat{T}_{HT,\phi_x} = T_{\phi_x}, \; E\hat{T}_{HT,q\phi_x^2} = T_{q\phi_x^2}, \; E\hat{T}_{HT,q\phi_x\phi_y} = T_{q\phi_x\phi_y} \;\; .$$

Using the Taylor linearization method, we derive a linear approximation of the function $\hat{T}_{FS}$.

The expansion of this estimator in a Taylor series up to the first order terms at the point $\left(\hat{T}_{HT}, \hat{T}_{HT,\phi_x}, \hat{T}_{HT,q\phi_x^2}, \hat{T}_{HT,q\phi_x\phi_y}\right) = \left(T, (1+\varphi^2)T_{\phi_x}, T_{q\phi_x^2}, T_{q\phi_x\phi_y}\right)$ is

$$\hat{T}_{FS}^{(linear)} = \frac{1}{1+\varphi^2}\left[\hat{T}_{HT} - B\left(\hat{T}_{HT,\phi_x} - (1+\varphi^2)T_{\phi_x}\right)\right]. \tag{20}$$

The approximate variance of the estimator $\hat{T}_{FS}$ is equal to the exact variance of $\hat{T}_{FS}^{(linear)}$:

$$AV\left(\hat{T}_{FS}\right) = V\left(\hat{T}_{FS}^{(linear)}\right).$$

By calculating the variance $V\left(\hat{T}_{FS}^{(linear)}\right)$ and taking into account equality (18), we get the expression of $AV\left(\hat{T}_{FS}\right)$.

The approximate variance of the estimator $\hat{T}_{lin}$ is obtained similarly, using an expansion of $\hat{T}_{lin}$ in a Taylor series at the point

$$\left(\hat{T}_{HT}, \hat{T}_{HT,\phi_x}, \hat{T}_{HT,q\phi_x^2}, \hat{T}_{HT,q\phi_x\phi_y}\right) = \left(T, \frac{(1-\alpha)(1+\varphi^2)}{1-\alpha+\beta\varphi^2}T_{\phi_x}, T_{q\phi_x^2}, T_{q\phi_x\phi_y}\right). \qquad \square$$

Remark 3. The estimator of the variance of the estimator $\hat{T}_{DS}$ can be defined as follows. First, we write

$$AV\left(\hat{T}_{DS}\right) = V\left(\sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}\left(\phi(\mathbf{y}_i, \mathbf{y}_j) - B\phi(\mathbf{x}_i, \mathbf{x}_j)\right)\right) = V\left(\hat{T}_{HT,\phi_y - B\phi_x}\right),$$

where $\hat{T}_{HT,\phi_y - B\phi_x}$ is the Horvitz-Thompson estimator of the total of the variable $\phi_y - B\phi_x$ defined on the population $\mathbf{U}^*$ and taking values $\phi_{y1} - B\phi_{x1}$, $\phi_{y2} - B\phi_{x2}$, ..., $\phi_{yN^*} - B\phi_{xN^*}$. Here $\phi_{ym} = \phi(\mathbf{y}_i, \mathbf{y}_j)$, $\phi_{xm} = \phi(\mathbf{x}_i, \mathbf{x}_j)$ for a pair of indices $m = (ij)$. Then, using Result 2.8.1 from (Särndal, Swensson and Wretman, 1992, p. 43), we get the more convenient expression of the approximate variance

$$AV\left(\hat{T}_{DS}\right) = V\left(\hat{T}_{HT,\phi_y - B\phi_x}\right) = \sum_{r=1}^{N^*}\sum_{m=1}^{N^*}\left(\pi^*_{rm} - \pi^*_r\pi^*_m\right)\frac{\phi_{yr} - B\phi_{yr}}{\pi^*_r}\cdot\frac{\phi_{ym} - B\phi_{ym}}{\pi^*_m},$$

where π^*_m and π^*_{rm} are the first and second order inclusion probabilities over the synthetic population $\mathbf{U}^*$. In fact, π^*_m and π^*_{rm} coincide with the second and fourth order inclusion probabilities over the population $\mathbf{U}$: $\pi^*_m = \pi_{ij}$ and $\pi^*_{rm} = \pi_{ijkl}$ for $m = (ij)$ and $r = (kl)$, where π_{ij} and π_{ijkl} are assumed to be strictly positive (see Subsection 2.1).

As the estimator of the variances $V\left(\hat{T}_{DS}\right)$ and $V\left(\hat{T}_{HT,\phi_y - B\phi_x}\right)$, we take

$$\hat{V}\left(\hat{T}_{DS}\right) = \hat{V}\left(\hat{T}_{HT,\phi_y - B\phi_x}\right) = \sum_{r\in\mathbf{s}^*}\sum_{m\in\mathbf{s}^*}\left(1 - \frac{\pi^*_r\pi^*_m}{\pi^*_{rm}}\right)\frac{\phi_{yr} - \hat{B}\phi_{yr}}{\pi^*_r}\cdot\frac{\phi_{ym} - \hat{B}\phi_{ym}}{\pi^*_m},$$

where $\hat{B} = \hat{T}_{HT,q\phi_x\phi_y} \Big/ \hat{T}_{HT,q\phi_x^2}$.

The estimators of the variances $V\left(\hat{T}_{FS}\right)$ and $V\left(\hat{T}_{lin}\right)$ can be derived in a similar way.

Alternatively, the replication methods, such as the jackknife, bootstrap and balanced halfsamples (see e.g. Särndal, Swensson and Wretman, 1992), can be used for estimating the variances of the estimators $\hat{T}_{DS}$, $\hat{T}_{FS}$ and $\hat{T}_{lin}$.

3.2. Comparison of variances

Next, we will compare the variances of estimators by analyzing their differences. Let us start from the difference $AV\left(\hat{T}_{DS}\right) - V\left(\hat{T}_{HT}\right)$. Using equality (18), it can be easily expressed as follows

$$AV\left(\hat{T}_{DS}\right) - V\left(\hat{T}_{HT}\right) = B^2V\left(\hat{T}_{HT,\phi_x}\right) - 2BC\left(\hat{T}_{HT}, \hat{T}_{HT,\phi_x}\right).$$

Thus, $AV\left(\hat{T}_{DS}\right) \le V\left(\hat{T}_{HT}\right)$, if

$$\begin{aligned} \rho\left(\hat{T}_{HT}, \hat{T}_{HT,\phi_x}\right) &\ge \frac{1}{2} B \sqrt{V\left(\hat{T}_{HT,\phi_x}\right) \Big/ V\left(\hat{T}_{HT}\right)}, \text{ as } B > 0, \text{ or} \\ \rho\left(\hat{T}_{HT}, \hat{T}_{HT,\phi_x}\right) &\le \frac{1}{2} B \sqrt{V\left(\hat{T}_{HT,\phi_x}\right) \Big/ V\left(\hat{T}_{HT}\right)}, \text{ as } B < 0, \end{aligned} \tag{21}$$

where $\rho\left(\hat{T}_{HT}, \hat{T}_{HT,\phi_x}\right)$ is the correlation coefficient between $\hat{T}_{HT}$ and $\hat{T}_{HT,\phi_x}$.

A negative value of $B = T_{q\phi_x\phi_y} \Big/ T_{q\phi_x^2}$ may occur if the range of the function $\phi(\bullet,\bullet)$ is **R**, e.g. when T is a covariance between two study variables.

In the case of simple random sampling without replacement, inequalities (21) reduce to

$$\begin{aligned} \rho\left(\phi_y, \phi_x\right) &\ge \frac{1}{2} B \sqrt{S_{\phi_x}^2 \Big/ S_{\phi_y}^2}, \text{ as } B > 0, \\ \rho\left(\phi_y, \phi_x\right) &\le \frac{1}{2} B \sqrt{S_{\phi_x}^2 \Big/ S_{\phi_y}^2}, \text{ as } B < 0, \end{aligned}$$

where $\rho\left(\phi_y, \phi_x\right)$ is the correlation coefficient between the variables $\phi_y : \phi_{y1}, \phi_{y2}, ..., \phi_{yN^*}$ and $\phi_x : \phi_{x1}, \phi_{x2}, ..., \phi_{xN^*}$, defined on the population $\mathbf{U}^*$. The notation $S_{\phi_y}^2$, $S_{\phi_x}^2$ is used to denote variances of the variables ϕ_y and ϕ_x, respectively.

Comparison of calibrated and penalized calibration estimators. Equality (19) shows that the approximate variance of the penalized estimator $\hat{T}_{FS}$ is lower than that of the calibrated estimator $\hat{T}_{DS}$. Even more, according to the lines of Farrell and Singh (2002), the approximate mean square of the estimator $\hat{T}_{FS}$ is minimized when

$$\varphi^2 = AV\left(\hat{T}_{DS}\right) \Big/ \left(T - BT_{\phi_x}\right)^2, \tag{22}$$

and it equals

$$AMSE_{\min}\left(\hat{T}_{FS}\right) = \frac{1}{1+\varphi^2} AV\left(\hat{T}_{DS}\right).$$

One can easily verify this equality using the same expansion (20) of $\hat{T}_{FS}$ in a Taylor series.

Comparison of the calibrated estimators $\hat{T}_{DS}$ and $\hat{T}_{lin}$. First, we derive the optimal values of coefficients α and β which minimize the approximate

variance of the estimator $\hat{T}_{lin}$ under the condition $AE\left(\hat{T}_{lin}\right)=T$ that indicates the approximate unbiasedness of the estimator. Denote the optimal values of α and β by $\alpha_{\min}$ and $\beta_{\min}$, respectively.

Let us define the Lagrange function

$$\Lambda^{*}=AV\left(\hat{T}_{lin}\right)-\lambda^{*}\left(AE\left(\hat{T}_{lin}\right)-T\right).$$

By solving the equations

$$\frac{\partial\Lambda^{*}}{\partial\alpha}=0,\ \frac{\partial\Lambda^{*}}{\partial\beta}=0 \text{ and } AE\left(\hat{T}_{lin}\right)=T,$$

we find

$$\begin{aligned}\alpha_{\min}&=\frac{AV\left(\hat{T}_{DS}\right)-C\left(\hat{T}_{HT}-B\hat{T}_{HT,\phi_x},\hat{T}_{HT}\right)}{B^2V\left(\hat{T}_{HT,\phi_x}\right)},\\ \beta_{\min}&=\frac{V\left(\hat{T}_{HT}\right)-C\left(\hat{T}_{HT}-B\hat{T}_{HT,\phi_x},\hat{T}_{HT}\right)}{B^2V\left(\hat{T}_{HT,\phi_x}\right)}.\end{aligned}\tag{23}$$

The second derivatives test for critical points shows that the values $\alpha_{\min}$ and $\beta_{\min}$ satisfy the condition of the minimal approximate variance which is equal to

$$\begin{aligned}AV_{\min}\left(\hat{T}_{lin}\right)&=\frac{AV\left(\hat{T}_{DS}\right)V\left(\hat{T}_{HT}\right)-C^2\left(\hat{T}_{HT}-B\hat{T}_{HT,\phi_x},\hat{T}_{HT}\right)}{AV\left(\hat{T}_{DS}\right)+V\left(\hat{T}_{HT}\right)-2C\left(\hat{T}_{HT}-B\hat{T}_{HT,\phi_x},\hat{T}_{HT}\right)}\\&=\frac{AV\left(\hat{T}_{DS}\right)V\left(\hat{T}_{HT}\right)\left(1-\rho^2\left(\hat{T}_{HT}-B\hat{T}_{HT,\phi_x},\hat{T}_{HT}\right)\right)}{B^2V\left(\hat{T}_{HT,\phi_x}\right)},\end{aligned}$$

where $\rho\left(\hat{T}_{HT}-B\hat{T}_{HT,\phi_x},\hat{T}_{HT}\right)$ is the correlation coefficient between the estimators $\hat{T}_{HT,\phi_y-B\phi_x}=\hat{T}_{HT}-B\hat{T}_{HT,\phi_x}$ and $\hat{T}_{HT}$.

The difference

$$AV_{\min}\left(\hat{T}_{lin}\right)-AV\left(\hat{T}_{DS}\right)=-\frac{\left(AV\left(\hat{T}_{DS}\right)-C\left(\hat{T}_{HT}-B\hat{T}_{HT,\phi_x},\hat{T}_{HT}\right)\right)^2}{B^2V\left(\hat{T}_{HT,\phi_x}\right)}$$

is negative or is equal to zero. It means that the approximate variance of the estimator $\hat{T}_{lin}$ is not higher than that of the calibrated estimator $\hat{T}_{DS}$.

Note that $\alpha_{\min} + \beta_{\min} = 1$. Therefore, under such a setting of optimal coefficients, the estimator $\hat{T}_{FS}$ is not included into the linear combination (14), and, consequently, $\hat{T}_{lin} = \alpha_{\min}\hat{T}_{HT} + \beta_{\min}\hat{T}_{DS}$.

Comparison of the calibrated estimators $\hat{T}_{FS}$ and $\hat{T}_{lin}$. The difference between the approximate variances $AV_{\min}\left(\hat{T}_{lin}\right)$ and $AV\left(\hat{T}_{FS}\right)$ can be expressed as follows:

$$AV_{\min}\left(\hat{T}_{lin}\right) - AV\left(\hat{T}_{FS}\right) = \frac{AV\left(\hat{T}_{DS}\right)V\left(\hat{T}_{HT}\right)}{-\left(AV\left(\hat{T}_{DS}\right) + V\left(\hat{T}_{HT}\right) - 2C\left(\hat{T}_{HT} - B\hat{T}_{HT,\phi_x}, \hat{T}_{HT}\right)\right)\left(1+\varphi^2\right)^2}$$

$$\times\left[\rho^2\left(\hat{T}_{HT} - B\hat{T}_{HT,\phi_x}, \hat{T}_{HT}\right)\left(1+\varphi^2\right)^2 + \frac{\left(AV\left(\hat{T}_{DS}\right) - 2C\left(\hat{T}_{HT} - B\hat{T}_{HT,\phi_x}, \hat{T}_{HT}\right)\right)}{V\left(\hat{T}_{HT}\right)} - \varphi^2\left(2+\varphi^2\right)\right].$$

Since

$$AV\left(\hat{T}_{DS}\right) + V\left(\hat{T}_{HT}\right) - 2C\left(\hat{T}_{HT} - B\hat{T}_{HT,\phi_x}, \hat{T}_{HT}\right) = B^2V\left(\hat{T}_{HT,\phi_x}\right),$$

the inequality $AV_{\min}\left(\hat{T}_{lin}\right) - AV\left(\hat{T}_{FS}\right) \le 0$ is equivalent to that

$$\rho^2\left(\hat{T}_{HT} - B\hat{T}_{HT,\phi_x}, \hat{T}_{HT}\right) \ge 1 - \frac{B^2V\left(\hat{T}_{HT,\phi_x}\right)}{\left(1+\varphi^2\right)^2V\left(\hat{T}_{HT}\right)}. \tag{24}$$

Under condition (24) the approximate variance of the estimator $\hat{T}_{lin}$ is not higher than that of the penalized calibration estimator $\hat{T}_{FS}$.

In the case of simple random sampling without replacement, inequality (24) reduces to

$$\rho^2\left(\phi_y - B\phi_x, \phi_y\right) \ge 1 - \frac{B^2S^2_{\phi_x}}{(1+\varphi^2)^2S^2_{\phi_y}}.$$

4. Simulation study

4.1. Simulation setup

The simulation study is performed to observe the efficiency of the Horvitz-Thompson type and calibrated estimators of the finite population variance and covariance. For that purpose, we consider a subset (of size 300) of the real

population from the Lithuanian Enterprise Survey. The study variables y and z are the profit of an enterprise in a different time period, whereas the values of auxiliary variables a and b are the numbers of employees of the enterprise at the same periods. The correlation coefficients between the study and auxiliary variables are $\rho(y,a)=0.81$, $\rho(z,b)=0.90$, $\rho(y,b)=0.63$ and $\rho(z,a)=0.60$. The relationship between the variables ϕ_y and ϕ_x is also strong enough, because $\rho(\phi_y,\phi_x)=0.64$ (when the estimated parameter is a finite population variance) and $\rho(\phi_y,\phi_x)=0.88$ (in the case of estimation of a finite population covariance).

The population is stratified into two strata by the size of the survey variable y. The stratified simple random sample is used as a sample design. The sample size $n=100$ is allocated to strata, using Neyman's optimal allocation. $M=10000$ samples were drawn and for each of them the estimators $\hat{S}^2_{HT}$, $\hat{S}^2_{DS}$, $\hat{S}^2_{FS}$ and $\hat{S}^2_{lin}$ of the finite population variance S^2_y, and the estimators $\hat{C}_{HT}$, $\hat{C}_{DS}$, $\hat{C}_{FS}$ and $\hat{C}_{lin}$ of the finite population covariance $C(y,z)$ were computed.

The uniform weights $q_{ij}=1$ were used for the calibrated estimators. A value of the parameter φ was calculated using formula (22). The coefficients α and β that appear in the expression of $\hat{T}_{lin}$, were defined in the following way. We give here an explanation in the general case, where the parameter T is any quadratic function. First, we write

$$\begin{aligned} V\left(\hat{T}_{lin}\right) &= \alpha^2 V\left(\hat{T}_{HT}\right)+\beta^2 V\left(\hat{T}_{DS}\right)+(1-\alpha-\beta)^2 V\left(\hat{T}_{FS}\right)+2\alpha\beta C\left(\hat{T}_{HT},\hat{T}_{DS}\right) \\ &\quad +2\alpha(1-\alpha-\beta)C\left(\hat{T}_{HT},\hat{T}_{FS}\right)+2\beta(1-\alpha-\beta)C\left(\hat{T}_{DS},\hat{T}_{FS}\right), \\ E\hat{T}_{lin} &= \alpha E\hat{T}_{HT}+\beta E\hat{T}_{DS}+(1-\alpha-\beta)E\hat{T}_{FS}. \end{aligned} \tag{25}$$

Then, by replacing in (25) true values of variances $V\left(\hat{T}_{HT}\right)$, $V\left(\hat{T}_{DS}\right)$, $V\left(\hat{T}_{FS}\right)$, expectations $E\hat{T}_{HT}$, $E\hat{T}_{DS}$, $E\hat{T}_{FS}$, and covariances between the estimators with the empirical ones, we minimize the empirical mean square error $MSE_{emp}\left(\hat{T}_{lin}\right)=V_{emp}\left(\hat{T}_{lin}\right)+\left(E_{emp}\hat{T}_{lin}-T\right)^2$ subject to the constraint $E_{emp}\hat{T}_{lin}=T$. The solution of this optimization problem ($\widetilde{\alpha}_{\min}$, $\widetilde{\beta}_{\min}$) is used in (14) for calculating estimates. This choice of the coefficients α and β often leads to a little bit more accurate estimates of $\hat{T}_{lin}$ as compared to those which are computed using $\alpha_{\min}$ and $\beta_{\min}$, defined by (23). Thus, using our data, we obtain

$\widetilde{\alpha}_{\min} = 0.30$, $\widetilde{\beta}_{\min} = 0.81$ ($\widetilde{\gamma}_{\min} = 1 - \widetilde{\alpha}_{\min} - \widetilde{\beta}_{\min} = -0.11$), if the parameter T is a population variance, and the values $\widetilde{\alpha}_{\min} = 0.09$, $\widetilde{\beta}_{\min} = 0.69$ ($\widetilde{\gamma}_{\min} = 0.22$) were used in the case of estimation of a finite population covariance.

4.2. Simulation results

The empirical relative bias (RB), variance (V), mean square error (MSE), and the coefficient of variation (cv) have been calculated for each estimator (see Tables 1 and 2). For any estimator $\hat{\theta}$ of the finite population parameter θ, all these characteristics of accuracy are defined by the following equations:

$$RB(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M}\frac{\hat{\theta}_i - \theta}{\theta}, \quad V(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\theta}_i - \frac{1}{M}\sum_{j=1}^{M}\hat{\theta}_j\right)^2,$$

$$MSE(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\theta}_i - \theta)^2, \quad cv(\hat{\theta}) = \sqrt{Var(\hat{\theta})}\cdot\left(\frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i\right)^{-1},$$

where $\hat{\theta}_i$ is the estimate of parameter θ, computed using data of the ith simulated sample.

Table 1. The main estimated characteristics of accuracy for the estimators of the finite population variance

Estimator	RB	$V \times 10^{-17}$	$MSE \times 10^{-17}$	cv
$\hat{S}^2_{HT}$	-0.0039	5.27	5.40	0.0991
$\hat{S}^2_{DS}$	0.0070	4.11	4.13	0.0855
$\hat{S}^2_{FS}[\varphi^2 = 0.09]$	-0.0053	3.95	3.96	0.0849
$\hat{S}^2_{lin}$	0.0037	3.64	3.64	0.0808

Table 2. The main estimated characteristics of accuracy for the estimators of the finite population covariance

Estimator	RB	$V \times 10^{-13}$	$MSE \times 10^{-13}$	cv
$\hat{C}_{HT}$	-0.0015	11.50	13.82	0.1752
$\hat{C}_{DS}$	0.0115	3.61	3.67	0.0899
$\hat{C}_{FS}[\varphi^2 = 0.10]$	-0.0066	3.32	3.34	0.0878
$\hat{C}_{lin}$	0.0010	3.27	3.27	0.0864

Due to quite a high correlation between the variables ϕ_y and ϕ_x, the calibrated estimators of population variance and covariance are much more accurate compared to the Horvitz-Thompson estimators of the same parameters. More significant differences between the characteristics of accuracy of the calibrated and Horvitz-Thompson estimators are observed in the case of estimation of the population covariance (see Table 2), where the coefficient of variation of the calibrated estimators $\hat{C}_{DS}$, $\hat{C}_{FS}$, $\hat{C}_{lin}$ is approximately two times lower than that of the only design-based estimator $\hat{C}_{HT}$. The variances and mean square errors differ about four times.

The estimators $\hat{S}^2_{lin}$ and $\hat{C}_{lin}$ which are the linear combinations of the estimators $\hat{S}^2_{HT}, \hat{S}^2_{DS}$, $\hat{S}^2_{FS}$ and $\hat{C}_{HT}$, $\hat{C}_{DS}$, $\hat{C}_{FS}$, respectively, outperform all the estimators from the corresponding group. The empirical analogs of the terms included into (24) do not satisfy this inequality. Thus, it seems that the approximate variances of the estimators $\hat{S}^2_{lin}$ and $\hat{C}_{lin}$ are higher than that of the corresponding penalized estimators $\hat{S}^2_{FS}$ and $\hat{C}_{FS}$, although the behaviour of empirical variances is contrary. The reason for that could be due to the linearization when only the first order Taylor approximations of the estimators are used for calculating approximate variances, and the remainder terms of a Taylor expansion are neglected.

Comparing the penalized calibration estimators $\hat{S}^2_{FS}$ and $\hat{C}_{FS}$ to the corresponding calibrated estimators $\hat{S}^2_{DS}$ and $\hat{C}_{DS}$, we note that not only the variance, but also the mean square error of the penalized estimators is lower than the variance of the calibrated estimators, as it is in the case of the approximate variance and mean square error (see Section 3).

5. Conclusion

Some of the estimators of finite population parameters can be treated as they are of quite a good quality in the sense of a small variance or small bias, but other characteristics of accuracy (e.g. mean square error) may not satisfy the survey statisticians and practitioners. In our case, the Horvitz-Thompson estimator $\hat{T}_{HT}$ is unbiased, but its variance may be relatively large. The calibrated estimator $\hat{T}_{DS}$ is preferable because of its lower variance (especially if the variables ϕ_y and ϕ_x

are well correlated), although it is slightly biased. Of course, the small bias has a minor impact on the results. The penalized calibration estimator $\hat{T}_{FS}$ has to be used carefully because an increase in the penalty (φ) may have a negative impact on the bias. The best properties of the estimators $\hat{T}_{HT}$, $\hat{T}_{DS}$ and $\hat{T}_{FS}$ are reflected by that of the estimator $\hat{T}_{lin}$. For some sets of coefficients α, β and γ, it is unbiased and may have the lowest variance among the estimators discussed in this paper.

REFERENCES

Deville, J.C. and Särndal, C. E., 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87, pp.376–382.

Farrell, P. and Singh, S., 2002. Penalized chi square distance function in survey sampling. *ASA Proceedings,* pp.963–968.

Guggemos, F. and Tillé, Y., 2010. Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference,* 140, pp.3199–3212.

Plikusas, A. and Pumputis, D., 2007. Calibrated estimators of the population covariance. *Acta Applicandae Mathematicae*, 97, pp.177–187.

Plikusas, A. and Pumputis, D., 2010. Estimation of the finite population covariance using calibration. *Nonlinear Analysis: Modelling and Control*, 15(3), pp.325–340.

Särndal, C.E., 2007. The calibration approach in survey theory and practice. *Survey Methodology*, 33(2), pp.99–119.

Särndal, C.E. Swensson, B. and Wretman, J., 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, S., 2003. On Farrell and Singh's penalized chi square distance functions in survey sampling. *SCC Proceedings,* pp.173–178.

Singh, S. Horn, S. Chowdhury, S. and Yu, F., 1999. Calibration of the estimators of variance. *Austral. & New Zealand J. Statist.,* 41(2), pp.199–212.

Sitter, R.R. and Wu, C., 2002. Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association,* 97(458), pp.535–543.

Yates, F. and Grundy, P., 1953. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, 15(2), pp.253–261.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 331—344

FINITE MIXTURES MODEL APPROACH TO SENSITIVE QUESTIONS IN SURVEYS

Shcherbina Artem[1], **Maiboroda Rostyslav**[2]

ABSTRACT

Observations from mixtures of different subpopulations are common in biological and sociological studies. We consider the case, when the observations are taken from a set of groups containing subjects, which belong to different subpopulations. Proportion of each subpopulation in a group is known and can vary from group to group. Our aim is to estimate the means of an observed variable for subjects, which belong to each subpopulation. In this paper we consider the case, when subpopulations are defined by answers on so called "sensitive questions". We consider some parametric and nonparametric estimates of the subpopulation means, such as weighted means, maximum likelihood and weighted least squares estimates. Finite sample properties of these estimates are analyzed. Mean square errors of the estimates are compared on simulated data. Some asymptotic results are also given.

Key words: anonymous survey; sensitive questions; maximum likelihood; weighted mean; weighted least squares.

1. Introduction

Problems of anonymous survey data analysis arise in many sociologic studies. E.g. anonymous surveys are usually used to avoid inadequate answers on so called "sensitive questions" (see Kerkvliet, 1994; Ong & Weiss, 2000). For a recent review of some well-known techniques to treat sensitive questions statistics the reader is addressed to E. Coutts & B. Jann (2011) and references herein. Statistical inference by voting data is another familiar example. In this paper we discuss the case when the results of an anonymous survey are used for inference together with some non-anonymous information on its' respondents.

[1] National Taras Shevchenko University of Kyiv, Ukraine.
E-mail: artshcherbina@gmail.com
[2] National Taras Shevchenko University of Kyiv, Ukraine. E-mail: mre@univ.kiev.ua

Let us consider the motivating example. Suppose that a survey was held in which first year university students were asked the question "Have you ever cheated on a school exam? (Yes/No)" Since this question is sensitive, the survey was held anonymously in different academic groups. As a result, the set of numbers of cheaters (N_{i1}) and non-cheaters (N_{i2}) was obtained for the groups $i = 1,\ldots, K$. The researcher would like to analyze the influence of cheating on the students' school marks. Say, it may be interesting to estimate and compare the mean marks in math for cheaters (μ_1) and non-cheaters (μ_2). Here marks of the students can be obtained from University journals. Thus we don't need to include such questions in the survey.

The simplest way to estimate μ_l is to use the following regression model for the mean marks T_i over the i-th group:

$$T_i \approx p_{i1}\mu_1 + p_{i2}\mu_2 ,$$

where $p_{il} = N_{il} /(N_{i1} + N_{i2})$ is the proportion of cheaters ($l = 1$) or non-cheaters ($l = 2$) in the i-th group. Ordinary least squares estimates (OLSE) for μ_l are

$$\hat{\mu}_l = \frac{1}{K}\sum_{i=1}^{K} a_{il} T_i \qquad (1)$$

where a_{il} are the minimax weights described in Section 3.1. These estimates are unbiased and consistent if p_{i1} is not constant for all $i = 1,...,K$.

But using this approach we drop a good deal of information on the structure of our data which is a mixture of two subpopulations (of cheaters and non-cheaters). Taking this into account one can construct more accurate estimates for μ_{il}. E.g., if some parametric model is imposed on the distributions of the marks of cheaters and non-cheaters, then maximum likelihood estimates are available. Such parametric approach is natural if the considered marks may attain only a small number of fixed values. The simplest case is a binary mark (success=1/failure=0) whose distribution is determined by the probability of success. The ML estimates of such probabilities μ_l for both subpopulations are given in Section 3.3. Note that these estimates can be consistent even when p_{i1} is constant. The ML estimates are asymptotically efficient under suitable assumptions (cf. Borovkov, 1998, Shcherbina 2011a).

On the other hand, the ML estimates can be inadequate if no good parametric model is known for the observed variable distribution. Therefore we developed three new non-parametric estimates for μ_l :

- weighted means of the form (1) with adaptive choice of the weights a_{il} ;
- estimates based on an approximate likelihood (weighted least squares) involving both linear and quadratic statistics from the data;

- estimates in which the ML approach is utilized for the CDF estimation over subpopulations of cheaters and non-cheaters. The estimates for μ_l then are derived as integrals by the CDFs' estimates.

Of course, the proposed estimates may be used not only for the cheating effects analysis, but also for statistical inference by any anonymous survey data.

