STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 15, Number 1, Winter 2014

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Beata Witek: Secretary
b.witek@stat.gov.pl. Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec: 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, Winter 2014
Vol. 15, No. 1, pp. 1–5

FROM THE EDITOR

At the outset of this issue I would like to announce some novelties in the journal, with respect to its issuing and editing. First, from this issue on, *Statistics in Transition new series (SiTns)* will be printed four times a year, just becoming a regular quarterly journal. This will give us an opportunity to be more responsive to systematically growing number of submitted papers, and to earn an extra point in certain indexation bases. Currently, in addition to several bases monitoring our publications, we are under consideration by SCOPUS. The second announcement is about preparing the journal's special issue, envisaged as a thematic collection of articles devoted to the subjective well-being, with preference given to papers presenting current research on this topic in public (government) statistics. In view of the fact that arranging for such a thematic collection of articles is a challenging task we have invited Prof. Graham Kalton (who also inspired us with this idea) to act as a *Guest (co)Editor* of such a topical collection of papers. The formal call for papers will be published in the next issue (scheduled for July).

ACKNOWLEDGEMENTS AND NOMINATIONS

With this issue some important personal changes take place. First of all, on behalf of the Editorial Office and also on behalf of the Co-Chairmen of the Editorial Board - **Prof. Janusz Witkowski**, the President of the Central Statistical Office and of **Prof. Czesław Domański**, the President of the Polish Statistical Association - I would like to express our gratefulness to the long-serving (since Autumn 2007) members of the journal's Editorial Board: Prof. Prof. **Walenty Ostasiewicz, Tomasz Panek, Jan Paradysz, Mirosław Szreder,** and Mr. **Wiesław Łagodziński.** Our excellent collaboration with each of these prominent researchers and scholars to a great extent contributed to the constantly rising recognition and overall quality of the journal. I deeply appreciate having chance to collaborate with so prominent experts, and hope to have the privilege to continue such a fruitful collaboration in the future.

At the same time, I am extremely pleased to announce nomination of the five world renowned experts - previous members of our Associate Editors panel - to resume the membership of the Editorial Board: **Sir Anthony B. Atkinson,** University of Oxford, UK, **Prof. Malay Ghosh**, University of Florida, USA, **Prof. Graham Kalton**, Westat, and the University of Maryland, USA,

Prof. Mirosław Krzyśko, Adam Mickiewicz University in Poznań, Poland, and **Prof. Janusz L. Wywiał**, University of Economics in Katowice, Poland.

We welcome new members of the Editorial Board and look forward to so exceptional opportunity to collaborate with frontiers in their fields.

AN OVERVIEW OF THE CONTENTS

As regards the contents of this issue, it starts with five papers devoted to different issues in *Sampling Methods and Estimation.* **Tomasz Bąk's** paper, ***Triangular Method of Spatial Sampling,*** presents a new adaptive method of spatial sampling, starting with developing a theory of this method, followed by discussion of the benefits of decreased size of a sample due to the employment of this method in sampling of natural area units. Initial sampling of the first three elements is described and density of sampling at the initial stage is obtained by the Monte Carlo method. The density is defined on the basis of the logarithm of inverse square of the Euclidean distance function and a simulation of the triangular method of spatial sampling is finally conducted. An example is given for sampling forest areas in research on approximating the ability of trees to absorb carbon dioxide. The triangular method of spatial sampling is used at the strata sampling stage, and density of sampling in the simulated forest is obtained using Monte Carlo method.

Ashok V. Dorugade introduces a new estimator - ***A Modified Two-Parameter Estimator in Linear Regression*** - envisaged as an alternative to the OLS estimator for the vector of parameters in the linear regression model in the case when multicollinearity is present in the data. The properties of the proposed estimator are discussed along with its performance in terms of the matrix mean square error criterion. A new two-parameter estimator (NTP), an almost unbiased two-parameter estimator (AUTP), and other well-known estimators are being discussed. A numerical example and simulation study are conducted to illustrate the superiority of the proposed estimator.