The estimates are based on the theory of finite mixture models with varying mixing proportions (see Maiboroda, 1996; Maiboroda & Sugakova, 2008).

The rest of the paper is organized as follows. In Section 2 a formal description of the model is presented. In Section 3 we introduce the estimates and discuss their asymptotic properties. In Section 4 performance of the estimates is compared on simulated data. Concluding remarks are placed in Section 5.

2. Model description

Assume that the considered population U consists of two subpopulations U_1 or U_2. Some sample is taken from U at random. This sample is divided into K groups of subjects (sub-samples). Numbers N_{i1} and N_{i2} of the subjects from U_1 and U_2 in group i are known. Then $N_i = N_{i1} + N_{i2}$ is the total number of subjects in the i-th group. Let X_{ij} be the observed variable X of subject j from the group i. The variables X_{ij} are modelled as generated by a probability model, namely as independent random variables with distribution F_k for all subjects from the subpopulation U_k, $k = 1,2$.

To analyse such data a finite mixture model (FMM) may be used. In our case it is of the following form:

$$\mathrm{P}(X_{ij} \in A) = p_{i1}F_1(A) + p_{i2}F_2(A),$$

where A is any measurable subset of the observations space, p_{il} are the probabilities to observe a unit from the l-th subpopulation for subjects from the i-th group (the mixing probability, the concentration of the i-th component in the mixture):

$$p_{il} = \frac{N_{il}}{N_i}, \quad i = 1,2,\ldots,K, \quad l = 1,2.$$

For recent results on FMMs see McLachan and Pell (2000).

Finite mixtures with varying concentrations were considered in Maiboroda (1996), Maiboroda & Sugakova, (2008) for independent observations.

The observations discussed in this paper are dependent since the numbers N_{ij} of subjects belonging to each subpopulation are known. The best results in the

estimation will be achieved if the concentrations p_{il} are widely distributed on $[0,1]$.

Let $\mu=(\mu_1,\mu_2)$ be the mean values and $\sigma^2=(\sigma_1^2,\sigma_2^2)$ be the variances of the distributions F_1 and F_2.

3. Estimation

We want to estimate the parameter μ by observed characteristics X_{ij} and subpopulations sizes N_{i1} and N_{i2} in groups $i=1,...,K$.

3.1. Weighted means

Consider the mean value of X in i-th group $T_i=\frac{1}{N_i}\sum_{j=1}^{N_i}X_{ij}$. The expectation of T_i is

$$\mathrm{E}\,T_i=\mathrm{E}\frac{1}{N_i}\sum_{j=1}^{N_i}X_{ij}=\frac{1}{N_i}(N_{i1}\mu_1+N_{i2}\mu_2)=p_{i1}\mu_1+p_{i2}\mu_2,\quad i=1,2,\ldots,K.$$

Consider the following estimate of the mean value of the l-th subpopulation:

$$\hat{\mu}(a_l)=\frac{1}{K}\sum_{i=1}^{K}a_{il}T_i,$$

where $a_l=(a_{1l},a_{2l},\ldots,a_{Kl})$ is some coefficients vector.

It is readily seen, that the estimate $\hat{\mu}(a_l)$ for μ_l is unbiased if the following conditions hold:

$$\frac{1}{K}\sum_{i=1}^{K}a_{il}p_{im}=I_{\{m=l\}},\quad m=1,2. \tag{2}$$

By straightforward calculations we get

Proposition 1. The variance of the estimate $\hat{\mu}(a_l)$ is

$$\mathrm{D}\,\hat{\mu}(a_l)=\frac{1}{K^2}\sum_{i=1}^{K}a_{il}^2d_i,\quad d_i=\frac{1}{N_i}\left(p_{i1}\sigma_1^2+p_{i2}\sigma_2^2\right),\quad i=1,2,\ldots,K.$$

The best coefficients a_l minimize the variance of $\hat{\mu}(a_l)$ under (2). Coefficients a_l that satisfy (2) and minimize the sum $\sum_{i=1}^{K}a_{il}^2\hat{d}_i$ are

$$a_{i1}(\hat{d})=\frac{\left(r_0(\hat{d})-r_1(\hat{d})\right)p_{i1}+r_2(\hat{d})-r_1(\hat{d})}{\hat{d}_i\left(r_0(\hat{d})r_2(\hat{d})-r_1^2(\hat{d})\right)},\quad a_{i2}(\hat{d})=\frac{r_2(\hat{d})-r_1(\hat{d})p_{i1}}{\hat{d}_i\left(r_0(\hat{d})r_2(\hat{d})-r_1^2(\hat{d})\right)},$$

where $r_j(\hat{d})$ are weighted empirical moments of subpopulations proportions:

$$r_j(\hat{d}) = \frac{1}{K}\sum_{i=1}^{K}\frac{p_{i1}^j}{\hat{d}_i}, \quad j = 0,1,2.$$

The equality $r_0(\hat{d})r_2(\hat{d}) - r_1^2(\hat{d}) = 0$ is possible only when all proportions p_i are equal. In this case the weighted means cannot be used for estimation of the mean value.

Since the values d_i depend on the unknown parameter σ^2, we have to replace the vector $d = (d_1, d_2, \ldots, d_K)$ by some estimate $\hat{d} = \left(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_K\right)$. Taking $\hat{d}$ as the vector of units $I_K = (1,1,\ldots,1)$, we obtain minimax weights $a_l\left(I_K\right)$ introduced in Maiboroda (1996). On general theory of minimax estimation see section 2.21 in Borovkov (1998).

Although minimax weights do not minimize the variance of $\hat{\mu}(a_l)$, they are not too bad. The following theorem establishes conditions of consistency and asymptotic normality of estimates with minimax coefficients.

Theorem 1. Let there exist $C > 0$, such that $r_{0,K}\left(I_K\right)r_{2,K}\left(I_K\right) - r_{1,K}^2\left(I_K\right) > C$ for all K. Then the estimate $\hat{\mu}_l\left(a_l\left(I_K\right)\right)$ is consistent and distributions of the normalized estimate

$$\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}\left(a_l\left(I_K\right)\right)}}\left(\hat{\mu}\left(a_l\left(I_K\right)\right) - \mu_l\right)$$

converge weakly to the standard normal distribution as $K \to \infty$.

The proof of this theorem is based on the Central Limit Theorem.

As we shall see now, the variances σ_l^2 in the formula for $D\hat{\mu}(a_i)$ can be estimated by the same way as μ_l, so the asymptotic normality can be used for asymptotic confidence intervals construction. The same is true for other estimates considered below.

To improve the weighted means performance we may use the adaptive approach. To do this we use the weighed means with minimax weights as pilot estimates for parameter σ^2. E.g. the estimate for the l-th subpopulation is

$$\hat{\sigma}_{l,K}^2 = \frac{1}{K}\sum_{i=1}^{K} a_{il}\left(I_K\right)\frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}^2 - \hat{\mu}^2\left(a_l\left(I_K\right)\right).$$

Then we can estimate the vector d by $\hat{d}_K = \left(\hat{d}_{1,K}, \hat{d}_{2,K}, \ldots, \hat{d}_{K,K}\right)$ where $\hat{d}_{i,K} = \left(p_{i1}\hat{\sigma}_1^2 + p_{i2}\hat{\sigma}_2^2\right)\big/N_i$ and use $r_j\left(\hat{d}_K\right)$ to derive coefficients $a_l\left(\hat{d}_K\right)$. The

adaptive estimate for μ_l is now $\hat{\mu}\left(a_l\left(\hat{d}_K\right)\right)$. On efficiency of adaptive estimates in mixture models see Maiboroda (1999).

A valuable property of the adaptive estimate is its asymptotic normality with the same asymptotic variance as for the estimate with best coefficients $a_l\left(d_K\right)$.

Theorem 2. Under the conditions of Theorem 1, the estimate $\hat{\mu}\left(a_l\left(\hat{d}_K\right)\right)$ is consistent and distributions of the normalized estimate

$$\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}(a_l(d_K))}}\left(\hat{\mu}\left(a_l\left(\hat{d}_K\right)\right)-\mu_l\right)$$

converge weakly to the standard normal distribution as $K\to\infty$.

The proof is based on the following two observations:

(i) $\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}(a_1(d_K))}}(\hat{\mu}(a_1(d_K))-\mu_1)$ converge weakly to the standard normal distribution as $K\to\infty$ by the Central Limit Theorem.

(ii) $\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}(a_1(d_K))}}(\hat{\mu}(a_1(d_K))-\hat{\mu}(a_1(\hat{d}_K))\to 0$ in probability. It can be shown by applying lemma 3.2.1 from Maiboroda & Sugakova, (2008).

Complete proofs of Theorems 1 and 2 can be found in Shcherbina (2011).

Hence, the adaptive weights allow one to build estimates with the same asymptotic quality, as with the best coefficients.

But such weights should be used carefully for small samples. Estimation of the subsamples' variances can introduce additional variability. That is why the simple minimax coefficients sometimes perform better.

3.2. Weighted least squares

Consider the statistics $S_i=\left(\sum_{j=1}^{N_i}X_{ij},\sum_{j\neq k}X_{ij}X_{ik}\right)$. They are independent random vectors for $i=1,...,K$ with the following mathematical expectations and covariance matrices:

$$f_i(\mu)=\mathrm{E}\,S_i=f\left(N_{i1},N_{i2},\mu\right),$$
$$\Sigma_i=\mathrm{Cov}\,S_i=\Sigma\left(N_{i1},N_{i2},\mu,\sigma^2\right).$$

The elements of the function f are the following:

$$f_1\left(n_1,n_2,\mu\right)=n_1\mu_1+n_2\mu_2,\quad f_2\left(n_1,n_2,\mu\right)=n_1\left(n_1-1\right)\mu_1^2+2n_1n_2\mu_1\mu_2+n_2\left(n_2-1\right)\mu_2^2.$$

The matrix Σ is symmetric with the following elements:

$$\Sigma_{11}\left(n_1,n_2,\mu,\sigma^2\right)=n_1\sigma_1^2+n_2\sigma_2^2,$$
$$\Sigma_{12}\left(n_1,n_2,\mu,\sigma^2\right)=2n_1n_2\mu_1\sigma_1^2+2n_1n_2\mu_2\sigma_2^2+2n_1\left(n_1-1\right)\mu_1\sigma_1^2+2n_2\left(n_2-1\right)\mu_2\sigma_2^2,$$
$$\Sigma_{22}\left(n_1,n_2,\mu,\sigma^2\right)=4n_1\left(n_1-1\right)^2\mu_1^2\sigma_1^2+4n_2\left(n_2-1\right)\mu_2^2\sigma_2^2+4n_1^2n_2\mu_1^2\sigma_2^2+4n_1n_2^2\mu_2^2\sigma_1^2$$
$$+8n_1n_2\mu_1\mu_2\left(\left(n_1-1\right)\sigma_1^2+\left(n_2-1\right)\sigma_2^2\right)+2\left(n_1\sigma_1^2+n_2\sigma_2^2\right)^2-2n_1\sigma_1^4-2n_2\sigma_2^4.$$

The matrix Σ depends on μ and σ^2. Here we estimate them with weighed means estimates as described above, compute matrices $\Sigma_i=\Sigma\left(N_{i1},N_{i2},\hat{\mu}^2\left(a_l\left(I_K\right)\right),\hat{\sigma}^2_{l,K}\right)$ and then use them as fixed. Consider the generalized least squares criterion

$$\sum_{i=1}^{K}\left(S_i-f_i\left(\mu\right)\right)\Sigma_i^{-1}\left(S_i-f_i\left(\mu\right)\right)^T\to\min. \quad (3)$$

Denote the solution of (3) by $\hat{\mu}_{LS}$. This estimator can perform better than linear and adaptive estimates since it uses second order sums.

Note that this LS criterion is approximately equal to the log-likelihood when the sizes N_i of the groups are large due to the asymptotic normality of S_i as $N_i\to\infty$.

In practice it can be difficult to minimize (3). By differentiating (3) with respect to μ we get the following estimating equations (see Heyde, 1997):

$$\sum_{i=1}^{K}\frac{\partial f_i}{\partial\mu}\left(\mu\right)A_i\left(S_i-f_i\left(\mu\right)\right)^T=0.$$

They can be solved by a Newton-type algorithm (see Small & Wang, 2003).

3.3. Maximum likelihood for parametric case

Let the characteristic X be binary, i.e. attains only two values 1 (success) or 0 (failure) with probabilities of success μ_1 and μ_2 for subjects from the first and second subpopulations. The statistic $Z_i=\left(N_{i1},N_{i2},\sum_{j=1}^{N_i}X_{ij}\right)$ is then sufficient for estimating the parameter μ by the observations from i-th group. The density function of Z_i is

$$f_i(x,q)=\mathrm{P}\left(\sum_{j=1}^{N_i}X_{ij}=x\,|\,\mu=q\right)=\sum_{k=0\vee\left(x-N_{i2}\right)}^{x\wedge N_{i1}}\binom{N_{i1}}{k}q_1^k(1-q_1)^{N_{i1}-k}\binom{N_{i2}}{x-k}q_2^{x-k}(1-q_2)^{N_{i2}-x+k}$$

Then, the maximum likelihood estimate is defined as

$$\hat{\mu}_{ML}=\arg\max_{q\in[0,1]^2}\sum_{i=1}^{K}\ln f_i\left(\sum_{j=1}^{N_i}X_{ij},q\right).$$

The likelihood function can have several points of global maxima. In that case we can select either of them.

To establish asymptotic properties of this estimate we will treat subpopulation sizes in groups (N_{i1}, N_{i2}) as independent random vectors with unknown distribution G. This approach is analogous to the use of structural regression models in which regressors are considered as random variables. It doesn't affect the search of the likelihood extremal points, but simplify the asymptotic considerations, since the law of large numbers is now in force for empirical means of N_{ik} and functions from them.

We have to distinguish two cases:

There are groups with unequal subpopulation sizes, i.e. $\mathrm{P}(N_{i1} \neq N_{i2}) > 0$.

All the groups have equal subpopulations sizes, $N_{i1} = N_{i2}$ *a.s.*

In the first case the likelihood function have unique maximum. The following theorem establishes consistency and asymptotic normality of the maximum likelihood estimate in the first case.

Theorem 3. If the domain sizes have finite first moment $EN_i < \infty$, and $\mathrm{P}(N_{i1} \neq N_{i2}) > 0$ then the maximum likelihood estimate $\hat{\mu}_{ML}$ is consistent as $K \to \infty$. If the domain sizes have finite second moment $EN_i^2 < \infty$ and there is no constant $C > 0$ such that $N_{i1} = CN_{i2}$ *a.s.*, then the maximum likelihood estimate $\hat{\mu}_{ML}$ is asymptotically normal.

In the second case the likelihood function becomes symmetric with respect to the interchange of q_1 and q_2. That means that we can estimate parameters q_1 and q_2 up to a permutation only. Such data arise sometimes in genomic studies, see Scherbina (2011a). Let us constrict the parametric space to $\mu \in \{(q_1, q_2) \mid q_1 \le q_2\}$ and use the following maximum likelihood estimate

$$\hat{\mu}^*_{ML} = \arg\max_{0 \le q_1 \le q_2 \le 1} \sum_{i=1}^{K} \ln f_i\left(\sum_{j=1}^{N_i} X_{ij}, q\right).$$

Theorem 4. Assume that $\mu_1 \le \mu_2$. If the subpopulations sizes have finite first moment $EN_i < \infty$, then the maximum likelihood estimate $\hat{\mu}^*_{ML}$ is consistent as $K \to \infty$. If the domain sizes have finite second moment $EN_i^2 < \infty$ and $\mu_1 < \mu_2$, then $\hat{\mu}^*_{ML}$ is asymptotically normal.

Proofs of Theorems 3 and 4 are based on general theory of ML estimates (see Borovkov, 1998) and are presented in Shcherbina (2011a).

3.4. CDFS based estimates

Although the binomial distribution for X_{ij} described in the previous subsection is very restrictive, this technique can be used for arbitrary distributed variables. Let us fix any $x \in \mathbb{R}$ and replace each value X_{ij} with the indicator $I_{\{X_{ij}<x\}}$. We obtain a sample of binary variables and may use estimates for probabilities of success for two subpopulations described in Section 3.3. Clearly, these probabilities are the cumulative density functions $F_1(x)$ and $F_2(x)$ for first and second subpopulations. Thereby we get estimates $\hat{F}_1(x)$ and $\hat{F}_2(x)$ for them. Since they are not usual empirical CDFs, they can be not monotone. Despite of that, we can estimate μ by

$$\hat{\mu}_{CDF} = \left(\int_{\mathbb{R}} x d\hat{F}_1(x), \int_{\mathbb{R}} x d\hat{F}_2(x) \right).$$

To estimate means we calculate functions $\hat{F}_l(x)$ only at sample points $\{X_{ij}\}$. Let $\{X_i^*\}_{i=1}^N$ be the ordered sequence of observed characteristics, where N is the total sample size. We take $\hat{F}_l(x)$ as peacewise constant with jumps $\hat{F}_l(X_{i+1}^*) - \hat{F}_l(X_i^*)$ in points X_i^* for $i = 1,\ldots,N-1$ and jump $1 - \hat{F}_l(X_N^*)$ in point X_N^*. Now, the integrals can be easily computed:

$$\int_{\mathbb{R}} x d\hat{F}_l(x) = \sum_{i=1}^{N-1} X_i^* \left(\hat{F}_l(X_{i+1}^*) - \hat{F}_l(X_i^*) \right) + X_N^* \left(1 - \hat{F}_l(X_N^*) \right).$$

4. Simulation results

This section provides some simulation studies. Performance of the estimates considered in the previous section is compared on artificial data. We try different distributions for characteristic X, different subpopulation sizes and numbers of groups. Each iteration consists of the following steps:

1. Subpopulation sizes in groups N_{i1} and N_{i2} are generated from some distribution.

2. For each group N_{i1} and N_{i2} independent random variables are generated with distributions that correspond to the first and second subpopulations. Thus, we get the sample $\{X_{ij}, i = 1,2,\ldots,K,\ j = 1,2,\ldots,N_i\}$.

3. Estimates described in the previous chapter are computed.

Then, the mean square errors (MSE) of the estimates are computed. They are multiplied by K in plots.

Example 1. The variable X has Binomial distribution with parameters $\mu_1 = 0.2$ and $\mu_2 = 0.5$. Subpopulation sizes (N_{i1}, N_{i2}) can be $(1,2)$ or $(2,1)$ with equal probability. Number of groups K varies from 10 to 200. Number of simulations is equal 5000.

In the next table the correspondences between estimates and type of lines in plots are shown.

Estimate	Type of lines in plots
Weighted means	Dotted
Weighted least squares	Dashed
Maximum likelihood	Solid

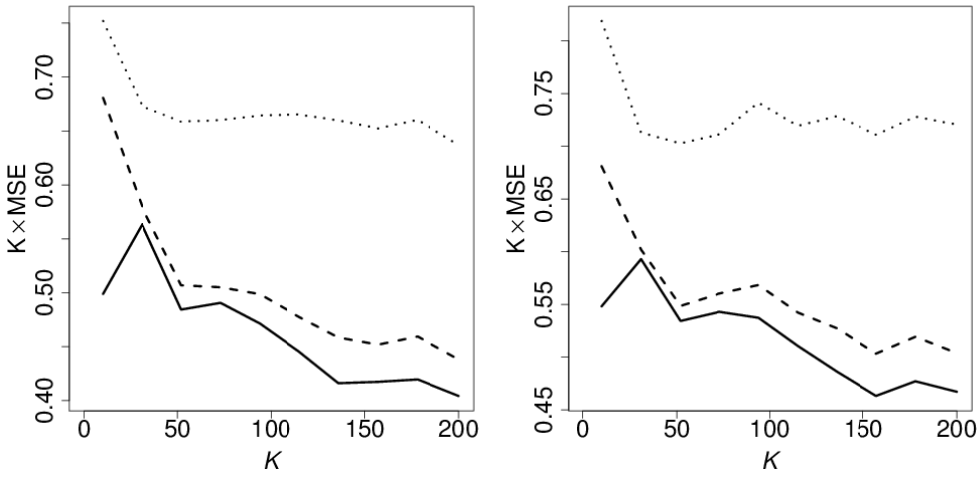


Figure 1. Normalized MSE of estimates for μ_1 and μ_2 for different number of groups K

Example 2. Consider the quality of CDF estimation described in Section 3.4. We took exponential distribution with intensity 1 on U_1 and standard normal distribution on U_2. Group sizes N_i are equal to 5. Size of the first subpopulation N_i^1 is uniformly distributed on $\{1,2,3,4\}$. Number of groups K varies from 20 to 200.

In the Figure 2 the CDFs (thin lines) and their estimates (solid and dashed lines) are represented for both subpopulations. As we can see, the deviations from the true values decrease rapidly when K increases. Although the estimated CDFs are not monotonous, they can be directly used in the estimator $\hat{\mu}_{CDF}$.

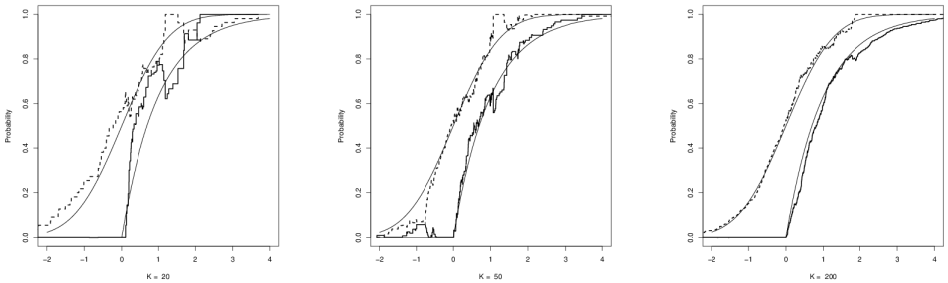


Figure 2. Estimation of CDF by maximum likelihood approach

Example 3. The variable X has Student t-distribution with 3 degrees of freedom and normal distribution with mean 2 and variance 4. Group sizes N_i are equal to 5. Size of the first subpopulation N_i^1 is uniformly distributed on $\{1,2,3,4\}$. Number of groups K varies from 10 to 200. Number of simulations equals to 1000.

In the next table the correspondences between estimates and type of lines in plots are shown.

Estimate	**Type of lines in plots**
Weighted means	Dotted
Weighted least squares	Dashed
CDF' based	Solid

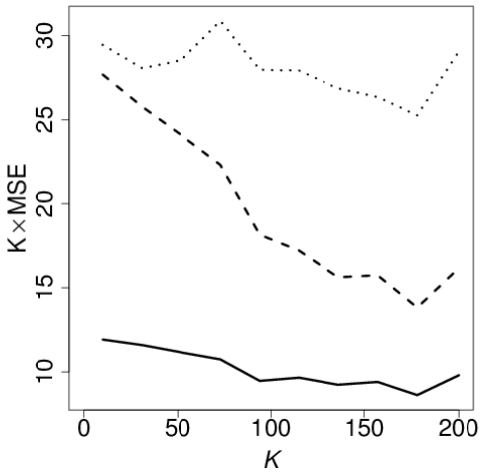


Figure 3. Normalized MSE of estimates for μ_1 for different number of groups K

5. Concluding remarks

The simulation studies indicate that the maximum likelihood estimates are the best in the parametric case, although the weighted least squares estimate is also quite good. Presence of the second order sums in weighted least squares estimates make them much better than the minimax estimates. In the nonparametric case the technique based on CDF estimation seems very promising. It shows the best performance in comparison with the other estimates. Weighted means estimate can be used as pilot estimates of means and variances or as a starting point in numerical calculations of the other estimates.

On the other hand, some caveats should be made about application of these estimates to real data analysis. One expects them to perform satisfactory only if the data satisfy the mixture model on which the estimates are based. Say, the division of the subjects into groups must not depend on their observed variable X. The values of X for subjects belonging to the same group must be independent given the subpopulations distribution in this group. There must be no outer factors shifting the distributions of X for subjects of the same population in different groups. So, a careful choice of the survey design is needed for efficient work of the proposed estimates.

Say, in our cheating example we may expect independence of the first year students' previous school marks and their distribution by academic groups if the groups were formed in random by the Dean's office. But it is not the case when we analyze second year students' current marks, since they may be dependent on their common experience which is different in different groups. In this case one needs more ingenious experiment design, e.g. using surveying of random groups of students attending a gym.

Despite these restrictions we hope that the proposed estimates will be useful in sensitive questions statistics and other problems connected with merging anonymous and non-anonymous surveys information.

Aknowlegements

The authors are thankful to the Referee and Editor for the fruitful discussion.

REFERENCES

BOROVKOV A.A. (1998). *Mathematical statistics,* Gordon and Breach Science Publishers, Amsterdam.

CHRISTOPHER C. HEYDE (1997) *Quasi-Likelihood And Its Application: A General Approach to Optimal Parameter Estimation,* Springer.

CHRISTOPHER G. SMALL, JINFANG WANG, (2003), *Numerical Methods for Nonlinear Estimating Equations*, Oxford.

COUTTS E. & JANN B. (2011), *Sensitive Questions in Online Surveys:Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)*, Sociological Methods & Research, 40, 169-193.

KERKVLIET J. (1994) Cheating *by economics students: A comparison of survey results.* The Journal of Economic Education, Vol. 25, No. 2, p.121-133.

MAIBORODA R. (1996) *Estimates for distributions of components of mixtures with varying concentrations.* Ukrainian Mathematical Journal, 48(4), 618-622.

MAIBORODA R. (1999) A*n asymptotically effective probability estimator constructed from observations of a mixture.* Theory Probab. Math. Stat. 59, 121-128

MAIBORODA R. & SUGAKOVA O. (2008) *Estimation and classification by observations from mixtures*, Kyiv University Publishers, Kyiv (in Ukrainian).

McLACKLAN G. J., PEEL D. (2000). *Finite Mixture Models*, Wiley, New York.

ONG A.D. and WEISS D.J. *The Impact of Anonymity on Responses to Sensitive Questions. Journal of Applied Social Psychology*. Volume 30, Issue 8, p. 1691–1708.

SHCHERBINA A. (2011) *Mean value estimation in the model of mixture with varying concentrations*. Teor. Imovir. Ta Matem. Statyst., No. 84, pp. 142-154. (In Ukrainian, English translation to appear in Theory Probab. Math. Stat).

SHCHERBINA A. (2011a) *Estimation of parameters of binomial distribution in mixture model*. Teor. Imovir. Ta Matem. Statyst. (In Ukrainian, English translation to appear in Theory Probab. Math. Stat).

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 345—356

PERCENTILE-ADJUSTED ESTIMATION OF POVERTY INDICATORS FOR DOMAINS UNDER OUTLIER CONTAMINATION

Ari Veijanen[1], **Risto Lehtonen**[2]

ABSTRACT

Traditional estimation of poverty and inequality indicators, such as the Gini coefficient, for regions does not currently use auxiliary information or models fitted to income survey data. A predictor-type estimator constructed from ordinary mixed model predictions is not necessarily useful, as the predictions have too small spread for estimation of income statistics. Ordinary bias corrections are aimed at correcting the expectation of predictions, but poverty indicators would not be affected at all by a correction involving multiplication of predictions. We need a method improving the shape of the distribution of predictions, as poverty indicators describe differences of income between people. We therefore introduce a transformation bringing the percentiles of transformed predictions closer to the percentiles of sample values. The experiments show that the transformation results in smaller MSE of a predictor. If unit-level data from population are not available, the marginal domain frequencies of qualitative auxiliary variables can be successfully incorporated into a new calibration-based predictor-type estimator. The results are based on design-based simulation experiments where we use a population generated from an EU-wide income survey. The study is a part of the AMELI project funded by the European Union under the Seventh Framework Programme for research and technological development (FP7).

Key words: small area estimation; poverty indicator; income data; bias correction; auxiliary information; mixed model; prediction.

1. Introduction

Regional income statistics have received a lot of attention in recent years. Thresholds for a poverty indicator have been used in regional allocation of resources (e.g. Zaslavsky and Schirm, 2002). The European Union conducts regular income surveys (Statistics on income and living conditions, SILC, see e.g.

1 Statistics Finland. E-mail: ari.veijanen@stat.fi
2 University of Helsinki. E-mail: risto.lehtonen@helsinki.fi

Clemenceau, Museux and Bauer, 2006) yielding information about income and social attributes of households. In the AMELI project (Advanced Methodology for European Laeken Indicators), we studied the estimation of indicators on poverty and social exclusion (the so-called Laeken indicators) including poverty gap, quintile share ratio, and the Gini coefficient. These statistics are nonlinear, so methods designed for estimation of domain totals, such as GREG (Särndal, Swensson and Wretman, 1992; Lehtonen and Veijanen, 2009) or EBLUP (empirical best linear unbiased predictor) cannot be applied. In this paper, we introduce new methods for estimation of poverty indicators in regions (small areas) and other domains.

The equivalized income constitutes the key variable underlying the monetary poverty indicators. It is defined as a household's total disposable income divided by its "equivalent size", to take account of the composition of the household (European Commission, 2003). Equivalization is made on the basis of the OECD modified scale, which assigns weight 1.0 for the first adult, 0.5 for every additional person aged 14 or over, and 0.3 for every child under 14. The equivalized income is attributed to each household member including children, and the poverty indicators are defined for unit-level data composed of persons.

We compare currently widely used “default” estimators of poverty indicators with corresponding predictor-type estimators. A design-based default estimator incorporates design weights and sample values but does not use auxiliary data or modelling. It is a so-called direct estimator, as observations from other domains do not contribute to a domain estimate (Lehtonen and Veijanen, 2009, p. 223). The direct estimator probably has small design bias, but its variance can be large in domains with a small sample size. In an indirect predictor-type estimator based on unit-level auxiliary information, predictions are plugged into the formula of the poverty indicator defined at the population level. The estimator is expected to have small variance but it may be seriously design biased.

As the equivalized income is approximately distributed as lognormal, a model is often fitted to log-transformed data and the fitted values are back-transformed to the original scale. In estimation of domain totals, this should be followed by a bias correction, such as RAST (Ratio Adjusted by Sample Total; Chambers and Dorfman, 2003; Fabrizi, Ferrante and Pacei, 2007). Chandra and Chambers (2011) discuss non-linear transformations and introduce more accurate bias corrections. However, the quintile share especially is determined by the shape of the income distribution, not its average. As the distribution of predictions is concentrated around the average, they yield too large quintile shares or too small Gini coefficients and poverty gaps. We define a transformation that brings the percentiles of transformed predictions closer to the percentiles of sample values. Our technique is model-based but it aims to correct for design bias with the use of design weights, so it is comparable to design consistent pseudo synthetic, pseudo EBLUP and pseudo EBP (empirical best predictor) type approaches (Rao 2003; You and Rao, 2002; Jiang and Lahiri, 2006).

Predictions can be avoided altogether by simulating unknown observations from their conditional distribution given the sample values and auxiliary data (Molina and Rao, 2010). However, in our simulations deviations from the

assumed model resulted in large bias. Estimating quantiles of the income distribution by M-quantile regression (Chambers and Tzavidis, 2006) might also provide an alternative basis for small area estimation of poverty indicators.

This article is organized as follows. Section 2 contains notation and basic definitions. Section 3 introduces the technique of transforming the predictions. As unit-level auxiliary data are not always available, we introduce in Section 4 new frequency-calibrated estimators incorporating known marginal totals of auxiliary variables. In Section 5, we evaluate the design bias and accuracy of the estimators by Monte Carlo experiments using a synthetic Amelia population, which is based on the EU-wide SILC survey (Alfons et al., 2011). Short discussion is in Section 6.

2. Definitions

2.1. Domain models

The fixed and finite population of interest is denoted $U = \{1, 2, ..., k, ..., N\}$, where *k* refers to a unit's label. The population is divided into subsets called *domains* by region, for example. Each domain U_d, indexed by *d*, has N_d elements. The sample *s* is composed of corresponding subsets s_d of size n_d. The domains are called unplanned unless the sampling design is stratified by domains (Lehtonen and Veijanen, 2009, p. 222). Design weights, inverses of inclusion probabilities, are denoted by a_k.

To account for differences between domains, a linear mixed model incorporates domain-specific random effects $u_d \sim N(0, \sigma_u^2)$. The model is given by

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k, k \in U_d, \varepsilon_k \sim N(0, \sigma^2).$$

The random effects may also be associated with aggregates of domains. The parameters $\boldsymbol{\beta}$, σ_u^2 and σ^2 are first estimated from the data by using ML or REML methods, and the values of the random effects are then predicted. This yields predictions $\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d, k \in U_d$. Typically, design weights are not incorporated in the estimation of the model.

2.2. Poverty indicators

The *S20/S80 ratio,* or *quintile share ratio*, compares the average equivalized incomes in the poorest and the richest quintile. Each quintile contains 20 % of people; in the design-based case accounting for 20 % of design weights. The *default* (direct) *estimators* of the first (S20) and the fifth quintile average (S80) are Hájek estimators, that is, weighted domain averages involving design weights.

The direct quintile share estimate is the ratio of S20 to S80. The predictor-type estimator of quintile share in a domain is the ratio of averages of predictions in the first and fifth quintiles.

Consider ordered equivalized incomes y in a domain U_d. The *Lorenz curve* L_d describes how the first k ($k=1,2,...,N_d$) persons' proportion of the total income depends on their numerical proportion. The curve is approximated by a line between points

$$L_d\left(\frac{k}{N_d}\right)=\frac{\sum_{i\le k;i\in U_d} y_{(i)}}{\sum_{t\in U_d} y_t} \quad (y_{(1)}\le y_{(2)}\le ...\le y_{(N_d)}) \tag{1}$$

The *Gini coefficient* G_d is defined as

$$G_d = 1-2\int_0^1 L_d(x)dx\,. \tag{2}$$

A direct estimator of the Lorenz curve is defined in ordered sample: for unit k,

$$L_{HT;d}\left(\frac{\sum_{i\le k;i\in s_d} a_i^s}{\sum_{t\in s_d} a_t}\right)=\frac{\sum_{i\le k;i\in s_d} a_i^s y_{(i)}}{\sum_{t\in s_d} a_t y_t} \quad (y_{(1)}\le y_{(2)}\le ...\le y_{(n_d)}),$$

where a_i^s denotes the design weight of the ith element in the ordered sample. The predictor-type estimator of the Lorenz curve is simply obtained by plugging the ordered predicted incomes into (1). The *default* (direct) estimator and the predictor of the Gini coefficient are then defined as integrals similar to (2).

Poverty gap in a domain describes the difference between the at-risk-of-poverty threshold r (typically 60 % of the median income in the whole country) and the median income $m_d(r)=Md\{y_k; y_k\le r; k\in U_d\}$ below the threshold:

$$g_d=\left(r-m_d(r)\right)/r\,. \tag{3}$$

The *default* (direct) *estimator* of the poverty gap is calculated as a similar ratio involving estimated threshold $\hat{r}$ and the design-based estimate of median income below the threshold obtained from estimated distribution function

$$\hat{F}_{HT}(y;\hat{r})=\frac{\sum_{k\in p_d} a_k I\{y_k\le y\}}{\sum_{k\in p_d} a_k};\; p_d=s_d\cap\{k: y_k\le\hat{r}\}\,.$$

The predictor contains $\hat{r}$ and predictions' median $\hat{m}_{pred;d}(\hat{r})=Md\{\hat{y}_k;\hat{y}_k\le\hat{r};k\in U_d\}$.

3. Percentile-adjusted predictors

The distribution of the equivalized income *y* is skewed with a long right tail. A candidate for a distributional assumption is the lognormal distribution, but it is often not realistic. In any case, after logarithmic transformation a model fits the data better. For pragmatic reasons, a model is fitted to $z_k = \log(y_k + 1)$, and the fitted values $\hat{z}_k$ are back-transformed to $\hat{y}_k = \exp(\hat{z}_k) - 1$. To correct for bias in estimates of domain totals, Chambers and Dorfman (2003) and Fabrizi, Ferrante and Pacei (2007) calculate a RAST multiplier c_R ensuring that the weighted sample sum of $c_R \hat{y}_k$ equals the corresponding sum of original incomes. However, the poverty indicators are not affected at all by multiplication amounting to a change in currency. Correction for shape of distribution is also required. We correct for bias and spread of predictions $\hat{y}_k = \exp(\hat{z}_k) - 1$ ($k \in U_d$) by a nonlinear transformation that brings the distribution of predictions closer to the distribution of observed values y_k ($k \in s_d$) in terms of percentiles, denoted by $\hat{p}_{cd}$ and p_{cd}, respectively. The percentiles p_{cd} of sample values are obtained from the estimated cumulative distribution function

$$\hat{F}_{HT;d}(y) = \frac{1}{\hat{N}_d} \sum_{k \in s_d} a_k I\{y_k \le y\};\ \hat{N}_d = \sum_{k \in s_d} a_k .$$

Our goal is to obtain transformed predictions $\tilde{y}_k = e^{\alpha_d} \hat{y}_k^{\gamma}$ whose percentiles, denoted $\tilde{p}_{cd}$, are close to p_{cd} on logarithmic scale. The using of logarithms reduces the influence of large percentiles. To avoid unstable estimates in the smallest domains, we pooled the percentile data from all domains and minimized

$$\sum_d \sum_{c=1}^{C} \left(\log(\tilde{p}_{cd}) - \log(p_{cd})\right)^2 = \sum_d \sum_{c=1}^{C} \left(\alpha_d + \gamma \log(\hat{p}_{cd}) - \log(p_{cd})\right)^2 ,$$

where C=50 for poverty gap and C=99 for other indicators. The *percentile-adjusted*, or *p-adjusted*, predictions involve OLS estimates of parameters α_d and γ:

$$\log(\tilde{y}_k) = \hat{\alpha}_d + \hat{\gamma} \log(\hat{y}_k)\ (k \in U_d) . \tag{4}$$

We experimented also with a mixed model fitted to the percentile data but it did not yield significantly different results.

About 1.5 % of the people in the samples of our experiments had zero equivalized income. To incorporate zero predictions in the predictors we replaced a roughly identical proportion of the smallest predictions in each domain by zero. The transformation (4) was applied only to the positive predictions, with percentiles $\hat{p}_{cd}$ and p_{cd} calculated from positive predictions and sample values. Occurrence of zeroes could also be described by a model that includes a logistic component (Karlberg, 2000).

4. Frequency-calibrated predictors

We develop here a new method that may be feasible in situations where only aggregate-level qualitative auxiliary data are available. Suppose only the domain sizes and domain totals of qualitative auxiliary variables are known. Then we cannot access the values of $\mathbf{x}_k = \left(x_{1k}, x_{2k}, ..., x_{pk}\right)$ for every unit. However, for the calculation of a predictor, it is actually enough to know the population frequencies of distinct values of $\mathbf{x}_k$, that is, the frequencies of cells in the crosstabulation of the **x**-variables. We propose a method of estimating these frequencies.

Consider domain *d*. Let us denote the set of distinct values of $\mathbf{x}_k$, $k \in s_d$, by $X_d = \left\{z_1, z_2, ..., z_m\right\}$. A direct estimate of the domain frequency of $z \in X_d$ is

$$\hat{n}_z = \sum_{k \in s_d} a_k I\left\{\mathbf{x}_k = z\right\}.$$

These do not, in general, sum up to the known marginal totals t_d. This requirement is formulated as a calibration equation

$$\sum_{k \in U_d} \mathbf{x}_k = \sum_{z \in X_d} n_z z = t_d. \tag{5}$$

Calibration (e.g. Singh and Mohl, 1996; Särndal, 2007) yields new frequencies $\tilde{n}_z$ that are close to the $\hat{n}_z$ and also satisfy the calibration equations. We measure the distance of $\tilde{n} = (\tilde{n}_z; z \in X_d)$ to $\hat{n} = (\hat{n}_z; z \in X_d)$ by chi-squared distance

$$\sum_{z \in X_d} \frac{1}{\hat{n}_z}\left(\hat{n}_z - \tilde{n}_z\right)^2.$$

This distance is minimized subject to the calibration equations (5) by

$$\tilde{n}_z = \hat{n}_z\left(1 + \lambda_d' z\right), \tag{6}$$

where the Lagrange multiplier λ_d is

$$\lambda_d = \left(t_d - \sum_{z \in X_d} \hat{n}_z z\right)\left(\sum_{z \in X_d} \hat{n}_z z z'\right)^{-1}.$$

To avoid singular matrices, we excluded from each $z \in X_d$ the indicator variables of classes that did not appear in the sample domain. Moreover, if two variables had identical values in the domain, the latter variable was removed. Corresponding modifications were made in the vector t_d. If the algorithm still failed due to linear dependencies, for example, we used the initial estimates $\hat{n}_z$. This occurred rarely. Unfortunately, about 10 % of the $\tilde{n}_z$ were negative in our simulations. We replaced them by zero. After this, the calibration equations do not necessarily hold. Negative estimates might be avoided by the approach of

Singh and Mohl (1996) in taking into account range restrictions of calibrated weights.

The vector of predictions in the domain is finally obtained by repeating $\tilde{n}_z$ times the fitted value associated with each $z \in X_d$ (using rounded frequency estimates). Transformation (4) may then be applied. We call the resulting predictor a *frequency-calibrated*, or an *n-calibrated* predictor.

5. Monte carlo simulation experiments

5.1. Introduction

We present numerical results from design-based Monte Carlo simulation experiments with synthetic Amelia data set constructed by Alfons et al. (2011) to mimic the regional and demographic variation of income statistics in European Union. It contains about ten million persons. We applied SRSWOR (n = 2000). As domains we used 40 regions (variable DIS) or, in the case of poverty gap, 110 demographic population subsets defined by age class, gender and NUTS2 region. The domains were classified to minor, medium and major domains by expected sample size with class boundaries at 45 and 55 units for the DIS regions and at 20 and 30 units for the demographic domains. Our models fitted to equivalized income variable EDI2 incorporated age class and gender with interactions, attained education level (ISCED), activity (working, unemployed, retired, or otherwise inactive) and degree of urbanisation of residence (three classes). The mixed models, which had random intercepts associated with districts (DIS) or NUTS2 regions (in the case of poverty gap), were fitted by R package nlme using ML without design weights.

In contamination experiments, outliers were created in each sample without modifying the population. One percent of sampled persons were declared as outliers, chosen completely at random, and a normally distributed value from $N(500000,10000^2)$ was added to the personal cash or near-cash income of an outlier. We evaluated the estimates by absolute relative bias (ARB, absolute average error divided by the true value) and the relative root mean squared error

$$RRMSE = \frac{1}{\theta_d}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d)^2} \ .$$

5.2. Results

The results for quintile share, Gini coefficient and poverty gap are in tables 1, 2 and 3, respectively. Domain size is the most important factor affecting accuracy of estimation in a domain. In general, ARB and RRMSE were largest in small

domains. With a direct estimator and small samples, the estimates vary greatly, and show too large disparities between domains.

Table 1. Results with quintile share in regions (DIS).

Estimator	*ARB (%)*				*RRMSE (%)*			
	Expected domain sample size				*Expected domain sample size*			
	minor	*medium*	*major*	*all*	*minor*	*medium*	*major*	*all*
No contamination								
Direct	4.9	4.6	3.4	4.4	43.5	41.7	38.5	41.3
Ordinary predictor	110.2	125.8	129.6	122.2	113.7	129.4	133.2	125.7
p-adjusted predictor	12.3	8.6	5.7	8.9	16.0	13.6	11.4	13.7
n-calibrated predictor	11.1	13.3	10.6	11.9	31.3	29.6	25.9	29.1
Contaminated								
Direct	7.9	9.1	10.8	9.2	43.8	41.8	39.3	41.7
p-adjusted predictor	14.3	8.5	5.7	9.5	18.1	14.2	12.2	14.8
n-calibrated predictor	10.9	10.3	7.0	9.6	30.6	27.7	23.7	27.5

Table 2. Results for Gini coefficient in regions (DIS).

Estimator	*ARB (%)*				*RRMSE (%)*			
	Expected domain sample size				*Expected domain sample size*			
	minor	*medium*	*major*	*all*	*minor*	*medium*	*major*	*all*
No contamination								
Direct	2.6	2.1	1.6	2.1	12.9	11.6	10.4	11.7
Ordinary predictor	21.7	23.3	23.7	22.9	22.6	24.0	24.4	23.7
p-adjusted predictor	10.7	8.4	7.7	8.9	11.5	9.3	8.7	9.8
n-calibrated predictor	6.2	4.3	3.7	4.7	11.3	9.1	7.8	9.4
Contaminated								
Direct	7.8	8.7	9.8	8.7	21.5	20.6	19.9	20.7
p-adjusted predictor	12.7	10.3	9.6	10.8	13.4	11.1	10.5	11.6
n-calibrated predictor	7.8	6.0	5.5	6.4	12.5	10.1	9.0	10.5

Table 3. Results of poverty gap in domains defined by age class, gender and NUTS2 region.

Estimator	*ARB (%)*				*RRMSE (%)*			
	Expected domain sample size				*Expected domain sample size*			
	minor	*medium*	*major*	*all*	*minor*	*medium*	*major*	*all*
No contamination								
Direct	6.5	3.3	2.4	5.5	51.8	43.8	38.5	48.8
Ordinary predictor	36.9	42.5	39.4	38.2	46.7	46.8	43.2	46.4
p-adjusted predictor	18.1	23.8	22.0	19.6	44.4	37.8	30.7	41.6
n-calibrated predictor	14.1	18.4	20.8	15.7	63.0	51.2	41.4	58.4
Contaminated								
Direct	6.2	3.0	2.4	5.2	51.6	43.7	38.4	48.7
p-adjusted predictor	17.1	23.0	21.3	18.7	44.4	37.6	30.4	41.6
n-calibrated predictor	14.3	17.9	20.1	15.7	63.3	51.3	41.5	58.6

The ordinary predictor was usually so biased that it had even larger RRMSE than the direct estimator. The percentile-adjusted predictor and the n-calibrated predictor benefitted greatly from the transformation (4) bringing the distribution of predictions closer to the distribution of observations. Both bias and RRMSE decreased. The percentile-adjusted predictor had smaller RRMSE than the direct estimator in every domain size class. Moreover, these predictors were more robust to contamination than the direct estimator, although in the case of poverty gap all estimators were robust as the median is not affected by outliers.

The frequency-calibrated estimator (Eq. 6) was not usually as accurate as the p-adjusted predictor. This was expected, as the frequency-calibrated predictor has no access to unit-level information. The n-calibrated predictor was better than the direct estimator of quintile share and Gini coefficient, but not better than the direct estimator of poverty gap. The estimator appears to have similar robustness properties as the percentile-adjusted predictor, since the transformation (4) was applied.

According to our experiments, the method of Molina and Rao (2010) seems to be sensitive to deviations from assumed lognormal distribution of income (results not shown here). In our data, the right tail of the income distribution was "thinner" than expected under the distributional assumption. The domain maxima of the simulated income were often at least ten times larger than in the data and the estimated quintile shares were about 10 % of true values.

6. Discussion

The percentile-adjusted predictors incorporating transformed predictions yielded substantial improvements over current default method and ordinary domain predictor. When the transformation incorporates percentiles of observations up to the 99th percentile, rare outliers occurring with frequency smaller than one percent do not affect the percentile-adjusted predictor too much. The breakdown point of the estimator can probably be adjusted by changing the range of percentage points used in the transformation. The frequency-calibrated estimator (Eq. 6) performed surprisingly well. The estimation of mean square error with bootstrap appears feasible, although time-demanding, as it is necessary to fit the mixed model to each bootstrap sample.

Acknowledgments

This work was completed as a part of the AMELI project, which was supported by European Commission funding from the Seventh Framework Programme for research and technological development. Statistics Finland supported the work of the first author.

REFERENCES

ALFONS, A., TEMPL, M., FILZMOSER, P., KRAFT, S., HULLIGER, B, KOLB, J.-P. and MUNNICH, R., 2011. *Report on outcome of the simulation study.* (Deliverable 6.2 of WP6) [online] Trier: Project AMELI. Available at: <http://svn.uni-trier.de/AMELI>.

CHAMBERS, R. L. and DORFMAN, A. H., 2003. Transformed variables in survey sampling, Working paper M03/21, Southampton Statistical Sciences Research Institute.

CHAMBERS, R. and TZAVIDIS, N., 2006. M-quantile models for small area estimation. *Biometrika*, 93, pp. 255-268.

CHANDRA, H. and CHAMBERS, R., 2011. Small area estimation under transformation to linearity. *Survey Methodology*, 37, pp. 39-51.

CLEMENCEAU, A., J., MUSEUX M. and BAUER, M., 2006. *EU - SILC (Community statistics on Income and Living Conditions): issues and challenges.* [online]. Available at:
<http://www.stat.fi/eusilc/clemenceau_museux_bauer_rev2.pdf>.

European Commission (EUROSTAT Working Group statistics on income, poverty and social exclusion), 2003. *Laeken indicators. Detailed calculation methodology* (DOC. E2/IPSE/2003). Luxembourg.

FABRIZI, E., FERRANTE, M. R. and PACEI, S., 2007. Comparing alternative distributional assumptions in mixed models used for small area estimation of income parameters. *Statistics in Transition*, 8, pp. 423-439.

JIANG, J. and LAHIRI, P., 2006. Mixed model prediction and small area estimation. *Sociedad de Estadistica e Investigacion Operative Test*, 15, pp. 1-96.

KARLBERG, F., 2000. Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, pp. 229-241.

LEHTONEN, R. and VEIJANEN, A., 2009. Design-based methods of estimation for domains and small areas. In: C. R. Rao and Pfeffermann, D., eds. *Handbook of statistics, vol. 29(B). Sample surveys: theory, methods and inference*, pp. 219-249. Oxford: Elsevier.

MOLINA, I. and RAO, J. N. K., 2010. Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38, pp. 369-385.

RAO, J. N. K., 2003. *Small area estimation*. New York: John Wiley & Sons.

SINGH, A. C. and MOHL, C. A., 1996. Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, pp. 107-115.

SÄRNDAL, C.-E., 2007. The calibration approach in survey theory and practice. *Survey Methodology*, 33, pp. 99-119.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J., 1992. *Model assisted survey sampling*. New York: Springer-Verlag.

YOU, Y. and RAO, J. N. K., 2002. A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. *The Canadian Journal of Statistics,* 30, pp. 431-439.

ZASLAVSKY, A. M. and SCHIRM, A. L., 2002. Interactions between survey estimates and federal funding formulas. *Journal of Official Statistics*, 18, pp. 371-391.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 357—380

IMPROVED ESTIMATORS OF COEFFICIENT OF VARIATION IN A FINITE POPULATION

V. Archana[1], K. Aruna Rao[2]

ABSTRACT

Coefficient of Variation (C.V) is a unitless measure of dispersion. Hence it is widely used in many scientific and social investigations. Although a lot of work has been done concerning C.V in the infinite population models, it has been neglected in the finite populations. Many areas of applications of C.V involves the finite populations like the use in official statistics and economic surveys of the World Bank. This has motivated us to propose six new estimators of the population C.V. In finite population studies regression estimators are widely used and the idea is exploited to propose the new estimators. Three of the proposed estimators are the regression estimators of the C.V for the study variable while the other three estimators makes use of the regression estimators of population mean and variance to estimate the ratio $\frac{\sigma_y}{\bar{Y}}$, the population C.V for the study variable. The bias and mean square error (MSE) of these estimators were derived for the simple random sampling design. The performance of these estimators is compared using two real life data sets. The simulation is carried out to compare the estimators in terms of coverage probability and the length of the confidence interval. The small sample comparison indicates that two of the proposed estimators perform better than the sample C.V. The regression estimator using the information on the Population C.V of the auxiliary variable emerges as the best estimator.

Key words: Model based comparison; Coefficient of Variation; Simple Random Sampling; Regression estimator; Mean Square Error; Confidence interval.

[1] Dept. of Statistics, Mangalore University, Mangalagangothri, Konaje, Karnataka,India. E-mail: archana_stkl@yahoo.com.
[2] Dept. of Statistics, Mangalore University, Mangalagangothri, Konaje, Karnataka,India. E-mail: arunaraomu@yahoo.com.

1. Introduction

During the last few years, the Coefficient of variation (C.V) has received the attention of many statisticians, although it was used as a measure of variation by scientists in other disciplines. The C.V is a relative measure of dispersion and is unitless. Thus, it facilitates the comparison of variability measured in different units. Although some investigators prefer the use of standard deviation to co-efficient of variation it is difficult in many instances to draw meaningful conclusions from standard deviation as it is an absolute measure. The coefficient of variation expressed in percentages indicates quickly the extent of variability present in the data. Some of the specific examples include the study of rainfall (Anantha Krishnan and Soman (1989), Business and Engineering (De, Ghosh and Wells (1996)). The C.V is also common in applied probability fields such as renewal theory, queuing theory and reliability theory. The C.V is also used in multiple time scales and the life time (Kordonsky and Gerts bakh.(1997)).

The research on C.V dates back to the work of McKay (1932), Pearson(1932), and Fieller(1932) where they have studied a numerical approximation to the distribution of the sample C.V (in the case of normality). Later on it was extended by Hendrick and Robey (1936) and Koopmans et al.(1964). Nairy and Rao (2003) and the references cited there discusses the various tests for testing the equality of C.V's of independent normal distributions. The research work on the C.V of the normal distribution is fast growing and one of the recent references is that of Mahmoudvand and Hassani (2007) who proposed two new confidence intervals for the C.V in a normal distribution.

Compared to the research work on C.V of the normal distribution, research on C.V of a finite population is of recent origin. The estimation of C.V in finite population was initially discussed by Das and Tripathi (1981a, b). Since then various researchers have attempted the estimation of C.V which include the works of Rajyaguru and Gupta (2002, 2006), Tripathi et al.(2002) , Patel and Shah (2009), among others. Following the idea of Srivastava (1971, 80) and Das and Tripathi (1980), Tripathi et al.(2002) constructed a general class of estimators of C.V. This class of estimators is a hybrid class in the sense that a regression type of estimators is used to construct a general class of ratio estimators of C.V. The ratio/product and regression estimators of C.V constructed from the sample C.V are members of this class. They also obtained an optimum estimator belonging to this class. In this paper we derive a general expression for the bias and MSE of the regression estimator of any parameter of interest θ_y using information on any parameter η_x of the auxiliary variable. The general expression for the MSE indicates that if we construct a regression estimator using any other estimator including a hybrid estimator of θ_y then the regression estimator thus obtained is more efficient than the hybrid estimator. The focal point of this paper is to compare the performance of seven regression/regression type estimators of C.V

constructed from the sample C.V. In this comparison we have not included the optimum estimators of Tripathi et al.(2002). The reason for this is that, as indicated previously, we can always construct a more efficient regression estimator using this optimum estimator. Such estimators becomes complex in nature compared to the regression estimators based on sample C.V. Tripathi et al.(2002) compared the asymptotic performance of 22 estimators which includes the regression and regression type of estimators using two real life datasets. Patel and Shah (2009) compared the small sample MSE of five estimators of C.V which do not include the simple regression type of estimators based on sample C.V. In the last four decades a lot of papers have appeared on the regression estimators of other parameter of interest (like mean and variance). For some of these works see the references cited in Sahoo et al.(2003), Verma (2008) and Pradhan (2010). This has motivated us to undertake a comprehensive comparison of the regression/regression type of estimators constructed from sample C.V.

As a first step in this direction, we have proposed estimators of finite population C.V when the underlined sampling scheme is Simple Random Sample (with or without replacement). The first estimator is the sample C.V. Six new estimators are also proposed in the paper. Three of them are the regression estimators using the information on population C.V, mean and variance of an auxiliary variable. Other three estimators are ratio type regression estimators, where regression estimators are used for the estimation of population mean, and variance using information on the auxiliary variable. Bias and mean square error (MSE) of these estimators are derived to the order of $O(n^{-1})$. Since simulations based on a real life data setting cannot cover a wide variety of complexities regarding the performance of the estimators, we have resorted to model based comparison of the regression estimators. Using bivariate normal distribution the performance of the estimators is compared using i) small sample MSE, ii) coverage probability and iii) average length of the confidence interval. Extensive simulation is carried out covering a wide range of the correlation co-efficient between the study and auxiliary variable and various choice of the C.V of the auxiliary variable. Two of the six new estimators of the C.V perform better compared to the sample C.V. The regression estimator using the information on the C.V of the auxiliary variable emerges as the best estimator.

The organization of the paper is as follows. Section 2 presents the general expressions for bias and MSE of the regression estimators to the order of $O(n^{-1})$. The results are used to derive the bias and MSE of regression estimators of C.V constructed using sample C.V under simple random sampling. Comparisons of the asymptotic performance of the seven estimators are considered in section 3 using two real life data-sets. Section 4 deals with the small sample performance of these estimators. The final conclusions are presented in section 5.

2. General expression for the bias and MSE of the regression estimators.

Theorem 2.1: Let θ_y be the parameter of interest of the study variable 'y' to be estimated and let η_x denote a parameter of the auxiliary variable 'x'. Let $\hat{\theta}_y$ and $\hat{\eta}_x$ be their unbiased estimators then the regression estimator of θ_y is given by

$\hat{\theta}_{y_{\mathrm{Re}g}} = \hat{\theta}_{\mathrm{y}} + \hat{\beta}(\eta_{\mathrm{x}} - \hat{\eta}_x)$, where β denotes the regression co-efficient of $\hat{\theta}_y$ on $\hat{\eta}_x$ and is given by,

$$\beta = \frac{\mathrm{Cov}(\hat{\theta}_{\mathrm{y}}, \hat{\eta}_x)}{V(\hat{\eta}_x)}$$

and $\hat{\beta}$ denotes an asymptotically unbiased estimator of β then the bias and MSE of the regression estimator $\hat{\theta}_{y_{\mathrm{Re}g}}$ to the order of $O(\frac{1}{n})$ is given by

$$B(\hat{\theta}_{y_{\mathrm{Re}g}}) = -\mathrm{Cov}(\hat{\eta}_{\mathrm{x}}, \hat{\beta}) + o\left(\frac{1}{n}\right) \tag{2.1}$$

and

$$M(\hat{\theta}_{y_{\mathrm{Re}g}}) = \mathrm{V}(\hat{\theta}_{\mathrm{y}})(1-\rho^2) + o\left(\frac{1}{n}\right) \tag{2.2}$$

Proof: Using the ε - approach we have,

Let us take $\hat{\theta}_y = \theta_y(1+\varepsilon_1)$, $\hat{\eta}_{\mathrm{x}} = \eta_x(1+\varepsilon_2)$ and $\hat{\beta} = \beta(1+\varepsilon_3)$

then

$$\begin{aligned}\hat{\theta}_{y_{\mathrm{Re}g}} &= \hat{\theta}_y + \hat{\beta}(\eta_x - \hat{\eta}_x) \\ &= \theta_{\mathrm{y}}(1+\varepsilon_1) + \beta(1+\varepsilon_3)(\eta_x - \eta_x(1+\varepsilon_2)) \\ &= \theta_{\mathrm{y}}(1+\varepsilon_1) - \beta\eta_x(1+\varepsilon_3)\varepsilon_2\end{aligned}$$

Now, if we take expectations on both sides we get

$$B(\hat{\theta}_{y_{\mathrm{Re}g}}) = E(\hat{\theta}_{y_{\mathrm{Re}g}} - \theta_y) = -\beta\eta_x E(\varepsilon_2\varepsilon_3) \qquad \because E(\varepsilon_1) = 0 \text{ and } E(\varepsilon_2) = 0$$

$$= -\beta\eta_x \frac{Cov(\hat{\eta}_x, \hat{\beta})}{\beta\eta_x}$$

$$= -Cov(\hat{\eta}_x, \hat{\beta})$$

Similarly the variance of $\hat{\theta}_{y_{Reg}}$ is given by

$$M(\hat{\theta}_{y_{Reg}}) = E(\hat{\theta}_{y_{Reg}} - \theta_y)^2 = \theta_y{}^2 \mathrm{V}(\varepsilon_1) + \beta^2 \eta_x^2 \mathrm{V}(\varepsilon_2) - 2\theta_y \beta \eta_x Cov(\varepsilon_1, \varepsilon_2)$$

$$\because \ \mathrm{V}(\varepsilon_2 \varepsilon_3) \text{ is of order } O\left(\frac{1}{n^2}\right).$$

$$= \theta_y{}^2 \frac{\mathrm{V}(\hat{\theta}_y)}{\theta_y{}^2} + \beta^2 \eta_x^2 \frac{\mathrm{V}(\hat{\eta}_x)}{\eta_x^2} - 2\theta_y \beta \eta_x \frac{Cov(\hat{\theta}_y, \hat{\eta}_x)}{\theta_y \eta_x}$$

$$= \mathrm{V}(\hat{\theta}_y) - 2\beta Cov(\hat{\theta}_y, \hat{\eta}_x) + \beta^2 V(\hat{\eta}_x)$$

But $\beta = \dfrac{\mathrm{Cov}(\hat{\theta}_y, \hat{\eta}_x)}{V(\hat{\eta}_x)}$. If we substitute β in the above equation we get

$$\mathrm{M}(\hat{\theta}_{y_{Reg}}) = \mathrm{V}(\hat{\theta}_y) - \frac{2Cov^2(\hat{\theta}_y, \hat{\eta}_x)}{\mathrm{V}(\hat{\eta}_x)} + \frac{Cov^2(\hat{\theta}_y, \hat{\eta}_x)}{\mathrm{V}(\hat{\eta}_x)}$$

$$= \mathrm{V}(\hat{\theta}_y) - \frac{Cov^2(\hat{\theta}_y, \hat{\eta}_x)}{\mathrm{V}(\hat{\eta}_x)}$$

$$= \mathrm{V}(\hat{\theta}_y)(1 - \rho^2)$$

Hence, the proof follows.

Remark:

1. The expression for MSE to the order of $O(\frac{1}{n})$ does not change if we replace the unbiased estimators $\hat{\theta}_y$ and $\hat{\eta}_x$ by their asymptotically unbiased estimators of θ_y and η_x.

2. From the expression of MSE in (2.2), it becomes clear that if a regression estimator is constructed from an optimum estimator belonging to another class of estimators, this regression estimator is more efficient compared to the optimum estimator, although the decrease in the MSE may not be substantial.

2.2 SRSWR

In the sequel, we present the new estimators of C.V along with Bias and MSE.

2.2.1 Usual estimators $\left(\hat{\theta}_{y_1}\right)$:-

In theory of sampling it is customary to denote the study variable by 'y' and the auxiliary variable 'x'. Let $\overline{Y}$ and $\sigma_y{}^2$ denote the population mean and population variance for the study variable.

In the following, σ_y^2 is also denoted σ_{yy} so as to generalize the notations for the higher order moments of the study and auxiliary variables. The primary focus of interest is to estimate $\theta_y = \frac{\sigma_y}{\bar{Y}}$. The usual estimator is obtained by using the sample mean and sample standard deviation as an estimators of the denominator and the numerator respectively. It is given by

$$\hat{\theta}_{y_1} = \frac{s_y}{\bar{y}}. \tag{2.3}$$

where $\bar{y}$ is the sample mean and s_y^2 is the sample variance and are given by

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}.$$

Further, the bias and MSE of $\hat{\theta}_{y_1}$ are given by

$$bias\left(\hat{\theta}_{y_1}\right) = E\left(\hat{\theta}_{y_1} - \theta_{y_1}\right) = \left\{ -\frac{1}{8n}\frac{\sigma_{yyyy}}{\bar{Y}\sigma_y^3} + \frac{1}{8n}\frac{(n-3)}{(n-1)}\frac{(\sigma_{yy})^2}{\bar{Y}\sigma_y^3} + \frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^3 - \frac{1}{2n}\frac{\sigma_{yyy}}{\sigma_y \bar{Y}^2} \right\} + O\left(\frac{1}{n^2}\right), \tag{2.4}$$

and

$$MSE\left(\hat{\theta}_{y_1}\right) = E\left(\hat{\theta}_{y_1} - \theta_{y_1}\right)^2 = \left\{ \frac{1}{4n}\frac{\sigma_{yyyy}}{\sigma_y^2 \bar{Y}^2} - \frac{(n-3)}{4n(n-1)}\frac{(\sigma_{yy})^2}{\sigma_y^2 \bar{Y}^2} + \frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^4 - \frac{1}{n}\left(\frac{\sigma_{yyy}}{\bar{Y}^3}\right) \right\} + O\left(\frac{1}{n^2}\right). \tag{2.5}$$

2.2.2. Estimator $\left(\hat{\theta}_{y_2}\right)$

Let $\hat{\bar{Y}}_R$ denote the regression estimator of $\bar{Y}$. It is given by

$$\hat{\bar{Y}}_R = \bar{y} + b_1\left(\bar{X} - \bar{x}\right). \tag{2.6}$$

where b_1 is the estimator of B_1, the regression co-efficient of $\bar{y}$ on $\bar{x}$ and is given by,

$$B_1 = \frac{E(\bar{x} - \bar{X})(\bar{y} - \bar{Y})}{E(\bar{x} - \bar{X})^2} = \frac{\sigma_{xy}}{\sigma_{xx}}. \tag{2.7}$$

The estimator b_1 is obtained by substituting the corresponding sample moments in (2.6).

Using this regression estimator of $\overline{Y}$, the second estimator of $\hat{\theta}_{y_1}$ is given by

$$\hat{\theta}_{y_2} = \frac{s_y}{\overline{y} + b_1\left(\overline{X} - \overline{x}\right)}. \tag{2.8}$$

The bias and MSE of $\hat{\theta}_{y_2}$ are given by

$$bias\left(\hat{\theta}_{y_2}\right) = \left\{ -\frac{1}{8n}\frac{\sigma_{yyyy}}{{\sigma_y}^3\overline{Y}} + \frac{(n-3)}{8n(n-1)}\frac{\left(\sigma_{yy}\right)^2}{{\sigma_y}^3\overline{Y}} + \frac{1}{n}\left(\frac{\sigma_y}{\overline{Y}}\right)^3 + \frac{{B_1}^2{\sigma_x}^2\sigma_y}{n\overline{Y}^3} \right.$$
$$\left. -\frac{1}{2n}\frac{\sigma_{yyy}}{\sigma_y\overline{Y}^2} + \frac{B_1}{2n}\frac{\sigma_{xyy}}{\sigma_y\overline{Y}^2} - \frac{2B_1}{n}\frac{\sigma_{xy}\sigma_y}{\overline{Y}^3} \right\} + O\left(\frac{1}{n^2}\right), \tag{2.9}$$

and

$$MSE\left(\hat{\theta}_{y_2}\right) = \frac{{B_1}^2}{n}\frac{{\sigma_x}^2{\sigma_y}^2}{\overline{Y}^4} + \frac{1}{n}\left(\frac{\sigma_y}{\overline{Y}}\right)^4 + \frac{1}{4n}\frac{\sigma_{yyyy}}{{\sigma_y}^2\overline{Y}^2} - \frac{(n-3)}{4n(n-1)}\frac{\left(\sigma_{yy}\right)^2}{{\sigma_y}^2\overline{Y}^2} - \frac{1}{n}\frac{\sigma_{yyy}}{\overline{Y}^3}$$
$$\left. + \frac{B_1}{n}\frac{\sigma_{xyy}}{\overline{Y}^3} - \frac{2B_1}{n}\frac{\sigma_{xy}{\sigma_y}^2}{\overline{Y}^4} \right\} + O\left(\frac{1}{n^2}\right). \tag{2.10}$$

2.2.3 Estimator $\left(\hat{\theta}_{y_3}\right)$.

Standard textbooks in theory of sampling do not discuss the estimations of the population variance. Thus, regression estimators for the estimation of σ_y^2 (σ_{yy}) are not available in these textbooks. Following the discussion of the regression estimators for the population mean, we propose the regression estimator of ${\sigma_{Y_R}}^2$ as

$${\hat{\sigma}_{Y_R}}^2 = {s_y}^2 + b_2\left({\sigma_x}^2 - {s_x}^2\right). \tag{2.11}$$

where b_2 is the estimator of B_2 , the regression coefficient of s_y^2 on s_x^2 which is given by

$$B_2 = \frac{E\left(s_{yy} - \sigma_{yy}\right)\left(s_{xx} - \sigma_{xx}\right)}{E\left(s_{xx} - \sigma_{xx}\right)^2} = \frac{\sigma_{xxyy} - \sigma_{xx}\sigma_{yy}}{\sigma_{xxxx} - \left(\sigma_{xx}\right)^2} + O\left(\frac{1}{n}\right). \tag{2.12}$$

The estimator b_2 is obtained by substituting the corresponding sample moments in (2.11).

Using this, the estimator $\hat{\theta}_{y_3}$ is proposed and is given by

$$\hat{\theta}_{y_3} = \frac{\left[s_y{}^2 + b_2\left(\sigma_x{}^2 - s_x{}^2\right)\right]^{\frac{1}{2}}}{\bar{y}}. \tag{2.13}$$

Further, bias and MSE for $\hat{\theta}_{y_3}$ are given as

$$bias\left(\hat{\theta}_{y_3}\right) = \left\{-\frac{1}{8n}\frac{\sigma_{yyyy}}{\sigma_y{}^3\bar{Y}} + \frac{(n-3)}{8n(n-1)}\frac{(\sigma_{yy})^2}{\sigma_y{}^3\bar{Y}} - \frac{B_2{}^2}{8n}\frac{\sigma_{xxxx}}{\sigma_y{}^3\bar{Y}} + \frac{B_2{}^2}{8n}\frac{(n-3)}{(n-1)}\frac{(\sigma_{xx})^2}{\sigma_y{}^3\bar{Y}}\right. \tag{2.14}$$
$$\left. - \frac{1}{2n}\frac{\sigma_{yyy}}{\sigma_y\bar{Y}^2} + \frac{B_2}{2n}\frac{\sigma_{xxy}}{\sigma_y\bar{Y}^2} + \frac{B_2}{4n}\frac{\sigma_{xxyy}}{\sigma_y{}^3\bar{Y}} - \frac{B_2}{4n}\frac{\sigma_{xx}\sigma_{yy}}{\sigma_y{}^3\bar{Y}}\right\} + O\left(\frac{1}{n^2}\right),$$

and

$$MSE\left(\hat{\theta}_{y_3}\right) = \left\{\frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^4 + \frac{1}{4n}\frac{\sigma_{yyyy}}{\bar{Y}^2\sigma_y{}^2} - \frac{(n-3)}{4n(n-1)}\frac{(\sigma_{yy})^2}{\bar{Y}^2\sigma_y{}^2} + \frac{1}{4n}\frac{\sigma_{xxxx}B_2{}^2}{\sigma_y{}^2\bar{Y}^2}\right.$$
$$\left. - \frac{1}{4n}\frac{(n-3)}{(n-1)}\frac{(\sigma_{xx})^2 B_2{}^2}{\sigma_y{}^2\bar{Y}^2} - \frac{1}{n}\frac{\sigma_{yyy}}{\bar{Y}^3} + \frac{1}{n}\frac{\sigma_{xxy}B_2}{\bar{Y}^3} - \frac{B_2}{2n}\frac{\sigma_{xxyy}}{\sigma_y{}^2\bar{Y}^2} + \frac{B_2}{2n}\frac{\sigma_{xx}\sigma_{yy}}{\sigma_y{}^2\bar{Y}^2}\right\} + O\left(\frac{1}{n^2}\right). \tag{2.15}$$

2.2.4 Estimator $\left(\hat{\theta}_{y_4}\right)$

This estimator is obtained by using regression estimators of σ_{yy} and $\bar{Y}$ respectively and is given by

$$\hat{\theta}_{y_4} = \frac{\left[s_y{}^2 + b_2\left(\sigma_x{}^2 - s_x{}^2\right)\right]^{\frac{1}{2}}}{\left[\bar{y} + b_1\left(\bar{X} - \bar{x}\right)\right]}. \tag{2.16}$$

The expressions for the bias and MSE of $\hat{\theta}_{y_4}$ are given by

$$bias\left(\hat{\theta}_{y_4}\right) = \left\{\frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^3 + \frac{B_1{}^2\sigma_x{}^2\sigma_y}{n\bar{Y}^3} - \frac{B_2{}^2}{8n}\frac{\sigma_{xxxx}}{\sigma_y{}^3\bar{Y}} + \frac{B_2{}^2}{8n}\frac{(n-3)}{(n-1)}\frac{(\sigma_{xx})^2}{\sigma_y{}^3\bar{Y}} - \frac{1}{8n}\frac{\sigma_{yyyy}}{\sigma_y{}^3\bar{Y}} + \frac{1}{8n}\frac{(n-3)}{(n-1)}\frac{(\sigma_{yy})^2}{\sigma_y{}^3\bar{Y}}\right.$$
$$\left. - \frac{2B_1\sigma_{xy}\sigma_y}{n\bar{Y}^3} - \frac{1}{2n}\frac{\sigma_{yyy}}{\sigma_y\bar{Y}^2} + \frac{B_2}{2n}\frac{\sigma_{xxy}}{\sigma_y\bar{Y}^2} + \frac{B_1}{2n}\frac{\sigma_{xyy}}{\sigma_y\bar{Y}^2} - \frac{B_1B_2}{2n}\frac{\sigma_{xxx}}{\sigma_y\bar{Y}^2} + \frac{B_2}{4n}\frac{\sigma_{xxyy}}{\sigma_y{}^3\bar{Y}} - \frac{B_2}{4n}\frac{\sigma_{xx}\sigma_{yy}}{\sigma_y{}^3\bar{Y}}\right\} + O\left(\frac{1}{n^2}\right), \tag{2.17}$$

and

$$MSE\left(\hat{\theta}_{y_4}\right)=E\left(\hat{\theta}_{y_4}-\theta_{y_4}\right)^2=\left\{\frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^4+\frac{\sigma_x^2\sigma_y B_1^2}{n\bar{Y}^4}+\frac{1}{4n}\frac{\sigma_{yyyy}}{\sigma_y^2\bar{Y}^2}-\frac{(n-3)}{4n(n-1)}\frac{(\sigma_{yy})^2}{\sigma_y^2\bar{Y}^2}+\frac{B_2^2}{4n}\frac{\sigma_{xxxx}}{\sigma_y^2\bar{Y}^2}-\frac{(n-3)}{4n(n-1)}\frac{(\sigma_{xx})^2B_2^2}{\sigma_y^2\bar{Y}^2}\right.$$

$$\left.-\frac{2B_1}{n}\frac{\sigma_{xy}\sigma_y^2}{\bar{Y}^4}-\frac{1}{n}\frac{\sigma_{yyy}}{\bar{Y}^3}+\frac{B_2}{n}\frac{\sigma_{xxy}}{\bar{Y}^3}+\frac{B_1}{n}\frac{\sigma_{xyy}}{\bar{Y}^3}-\frac{B_1B_2}{n}\frac{\sigma_{xxx}}{\bar{Y}^3}-\frac{B_2}{2n}\frac{\sigma_{xxyy}}{\sigma_y^2\bar{Y}^2}+\frac{B_2}{2n}\frac{\sigma_{xx}\sigma_{yy}}{\sigma_y^2\bar{Y}^2}\right\}+O\left(\frac{1}{n^2}\right). \tag{2.18}$$

2.2.5. Regression estimator $\left(\hat{\theta}_{y_5}\right)$

The preceding 3 estimators for population C.V consists of estimating the ratio $\frac{\sigma_y}{\bar{Y}}$ by using improved estimators for the numerator and denominator. We now propose regression estimator for this ratio. The basic logic is the same as in the case of estimation of the mean and some details are omitted. The regression estimator $\hat{\theta}_{y_5}$ is useful when we have the knowledge on the population mean $\bar{X}$ of the auxiliary variable. It is given by

$$\hat{\theta}_{y_5}=\frac{s_y}{\bar{y}}+b_3\left(\bar{X}-\bar{x}\right). \tag{2.19}$$

where b_3 is the estimate of B_3 which is given by

$$B_3=\frac{Cov\left(\frac{s_y}{\bar{y}},\bar{x}\right)}{V\left(\bar{x}\right)}=\frac{\frac{\sigma_{xyy}}{2\bar{Y}\sigma_y}-\frac{\sigma_{xy}\sigma_y}{\bar{Y}^2}}{\sigma_{xx}}+O\left(\frac{1}{n}\right). \tag{2.20}$$

Further, the bias and MSE of $\hat{\theta}_{y_5}$ are given by

$$bias\left(\hat{\theta}_{y_5}\right)=\left\{\frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^3-\frac{1}{8n}\frac{\sigma_{yyyy}}{\sigma y^3\bar{Y}}+\frac{(n-3)}{8n(n-1)}\frac{(\sigma_{yy})^2}{\sigma_y^3\bar{Y}}-\frac{1}{2n}\frac{\sigma_{yyyy}}{\bar{Y}^2}\right\}+O\left(\frac{1}{n^2}\right), \tag{2.21}$$

and

$$MSE\left(\hat{\theta}_{y_5}\right)=\left\{\frac{1}{n}\sigma_x^2B_3^2+\frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^4+\frac{1}{4n}\frac{\sigma_{yyyy}}{\sigma_y^2\bar{Y}^2}-\frac{(n-3)}{4n(n-1)}\frac{(\sigma_{yy})^2}{\sigma_y^2\bar{Y}^2}+\frac{2}{n}\frac{\sigma_{xy}B_3\sigma_y}{\bar{Y}^2}-\frac{1}{n}\frac{\sigma_{xyy}B_3}{\sigma_y\bar{Y}}-\frac{1}{n}\frac{\sigma_{yyy}}{\bar{Y}^3}\right\}+O\left(\frac{1}{n^2}\right). \tag{2.22}$$

2.2.6. Regression estimator ($\hat{\theta}_{y_6}$)

This estimator is useful when the information for the variance of the auxiliary variable is known and is given by,

$$\hat{\theta}_{y_6} = \frac{s_y}{\bar{y}} + b_4\left(\sigma_x{}^2 - s_x{}^2\right). \tag{2.23}$$

where b_4 is the estimate of B_4 which is given by

$$B_4 = \frac{Cov\left(\frac{s_y}{\bar{y}}, s_x{}^2\right)}{Var\left(s_x{}^2\right)} = \left\{\frac{\frac{\sigma_{xxyy}}{2\bar{Y}\sigma_y} - \frac{\sigma_{xx}\sigma_{yy}}{2\bar{Y}\sigma_y} - \frac{\sigma_{xxy}\sigma_y}{\bar{Y}^2}}{\sigma_{xxxx} - \left(\sigma_{xx}\right)^2}\right\} + O\left(\frac{1}{n}\right). \tag{2.24}$$

Further, the bias and MSE for $\hat{\theta}_{y_6}$ are given by

$$bias\left(\hat{\theta}_{y_6}\right) = \left\{-\frac{1}{8n}\frac{\sigma_{yyyy}}{\bar{Y}\sigma_y{}^3} + \frac{(n-3)}{8n(n-1)}\frac{\left(\sigma_{yy}\right)^2}{\bar{Y}\sigma_y{}^3} + \frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^3 - \frac{1}{2n}\frac{\sigma_{yyy}}{\bar{Y}^2\sigma_y}\right\} + O\left(\frac{1}{n^2}\right), \tag{2.25}$$

and

$$MSE\left(\hat{\theta}_{y_6}\right) = \left\{\frac{1}{4n}\frac{\sigma_{yyyy}}{\sigma_y{}^2\bar{Y}^2} - \frac{(n-3)}{4n(n-1)}\frac{\left(\sigma_{yy}\right)^2}{\sigma_y{}^2\bar{Y}^2} + \frac{B_4{}^2}{n}\sigma_{xxxx} - \frac{(n-3)}{n(n-1)}B_4{}^2\left(\sigma_{xx}\right)^2 + \frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^4 - \frac{B_4}{n}\left(\frac{\sigma_{xxyy}}{\sigma_y\bar{Y}}\right) + \frac{B_4}{n}\frac{\sigma_{xx}\sigma_{yy}}{\sigma_y\bar{Y}} - \frac{1}{n}\frac{\sigma_{yyy}}{\bar{Y}^3} + \frac{2}{n}\frac{\sigma_{xxy}\sigma_y B_4}{\bar{Y}^2}\right\} + O\left(\frac{1}{n^2}\right). \tag{2.26}$$

2.2.7. Regression estimator $\left(\hat{\theta}_{y_7}\right)$

This regression estimator of the population C.V of the study variable uses the population C.V of the auxiliary variable. In many instances, although the information on population mean or variance of the auxiliary variable is not known, it is likely that the information on the population C.V of the auxiliary variable may be known. This is especially true with respect to sampling of forests, agricultural fields etc. The estimator is given by

$$\hat{\theta}_{y_7} = \frac{s_y}{\bar{y}} + b_5\left(\frac{\sigma_x}{\bar{X}} - \frac{s_x}{\bar{x}}\right), \tag{2.27}$$

where b_5 is the estimate of B_5 which is given by

$$B_5 = \frac{Cov\left(\frac{s_y}{\bar{y}}, \frac{s_x}{\bar{x}}\right)}{Var\left(\frac{s_x}{\bar{x}}\right)}$$

$$= \left\{ \frac{\frac{\sigma_{xxyy}}{4\bar{X}\bar{Y}\sigma_x\sigma_y} - \frac{\sigma_{xx}\sigma_{yy}}{4\bar{X}\bar{Y}\sigma_x\sigma_y} - \frac{\sigma_{xyy}\sigma_x}{2\sigma_y\bar{Y}\bar{X}^2} - \frac{1}{2}\frac{\sigma_{xxy}\sigma_y}{\sigma_x\bar{X}\bar{Y}^2} + \frac{\sigma_{xy}\sigma_x\sigma_y}{\bar{X}^2\bar{Y}^2}}{\frac{1}{4}\frac{\sigma_{xxxx}}{{\sigma_x}^2\bar{X}^2} - \frac{(\sigma_{xx})^2}{4\bar{X}^2{\sigma_x}^2} + \left(\frac{\sigma_x}{\bar{X}}\right)^4 - \frac{\sigma_{xxx}}{\bar{X}^3}} \right\} + O\left(\frac{1}{n}\right), \tag{2.28}$$

The expressions for the bias and MSE of $\hat{\theta}_{y_7}$ are given by

$$bias\left(\hat{\theta}_{y_7}\right) = \left\{ \frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^3 - \frac{1}{n}\left(\frac{\sigma_x}{\bar{X}}\right)^3 B_5 - \frac{1}{8n}\frac{\sigma_{yyyy}}{{\sigma_y}^3\bar{Y}} + \frac{1}{8n}\frac{(n-3)(\sigma_{yy})^2}{(n-1){\sigma_y}^3\bar{Y}} \right.$$
$$\left. + \frac{B_5}{8n}\frac{\sigma_{xxxx}}{{\sigma_x}^3\bar{X}} - \frac{(n-3)}{8n(n-1)}\frac{(\sigma_{xx})^2}{{\sigma_x}^3\bar{X}}B_5 - \frac{1}{2n}\frac{\sigma_{yyy}}{\sigma_y\bar{Y}^2} + \frac{1}{2n}\frac{\sigma_{xxx}}{\sigma_x\bar{X}^2}B_5 \right\} + O\left(\frac{1}{n^2}\right), \tag{2.29}$$

and

$$MSE\left(\hat{\theta}_{y_7}\right) = \left\{ \frac{1}{n}\left(\frac{\sigma_y}{\bar{Y}}\right)^4 + \frac{1}{n}\left(\frac{\sigma_x}{\bar{X}}\right)^4 B_5^2 + \frac{1}{4n}\frac{\sigma_{yyyy}}{\sigma_y^2\bar{Y}^2} - \frac{(n-3)}{4n(n-1)}\frac{(\sigma_{yy})^2}{\sigma_y^2\bar{Y}^2} + \frac{B_5^2\sigma_{xxxx}}{4n\sigma_x^2\bar{X}^2} - \frac{(n-3)}{4n(n-1)}\frac{B_5^2(\sigma_{xx})^2}{\sigma_x^2\bar{X}^2} \right.$$
$$\left. - \frac{2B_5\sigma_{xy}\sigma_x\sigma_y}{n\bar{X}^2\bar{Y}^2} - \frac{1}{n}\frac{\sigma_{yyy}}{\bar{Y}^3} + \frac{B_5}{n}\frac{\sigma_{xyy}\sigma_y}{\sigma_x\bar{X}\bar{Y}^2} + \frac{B_5}{n}\frac{\sigma_{xyy}\sigma_x}{\sigma_y\bar{X}^2\bar{Y}} - \frac{B_5^2\sigma_{xxx}}{n\bar{X}^3} - \frac{B_5}{2n}\frac{\sigma_{xxyy}}{\sigma_x\sigma_y\bar{X}\bar{Y}} + \frac{B_5}{2n}\frac{\sigma_{xx}\sigma_{yy}}{\sigma_x\sigma_y\bar{X}\bar{Y}} \right\} + O\left(\frac{1}{n^2}\right). \tag{2.30}$$

The Bias and MSE of these estimators are derived to the order of $O(n^{-1})$ by the authors using Taylor series expansion and higher order moments of sample mean and variance. These moments to the order of $O(n^{-1})$ are derived by the author and are given in Appendix A for the case of SRSWR and Appendix B for SRSWOR. The expression for the MSE of $\hat{\theta}_{y_1}$ coincides with the expressions derived by Kendall and Stuart (1977: p248) for the infinite population model. Thus, we notice that to the order of $O(1)$, $\hat{\theta}_{y_1}$ is unbiased.

2.3. SRSWOR

In the case of SRSWR, the sample variance s_y^2 is an estimate of the population variance σ_y^2. However, for SRSWOR design s_y^2 is an estimate of $S_y^2 = \frac{1}{N-1}\sum\left(Y_i - \bar{Y}\right)^2$.

Therefore, we define the population C.V for the study variable (y) as $\theta_y = \frac{S_y}{\bar{Y}}$.

The population C.V for the auxiliary variable *'x'* is similarly defined.

2.3.1. Usual Estimator $\left(\hat{\theta}_{y_1}\right)$

Following the same discussion for the case of SRSWOR, the usual estimator is given by

$$\hat{\theta}_{y_1} = \frac{s_y}{\bar{y}}, \tag{2.31}$$

where $\bar{y}$ is the sample mean and s_y^2 is the sample variance respectively. To simplify notations, no separate subscript is used in the case of SRSWOR and the context will make it clear whether the reference is WR or WOR schemes.
Further, the bias and MSE of $\hat{\theta}_{y_1}$ are given by

$$bias\left(\hat{\theta}_{y_1}\right) = \left\{-\frac{1}{8}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{yyyy}}{\bar{Y}S_y^3} + \frac{1}{8}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\left(S_{yy}\right)^2}{\bar{Y}S_y^3} + \left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{S_y}{\bar{Y}}\right)^3 - \frac{1}{2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{yyy}}{S_y\bar{Y}^2}\right\} + O\left(\frac{1}{n^2}\right), \tag{2.32}$$

and

$$MSE\left(\hat{\theta}_{y_1}\right) = \left\{\frac{1}{4}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{yyyy}}{S_y^2\bar{Y}^2} - \frac{1}{4}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\left(S_{yy}\right)^2}{S_y^2\bar{Y}^2} + \left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{S_y}{\bar{Y}}\right)^4 - \left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{s_{yyy}}{\bar{Y}^3}\right)\right\} + O\left(\frac{1}{n^2}\right). \tag{2.33}$$

These expressions of bias and MSE are derived for the usual estimator in the case of SRSWOR by the authors and the expressions for MSE coincides with the expressions derived by Kendall and Stuart (1977) for the infinite population model, when $N \to \infty$.

2.3.2. Estimator $\left(\hat{\theta}_{y_2}\right)$

Using the regression estimator $\hat{\bar{Y}}_R = \bar{y} + b_1\left(\bar{X} - \bar{x}\right)$, the second estimator of θ_y is given by

$$\hat{\theta}_{y_2} = \frac{s_y}{\bar{y} + b_1\left(\bar{X} - \bar{x}\right)}. \tag{2.34}$$

The estimate of b_1 is B_1 which is given by $B_1 = \frac{S_{xy}}{S_{xx}}$. (2.35)

The bias and MSE of $\hat{\theta}_{y_2}$ is given by

$$bias\left(\hat{\theta}_{y_2}\right) = \left\{-\frac{1}{8}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{yyyy}}{{S_y}^3\bar{Y}} + \frac{1}{8}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\left(S_{yy}\right)^2}{{S_y}^3\bar{Y}} + \left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{S_y}{\bar{Y}}\right)^3 + \frac{{B_1}^2{S_x}^2S_y}{\bar{Y}^3}\left(\frac{1}{n}-\frac{1}{N}\right)\right.$$
$$\left. -\frac{1}{2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{yyy}}{S_y\bar{Y}^2} + \frac{B_1}{2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{xyy}}{S_y\bar{Y}^2} - \frac{2B_1}{\bar{Y}^3}\left(\frac{1}{n}-\frac{1}{N}\right)S_{xy}S_y\right\} + O\left(\frac{1}{n^2}\right), \tag{2.36}$$

and

$$MSE\left(\hat{\theta}_{y_2}\right) = \left\{\frac{{B_1}^2{S_x}^2{S_y}^2}{\bar{Y}^4}\left(\frac{1}{n}-\frac{1}{N}\right) + \left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{S_y}{\bar{Y}}\right)^4 + \frac{1}{4}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{yyyy}}{{S_y}^2\bar{Y}^2} - \frac{1}{4}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{\left(S_{yy}\right)^2}{{S_y}^2\bar{Y}^2}\right.$$
$$\left. -\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{yyy}}{\bar{Y}^3} + B_1\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{xyy}}{\bar{Y}^3} - 2B_1\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{xy}{S_y}^2}{\bar{Y}^4}\right\} + O\left(\frac{1}{n^2}\right). \tag{2.37}$$

Similarly 5 other estimators of θ_y are also proposed in the case of SRSWOR scheme and the expressions for the Bias and MSE to the order of $O(n^{-1})$ are derived by the authors. To save space those expressions are not presented here and can be obtained by the authors.

3. First order comparison of the performance of the estimators.

From the expressions derived for the bias and MSE it is difficult to identify the estimators which have smaller bias and MSE. Thus, we have considered two data-sets to compare the estimators. They correspond to high value of population C.V and low value of population C.V respectively. The results corresponding to the data sets are described below.

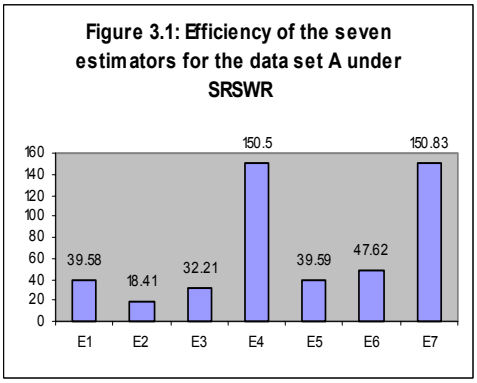
3.1 Data Set (A)

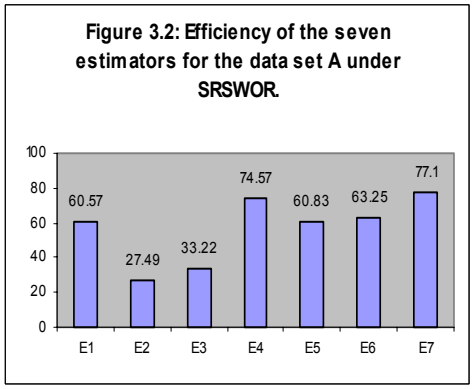
This data set corresponds to payment on motor insurance in 63 geographical regions of Sweden (Source: Swedish Committee on Analysis of Risk Premium in Motor Insurance). The payment for all the claims in thousands of Swedish Kronas is taken as the study variable and the number of claims is taken as the auxiliary variable. Population characteristics for the variables x and y are reported below.

The population C.V's for the variables x and y are 1.01 and 0.88 respectively. The values of skewness (β_1) and kurtosis $(\gamma_2 = \beta_2 - 3)$ for the variable x are 2.32 and 6.85, while for the variable y the respective values are 1.63 and 3.33 respectively. Thus, the distributions for the x and y variables are highly right skewed and peaked. The population correlation coefficient between x and y is 0.91.

The bias and efficiency for the 7 estimators of the C.V (□y) for SRSWR and WOR designs for a sample of size n=20 are represented diagrammatically in Fig(3.1) and Fig(3.2). To save space the table is not reported here. It is clear that for the SRSWR scheme, the estimators having maximum efficiency are the regression estimator θ_{y_7}, where the information on the population C.V of the auxiliary variable is used (Efficiency (E(θ_{y_7})) =150.83), followed by the regression estimator θ_{y_4} where the regression estimators of mean and variance are used to estimate □y with an efficiency of E(θ_{y_4})=150.50. The estimators having the least efficiency is $\theta_y = \frac{\sigma_y}{\bar{Y}}$ where improved estimator for the mean is used in the denominator of the ratio, (E(θ_{y_2})=18.41). The usual estimator θ_{y_1} ranks fifth (E(θ_{y_1})=39.58), when the estimators are ranked in terms of efficiency in the descending order.

Figure (3.1) reflects the comparative performance of the seven estimators for SRSWR scheme represented through bar diagram. A similar conclusion is obtained for the SRSWOR scheme (see Figure (3.2)).





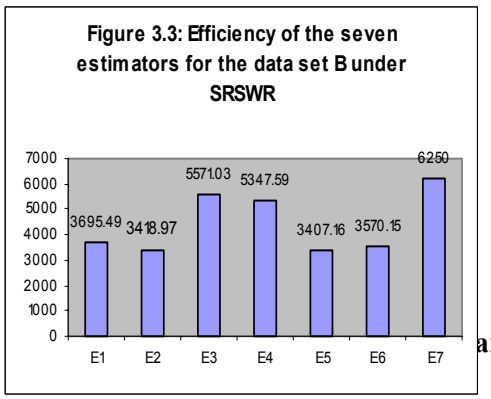
3.2. Data Set (B).

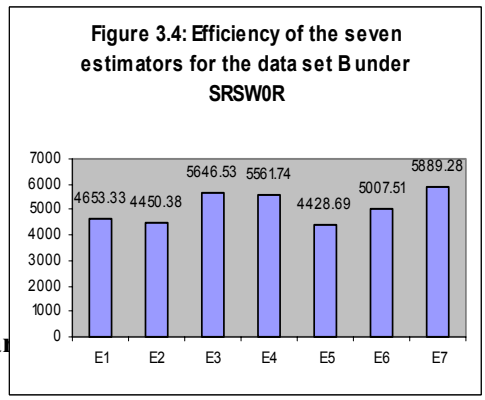
This real life data-set corresponds to cost of living index on grocery items and health care grouped by metropolitan areas of the United States (Source: Statistical Abstract of the United States, 120th edition). The cost of living index for health care is taken as the study variable and the cost living index for the grocery items is taken as the auxiliary variable. Population characteristics for the variables x and y are reported below.

The population C.V for the study variable 'y' is 0.10 and for the auxiliary variable 'x' is 0.074. The distributions of x and y variables are moderately right skewed. The distribution of variable x is mesokurtic $(\gamma_2 = \beta_2 - 3 = 0.047)$ and the distribution of variable y is slightly flat $(\gamma_2 = -0.30)$. The population correlation coefficient for the variables x and y is 0.84. The population size, N=45.

The bias and efficiency of the seven estimators of C.V (θ_y) are graphically represented using bar diagrams in Fig(3.3) and Fig(3.4) for the SRSWR and WOR designs. Since the population size is small the numerical calculation reported corresponds to a sample of size $n = 15$. It follows that for the SRSWR scheme, the maximum efficiency corresponds to the estimator $\hat{\theta}_{y_7}$ where the information on the C.V of the auxiliary variable is used (E($\hat{\theta}_{y_7}$)=6250.00), followed by the regression estimator $\hat{\theta}_{y_3}$ where improved estimator for the variance is used in the numerator of the ratio to estimate θ_y (E($\hat{\theta}_{y_3}$)=5571.03). The estimator having the least efficiency is $\hat{\theta}_{y_5}$ where information on the population mean $\overline{X}$ of the auxiliary variable are used (E($\hat{\theta}_{y_5}$)=3407.16), where as the usual estimator $\hat{\theta}_{y_1}$ ranks fourth, when the estimators are ranked in terms of efficiency in the descending order (E($\hat{\theta}_{y_1}$)=3695.49).

Figure (3.3) reflects the comparative performance of the estimators for SRSWR scheme. Proceeding in the same manner for SRSWOR, a similar conclusion is arrived (see Figure (3.4)).





In the preceding section we have compared the estimators via the asymptotic MSE. In the recent years the performance of the estimators are compared through the coverage probability of the confidence interval constructed from the estimator and the length of the confidence interval (Mahmoudvand and Hassani (2007)). Since the exact distributions of the estimators are difficult to tract analytically, we have carried out a simulation study to achieve the objective. Observations of size 'n' are generated from a bivariate normal distribution with parameters $\left(\mu_1,\sigma_1^2\right)$, $\left(\mu_2,\sigma_2^2\right)$ and ρ. For each sample the confidence interval is constructed using normal approximation to the distribution of the estimator. Each time the length of the confidence interval is also recorded using 10,000 simulations. The coverage probability and the average length of the confidence interval were recorded. Using the 10,000 simulated samples the MSE of the estimators are also computed. The comparison of the estimators through the average length of the confidence interval is valid only when they maintain the confidence level. Failure to maintain the confidence level only indicates that the normal approximation is not accurate to the sampling distribution of the estimators. In such cases the comparison is meaningful through the MSE.

The values of the C.V of the study variable used in the simulations were 0.1, 0.3, 0.5, 0.8, 1.0, and 2.0. For each fixed value of the study variable, a set of 4 values of C.V of the auxiliary variable are considered. They are 0.5, 1.0, 1.5 and 2 times the C.V of the study variable. The correlation co-efficient used in the simulation study are -0.9, -0.7, -0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5, 0.7, 0.9.

The sample sizes considered are n=100, 200. Only the confidence level used in the investigation=0.95.The total no. of configurations works out to be 6*4*11*2=528.

For each sample size and for a fixed value of C.V of the study variable the numerical values of the coverage probability are examined. In the present investigation a confidence interval is said to maintain the confidence level of 0.95, if the coverage probability exceeds 0.90 (approximately 5% error).

For each of the estimator at a fixed value of correlation co-efficient, the mean of the coverage probability is computed for the set of 4 values of C.V of the auxiliary variable, when in 3 or more cases the coverage probability exceeds 0.9. Whenever the confidence level is maintained the mean of the average length of the confidence interval is also obtained. When the estimators fail to maintain the confidence level attention is paid to the values of the MSE. After a careful scrutiny, two best estimators (satisfying the criterion of shortest length of the confidence interval/smaller MSE) are identified. They are $\hat{\theta}_4$ and $\hat{\theta}_7$. These are the best estimators for various configurations of the C.V of the auxiliary variable and the correlation co-efficient. For the other estimators, namely $\hat{\theta}_{y_2}$, $\hat{\theta}_{y_3}$, $\hat{\theta}_{y_5}$ (Regression estimator when information on mean of the auxiliary variable is used) and $\hat{\theta}_{y_6}$ (Regression estimator when information on variance of the auxiliary

variable is used) no consistent pattern regarding the performance either in terms of coverage probability or MSE is emerged for the various configurations and to save space the results are not reported here.

Table 4.1 presents the average coverage probability of the confidence interval and average length of the confidence interval for the ratio type regression estimators ($\hat{\theta}_{y_4}$), the regression estimator ($\hat{\theta}_{y_7}$) where the information on the population C.V of the auxiliary variable is used and the sample C.V ($\hat{\theta}_{y_1}$).

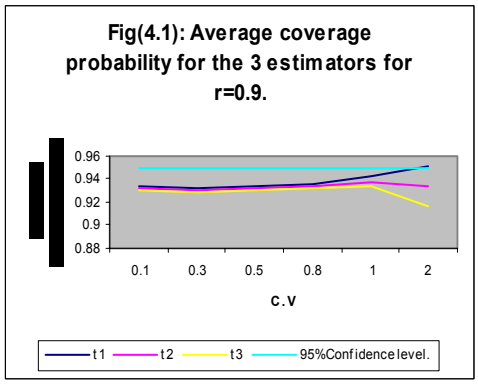
The regression estimators or the ratio type regression estimators do not maintain the coverage probability when the correlation co-efficient 'r' is low and thus the results are presented only when the correlation co-efficient is -0.9, -0.7, -0.5, 0.5, 0.7, 0.9. The result is presented in the table only when the sample size n=100 and the pattern of the results does not change for other sample size=200.

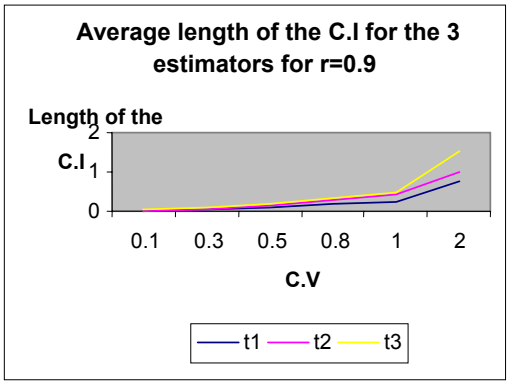
Table 4.1. Average coverage probability (in percentages) and average length of the C.V (in brackets) for α=0.05.

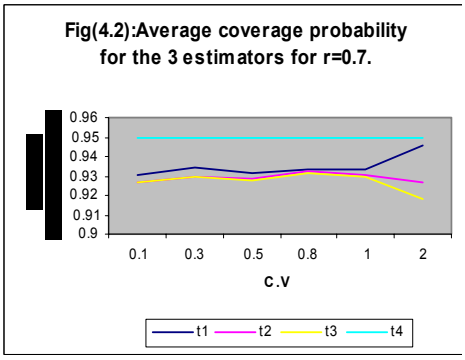
C.V of the study variable.		Correlation co-efficient (r) (n=100)					
		-0.9	-0.7	-0.5	0.5	0.7	0.9
0.1	θ_4	0.9355(0.0159)	0.9315(0.0239)	0.9320(0.0268)	0.9326(0.0261)	0.9302(0.0240)	0.9335(0.0165)
	θ_7	0.9347(0.0182)	0.9257(0.0247)	0.9284(0.0271)	0.9277(0.0268)	0.9271(0.0241)	0.9324(0.0167)
	θ_1	0.9342(0.0276)	0.9241(0.0281)	0.9273(0.0283)	0.9274(0.0287)	0.9269(0.0281)	0.9311(0.0271)
0.3	θ_4	0.9334(0.0505)	0.9327(0.0737)	0.9313(0.0815)	0.9341(0.0819)	0.9343(0.0754)	0.9318(0.0513)
	θ_7	0.9313(0.0764)	0.9293(0.0839)	0.9271(0.0848)	0.9276(0.0821)	0.9298(0.0819)	0.9299(0.0631)
	θ_1	0.9305(0.0895)	0.9278(0.0892)	0.9257(0.0883)	0.9261(0.0875)	0.9292(0.0883)	0.9287(0.0892)
0.5	θ_4	0.9342(0.0897)	0.9344(0.1346)	0.9341(0.1569)	0.9338(0.1553)	0.9319(0.1376)	0.9337(0.0912)
	θ_7	0.9322(0.1592)	0.9314(0.1603)	0.9318(0.1638)	0.9290(0.1597)	0.9283(0.1469)	0.9323(0.1343)
	θ_1	0.9305(0.1671)	0.9287(0.1678)	0.9307(0.1680)	0.9288(0.1686)	0.9278(0.1680)	0.9309(0.1673)
0.8	θ_4	0.9349(0.1620)	0.9344(0.2516)	0.9330(0.2909)	0.9326(0.2929)	0.9334(0.2526)	0.9359(0.1676)
	θ_7	0.9333(0.3081)	0.9330(0.3099)	0.9326(0.3106)	0.9309(0.2989)	0.9323(0.2643)	0.9341(0.3026)
	θ_1	0.9307(0.3253)	0.9294(0.3248)	0.9299(0.3256)	0.9287(0.3258)	0.9311(0.3253)	0.9323(0.3252)
1.0	θ_4	0.9376(0.2224)	0.9379(0.3696)	0.9364(0.4280)	0.9330(0.4223)	0.9329(0.3643)	0.9418(0.2278)
	θ_7	0.9370(0.4458)	0.9315(0.4596)	0.9316(0.4636)	0.9304(0.4323)	0.9302(0.3965)	0.9376(0.4478)
	θ_1	0.9329(0.4762)	0.9304(0.4795)	0.9294(0.4786)	0.9301(0.4780)	0.9292(0.4786)	0.9335(0.4769)
2.0	θ_4	0.9594(0.7388)	0.9593(1.1144)	0.9498(1.3842)	0.9398(1.4031)	0.9458(1.2025)	0.9516(0.7778)
	θ_7	0.9343(0.9945)	0.9275(1.4393)	0.9219(1.4998)	0.9202(1.4893)	0.9268(1.4122)	0.9341(0.9987)
	θ_1	0.9193(1.5441)	0.9200(1.5481)	0.9189(1.5463)	0.9176(1.5470)	0.9179(1.5471)	0.9161(1.5468)

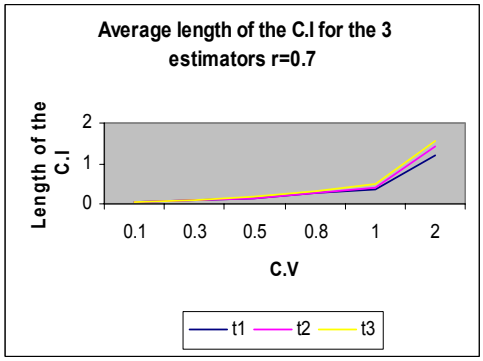
From the table it is clear that for the 2 estimators the coverage probability did not change either with the values of the correlation coefficient or with the values of C.V of the study variable. However, the length of the confidence interval for both the estimators steadily increases with the values of C.V of the study variable. When C.V=0.1, the average length of the confidence interval for the two estimators ($\hat{\theta}_{y_4}$ and $\hat{\theta}_{y_7}$) were respectively (0.0159, 0.0182), while for C.V=1.0, the respective values are (0.2224, 0.4458). The ratio of the length of the confidence interval to the value of C.V is approximately 16% when C.V=0.1, while it increased to 22% when C.V=1.0 for $\hat{\theta}_4$ and for $\hat{\theta}_7$ it is 18% for C.V=0.1 to 44% for C.V=1.0. For the sample C.V the average length of the confidence interval is larger compared to the other estimators. The ratio of the length of the confidence interval to the value of C.V for this estimator is 28% when CV=0.1, while it increased to 47% when C.V=1.0.

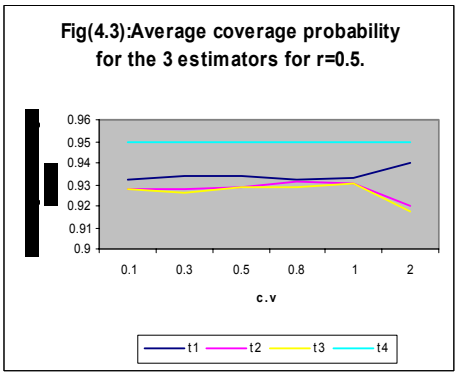
Fig (4.1), (4.2) and (4.3) represents the average length of confidence interval (right side) and average coverage probability (left side) versus C.V of the study variable for the 3 estimators.

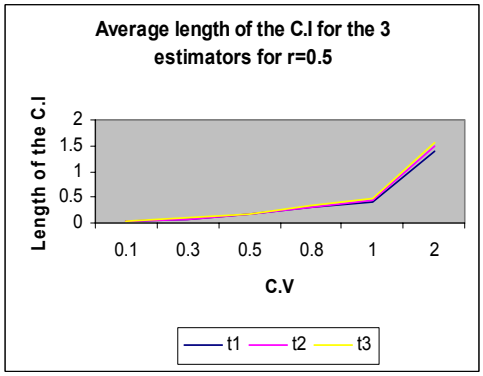












To obtain more accurate results we have compared the different estimators through their mean square error (MSE's). Table (4.2) presents the average mean square error of the ratio type regression estimator ($\hat{\theta}_{y_4}$), regression estimator ($\hat{\theta}_{y_7}$) where the information on the population C.V of the auxiliary variable is used and the sample C.V ($\hat{\theta}_{y_1}$) for low correlation co-efficient.

The MSE's of the ratio type regression estimator or the regression estimator increases steadily and it is observed that the estimators $\hat{\theta}_4$ and $\hat{\theta}_7$ has got the minimum MSE values when compared to $\hat{\theta}_1$ for the various configurations and when the correlation co-efficient is high. To save space the results are not reported when the correlation co-efficient is high. The results are presented only when the correlation co-efficient is -0.3, -0.1, 0, 0.1, 0.3 and also for the sample size n=100. The pattern of the results does not change for other sample size=200.

Table (4.2). Average MSE of the 3 estimators

C.V of the study variable.		Correlation co-efficient (r) (n=100)				
		-0.3	-0.1	0	0.1	0.3
0.1	θ_4	0.5126* 10^{-4}	0.5315* 10^{-4}	0.5219* 10^{-4}	0.4972* 10^{-4}	0.5185* 10^{-4}
	θ_7	0.5143* 10^{-4}	0.5362* 10^{-4}	0.5226* 10^{-4}	0.4978* 10^{-4}	0.5178* 10^{-4}
	θ_1	0.5109* 10^{-4}	0.5288* 10^{-4}	0.5161* 10^{-4}	0.4908* 10^{-4}	0.5133* 10^{-4}
0.3	θ_4	0.5616* 10^{-3}	0.5459* 10^{-3}	0.5487* 10^{-3}	0.5495* 10^{-3}	0.5249* 10^{-3}
	θ_7	0.5672* 10^{-3}	0.5463* 10^{-3}	0.5496* 10^{-3}	0.5510* 10^{-3}	0.5269* 10^{-3}
	θ_1	0.5596* 10^{-3}	0.5408* 10^{-3}	0.5423* 10^{-3}	0.5451* 10^{-3}	0.5209* 10^{-3}

Table (4.2). Average MSE of the 3 estimators (cont)

C.V of the study variable.		Correlation co-efficient (r) (n=100)				
		-0.3	-0.1	0	0.1	0.3
	θ_4	0.0020	0.0018	0.0017	0.0019	0.0021
	θ_7	0.0021	0.0019	0.0018	0.0021	0.0024
0.5	$\hat{\theta}_1$	0.0019	0.0017	0.0016	0.0018	0.0019
	θ_4	0.0071	0.0078	0.0075	0.0073	0.0077
	θ_7	0.0072	0.0079	0.0076	0.0074	0.0078
0.8	$\hat{\theta}_1$	0.0069	0.0074	0.0072	0.0071	0.0075
	θ_4	0.0145	0.0153	0.0159	0.0164	0.0164
	θ_7	0.0154	0.0155	0.0161	0.0167	0.0169
1.0	$\hat{\theta}_1$	0.0138	0.0148	0.0156	0.0156	0.0154
	θ_4	0.2147	0.2582	0.3082	0.2432	0.2259
	θ_7	0.2286	0.2610	0.3085	0.2492	0.2289
2.0	$\hat{\theta}_1$	0.1962	0.2430	0.3044	0.2357	0.2149

From the table which is reported here the conclusion that can be drawn is that for the 2 estimators the MSE of the estimator did not change either with the values of the correlation co-efficient or with the values of C.V of the study variable. However, the MSE for both the estimators steadily increases with the values of C.V of the study variable. When C.V=0.1, the average MSE for the two estimators ($\hat{\theta}_{y_4}$ and $\hat{\theta}_{y_7}$) were respectively (0.00005126, 0.00005143), while for C.V=1.0 the respective values of MSE are (0.0145, 0.0154). For the sample C.V, the respective values of average MSE are 0.00005109 for C.V=0.1 and 0.0138 for C.V=1.0. Thus, we notice that in the case of low correlation co-efficient, sample C.V performs better than the other 2 estimators.

5. Conclusions.

From the comparison of the asymptotic MSE's of the various estimators and the small sample comparison of the average length of the confidence interval, the conclusion that can be drawn is that the best estimators of C.V are the ratio type regression estimator, namely $\hat{\theta}_{y_4}$ and the regression estimator $\hat{\theta}_{y_7}$ where the information on population C.V of the auxiliary variable is used, when the auxiliary variable is correlated with the study variable. When there is low correlation the sample C.V emerges as the best estimator. In the estimation of population mean, regression estimator using the mean of the auxiliary variable

emerges as the best estimator irrespective of the value of correlation co-efficient (see Murthy(1967)). However, in the estimation of population C.V, regression estimators performs well only when the correlation co-efficient is moderate or large. Among the two (regression / ratio type regression) estimators, the regression estimator using information on population C.V uses less information than the ratio type regression estimator. In many instances it is likely that information on population C.V of the auxiliary variable is available, while the individual values of population mean and variance are not available. Thus, we recommend the regression estimator for use when an auxiliary variable is properly chosen so as to be correlated with the study variable. In the absence of such auxiliary variable, it is safe to use sample C.V as the estimator.

Acknowledgements.

This paper is part of the first author's PhD work under the guidance of second author. For this paper, the author received the 'Jan Tinbergen Award' for the year 2007 and was presented during the 56th session of ISI, in Lisbon, Portugal.

Appendix A

Population Moments in SRSWR

$$E\left(\bar{x}-\bar{X}\right)=0$$

$$E\left(\bar{y}-\bar{Y}\right)=0$$

$$E\left(\bar{x}-\bar{X}\right)^2=\frac{1}{n}\sigma_{xx}$$

$$E\left(\bar{y}-\bar{Y}\right)^2=\frac{1}{n}\sigma_{yy}$$

$$E\left(\bar{x}-\bar{X}\right)\left(\bar{y}-\bar{Y}\right)=\frac{1}{n}\sigma_{xy}$$

$$E\left(s_x{}^2-\sigma_x{}^2\right)=0$$

$$E\left(s_y{}^2-\sigma_y{}^2\right)=0$$

$$E\left(s_x{}^2-\sigma_x{}^2\right)^2=\frac{1}{n}\sigma_{xxxx}-\frac{(n-3)}{n(n-1)}\left(\sigma_{xx}\right)^2+O\left(\frac{1}{n}\right)$$

$$E\left(s_y{}^2-\sigma_y{}^2\right)^2=\frac{1}{n}\sigma_{yyyy}-\frac{(n-3)}{n(n-1)}\left(\sigma_{yy}\right)^2+O\left(\frac{1}{n}\right)$$

$$E\left(s_x^2-\sigma_x^2\right)\left(\bar{x}-\bar{X}\right)=\frac{1}{n}\sigma_{xxx}+O\left(\frac{1}{n}\right)$$

$$E\left(s_x^2-\sigma_x^2\right)\left(\bar{y}-\bar{Y}\right)=\frac{1}{n}\sigma_{xxy}+O\left(\frac{1}{n}\right)$$

$$E\left(s_y^2-\sigma_y^2\right)\left(\bar{x}-\bar{X}\right)=\frac{1}{n}\sigma_{xyy}+O\left(\frac{1}{n}\right)$$

$$E\left(s_x^2-\sigma_x^2\right)\left(\bar{x}-\bar{X}\right)=\frac{1}{n}\sigma_{xxx}+O\left(\frac{1}{n}\right)$$

$$E\left(s_x^2-\sigma_x^2\right)\left(\bar{y}-\bar{Y}\right)=\frac{1}{n}\sigma_{xxy}+O\left(\frac{1}{n}\right)$$

$$E\left(s_y^2-\sigma_y^2\right)\left(\bar{x}-\bar{X}\right)=\frac{1}{n}\sigma_{xyy}+O\left(\frac{1}{n}\right)$$

$$E\left(s_y^2-\sigma_y^2\right)\left(\bar{y}-\bar{Y}\right)=\frac{1}{n}\sigma_{yyy}+O\left(\frac{1}{n}\right)$$

$$E\left(s_x^2-\sigma_x^2\right)\left(s_y^2-\sigma_y^2\right)=\frac{1}{n}\sigma_{xxyy}-\frac{1}{n}\sigma_{xx}\sigma_{yy}+O\left(\frac{1}{n}\right)$$

Note : $\sigma_{yy}=\frac{1}{N}\sum_{i}^{N}\left(Y_i-\bar{Y}\right)^2$ $\qquad$ $\sigma_{xy}=\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\left(y_i-\bar{Y}\right)$

$$\sigma_{yyyy}=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^4 \qquad \sigma_{yyy}=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^3$$

$$\sigma_{xxy}=\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2\left(Y_i-\bar{Y}\right)$$

In the similar manner the other moments are defined. To save space the expressions are not given.

Appendix B

For the case of SRSWOR, the expressions for the moments are defined in a similar manner as in Appendix A. But Population variance σ_y^2 (σ_{yy}) is replaced by S_y^2 (S_{yy}). To save space the expressions are not presented here.

REFERENCES:

AHMED, S.E (2002): Simultaneous estimation of Co-efficient of Variation. *Journal of Statistical Planning and Inference*. 104, *31-51*.

ANANTHAKRISHNAN, R. and SOMAN, M.K. (1989): Statistical distribution of daily rainfall and its association with the co-efficient of variation of rainfall series. *International Journal of Climatology* 9, *485-500*.

BREUNIG, R.(2001): An almost unbiased estimator of the co-efficient of variation. *Economics Letters*, 70(1), *15-19.*

BRUS, D.J and RIELE W.J. Te (2001): Design-based regression estimators for spatial means of soil properties, *Geoderma*, 104, *257-279*.

CHATURVEDI, A. and RANI, U(1996): Fixed width confidence interval estimation of Inverse co-efficient of variation in normal population. *Microelectronics and Reliabiliy*, 36, *1305-1308*.

DAS, A.K. and TRIPATHI, T.P. (1981 a): Sampling strategies for coefficient of variation using knowledge of the mean using an auxiliary character. *Tech. Rep. Stat. Math. 5/81*. ISI, Calcutta.

DAS, A.K. and TRIPATHI, T.P. (1981 b): A class of estimators for co-efficient of variation using knowledge on coefficient of variation of an auxiliary character. Paper presented at annual conference of *Ind. Soc. Agricultural Statistics. Held at New Delhi, India.*

DE, P., GHOSH, J.B., and WELLS, C.E.(1996): Scheduling to minimize the co-efficient of variation. *International Journal of Production Economics*, 44, *249-253*.

FIELLER, E.C. (1932): A numerical test of the adequacy of A. T. McKay's approximation. *J. Roy. Stat. Soc.* 95, pp. *699–702.*

HENDRICKS, W.A. and ROBEY, W.K.(1936): The sampling distribution of the Co-efficient of variation. *Annals. Math. Stat.* (7), *129-132*.

KENDALL, M. and STUART, A. (1977): *The advanced Theory of Statistics*, 1, 4th edition. London: Charles Griffin and Co.

Kordonsky, Kh.B and Gertsbakh. I (1997): Multiple Time Scales and Life time Coefficient of Variation: Engineering Applications. *Lifetime Data Analysis*. 3, *139-156*.

MAHMOUDVAND, R. and HASSANI, H. (2007): Two new confidence intervals for the C.V in a normal distribution. *Journal of Applied Statistics,* 36: 4,*429-442.*

MCKAY, A.T. (1932): Distribution of the Co-efficient of variation and the extended 't' distribution. J.Royal.Stat.Soc.95, *696-698.*

MUKHOPADHYAY, PARIMAL (2008): *Theory and Methods of Survey Sampling*, Second edition, Prentice Hall of India, New Delhi.

MURTHY M.N. (1967): *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

NAIRY, S.K. and RAO, A.K. (2003): Tests for Coefficients of variation of normal population. *Communications in Statistics-Simulation and Computation*, 32(3), *641-661*.

NG, C.K.(2006): Performance of three methods of interval estimation of the co-efficient of variation. *InterStat*.

PANICHKITKOSOLKUL, W. (2009): Improved confidence intervals for a Co-efficient of variation of a normal distribution. *Thailand statistician,* 7(2), *193-199.*

PATEL, P. A. and SHAH RINA. (2009): A Monte Carlo comparison of some suggested estimators of Co-efficient of variation in finite population. *Journal of Statistics sciences, 1(2), 137-147*.

PRADHAN, B.K. (2010): Regression type estimators of finite population variance under multiphase sampling, *Bull. Malays. Math. Sci. Soc*, 33(2), *345-353.*

PEARSON, E.S.(1932): Comparison of A.T. McKay's approximation with experimental sampling results. J.Royal. Stat. Soc. 95, *703-704*.

RAJYAGURU, A. and GUPTA, P.C.(2002): On the estimation of Co-efficient of variation for finite population-I, *Journal of Statistical research, Dhaka*.

RAO., A.K, and Bhatta, A.R.S. (1989): A note on test for Co-efficient of Variation. *Calcutta Statistical Association Bulletin*, 38, *151-152*.

SINGH,H.P, SINGH,S and KIM, J.M.(2010): Efficient use of auxiliary variables in estimating finite population variance in two-phase sampling. *Communications of the Korean Statistical Society*, 17(2), *165-181*.

SAHOO, L.N; SAHOO, R.K and SENAPATHI, S.C (2003): Predictive estimation of Finite Population mean using Regression type estimators in double sampling procedures. *Journal of Statistical Research*, 37, *291-296.*

SINGH, M. (1993): Behavior of sample Co-efficient of Variation drawn from several distributions. *Sankhya*. *B*. 55, *65 – 76*.

SRIVASTAVA, S.K.(1980): A class of estimators using auxiliary information in sample surveys. *Canad. J. Statist,* 8(2), *253-254.*

TRIPATHI, T.P. SINGH, H.P. and UPADHYAYA, L.N. (2002): A general method of estimation and its application to the estimation of co-efficient of variation. *Statistics in Transition,* 5(6), *887-909.*

VERMA, M.R. (2008): Approximately optimum stratification for ratio and regression methods of estimation. *Applied Mathematics Letters*, 21(2),*200-207*.

VERRIL, S. (2003): Confidence bounds for normal and log-normal distribution co-efficient of variation, *Research paper,EPL-RP-609,Madison,Wisconsin,US*.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 381—400

THE DYNAMIC MODEL OF BIRTH AND DEATH OF ENTERPRISES

Henryk Gurgul[1], **Paweł Zając**[1]

ABSTRACT

The aim of this article is to define the model describing the dynamics of bankruptcy and foundation of new enterprises. In the first part we try to answer what bankruptcy in law, economic and social sense is. It results from the overview of literature that bankruptcy is as natural as growth, and both of these contradictions are complementary. Moreover, an important inference is the need for improving the bankruptcy mechanism, because the more efficient it is, the healthier market surrounds us. On the basis of bankruptcy new firms emerge. We derived a procedure in order to forecast the number of new firms. The conclusion is that dynamic mathematical models may be a useful tool of prediction of a number of new firms founded.

Key words: bankruptcy of enterprises, foundation of enterprises, dynamics of foundation and bankruptcy, forecasts.

JEL Classification: C02, L25.

1. Introduction

In recent years, the increasing risk and uncertainty has been the most important feature of companies and the whole economy's activity. Various firms, their groups and the whole country's economy encounter various unpredicted, negative results of events and processes located often in distant parts of the globe.

The latest example of the aforementioned is the current financial crisis and the following economic crisis which have stricken first the United States in 2007, and different countries across the whole world a year later. It proves the strength of mutual connotations between companies and diverse economics in the era of globalization. The growing uncertainty in world economy is the result of sliding of the industrial civilization towards a knowledge-based economy. These radical,

[1] Department of Applications of Mathematics in Economics, AGH University of Science and Technology, Cracow, Poland. Email addresses: henryk.gurgul@gmail.com, pzajac@zarz.agh.edu.pl.

groundbreaking changes lead to the replacement of currently used models of economy by new ones, applying to investments, production, trade, education, management, work, employment and consumption. This process of changes is accompanied by essential social and moral alterations including the modifications in the existing model of the family. Various professions, jobs, firms but also different branches of economy are becoming more vulnerable to bankruptcy of which the main cause is the unprecedented pace of technological progress, bringing equally positive and destructive changes. As a consequence, more often the 'old' ideas, solutions and even companies are replaced by the 'new' ones. This process regards also economic theories, even those awarded in the past by the Nobel Prize. The rapid changes are accompanied by increasing lack of security, the results of which are clearly visible in companies situated in Poland. The uncertainty affecting Polish companies is additionally inflamed by transformations and the necessity of adapting to the European Union's regulations. It is also a significant obstruction in formulating long-term strategies concerning growth, which can provoke management errors, or even in extreme cases, bankruptcy.

Problems concerning the uncertainty and risk, despite frequent occurrence in many papers, are comparatively hardly ever recognized in economic literature. The situation is even worse when it comes to insolvency. Usually used while referring to business, insolvency refers to the company's inability to pay off its debts. Business insolvency is defined in two different ways: cash flow insolvency concerning inability to pay debts as they fall due and balance sheet insolvency meaning that liabilities exceed assets. Another term commonly used in this context is bankruptcy. In colloquial language, the lately mentioned terms are used as synonyms. In fact the difference between them is significant as bankruptcy is a law term and insolvency belongs to the economical terminology. The insolvent company is a company unable to continue an individual activity without outside help. The most frequent source of such situation is the lost of trust among trading partners, clients, essential impairment of financial indicators, inability in implementation of commitments, or even financing its own current activity. It can possibly lead towards arrestment of the company, and eventually to its end. Detailed reasons of such cases in Poland are described in Zdyb (2008).

Insolvency on the other hand (which is the result of either law regulations or the court judgment) can take place despite the economic reasons, leading to the cessation of company. Company's liquidation regardless of its size, sphere of interest, territory or trade partners is the source of confusion and, more importantly, distress on the market. The consequences of this phenomenon are often unpaid debts and taxes or undone services for recent orders. The worst social outcome of the aforementioned is redundancy of employees, which negatively affects the unemployment rate. Other results of bankruptcy are often unexploited resources, frustration and discontent among the fired employees, employers, but also investors and shareholders who devoted their savings, assets and ideas in order to increase production or improve service. In the economic

literature bankruptcy is described as harmful to the business entity and the surrounding form of destruction bringing seriously fatal consequences in a short period of time.

However, in the long run the positive elements which are the result of closure of the ineffective company which does not bring fair remuneration to the owners, may dominate. Especially, the threat of shutting down the company immobilizes the entrepreneurs, encouraging them to undertake effective actions and daunt the continuous borrowings.

The market itself usually does not have the ability to eliminate the unsuccessful entrepreneurs. Therefore, certain protection procedures from disastrous outcomes of their activity are described in the bankruptcy law regulations which help to minimize the negative results which may affect the surrounding as a result of insolvency of the debtor. The obligation of existence of such regulations comes from the fact that no restrictions describe the conditions of setting up any business activity. In particular, this problem applies to suitability of candidates to become entrepreneurs. Hence, certain percentage of them is destined to fail as a result of the lack of competence, management errors, exterior conditions or other unrelated circumstances.

2. Literature overview

Although the problems of bankruptcy and insolvency are still not satisfactorily described, there is a significant number of contributions in English for example by Bebchuk (2002), Hurst (1995), Schonfelder (2003) or in Polish literature among them those by Białasiewicz and Buczkowski (1996), Bratnicki (2001), Czajka (1999), Durlik (1998), Osbert-Pociecha (2004), Perechuda (1999) and Zimniewicz (2000) which deal with reasons and implications of these problems in economy as a whole and individuals.

In the opinion of the eminent Austrian economist Joseph Schumpeter (1982) failure of the company is a consequence of coexisting processes of destruction and creation. The same point of view is expressed by Foster and Kaplan (2001). According to their theory inspiration, directing and controlling the processes of creative obliteration is stimulated by the financial markets supplying and financially reinforcing the possibilities of the companies as well as withdrawing the support along with decreasing competitiveness of the company. The developing company is offered financial resources, although almost immediately after the appearance of the symptoms of reduction in competitiveness and regression this support is removed. In case of bankruptcy the company not only needs to deal with its current activity, but also with the mutually connected processes of destruction and creation.

Greiner and Schein (1988) (compare Koźmiński and Piotrowski (1999)) claim that flexibility of the company depends on the abilities and creativity of the owner. Further improvements are results of problems which company encounters

in its later existence and the methods it implies to overcome them. If the managements do not notice the problems soon enough, the company is certainly heading to an end. Undertaking the activity in the conditions of uncertainty, rapidly changing situation characterized by frequently occurring problems and new goals, considerably affects the immune to changes businesses' lives cycles which, as a result, are usually seriously shorten. The companies adherent to old, uncompetitive solutions are swiftly eliminated from the market. The lack of alterations inside the company leads to its closure.

Similar subjects lie in the area of interest of Handy (1995), the author of the so-called shape of 's' letter. In the core of his theory lies the idea that entrepreneurs should prepare the company for the crisis period already in the time of its prosperity, by provoking artificial symptoms and undertaking adequate countermeasures. Although such actions weaken the organization in the short term, eventually changes developed by the fake crisis strengthen and immunize the company.

According to Frederick et al. (1988) as well as Davies and Blomstrom (compare Majchrzak (2003)) the companies' activities are always connected with executing social functions, using the resources given to them by the society. If the organization stops responding to those functions by not using those resources in a way that is desired by the society, it starts heading towards failure. This situation is connected with retrieving property, financial and human resources which are indispensible for the proper functioning and existence of the company. Therefore, the elimination of the company is the result of social desire to improve the effectiveness of economic activity.

Apart from the aforementioned advantages of insolvency, as the removal of unprofitable firms and the protection of debtees against dishonest debtors, the bankruptcy based on the certain rules and regulations allows one to eliminate from the market the companies functioning on the verge of cost-effectiveness. Those rules give also the opportunity to eradicate encumbrances which are caused by the prolonged existence of such firms.

In many cases, certain inability to refrain from engaging social sources into businesses which are destined to fail from the very beginning, for example in Poland where it applies mainly to post-communist companies such as mines and shipyards, is well visible.

In practice bankruptcy is mainly a tool of control and protection of the market (Zedler (2003)). After the increase phase the company which encounters the changing circumstances that it cannot deal with is exposed to the pressure from the side of the market that may lead at first to its (bankruptcy) insolvency and then to bankruptcy proceedings protecting debtees, employees and the whole country. Moreover, as a result of this proceedings such company is either reorganized or disposed, which is suppose to protect and give the possibility of better usage of the social resources.

The interesting interpretation of the theory of development and bankruptcy of the firms is based on the biologic theory developed by Darwin (Encyklopedia Biologiczna (1998)).

Darwin wrote that all living beings without any exception, have the tendency to numerous growth in such a big extent that no environment, not even the whole area of the earth or the whole ocean are not able to accommodate offspring of one couple after some generations. The unavoidable result of this process is the continuous fight of existence. It appears that the similar tendency is exhibited by the companies heading towards expansion which leads to competitive fight resembling the one between biological beings.

This process is the realization of natural selection. Its result is the survival of the best individuals possessing the most developed ability to adapt to the surrounding. This theory corresponds also to the companies.

Since the late sixties of the twentieth century quantitative methods became commonly used to predict the risk of bankruptcy and the progenitor of such research was Altman (1968). In his work he used discrimination methods. By working on those methods he was able to divide the analyzed firms into two groups: the ones vulnerable to bankruptcy and those not. The grouping was done on the basis of financial indicators defined earlier for those companies. The following belonged to those markers (ratios): working capital to total assets ratio, retained earnings to total assets ratio (retained earnings - profits which have not been paid out in dividends and which can be re-invested in the business), EBIT (Earnings Before Interest and Taxes) to total assets ratio, market value of equity to book value of total liabilities and sales to total assets. Using these methods allowed Altman to correctly classify up to 95 percent of companies the year before their bankruptcy and 83 percent two years before.

Studies initiated by Altman and continued by others led to foundation of the American Bankruptcy Institute (ABI) in USA in 1982. It delivers to the Congress and public opinion expertises describing the cases concerning companies bankruptcy, analyzes its causes and results. It integrates people of various professions starting from attorneys, accountants, auctioneers, assignees, bankers, lenders to professors. ABI is engaged in numerous educational and research activities concerning bankruptcy in the USA. The institute possesses numerous empirical data related to insolvency and regularly cooperates with the media. Analysis concerning the future of companies is a significant part of Institute's works. It is important, because companies' surroundings and connections become more and more complicated, which provides difficulties in formulating long-term, evolutionary strategy. Although the Institute participates in researching past events as the source of prognoses, its studies of company's future are an essential part of the ABI's mission.

As it was emphasized by Matschke and Broesel (2007) the deciding factor of the company's worth and position on the market is not its past but the profits it can bring in the future. In this context, the plausible scenarios of the future development of the situation inside the company, branch or even the whole

economy are being created. The chances and risks of the expansion possibilities are being analyzed. The purpose is to obtain immediate evaluation of the influence of social and economic decisions basing on the company's results. The Institute concentrates on detecting symptoms threatening companies in order to warn them of insolvency. Only fast enough information may direct towards the restraint of the emerging threat. The Institute also examines cases of fraud in accountancy, for example the so-called creative accountancy.

Early researches on insolvency of Polish firms have been conducted using discrimination methods introduced by Altman. Hadasik (1998) conducted research on the basis of financial statements of 39 companies from 1991 to 1997. She estimated parameters of models containing from 4 to 7 variables. Efficiency of those models was high as it reached from 88,52 percent to 96,72 percent. The most important variables were: debt to equity ratio, accounts receivable turnover ratio, inventory turnover ratio, inventory profitability.

In his research Hołda (2001) in order to estimate discrimination function used 40 statements of companies which declared bankruptcy between 1993 and 1996 and 40 statements of different firms that did not experience insolvency at that time. Using this date he built a model whose efficiency was estimated by the author at 92,5 percent. It included the following variables: current liquidity ratio, debt to equity ratio, return on assets (ROA), accounts payable turnover ratio (average payment period) and total assets turnover ratio.

The results of research on the insolvency which uses discrimination methods can significantly differ from one another according to the country and time period, as the proneness to bankruptcy is different in various periods. In relation to those variables diverse indicators are important. Therefore, it is pointless to conduct researches using the same set of variables not only for companies from different countries, but also for different time periods within single company. It causes that results of even a wide range of papers are not comparable. Hence, some new attempts to form models are applied in order to allow to predict the bankruptcy.

The model which is illustrated here and used in the empirical examinations does not refer to financial indicators. It is a model describing biological phenomenon of creation and degradation of red cells developed by Ważewska-Czyżewska and Lasota (1976). We extended and adjusted this model in order to describe the process of insolvency and creation of companies.

3. The general model

Let $N(t,a)$ be a number of companies which in moment t are not older than a. Then, $N(t) = \lim_{a\to\infty} N(t,a)$ is a general number of companies at time t.

Function $n(t,a) = \frac{\partial}{\partial a} N(t,a)$ would express density of age distribution for firms. In short time periods $n(t,a)$ represents number of companies which in moment *t* are in age *a*.

This function fulfils the condition:

$$\int_0^\infty n(t,s)ds = N(t) \tag{1}$$

Companies which in the moment *t* were aged *a* are in the moment *t+h* at age *a+h*. Difference $n(t,a) - n(t+h, a+h)$ means number of firms at age *a* which went bankrupt in the time period $(t, t+h)$. Destruction intensity *i(t,a)* of companies at age *a* in moment *t* can be specified as:

$$i(t,a) = \lim_{h \to 0} \frac{n(t,a) - n(t+h, a+h)}{h}.$$

In that situation $\lambda(t,a) = \frac{i(t,a)}{n(t,a)}$ is the empirical probability that company which in the moment *t* is at the age *a* will bankrupt to moment *t+1*. $\lambda(t,a)$ will be called destruction coefficient.

The mean value theorem leads to:

$$n(t+h, a+h) - n(t,a) = h\frac{\partial}{\partial t} n(\bar{t}, \bar{a}) + h\frac{\partial}{\partial a} n(\bar{t}, \bar{a}), \quad \bar{t} \in (t, t+h), \quad \bar{a} \in (a, a+h),$$

and therefore

$$\frac{\partial n}{\partial t} + \frac{\partial n}{\partial a} = -\lambda n \tag{2}$$

This equation is used in theorems about creation and degradation of red cells. It was found by von Forster in 1959. It appears that (2) is only a consequence of the destruction's coefficient definition and can be applied to the description of creation and insolvency of companies.

If we assume that *a*=0, function *n(t, a)* can be interpreted as a number of firms which were created in the moment *t*:

$$p(t) = n(t,0). \tag{3}$$

Equation (2) with initial condition (3) allows one to calculate function n(t, a). In order to find reverse relation let us introduce the term of the stimulation

process of firms creation. Derivative $\frac{dp(t)}{dt}$ represents increment of the number of newly created companies in the time unit (for instance year). Quotient

$$S(t) = \frac{1}{p(t)} \frac{dp}{dt} \tag{4}$$

represents the increment rate of new firms in the time unit. We can now say more about the stimulation process of firms creation. It is known that change of the amount of companies is the impulse stimulating (or slowing down) the process of creating new ones.

Because our goal is to build possibly simple model, we assume that the stimulation process of firm's creation *S(t)* is proportional to the general amount of companies on the market in previous moments.

$$S(t) = -\frac{d}{dt} \gamma N(t-h), \tag{5}$$

where γ is proportional coefficient, and by h we understand the lag (time delay) with which, after changing the general number of companies, new firms are established. Equation (5) implicates that the decreasing amount of companies is related to the increasing number of new ones and that the increasing amount of companies is connected with the decrease of newly created number.

Combining (4) and (5) provides to

$$\frac{dp(t)}{dt} = -p(t) \frac{d}{dt} \gamma N(t-h), \tag{6}$$

which can be solved as:

$$p(t) = \rho e^{-\gamma N(t-h)} \tag{7}$$

where ρ is integration constant. Collating (1), (2), (3) and (7) gives us:

$$\left.\begin{aligned} &\frac{\partial n}{\partial t} + \frac{\partial n}{\partial a} = -\lambda n \\ &n(t,0) = p(t) \\ &p(t) = \rho \exp\{-\gamma \int_0^{\infty} n(t-h,a)da\} \end{aligned}\right\} \tag{8}$$

We can observe in that set of equation three coefficients λ, ρ, γ. Coefficient λ was mentioned before while we were revealing equation (2) and it represents the empirical probability that companies which in moment t were at

age a would bankrupt until moment t+1. Coefficient γ characterizes the stimulation of the firm establishing process. Its meaning implies from equation (6). It is the growth rate of companies caused by the unitary change of the general number of companies on the market. The meaning of coefficient ρ is connected with the requirement of new companies on the market. If the requirement is bigger, then coefficient ρ gets higher values. Later in this paper we will try to explain detailed character of this relation.

4. The stationary model solution

Let us consider the simplified (independent from time) problem. Because *n(t,a)*, *p(t)* and $\lambda(t,a)$ do not depend on time t in stationary model, let us put

$$n(t,a) = \overline{n}(a), \quad p(t) = \overline{p} \quad \lambda(t,a) = \overline{\lambda}(a).$$

Then, we get

$$\overline{n}(a) = \overline{p}\exp\{-\int_0^a \overline{\lambda}(s)ds\}, \tag{9}$$

and

$$\overline{p} = \rho\exp\left\{-\gamma\overline{p}\int_0^\infty \exp\left[-\int_0^a \lambda(s)ds\right]da\right\}. \tag{10}$$

Let us denote by $E(\sigma)$ solution of equation

$$\sigma E + e^{-E} = 0, \qquad \sigma > 0.$$

Using $E(\sigma)$ function we can denote the stationary solution of (10) as:

$$\overline{n}(a) = \frac{1}{\gamma c}E\left(\frac{1}{\rho\gamma c}\right)\exp\left\{-\int_0^a \overline{\lambda}(s)ds\right\}, \tag{11}$$

where

$$c = \int_0^\infty \exp\left\{-\int_0^a \lambda(s)ds\right\}da.$$

The formula for $\overline{\lambda}(a)$ is important to the upcoming search of stationary solutions. From the analytical point of view the Gompertz curve, well-known and used in the reliability theory, appears to be correct. The curve is in the form:

$$\overline{\lambda}(a) = Ke^{\alpha a}. \tag{12}$$

According to the reliability theory, constant K represents the coefficient of destruction for companies at the beginning of their existence. Constant α can be calculated as a natural logarithm from comparative destructive coefficient in the time unit.

$$\alpha = \ln \frac{\bar{\lambda}(a+1)}{\bar{\lambda}(a)}. \tag{13}$$

Constant α can be understood as a decomposability parameter.

Combining (9) and (12) we get

$$n_0(a) := \frac{n(a)}{n(0)} = \exp\{-\frac{K}{\alpha}[\exp(\alpha a) - 1]\}. \tag{14}$$

Function $n_0(a)$ will termed the normalized stationary solution.

Let us consider function $n_0(t)$ which comes into existence by changing its name of variable from a to t. $n_0(t)$ would be the curve presenting proportion of the amount of companies who survived until the moment t to original number of companies. The curve in our notation will be called the decomposition curve.

Table 1 epitomizes data regarding numbers of newly established firms in Małopolska in Poland and their yearly survival indicators between 1999 and 2007.

Table 1. Newly established firms in Małopolska in Poland and their yearly survival indicators between 1999 and 2007

New established firms	Survive indicators (percent)								
	after 1 year	after 2 years	after 3 years	after 4 years	after 5 years	after 6 years	after 7 years	after 8 years	after 9 years
1999 26240	2000 91.5	2001 84.7	2002 81.6	2003 79.1	2004 74.1	2005 70.1	2006 65.2	2007 62.0	2008 59.6
2000 25150	2001 89.1	2002 84.5	2003 80.0	2004 74.6	2005 70.0	2006 64.7	2007 60.8	2008 58.1	
2001 19434	2002 93.8	2003 89.6	2004 83.1	2005 77.3	2006 71.2	2007 66.6	2008 63.1		
2002 16735	2003 93.4	2004 84.2	2005 76.8	2006 69.4	2007 64.6	2008 63.1			
2003 20377	2004 83.8	2005 76.0	2006 68.7	2007 63.9	2008 59.2				
2004 18650	2005 82.6	2006 73.4	2007 67.5	2008 63.4					
2005 20564	2006 81.4	2007 71.2	2008 65.7						
2006 24367	2007 82.6	2008 71.4							
2007 24119	2008 85.1								

The family of functions (14) allows a good data estimation. Figure 1 shows data from the years 1999-2004 with estimated decomposition curves using the Least Square Method. Table 2 includes estimated values of parameters K and α. Computation have been made using R computer program.

Figure 1. Data from the years 1999-2004 with fitted decomposition curves

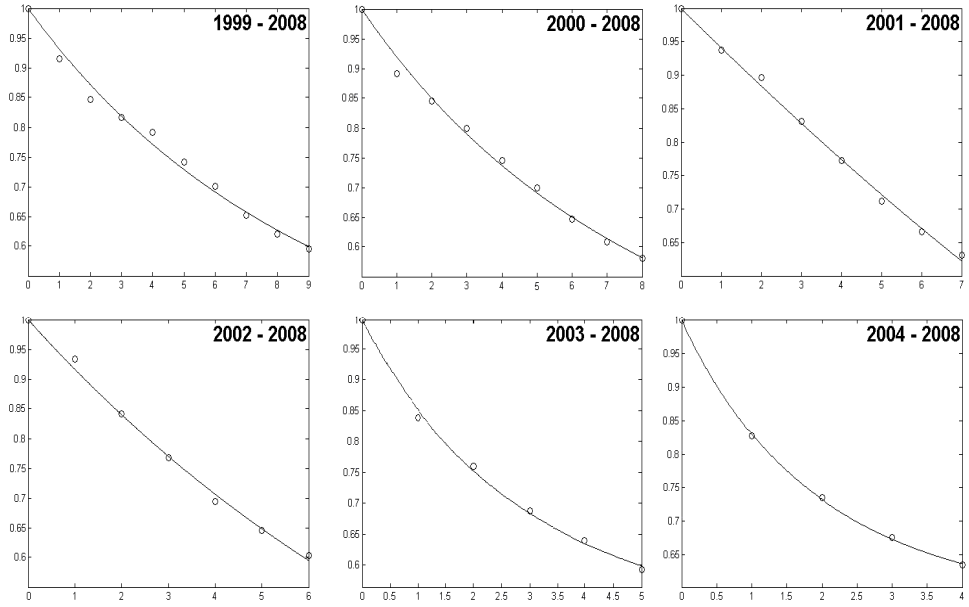


Table 2. Parameters K and α values computed using Least Square Method

	1999	**2000**	**2001**	**2002**	**2003**	**2004**
K	0.07256	0.086376	0.059350	0.087438	0.18182	0.227559
α	-0.0566	-0.06399	0.03684	-0.00342	-0.2559	-0.40269

As it can be observed parameters present different values for particular years. It is clear that values received for the period between 1999 and 2002 are similar, whereas since the year 2003 a significant change for both K and α took place. Having in mind that K applies to destructive coefficient, at the beginning the results from the years 1999 till 2002 should be treated as acceptable. In other cases, too high parameters are probably connected with too small number of observations.

The interpretation of the achieved results for K showed that about 7-8 percent of the companies finished their activity soon after the registration to the REGON

system. Values received for the parameter α are, as expected, negative. It means that we get an inequality $\overline{\lambda}(a+1) < \overline{\lambda}(a)$, what shows that the probability of the failure of the company is decreasing with time. According to these calculations the year 2001 seems to be the exception from this rule. On the figure presenting data from 2001 the linearity can be observed. Moreover, the disturbance developed probably on the basis of introducing the new GDP system can also be noticed.

Computation of K and α while accepting the correctness of the given model allows one to predict the survival rate of firms in the following years.

5. The reduced model

In this model our special attention will be focused on the behaviour of the general number of companies in time, the formula for *N(t)*. Let's define a new coefficient:

$$\mu = \frac{1}{N(t)} \int_0^\infty \lambda(t,a) n(t,a) da = \frac{\int_0^\infty \lambda(t,a) n(t,a) da}{\int_0^\infty n(t,a) da}. \tag{15}$$

The numerator tells us about the number of companies liquidated in the time unit and the denominator expresses the general amount of firms active in the time unit. The letter μ expresses empirical probability of bankruptcy announced in the time unit.

Integrating by a equation (2) in interval $[0,\infty)$ we receive:

$$\int_0^\infty \frac{\partial}{\partial t} n(t,a) da + \int_0^\infty \frac{\partial}{\partial a} n(t,a) da = -\int_0^\infty \lambda(t,a) n(t,a) da. \tag{16}$$

Considering equation (1) we can write the first component in the form:

$$\int_0^\infty \frac{\partial}{\partial t} n(t,a) da = \frac{\partial}{\partial t} \int_0^\infty n(t,a) da = \frac{\partial}{\partial t} N(t). \tag{17}$$

Assuming that all companies would go bankrupt with time ($\lim_{a\to\infty} n(t,a) = 0$) we can integrate the second component and get:

$$\int_0^\infty \frac{\partial}{\partial a} n(t,a) da = n(t,\infty) - n(t,0) = 0 - p(t) = -\rho e^{-\gamma N(t-h)}. \tag{18}$$

From previous equations (16), (17), (18) and the definition of μ (15) we finally obtain:

$$\frac{\partial}{\partial t}N(t) = -\mu N(t) + \rho e^{-\gamma N(t-h)}. \tag{19}$$

It is the wanted equation for the general amount of companies on the market, including four coefficients.

6. The experimental verification of the reduced model

In our computation we use data concerning Poland gained from the Department of Central Statistical Office in Cracow. Numbers categorized in sections (A-Q) are taken from the Statistical Classification of Economic Activities (Polska Klasyfikacja Działalności - PKD2004). Data used below include only private companies. Table 3 contains the area of interest of companies categorized by PKD2004 sections. For each section, the amount of newly created companies is given in table 4. Table 5 holds the general number of companies in Poland between 2003 and 2009.

Table 3. PKD2004 sections

PKD section	Area of companies activity
A	Agriculture, hunting and forestry
B	Fishing
C	Mining and quarrying
D	Manufacturing
E	Manufacturing and electricity, gas and water supply
F	Construction
G	Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods
H	Hotels and restaurants
I	Transport, storage and communications
J	Financial intermediation
K	Real estate, renting and business activities
L	Public administration and defence; compulsory social and health security
M	Education
N	Health and social work
O	Other community, social and personal service activities
P	Private households with employed persons
Q	Extra-territorial organisations and bodies

Table 4. Private business entity registered in Poland by PKD2004 sections

PKD section / Year	2003	2004	2005	2006	2007	2008	2009
A	9696	4878	4990	5172	4863	4748	5296
B	163	141	156	126	118	146	186
C	119	131	187	163	207	256	359
D	19401	19369	22926	25985	25391	24252	30 408
E	129	155	185	227	337	597	721
F	18086	19542	27456	39113	49833	57585	46912
G	86450	76045	86339	87585	80106	82495	100229
H	11110	9807	10641	10515	11900	12508	16137
I	13725	12884	13555	15899	18225	20289	20469
J	9823	9756	9739	11614	13433	15695	15179
K	43491	42862	46167	55207	48515	52164	59679
L	917	605	294	731	309	240	229
M	3847	3432	4392	5125	5538	6131	7728
N	6673	8260	8893	9262	10739	13955	17374
O	17609	16668	18550	26532	21842	23007	25975
P	2	5	8	5	3	6	7
Q	0	27	9	10	10	8	26

Table 5. General number of private companies registered in Poland by PKD2004 sections

PKD section/ Year	2003	2004	2005	2006	2007	2008	2009
A	99337	82883	86039	89011	91459	92954	93168
B	2027	2108	2037	1961	1955	1983	2007
C	1944	1998	2124	2190	2352	2544	2929
D	379731	375662	376043	373562	373680	370812	363885
E	2111	2177	2269	2338	2609	3176	3838
F	359569	354658	357169	366705	392112	424371	428235
G	1198978	1188549	1184713	1160369	1149307	1135963	1096819
H	111200	112188	113955	111853	112379	114274	117566
I	269035	262332	260748	258662	262696	268459	267999
J	127519	128890	129239	129480	132896	137018	130196
K	490884	506628	523701	542086	549543	585702	599189
L	14138	14719	14993	15718	16018	16213	16416
M	42242	43671	45788	47259	48998	51848	55595
N	138488	143915	149000	152729	158132	167059	176261
O	214876	221726	229979	243266	251877	260080	267925
P	192	180	179	20	20	27	31
Q	7	33	48	61	70	82	126

In the year 2003 reorganization in the Statistical Classification of Economic Activities was made. The reorganization caused logging out of companies which ended their activity before but have not been logged out from REGON system yet. For that reason our simulation is based on data from the years 2003-2009.

A significant role in the experimental verification of equation (19) is played by parameter h, which stands for interval between the death of the old company and the birth of the new one. Because we posses yearly data, we will consider only the case $h = 1$. Values of parameters ρ and γ are estimated using the nonlinear least squares method applied to the computer program R. Model (19) includes constant μ expressing empirical probability of bankruptcy announcement in the time unit. The mean value for years 2003-2009 of μ is taken to our computation. Applying the computed values of parameters to the model (19) allows us to predict the amount of newly created, as well as the general number of private companies in year 2010 (actual values are not available yet).

Table 6. Estimated values of parameters from the model (19) and predictions about year 2010

PKD section	ρ	γ	μ	Business entity prediction 2010	General number prediction 2010
A	5624.975	0.000001324436	0.043605	4972	94039
B	84.62812	-0.0002693095	0.072086	145	2008
C	7.140696	-0.001533788	0.04073	638	3427
D	554.2193	-0.00000440886	0.065337	27573	368972
E	22.97401	-0.001104265	0.046393	1592	5189
F	3.757.165	-0.00000626922	0.070624	55057	451411
G	0.01901632	-0.00000759622	0.077149	79	1018335
H	0.2324422	-0.00009622672	0.089583	19032	125367
I	58461.39	0.000004710672	0.059661	16542	268521
J	44.62851	-0.00004303992	0.081616	12113	131570
K	10178.98	-0000003.00623	0.061743	61659	622418
L	180625.6	0.0004020162	0.001499	246	16637
M	158.6318	-0.00007498634	0.069748	10254	61556
N	126.8506	-0.00002943065	0.033525	22708	192515
O	4352.547	-0.00000683107	0.049436	27139	281164
P	5.253919	-0.0007181747	0.698264	5	21
Q	21.20412	0.007387905	0.021459	8	131

While analyzing results of this prediction we can observe that the general number of private companies seems to be acceptable, as well as the predicted number of the newly created private business entities. Estimated values are acceptable and fit correctly to previous data.

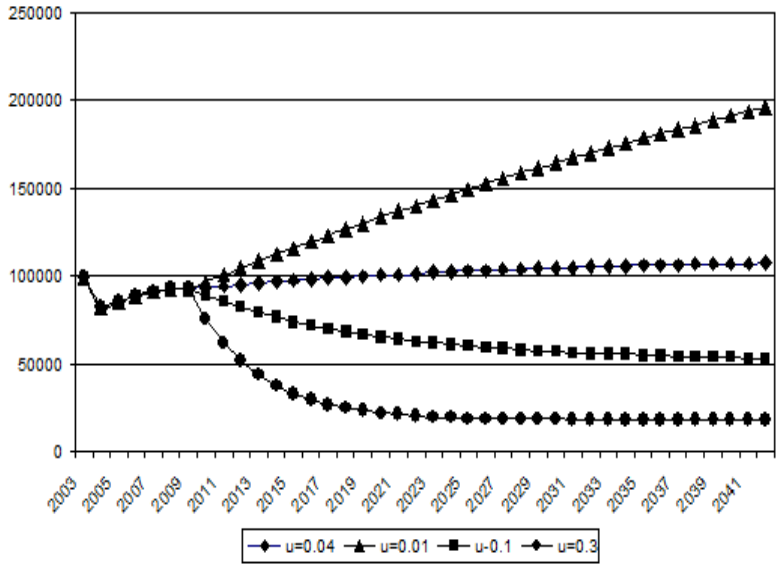
7. Long-term forecast

The limited sample size that is available is not sufficient to predict the number of registered companies in the long perspective. While examining data, it can be observed that in most cases in the given period of time in Poland the number of private companies in the REGON system, as well as of those recently registered in each section, increased. It may be the result of a considerable economic growth that has lately taken place in Poland. What we often get is the situation where we predict the constantly increasing trend. It leads to the conclusion that our model gets more interesting application in relation to more stable markets in which the decreasing number of companies leads to the increase in the amount of newly created ones and vice versa. This dependency was proved only for sections A, I, L and Q.

Let us take a closer look on the data from the A section. According to the model, the amount of private companies in the section A will constantly increase and is going to stabilize on the level of 111315.

Another application of the model (19) is a simulation of behaviour of the amount of companies related to the changing values of the parameter μ. Recall that μ represents empirical probability of bankruptcy announcement in the time unit. This kind of prediction can be applied to a situation where risk of bankruptcy is increasing (for example caused by global crisis) or decreasing (for example caused by decreasing taxes). Figure 2 illustrates long-term forecasts for different values of μ.

Figure 2. Long-term forecasts for section A for different values of μ



Our computations show that if empirical probability of bankruptcy decreases from 4.3 percent to 1 percent, the general amount of private companies in the section A in the long-term will increase to the level of 352613. In case the empirical probability of liquidation increases to 10 percent, the general number heads to 52473, and in case of probability 30 percent to 18301.

8. Final remarks

The creation and bankruptcy of companies is an inherent part of the market economy. Among the newly set up firms only those manage to endure which are flexible enough to respond in time to the changing circumstances. The remaining firms not meeting the demands of the market, by means of their insolvency release work force and resources, which can be and, in most of the cases, are used by newly created companies. The bankruptcy process of ineffective companies is harmful to employees, employers, creditors and related firms, but it improves innovativeness, creativity and efficiency of both recently started and already existing companies that have not failed. The consequence of this practice is the acceleration of the socio-economic development of different countries. However, this process of creative destruction should be monitored so as to maintain certain control over it. Therefore, the attempts of predicting insolvency of companies using financial indicators or certain models, including econometric ones, are of much significance in this context. The above presented author's model, which is a differential equation, applies to the last mentioned conception.

The undeniable advantage of the presented model is the fact that it is based on plausible, simple assumptions and it leads to interesting dependences. The verification of this model based on data from Poland proved correspondence of the model with the existing reality. Using the stationary model we can predict empirical probability of announcing bankruptcy for firms depending on their age. The used model allows one also to predict the future amount of newly created and the general number of companies on the market. As presented for data from the Statistical Classification of Economic Activities, the model can be applied to simulations of changes in the amount of companies for different empirical probability of companies death.

These introductory, encouraging results motivate to conduct further researches on the implementation of this model and creating its versions in order to respond to new scientific aims.

Further research ought to concentrate on the application of this model to other regions of Poland in order to compare obtained results. However, the main problem is too short time series of data available to particular provinces of Poland. Therefore, much value would be attributed to the application of this model to countries which have stable market economies, so that larger samples concerning births and deaths of companies can be extracted.

REFERENCES

AGIAKOGLU, C. & NEWBOLD, P., 1992, Empirical Evidence on Dickey–Fuller Type Tests, *Journal of Time Series Analysis*, 13, pp. 471–483.

ANWAR, M.S., DAVIES, S. & SAMPATH, R.K., 1996, Causality between Government Expenditures and Economic Growth: An Examination Using Cointegration Techniques, *Public Finance*, 51, pp. 166–184.

ALTMAN, E., 1968, Financial Ratios, Discriminant Analysis and the Prediction of the Corporate Bankruptcy, *The Journal of Finance*, 23, pp. 589-609.

BEBCHUK, L.A., 2002, Ex Ante Costs of Violating Absolute Priority in Bankruptcy, *The Journal of Finance*, 57, pp. 445-460.

BIAŁASIEWICZ, M. AND BUCZKOWSKI, T., 1996, Restrukturyzacja przedsiębiorstw i jej skutki na przykładzie niektórych przedsiębiorstw Szczecina, in: Przedsiębiorstwo w procesie transformacji, *Wydawnictwo Naukowe Uniwersytetu Szczecińskiego*.

BRATNICKI, M., 2001, Pułapki i problemy zarządzania strategicznego, in: Instrumenty Zarządzania we współczesnym przedsiębiorstwie, *Akademia Ekonomiczna w Poznaniu.*

CZAJKA, D., 1999, Przedsiębiorstwo w kryzysie, upadłość lub układ, *Wydawnictwo Zrzeszenia Prawników Polskich*, pp. 17-19.

DURLIK, I., 1998, Restrukturyzacja procesów gospodarczych: reengineering, teoria i praktyka, *Agencja Wydawnicza Placet*.

ENCYKLOPEDIA BIOLOGICZNA, 1998, Otałęga (Ed), *Opress*, pp. 341.

FOSTER, R. AND KAPLAN, S., 2001, Creative Destruction, *Financial Times*.

FREDERICK, W.C. AND DAVIS, K. AND POST, J.E., 1988, Corporate Social Responsibility and Business Ethics*, McGraw-Hill Publishing Company, New York, pp. 28-29.*

GREINER, L.E. AND SCHEIN, V.E., 1988, Power and Organization Development, *Addison-Wesley*.

HADASIK, P., 1988, Upadłość przedsiębiorstw w Polsce i metody jej prognozowania, *Zeszyty Naukowe – seria II*, 153, AE Poznań, pp. 81-91.

HANDY, CH., 1995, 'THE AGE OF PARADOX', HARVARD BUSINESS SCHOOL PRESS.

HOŁDA, A., 2001, Prognozowanie bankructwa jednostki w warunkach gospodarki polskiej z wykorzystaniem funkcji dyskryminacyjnej ZH, *Rachunkowość*, 5, pp. 306-310.

HURST, D.K., 1995, Crisis and renewal: meeting the challenge of organizational change, *Harvard Business School Press*, Boston.

KOŹMIŃSKI, A.K. AND PIOTROWSKI, W., 1999, Zarządzanie teoria i praktyka, *Wydawnictwo Naukowe PWN*.

MAJCHRZAK, J., 2003, Społeczny wymiar zarządzania przedsiębiorstwem. Mit czy konieczność?, in: Współczesne tendencje w zarządzaniu, *Zeszyty Naukowe*, 33, Wydawnictwo Akademii Ekonomicznej w Poznaniu.

MATSCHKE, J.M. AND BROESEL, G., 2007, Unternehmensbewertung, *Gabler-Verlag*.

OSBERT-POCIECHA, G., 2004, Twórcza destrukcja jako uwarunkowanie innowacyjnego rozwoju przedsiębiorstwa, *Prace Naukowe Akademii Ekonomicznej im. Oskara Langego we Wrocławiu*, 1045, pp. 273-278.

PERECHUDA, K., 1999, Metody zarządzania przedsiębiorstwem, *Wydawnictwo Akademii Ekonomicznej we Wrocławiu*.

SHONFELDER, B., 2003, Death or Survival. Post-Communist Bankruptcy Law in Action, *A Survey*, Freiberg, 2003, pp. 3.

SHUMPETER, J., 1982, The Theory of Economic Development: An inquiry into profits, capital, credit, interest and the business cycle, *Transaction Publisher*.

WAŻEWSKA-CZYŻEWSKA, M. AND LASOTA, A., 1976, Matematyczne problemy dynamiki układu krwinek czerwonych, *Roczniki Polskiego Towarzystwa Matematycznego*, Seria III. Matematyka Stosowana VI, pp. 24-40.

ZDYB, M., 2008, Jakie czynniki generują upadłości przedsiębiorstw w Polsce? Przyczyny upadłości przedsiębiorstw w Polsce, *Biuletyn E-rachunkowość*.

ZEDLER, F., 2003, Prawo upadłościowe i naprawcze–wprowadzenie, *Zakamycze*, Kraków.

ZIMNIEWICZ, K., 2000, Współczesne koncepcje i metody zarządzania, *Polskie Wydawnictwo Ekonomiczne*, Warszawa.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 401—412

IMPROVED SEPARATE RATIO EXPONENTIAL ESTIMATOR FOR POPULATION MEAN USING AUXILIARY INFORMATION

Rohini Yadav[1], Lakshmi N. Upadhyaya[1], Housila P. Singh[2] and S.Chatterjee[1]

ABSTRACT

This paper advocates the improved separate ratio exponential estimator for population mean $\bar{Y}$ of the study variable y using the information based on auxiliary variable x in stratified random sampling. The bias and mean squared error (MSE) of the suggested estimator have been obtained upto the first degree of approximation. The theoretical and numerical comparisons are carried out to show the efficiency of the suggested estimator over sample mean estimator, usual separate ratio and separate product estimator.

Key words: Study variable, auxiliary variable, stratified random sampling, separate ratio estimator, separate product estimator, bias and mean squared error.

1. Introduction

In sampling theory the use of the proper auxiliary information always increases the precision of an estimator. Stratification is one of the design tools which yield increased precision. Stratified sampling entails first dividing the whole population of N units into non-overlapping subpopulations of $N_1, N_2, ..., N_L$ units, respectively, called strata that together comprise the entire population, so that $N_1+N_2+...+N_L=N$ and then drawing an independent samples of size $n_1, n_2, ..., n_L$ from each stratum. If the sample in each stratum is a simple random sample, the whole procedure is described as stratified random sampling. We can stratify the population in such a manner that (i) within each

[1] Department of Applied Mathematics, Indian School of Mines, Dhanbad-826004, India.
E-mail: rohiniyadav.ism@gmail.com, lnupadhyaya@yahoo.com, schat_1@yahoo.co.in,
[2] School of Studies in Statistics, Vikram University, Ujjain-456 010, India.
E-mail: hpsujn@gmail.com

stratum there is as uniformity as possible and (ii) among various strata the difference are as great as possible.

To obtain the full benefits from stratification, the values of the N_h must be known. We use this technique because when we divide heterogeneous population into relatively more homogenous sub population, it reduces heterogeneity and hence increases precision of the estimator. This technique is also preferred because of its administrative convenient in carrying out the survey.

The ratio estimate of the population mean $\bar{Y}$ can be made in two ways. One is to make a separate ratio estimate of the total of each stratum and add these totals. An alternative estimate is derived from a single combined ratio. Many authors Kadilar and Cingi (2003), Singh and Vishwakarma (2006), Singh and Vishwakarma (2010), Koyuncu and Kadilar (2010) etc. have suggested the estimators of population parameters in stratified random sampling.

Let the population of size N is equally divided into L strata with N_h elements in the h^{th} stratum such that $N=\sum_{h=1}^{L} N_h$. Let n_h be the size of the sample drawn from h^{th} stratum of size N_h by using simple random sampling without replacement (SRSWOR) such that sample size $n=\sum_{h=1}^{L} n_h$. Let y and x be the study and the auxiliary variables, respectively, assuming values y_{hi} and x_{hi} for the i^{th} unit in h^{th} stratum.

Let $W_h=(N_h/N)$ be the stratum weight, $f_h=(n_h/N_h)$ be the sampling fraction, $\left[\bar{Y}_h=(1/N_h)\sum_{i=1}^{N_h} y_{hi},\ \bar{X}_h=(1/N_h)\sum_{i=1}^{N_h} x_{hi}\right]$ and $\left[\bar{y}_h=(1/n_h)\sum_{i=1}^{n_h} y_{hi},\ \bar{x}_h=(1/n_h)\sum_{i=1}^{n_h} x_{hi}\right]$ be the population means and sample means of the study variate y and the auxiliary variate x respectively. Our purpose is to estimate the population mean $\bar{Y}=\sum_{h=1}^{L} W_h\bar{Y}_h=\bar{Y}_{st}$ of the study variable y.

When the population mean $\bar{X}_h$ of the h^{th} stratum of the auxiliary variable x is known then the usual separate ratio, product and regression estimators for population mean $\bar{Y}$ are respectively given as

$$\bar{y}_{rs}=\sum_{h=1}^{L} W_h\left(\frac{\bar{y}_h}{\bar{x}_h}\bar{X}_h\right) \tag{1.1}$$

$$\bar{y}_{ps}=\sum_{h=1}^{L} W_h\left(\bar{y}_h \frac{\bar{x}_h}{\bar{X}_h}\right) \tag{1.2}$$

$$\bar{y}_{lrs}=\sum_{h=1}^{L} W_h\left[\bar{y}_h+b_h\left(\bar{X}_h-\bar{x}_h\right)\right] \tag{1.3}$$

where $b_h=\left(s_{yxh}/s_{xh}^2\right)$ is the sample regression coefficient of y on x of the h^{th} stratum, $s_{yxh}=\left\{1/\left(n_h-1\right)\right\}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)\left(x_{hi}-\bar{x}_h\right)$ is the sample covariance between y and x, $s_{yh}^2=\left\{1/\left(n_h-1\right)\right\}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2$ is the sample mean square/variance of y and $s_{xh}^2=\left\{1/\left(n_h-1\right)\right\}\sum_{i=1}^{n_h}\left(x_{hi}-\bar{x}_h\right)^2$ is the sample mean square/variance of x in the h^{th} stratum respectively.

We know that

$$\mathrm{Var}\left(\bar{y}_{st}\right)=\sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2 \tag{1.4}$$

where $S_{yh}^2=\left\{1/\left(N_h-1\right)\right\}\sum_{i=1}^{N_h}\left(y_{hi}-\bar{Y}_h\right)^2$ is the population mean square/variance of the study variate y.

The mean squared error of the estimators $\bar{y}_{rs}$, $\bar{y}_{ps}$ and $\bar{y}_{lrs}$ are respectively given by

$$\mathrm{MSE}\left(\bar{y}_{rs}\right)=\sum_{h=1}^{L} W_h^2\gamma_h\left[S_{yh}^2+R_h^2S_{xh}^2-2R_hS_{yxh}\right] \tag{1.5}$$

$$\mathrm{MSE}\left(\bar{y}_{ps}\right)=\sum_{h=1}^{L} W_h^2\gamma_h\left[S_{yh}^2+R_h^2S_{xh}^2+2R_hS_{yxh}\right] \tag{1.6}$$

$$\mathrm{Var}\left(\bar{y}_{lrs}\right)=\sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2\left(1-\rho_h^2\right) \tag{1.7}$$

where $\gamma_h=\left(\frac{1}{n_h}-\frac{1}{N_h}\right)$, $R_h=\left(\bar{Y}_h/\bar{X}_h\right)$ and $\rho_{hyx}=\left(S_{hyx}/S_{yh}S_{xh}\right)$.

In this paper, we suggested an improved separate ratio exponential estimator of the population mean $\bar{Y}$ of the study variable y using the supplementary information of the auxiliary variable x. The bias and mean squared error have been obtained upto the first degree of approximation.

2. An improved separate ratio exponential estimator

Motivated by Upadhyaya et al (2011), we suggested a separate ratio exponential estimator $t_{RS}^{(a)}$ of the population mean $\bar{Y}$ of the study variable y is defined as

$$t_{RS}^{(a)}=\sum_{h=1}^{L} W_h \bar{y}_h \exp\left[\frac{\bar{X}_h-\bar{x}_h}{\bar{X}_h+(a_h-1)\bar{x}_h}\right] \qquad (2.1)$$

To obtain the bias and mean square error (MSE) of the estimator $t_{RS}^{(a)}$ at (2.1), we write

$\bar{y}_h=\bar{Y}_h(1+e_{0h})$ and $\bar{x}_h=\bar{X}_h(1+e_{1h})$

such that $E(e_{0h})=E(e_{1h})=0$

and ignoring the finite population correction (fpc) term, we have

$$E(e_{0h}^2)=\frac{1}{\bar{Y}_h^2}\gamma_h S_{yh}^2,\quad E(e_{1h}^2)=\frac{1}{\bar{X}_h^2}\gamma_h S_{xh}^2,\quad E(e_{0h}e_{1h})=\frac{1}{\bar{Y}_h\bar{X}_h}\gamma_h S_{yxh} \qquad (2.2)$$

Expressing (2.1) in terms of e's, we have

$$t_{RS}^{(a)}=\sum_{h=1}^{L} W_h \bar{Y}_h(1+e_{0h})\exp\left[-\frac{e_{1h}}{a_h}\left\{1+\left(\frac{a_h-1}{a_h}\right)e_{1h}\right\}^{-1}\right]$$

$$=\sum_{h=1}^{L} W_h \bar{Y}_h(1+e_{0h})\left[1-\frac{e_{1h}}{a_h}\left\{1+\left(\frac{a_h-1}{a_h}\right)e_{1h}\right\}^{-1}+\frac{e_{1h}^2}{2a_h^2}\left\{1+\left(\frac{a_h-1}{a_h}\right)e_{1h}\right\}^{-2}-\ldots\right]$$

$$=\sum_{h=1}^{L} W_h \bar{Y}_h(1+e_{0h})\left[1-\frac{e_{1h}}{a_h}\left\{1-\left(\frac{a_h-1}{a_h}\right)e_{1h}+\left(\frac{a_h-1}{a_h}\right)^2 e_{1h}^2-\ldots\right\}+\frac{e_{1h}^2}{2a_h^2}\left\{1-2\left(\frac{a_h-1}{a_h}\right)e_{1h}+3\left(\frac{a_h-1}{a_h}\right)\right.\right.$$

$$=\sum_{h=1}^{L} W_h \bar{Y}_h(1+e_{0h})\left[1-\frac{e_{1h}}{a_h}+\frac{(a_h-1)}{a_h^2}e_{1h}^2+\frac{1}{2a_h^2}e_{1h}^2+\ldots\right]$$

$$=\sum_{h=1}^{L} W_h \bar{Y}_h \left[1+e_{0h}-\frac{e_{1h}}{a_h}-\frac{e_{0h}e_{1h}}{a_h}+\frac{e_{1h}^2}{a_h^2}\left(a_h-\frac{1}{2}\right)+\ldots \right]$$

Neglecting the terms of e's having power greater than two, we have

$$\left(t_{RS}^{(a)}-\bar{Y}\right)=\sum_{h=1}^{L} W_h \bar{Y}_h \left[e_{0h}-\frac{e_{1h}}{a_h}+\frac{e_{1h}^2}{a_h^2}\left(a_h-\frac{1}{2}\right)-\frac{e_{0h}e_{1h}}{a_h} \right] \tag{2.3}$$

Taking expectation on both sides of (2.3), we have the bias of $t_{RS}^{(a)}$ upto the first degree of approximation as

$$B\left(t_{RS}^{(a)}\right)=\sum_{h=1}^{L} W_h^2 \gamma_h \frac{1}{\bar{X}_h}\left[\frac{1}{a_h^2}\left(a_h^2-a_h+\frac{1}{2}\right)R_h S_{xh}^2+\frac{S_{yxh}}{a_h} \right] \tag{2.4}$$

Squaring both sides of (2.3) and neglecting the terms having power greater than two, we have

$$\left(t_{RS}^{(a)}-\bar{Y}\right)^2 \cong \left[\sum_{h=1}^{L} W_h^2 \gamma_h \left(e_{0h}-\frac{e_{1h}}{a_h} \right) \right]^2$$

$$=\sum_{h=1}^{L} W_h^2 \bar{Y}_h^2 \left(e_{0h}-\frac{e_{1h}}{a_h} \right)^2 + \sum_{h\neq h'=1}^{L} W_h W_{h'} \bar{Y}_h \bar{Y}_{h'} \left(e_{0h}-\frac{e_{1h}}{a_h} \right)\left(e_{0h'}-\frac{e_{1h'}}{a_{h'}} \right)$$

$$\left(t_{RS}^{(a)}-\bar{Y}_h\right)^2 =\sum_{h=1}^{L} W_h^2 \bar{Y}_h^2 \left(e_{0h}^2+\frac{e_{1h}^2}{a_h^2}-2\frac{e_{0h}e_{1h}}{a_h} \right) \tag{2.5}$$

[since sampling from stratum to stratum is independent from each other]

Taking expectation on both sides of (2.5), we have the mean squared error of $t_{RS}^{(a)}$ upto the first degree of approximation as

$$MSE\left(t_{RS}^{(a)}\right)=\sum_{h=1}^{L} W_h^2 \gamma_h \left[S_{yh}^2+\frac{R_h^2}{a_h^2}S_{xh}^2-2\frac{R_h}{a_h}S_{yxh} \right] \tag{2.6}$$

where $R_h=\left(\bar{Y}_h/\bar{X}_h\right)$

The MSE of $t_{RS}^{(a)}$ is minimized at

$$a_h=\left(\frac{R_h}{\beta_h}\right)=a_{h0} \quad (\text{say}) \tag{2.7}$$

Thus the resulting minimum MSE of $t_{RS}^{(a)}$ is given by

$$\min.MSE\left(t_{RS}^{(a)}\right)=\sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2\left(1-\rho_h^2\right) \qquad (2.8)$$

which is also equal to the variance of the separate regression estimator $\bar{y}_{lrs}=\sum_{h=1}^{L} W_h\left[\bar{y}_h+b_h\left(\bar{X}_h-\bar{x}_h\right)\right]$.

3. Efficiency comparison

<u>Case I</u>. When the scalar a_h does not coincides at its exact optimum value a_{h0}

From (1.4), (1.5), (1.6) and (2.6), we have

(i) $Var\left(\bar{y}_{st}\right)-MSE\left(t_{RS}^{(a)}\right)>0$ if

$$\left[\sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2-\sum_{h=1}^{L} W_h^2\gamma_h\left(S_{yh}^2+\frac{R_h^2}{a_h^2}S_{xh}^2-2\frac{R_h}{a_h}S_{yxh}\right)\right]>0$$

$$\text{Since } \sum_{h=1}^{L} W_h^2\gamma_h\left(\frac{R_h^2}{a_h^2}S_{xh}^2-2\frac{R_h}{a_h}S_{yxh}\right)<0 \qquad (3.1)$$

$$\text{Therefore, } \left(\frac{R_h^2}{a_h^2}S_{xh}^2-2\frac{R_h}{a_h}S_{yxh}\right)<0$$

$$a_h\geq\frac{R_h}{2\beta_h} \quad \text{and} \quad R_h\leq 0 \qquad (3.2)$$

(ii) $MSE\left(t_{rs}\right)-MSE\left(t_{RS}^{(a)}\right)>0$ if

$$\left[\sum_{h=1}^{L} W_h^2\gamma_h\left(S_{yh}^2+R_h^2S_{xh}^2-2R_hS_{yxh}\right)-\sum_{h=1}^{L} W_h^2\gamma_h\left(S_{yh}^2+\frac{R_h^2}{a_h^2}S_{xh}^2-2\frac{R_h}{a_h}S_{yxh}\right)\right]>0$$

$$\sum_{h=1}^{L} W_h^2\gamma_h\left[R_h^2S_{xh}^2-\frac{R_h^2}{a_h^2}S_{xh}^2-2R_hS_{yxh}+2\frac{R_h}{a_h}S_{yxh}\right]>0 \qquad (3.3)$$

$$\text{either} \quad R_h\left(1-\frac{1}{a_h}\right)>0 \quad \text{or} \quad \left[\left(1+\frac{1}{a_h}\right)R_hS_{xh}^2-2S_{yxh}\right]>0$$

$$\left.\begin{array}{l} \text{either} \quad 1< a_h <\left(\dfrac{R_h}{2\beta_h - R_h}\right) \\ \text{or} \quad \left(\dfrac{R_h}{2\beta_h - R_h}\right) < a_h < 1 \end{array}\right\} \tag{3.4}$$

or, equivalently

$$\min.\left\{1,\left(\frac{R_h}{2\beta_h - R_h}\right)\right\} < a_h < \max.\left\{1,\left(\frac{R_h}{2\beta_h - R_h}\right)\right\} \tag{3.5}$$

(iii) $MSE\left(t_{ps}\right) - MSE\left(t_{RS}^{(a)}\right) > 0$ if

$$\left[\sum_{h=1}^{L} W_h^2 \gamma_h \left(S_{yh}^2 + R_h^2 S_{xh}^2 + 2R_h S_{yxh}\right) - \sum_{h=1}^{L} W_h^2 \gamma_h \left(S_{yh}^2 + \frac{R_h^2}{a_h^2} S_{xh}^2 - 2\frac{R_h}{a_h} S_{yxh}\right)\right] > 0$$

$$\sum_{h=1}^{L} W_h^2 \gamma_h \left[R_h^2 S_{xh}^2 - \frac{R_h^2}{a_h^2} S_{xh}^2 + 2R_h S_{yxh} + 2\frac{R_h}{a_h} S_{yxh}\right] > 0 \tag{3.6}$$

$$\text{either} \quad R_h\left(1 + \frac{1}{a_h}\right) > 0 \quad \text{or} \quad \left[\left(1 - \frac{1}{a_h}\right) R_h S_{xh}^2 + 2S_{yxh}\right] > 0$$

$$\left.\begin{array}{l} \text{either} \quad -1< a_h <\left(\dfrac{R_h}{2\beta_h + R_h}\right) \\ \text{or} \quad \left(\dfrac{R_h}{2\beta_h + R_h}\right) < a_h < -1 \end{array}\right\} \tag{3.7}$$

or, equivalently

$$\min.\left\{-1,\left(\frac{R_h}{2\beta_h + R_h}\right)\right\} < a_h < \max.\left\{-1,\left(\frac{R_h}{2\beta_h + R_h}\right)\right\} \tag{3.8}$$

Case II. When the scalar a_h coincides at its exact optimum value a_{h0}

From (1.4), (1.5), (1.6) and (2.8), we have

(iv) $\mathrm{Var}\left(\bar{y}_{st}\right)$-min.$\mathrm{MSE}\left(t_{RS}^{(a)}\right) > 0$ if

$$\left[\sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2 - \sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2\left(1-\rho_h^2\right)\right]>0$$

Since $\sum_{h=1}^{L} W_h^2\gamma_h S_{yh}^2\rho_h^2 > 0$ (3.9)

Therefore, $\rho_h^2>0$ (3.10)

(v) $\mathrm{MSE}\left(t_{rs}\right)$-min.$\mathrm{MSE}\left(t_{RS}^{(a)}\right)>0$ if

$$\sum_{h=1}^{L} W_h^2\gamma_h\left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yxh} - S_{yh}^2\left(1-\rho_h^2\right)\right]>0 \quad (3.11)$$

$$\rho_h^2 > -R_h^2\frac{S_{xh}^2}{S_{yh}^2}\left(1-2\frac{\beta_h}{R_h}\right) \quad (3.12)$$

(vi) $\mathrm{MSE}\left(t_{ps}\right)$-$\mathrm{MSE}\left(t_{RS}^{(a)}\right)>0$ if

$$\sum_{h=1}^{L} W_h^2\gamma_h\left[S_{yh}^2 + R_h^2 S_{xh}^2 + 2R_h S_{yxh} - S_{yh}^2\left(1-\rho_h^2\right)\right]>0 \quad (3.13)$$

$$\rho_h^2 > -R_h^2\frac{S_{xh}^2}{S_{yh}^2}\left(1+2\frac{\beta_h}{R_h}\right) \quad (3.14)$$

4. Empirical study

To judge the merits of the proposed estimator over usual unbiased estimator $\bar{y}_{st}$, we considered a population data set whose description is given in the Table 4.1.

Table4.1. Data Statistics I [Source: Kadilar and Cingi (2005)]

In this data set, Y is the apple production amount and X is the number of apple trees in 854 villages of Turkey in 1999. The population information about this data set is given as:

$N_1=106$	$N_2=106$	$N_3=94$
$N_4=171$	$N_5=204$	$N_6=173$
$n_1=9$	$n_2=17$	$n_3=38$
$n_4=67$	$n_5=7$	$n_6=2$
$\bar{X}_1=24375$	$\bar{X}_2=27421$	$\bar{X}_3=72409$
$\bar{X}_4=74365$	$\bar{X}_5=26441$	$\bar{X}_6=9844$
$\bar{Y}_1=1536$	$\bar{Y}_2=2212$	$\bar{Y}_3=9384$
$\bar{Y}_4=5588$	$\bar{Y}_5=967$	$\bar{Y}_6=404$
$C_{x1}=2.02$	$C_{x2}=2.10$	$C_{x3}=2.22$
$C_{x4}=3.84$	$C_{x5}=1.72$	$C_{x6}=1.91$
$C_{y1}=4.18$	$C_{y2}=5.22$	$C_{y3}=3.19$
$C_{y4}=5.13$	$C_{y5}=2.47$	$C_{y6}=2.34$
$S_{x1}=49189$	$S_{x2}=57461$	$S_{x3}=160757$
$S_{x4}=285603$	$S_{x5}=45403$	$S_{x6}=18794$
$S_{y1}=6425$	$S_{y2}=11552$	$S_{y3}=29907$
$S_{y4}=28643$	$S_{y5}=2390$	$S_{y6}=946$
$\rho_1=0.82$	$\rho_2=0.86$	$\rho_3=0.90$
$\rho_4=0.99$	$\rho_5=0.71$	$\rho_6=0.89$
$\beta_2(x_1)=25.71$	$\beta_2(x_2)=34.57$	$\beta_2(x_3)=26.14$
$\beta_2(x_4)=97.60$	$\beta_2(x_5)=27.47$	$\beta_2(x_6)=28.10$
$\gamma_1=0.102$	$\gamma_2=0.049$	$\gamma_3=0.016$
$\gamma_4=0.009$	$\gamma_5=0.138$	$\gamma_6=0.006$
$\omega_1^2=0.015$	$\omega_2^2=0.015$	$\omega_3^2=0.012$
$\omega_4^2=0.04$	$\omega_5^2=0.057$	$\omega_6^2=0.041$

For the purpose of the efficiency comparison of the proposed estimator, we have computed the percent relative efficiencies (PREs) of the estimators with respect to the usual unbiased estimator $\bar{y}_{st}$ using the formula:

$$PRE\left(t,\bar{y}_{st}\right)=\frac{MSE\left(\bar{y}_{st}\right)}{MSE\left(t\right)}X100\,;\quad \text{where } t=\left(\bar{y}_{st},\ \bar{y}_{rs},\ \bar{y}_{ps} \text{ and } t_{RS}^{(a)}\right)$$

The findings are given in the Table 4.2.

Table4.2. PREs of the estimators $\left(\bar{y}_{st},\ \bar{y}_{rs},\ \bar{y}_{ps} \text{ and } t_{RS}^{(a)}\right)$ with respect to the usual unbiased estimator $\bar{y}_{st}$

S. No.	Estimator	$PRE\left(.,\ \bar{y}_{st}\right)$
1.	$\bar{y}_{st}$	100.00
2.	$\bar{y}_{rs}$	423.2052
3.	$\bar{y}_{ps}$	37.6085
4.	$t_{RS}^{(a)}$	629.0317

From Table 4.2, it is clear that the suggested improved separate ratio exponential estimator $t_{RS}^{(a)}$ is more efficient than the unbiased sample mean estimator $\bar{y}_{st}$, usual separate ratio estimator $\bar{y}_{rs}$ and the usual separate product estimator $\bar{y}_{ps}$.

5. Conclusion

It is observed that the suggested improved separate ratio exponential estimator $t_{RS}^{(a)}$ is more precise than $\bar{y}_{st}$, $\bar{y}_{rs}$ and $\bar{y}_{ps}$. Thus the use of proposed estimator is justified in practice.

Acknowledgements

The authors acknowledge the University Grants Commission, New Delhi, India for financial support in the project number F. No. 34-137/2008(SR). The authors are also thankful to Indian School of Mines, Dhanbad and Vikram University, Ujjain for providing the facilities to carry out the research work.

REFERENCES

KADILAR, C., CINGI, H. (2003): Ratio estimators in stratified random sampling. *Biometrical Journal,* 45, 2, 218-225.

KADILAR, C. and CINGI, H. (2005): A new estimator in stratified random sampling, *Commun. in Statist.: Theory and Meth.*, 34, 597-602.

KOYUNCU, N. and KADILAR, C. (2010): On the family of estimators of population mean in stratified random sampling, *Pak. J. Statist.,* vol. 26(2), 427-443.

SINGH, H.P. and VISHWAKARMA, G.K. (2006): An efficient variant of the product and ratio estimators in stratified random sampling, *Statistics in Transition*, 7(6), 1311-1325.

SINGH, H.P. and VISHWAKARMA, G.K. (2010): A general procedure for estimating the population mean in stratified sampling using auxiliary information. *Metron,* vol. LXVIII, n.1, pp 47-65.

UPADHYAYA, L.N., SINGH, H.P., CHATTERJEE, S. and YADAV, R. (2011): Improved ratio and product exponential type estimators. *Journal of Statistical Theory and Practice*, vol. 5, no. 2, pp. 285-302.

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 413—414

BOOK REVIEW

Tadeusz Walczak: Dictionary of Statistical Terms: English-Polish and Polish-English, Wydawnictwo C.H. Beck, Warszawa 2011

The Anglo-Polish and Polish-English ***Dictionary of Statistical Terms*** is addressed both to all who use the economic and statistical literature and methodological documents, as well as to research workers, university students and all other persons using statistical literature and statistical publications disseminated by different countries and international bodies. This publication may be an essential assistance for statistical texts translators.

The Dictionary consists of ***preface*** in Polish, ***biography***, and ***three annexes*** after Part I English–Polish: Annex 1– ***acronyms of statistical terms, computer science*** and general; Annex 2 – ***acronyms of institutions, associations and international organizations***; and Annex 3: - ***spellings of British and American of terms*** used in the Dictionary.

Preparing the ***Dictionary*** the author used a number of publications: statistical handbooks, methodological papers, statistical publications, documents published by international organizations, and other sources of lexicon character which are given in the biography.

The present version of the Dictionary is result of many years of the author's involvement in statistical terminology. In 1997 the author published the first edition of dictionary of English-Polish[1], commonly used mainly by employees of official statistics in Poland. I prepared a review of that Dictionary and published it in 1998 in "Wiadomości Statystyczne"[2]. This Dictionary was also used by the authors of the "ISI Multilingual Glossary of Statistical Terms" prepared by the International Statistical Institute[3] for supplementing the Polish terms.

---

[1] Tadeusz Walczak. *Słownik terminów statystycznych angielsko-polski,* wyd. GUS, Warszawa, 1997.

[2] J. Kordos: Review of: Tadeusz Walczak: *Słownik terminów statystycznych angielsko-polskich*, GUS, Warszawa,1997, 247 pages, *Wiadomości Statystyczne*, nr 3, February 1998, pp. 89-91.

[3] See: http://isi.cbs.nl/glossary.htm.

Prof. dr hab. Tadeusz Walczak **–** professor emeritus of Warsaw School of Economics. A vice President of the Central Statistical Office of Poland in 1972–1990, an Editor-in-Chief of "***Wiadomości Statystyczne***" (Statistical News) – a monthly journal of the Central Statistical Office of Poland since 1994. The Author of many methodological texts and statistical documents translations. A consultant for English versions of statistical publications issued by the Central Statistical Office of Poland.

Prepared by Jan Kordos, jan1kor2@aster.pl

STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 415—423

The ISI Satellite Conference on Improving Statistical Systems Worldwide - Building Capacity

Krakow, Poland, 18-19 August 2011

The ISI Satellite Conference on Improving Statistical Systems Worldwide - Building Capacity was held in Krakow, Poland from 18 to 19 August 2011, as a satellite of the 58th ISI Congress held in Dublin, Ireland, from 21 to 26 August 2011. The conference was held in the Collegium Magnum of the Jagelonian University. It brought together over 100 participants from countries of all continents andfourteen international institutions – statisticians, researchers and related experts – to share their experience and focus on the recent changes of policy regarding the work on improving statistics in developing countries.

The satellite conference was organized by the Polish Statistical Office in cooperation with the World Bank, African Development Bank, Paris 21 and the International Statistical Institute. It was financed by the World Bank, Polish Statistical Office and ISI.

Polish Statistical Office and the World Bank co-chaired the Organizing committee and the program committee. The Polish Statistical Office in Krakow chaired the local organizing committee.

The conference was opened by the President of the Polish Statistical Office Janusz Witkowski. The deputy mayor of Krakow Elżbieta Lęcznarowicz and the ISI representative Frederick Vogel spoke at the opening ceremony.

The first session of the conference was started by the World Bank representative Misha Belkindas who, in his key note address spoke about the role of statistics in the implementation of actions such as Millennium Development Goals, Marrakech Action Plan for Statistics and Dakar Declaration, which were undertaken for the economic and social development of countries. He discussed the current activities aimed at supporting developing countries in preparing their national strategies for the development of statistics that will improve the coordination and management of statistical systems. He stressed that now it is necessary to accede to implement strategic development plans and actions to ensure stability of statistical systems. This should result in better statistics and more efficient use of statistics. The role of statistics, apart from providing information, is also supporting decision making at various levels, which implies that statistical systems should not only respond quickly to changes, but also anticipate them.

During next presentations, representatives of Paris21 consortium, Afghanistan and India shared their experience in the development and implementation of statistical strategies and national development plans. Eric Bensel from Paris21 stressed that the gathering of statistical data for informing, monitoring and evaluation of national development plans, requires strategic planning in the form of the creation of the National Strategy for the Development of Statistics (NSDS). Experience shows that what determines the success of the strategy is the design stage, which includes the involvement of all participants (producers, users and funders), the integration of needs of the sector statistics, the inclusion of training and human resource development, as well as synchronization with the development processes of the country. Only a well-developed strategy will ensure its effective implementation. Abdul Rahman Ghafoori from Afghanistan pointed out, however, that a comprehensive cooperation, which takes into account the needs of all users in developing countries, is often impossible. In Afghanistan, after 30 years of war, a major achievement was the development of the 2010 Afghanistan National Statistical Plan, to provide an integrated system for collection and possession of reliable statistics by the Central Statistical Organization. A pragmatic approach "learning by doing" was introduced, with the focus on the implementation of several key actions. Coordination of financing is necessary but it still remains a challenge for the future. As for the statistical system in India, properly trained statistical staff is a challenge. Kuldeep K. Lamba elaborated on training of staff in his country (also with the assistance of the World Bank), starting with intensive training of new employees, through continuous training in methodology and statistical research and new technologies in National Academy of Statistical Administration, which was established specially for this purpose. There are also promotional activities and traineeship for students carried out to interest them in taking up employment in the statistics. Particular attention is focused on statistical capacity building at local level.

The subject of the next session (chaired by Wlodzimierz Okrasa) was the issue of partnership in the national statistical systems. Grace Bediako from Ghana emphasized that the major challenge for the national statistical system is decentralization of production and distribution of harmonized statistics. National Statistical Offices as coordinators in the field of statistics often have limited opportunities to ensure that the data made available by different companies are compatible with the required standards for dissemination. Strengthening the position of national statistical offices can be achieved, among others, by providing a competent, professional staff. Offices possessing highly qualified staff will be more reliable for data users than other data producers. To strengthen the national statistical system one can employ appropriately designed training and also, e.g. Virtual Statistical System (VSS). Grace Bediako also stated that in order to improve the effectiveness of the entire statistical system, in addition to changes in legislation, a fundamental change in governing independency of companies that collect statistical information is needed. A good opportunity to improve the coherence of national statistics is to develop national statistics development

strategies (NSDS) and national statistical systems (NSS). The adoption of the United Nations Fundamental Principles of Official Statistics also greatly increases the chances of achieving greater consistency and improve data quality.

The problem of building a partnership for the development of statistics in developing countries was also touched by Graham Eele from the World Bank. He recalled that in accordance with the Marrakech Action Plan, the effective development of statistical systems in low-income countries requires that the countries themselves determine the priorities and define their own path of development. Since 2004 significant progress has been observed in supporting countries and their preparation of strategies for the development of statistics. The challenge now is to find ways to implement these strategies into effect in a way that causes lasting improvement in productivity and utilization of statistics. An important part of this process is to build a national partnership. Strong and effective national partnership is part of a comprehensive sectoral approach and has been used successfully for many years in promoting development in areas such as health, education and agriculture. The World Bank funds, together with other partners, such an approach to the development of statistical systems in developing countries.

Technical assistance to developing countries such as training and staff is directed according to the needs or to individual countries or groups of countries. Coordination of statistical training in Africa was discussed during the next session by Dimitri Sanga of UNECE (United Nations Economic Commission for Africa). One of the major challenges facing the quality and timeliness of statistics to serve the development of African countries is inadequate and poorly educated human resources. In response to huge demand of these countries in the area of statistical training, many initiatives and actions aimed at improving the situation in this regard are undertaken. In order to coordinate activities, the interested parties agreed to combine statistical training initiatives in Africa. The created Group which is subject to the Statistical Commission for Africa, is the supreme body of statistics and statistical development in Africa. Since the recognition of the Group by the Statistical Commission for Africa in January 2010 it has contributed to effectively conduct many statistical training directed to African countries. The Group created the necessary conditions for the supply and demand for statistical training, as well as provided technical and financial support for partners through the exchange of information and best practices. The basis of the group's work is the preparation and implementation of the Statistical Training Programme for Africa under the Reference Regional Strategic Framework for Statistical Capacity Building in Africa.

The next speaker, Oliver Chinganya from AfDB (African Development Bank) presented the activities undertaken to build a sustainable statistics in Africa. Statistical systems of African countries are not able to meet the increasing demand for statistical data necessary for the implementation of many initiatives such as poverty reduction strategies or Millennium Development Goals. While there has been overall progress in the development of statistics in many countries,

there is still a deficit in the capability to provide users with high-quality statistical data. The difficulties relate to such key areas as the use of data, management, statistical and technical infrastructure, human and financial resources. The initiatives of international organizations related to the development of statistical capacity of African countries are focused more on management of development performance consisting in stepping up the efficiency of the use of data, increasing involvement of governments in the development of statistics, as well as development and increase of financial aid. In addition, in Africa currently a number of initiatives is being undertaken aimed at strengthening the role of national and regional statistical institutions by developing strategies, creating committees, organizing symposia, conferences, etc.

The idea of the Virtual Statistical System (https://www.virtualstatisticalsystem.org/) as a tool for teaching online, which is used to support institutions of developing countries in creating sustainable statistics was presented by Ronald Luttikhuizen from the World Bank. This system via the website provides employees with permanent access to a wide range of training materials. It enables training both at work and at home which ensures that it is durable and sustainable. The training with the use of VSS consists of ten training modules containing about 200 lessons in three levels. Participants of all levels of training who complete the course successfully will receive a certificate. The cost of VSS training with adequate number of participants may be less than 10% of the cost of traditional training.

The knowledge and experience gained by developing countries can shared and learned among themselves. Such cooperation, the co-called south-south cooperation, also applies to countries that have undergone transformation from a centrally managed economy to a market economy. Different models of cooperation in the field of statistics among developing countries have been presented by Pieter Everaers from Eurostat. He stressed that presently functioning models are based on cooperation in different areas (management, training, IT, statistics, sets of indicators) and different forms of cooperation (support / partnership / full cooperation).

Supporting the effective implementation of national strategies for statistical development in developing countries is not possible without the financial support of international and national organizations. The problem of efficiency and sustainability of such funding was raised by Frances Harper from the UK's Department for International Development. She stressed that in order for financial assistance to be effective, a close cooperation between all involved partners (funding organizations, governments, users of the data) is necessary. Pieter Dorst from the Netherlands also emphasized the role of effective cooperation, pointing to the special role and needs of the countries concerned, but also to the need to increase the responsibility of the benefiting countries. Financing the development of statistics in developing countries is necessary to achieve Millennium Development Goals, and the starting point for programming and financing development of statistics at the national level should be the formulation of national development plans.

In the session on cooperation of public statistics between countries that have undergone or are in transition(chaired by Marek Cierpiał-Wolan) representatives from Poland (Józef Oleński), Russia (Peter Dolgopolov) and Ukraine (Vadym Pischeiko) delivered their speeches. Józef Oleński elaborated on the synergistic effect of official statistics cooperation between developed countries, after the transformation, or those who are in transition. He stressed that cooperation based on mutual learning between partners gives the best results in building a developed statistical systems, and listed, as examples of good practice using the mutual experience of countries, the Polish cooperation in the field of cross-border statistics on the external border of the EU, social statistics and the use of information technology in employing administrative records.

Peter Dolgopolov spoke about the development of Russian official statistics and the role of international cooperation in the period of transition. Programs funded by the World Bank and the European Union allowed the Russian statistics to take advantage of international experience, conducting intensive training and projects, including regional ones, among the countries of the CIS. Through close cooperation with international organizations and foreign statistical offices the Russian statistics now meets international standards and begins to play a more active role in sharing and transferring knowledge and experience with other nations.

Vadym Pisheyko presented the actions taken by the Ukrainian statistics after gaining independence, which are related to its adaptation to international standards. Most of the works was funded by the World Bank under the project "Development of the state statistics for monitoring socio-economic transformation ", but the Ukrainian statistics also cooperates with other international organizations. Apart from collecting, processing and sharing of data, as well as strengthening the statistical infrastructure and dividing responsibilities between central, regional and local offices, these actions have resulted in developed principles of statistics, technical modernization, dissemination of European standards in regional statistics.

The last session of the conference concerned the problems of development of agricultural statistics in relation to the growing needs for high-quality data in this field. Bjorn Wold of the Statistical Office of Norway raised the issue of new challenges and opportunities facing the agricultural statistics in Africa. He pointed out that in the face of the observed rapid development of African agriculture statistical data from previous surveys are no longer sufficient. He stressed that the statistics are not kept pace with current information needs from a variety of reasons, including lack of marketing activities on the role of statistics, research being carried out not by the official statistics but by ministries not interested in incorporating their findings into the statistical system, and new technologies not being used to conduct research. Possible solutions to these problems can be found in close collaboration between the statisticians and the institutions conducting agricultural research, establishing research priorities, strengthening the leading role of official statistics, implementing in statistics the generally recommended secoral approach, providing statistics that allow for linking the results of research in agriculture with the economic conditions of the country (data from the national

accounts, consumption prices, etc.). It is necessary, of course, to implement new research techniques also to reduce research costs.

The FAO representative Naman Keita described the development of Agricultural Statistics in the context of the Global Strategy to Improve Agricultural and Rural Statistics. The Global Strategy to Improve Agricultural and Rural Statistics has been developed in response to the downward trend in the quantity and quality of data, as well as the lack of adequate statistics on agriculture in many countries. These data are necessary both at national and international level to take action on food security, sustainable agriculture development and climate change.

The key reasons for this were not enough human resources and the lack of technical and organizational capabilities for the collection, compilation, analysis and dissemination of agricultural statistics. The aim is to provide general principles and methodologies that will lead to a significant improvement of national and international statistics on food and agriculture needed to conduct analysis and decision making in this area. Implementation of the strategy will enable the provision of a minimum set of basic data using a methodology that will ensure the integration of agricultural statistics with the national statistical system. A key condition for effective implementation of global strategy is to specify in each country specific actions at national, regional and international levels to identify priority areas and ensure the information needs at different levels. Therefore, global and regional strategy implementation plans containing components for the methodological aspects of training and technical support are prepared simultaneously.

The conference ended with a panel discussion on future activities towards the development of statistics. Participants in the discussion emphasized the effects of implementing the Marrakesh Action Plan for Statistics. Most developing countries have already developed and implemented to varying degrees the National Statistics Development Strategies. Financing of development of statistics has increased substantially, national statistics to monitor Millennium Development Goals are more widely available and the development of statistics plays an increasingly important role in national development plans. However, further action within the next proposed action plan for statistics – Busan Action Plan – is necessary. What was stressed, with reference to the priorities of this plan, was the need for further funding to implement effective national statistics development strategies and to increase the role of statistics in decision making through promotion, dissemination, as well as exchange of knowledge and experience between countries.

This conference in one of the series of conferences on statistical capacity development organized in many parts of the World. The participants not only had a chance to exchange views and benefit from broad experience of the audience but also had an opportunity to meet with Polish statisticians many of whom participated in the conference. The participants, thanks to the organizers had an opportunity to tour Krakow and the famous salt mines in Wieliczka.

7th Conference Survey Sampling in Economic and Social Research in 60th anniversary of foundation of the Department of Statistics at the University of Economics in Katowice

The conference took place 18-20 September 2011 in Katowice and was organized by the Department of Statistics of the University of Economics in Katowice with the cooperation of the Department of Statistical Methods of the University of Łódź and Polish Statistical Association.

Forty-two persons from 9 countries, representing universities, statistical agencies and other institutions participated in the conference. Traditionally, due to the large number of foreign participants, the conference language was English. This year almost a half of the talks was presented by participants from Czech Republic, France, Great Britain, Greece, Lithuania, Netherland, Spain and USA. The invited papers were presented by:

Prof. Malay Ghosh (University of Florida, USA) Estimation of median incomes for small areas : A bayesian semiparametric approach,

Prof. Jean-Claude Deville (Ecole Nationale de la Statitique et de l'Analyse de l'Information, Laboratoire de Statistique d'Enquete, France) Uses of calibration and balanced sampling for the correction of non-response in surveys,

Prof. Parthasarathi Lahiri (University of Maryland, USA) Robust small area estimation,

Prof. Nicholas T. Longford (Universitat Pompeu Fabra, Spain) Policy-related small-area estimation.

Titles of contributed talks are presented in the appendix. It should be noticed that majority of Polish specialists in survey sampling participated in the conference including Professor Czesław Domański – The Chair of Polish Statistical Association and Professor Jan Kordos. The first session was opened by the Dean of the Faculty of Management Professor Krystyna Jędralska. Professor Andrzej St. Barczak the Chairman of the Presidium of the Committee of Statistics and Econometrics of Polish Academy of Science and chairmen of departments of different quantitative disciplines at the University of Economics in Katowice participated in the first session dedicated to the 60th anniversary of the Department of Statistics

The purpose of this conference was to stimulate research both in theoretical and methodological developments in survey sampling and related fields, and practical applications of the methods, including their potential usage in various research areas. Topics discussed during the conference included: estimation of population parameters based on complex samples, statistical inference based on incomplete data, small area estimation, sample size and cost optimization in survey sampling, sampling designs, statistical inference using auxiliary information, model-based

estimation, longitudinal surveys, practical implementations of sampling methods, sampling in statistical quality control, and sampling in auditing.

Program Committee of the conference consisted of: Professor Andrzej St. Barczak, Professor Czesław Bracha, Professor Czesław Domański, Professor Malay Ghosh, Professor Nicholas T. Longford, Professor Zdzisław Hellwig, Professor Jan Kordos (Chair), Professor Walenty Ostasiewicz, Professor Jan Paradysz, Professor Jan Steczkowski, Professor Jacek Wesołowski, Professor Janusz Wywiał. Members of Local Organizing Committee were: Professor Janusz Wywiał (Chair), Professor Józef Kolonko, Dr. Wojciech Gamrot and Dr. Tomasz Żądło (secretary).

Details including abstracts of all of the talks are available at the conference website *http://web.ue.katowice.pl/metoda*.

The Organizing Committee would like to thank sponsors of the conference: University of Economics in Katowice, Katowice City Hall and SPSS Poland.

Appendix - contributed talks:

1. Y. Berger *Variance estimation of hot-deck imputed estimator of change over time from repeated surveys*
2. Cz. Domański *The first Polish association of statisticians*
3. N. Farmakis *Sampling and Coefficient of Variation: Approximating Exponential Distributions*
4. W. Gamrot *Estimators for the Horvitz-Thompson statistic based on its posterior distribution*
5. A. Imiołek *Practical, statistical and economic aspects of using survey studies for identification of the key plant cultivation technology factors*
6. A. Jędrzejczak, J. Kubacki *The comparison of generalized variance function with other methods of precision estimation for Polish Household Budget Survey*
7. T. Klimanek *Estimation of unemployment in Wielkopolska region via spatial model with information about commutes*
8. R. Konarski *Application of Latent Class Analysis to Estimate Full Population Structure from Incomplete Data*
9. J. Kordos *Review of application of rotation methods in sample surveys in Poland. Theory and practice*
10. B. Kowalczyk *Estimation of net changes in the context of multipurpose rotating surveys*
11. A. Kozłowski *Usefulness of past data in sampling design for exit poll surveys*
12. D. Krapavickaitė *Successive sampling design in practice*
13. S. Krieg *Improvement of estimates for the Dutch Structural Business Survey by small area estimation*

14. Ch. Marc *Additional samples with balancing or overlapping conditions and given inclusion probabilities : theoretical approach and examples in the framework of PISA surveys*
15. O. Vilikus *Optimization of sample size and number of tasks per respondent in conjoint studies using simulated datasets*
16. J. Wesołowski *Rotation schemes and Chebyshev polynomials*
17. J. L. Wywiał *On distribution of Horvitz-Thompson statistic under the rejective sample*
18. T. Żądło *On accuracy of two predictors for spatially correlated longitudinal data*

Prepared by:
Tomasz Żądło
(Department of Statistics, University of Economics in Katowice).