Arkadiusz Kozłowski discusses possibilities for improving ***The Use of Non-Sample Information in Exit Poll Surveys in Poland*** - the quality and overall precision of the survey - through using the non-sample information more efficiently. Statistical methods aiming at incorporating the information about the relevant variables to the survey, both at the stage of selecting the sample of precincts and at the stage of forecasting election results are proposed. The presented approach is tested by simulation on the parliamentary election 2011 data. The results confirm the possibility of a significant increase in the effectiveness of estimates by choosing a more representative sample and by applying complex estimation of parameters.

Another paper aiming at using 'external' information in a more efficient way, ***An Improved Estimator for Population Mean Using Auxiliary Information in Stratified Random Sampling*** by **Malik S., Singh V. K.,** and **Singh R.** concentrates on the development of a new estimator for population mean $\bar{Y}$ of the study variable y, in the case of stratified random sampling. Using the information based on auxiliary variable x, a formula for the mean squared error (MSE) of the proposed estimator is derived up to the first order of approximation. An empirical study (a numerical example) demonstrates the efficiency of the suggested estimator over sample mean estimator, usual separate ratio, separate product estimator and other proposed estimators.

The next paper, ***A Modified Mixed Randomized Response Model*** by **Housila P. Singh** and **Tanveer A. Tarray** is devoted to the problem arising in survey research from the fact that people wish to hide some information from others, especially on the so-called sensitive issues. These include savings, the extent of their accumulated wealth, their history of intentional tax evasion and other illegal or unethical practices leading to earnings from clandestine sources, crimes, trade in contraband goods, susceptibility to intoxication, expenditures on addictions of various forms, homosexuality, and similar issues which are customarily disapproved of by society. Authors start with briefing on some methods for dealing with this kind of problem - Warner's (1965) survey technique that is known as randomized response (RR) technique and its revised version by Greenberg et al. (1971) for qualitative variables; various further modifications given by several researchers (Chaudhuri 2011, Kim and Warde, 2005), and by Nazuk and Shabbir, 2010) who presented mixed randomized response models using simple random sampling with replacement sampling scheme improving the privacy of respondents. Authors propose a modified mixed randomized response model to estimate the proportion of a qualitative sensitive variable, along with recommendations. It has been shown that the suggested randomized response model is always better than Kim and Warde's model while it is better than Nazuk and Shabbir's model under some realistic conditions. Supporting material for these results is also given in the paper.

The 'research article' section consists of **Jacek Białek's** paper on ***Application of the Original Price Index Formula to Measuring the CPI's Commodity Substitution Bias.*** It examines the possibility of applying the original price index formula to measuring the commodity substitution bias associated with the Consumer Price Index (CPI). The CPI bias values - calculated by using the original price index formula - is compared through simulation study with those calculated on the basis of some known, superlative price indices.

In the last, methodologically oriented, 'other articles' section some papers presented at the XXXII International Conference on Multivariate Statistical Analysis 2013 in Łódź are included.

Grażyna Dehnel's paper on ***Winsorization Methods in Polish Business Survey*** is devoted to one of the major problems involved in estimating information about economic activity across small domains due to excessively small sample size and incompleteness of data sources. In view of the fact that often it is not obvious whether the implementation of traditional estimation methods meets the desired requirements (assumptions about being free from bias or about variance), and given the pressure to produce accurate estimates at a low level of aggregation, or to substantially reduce sample size, the need to develop a more sophisticated approach to estimation seems to be inescapable. The aim of this study was to test the usefulness of *winsorization* methods in such a problem context in order to estimate economic statistics from the DG1 survey in a more efficient way. One of the conclusions states that the use of the winsorized estimation reduces estimator variance and the effect of outliers. Also, the winsorized estimator nearly always outperforms the expansion estimator in terms of MSE.

Daniel Kosiorowski, Dominik Mielczarek, Jerzy Rydlewski, Małgorzata Snarska discuss the problem of ***Sparse Methods for Analysis of Sparse Multivariate Data From Big Economic Databases.*** Authors present a new approach to *sparse* high-dimensional data sets meant as data which contain many zeros among coordinates of observations. Taking jointly the selected *sparse methods* recently proposed in multivariate statistics and kernel density framework for discrete data, they outline a general perspective for bringing out useful information from big economic databases. As a framework for considerations they use the so-called functional data analysis, which originates from Ramsay and Silverman works, and particularly, the functional principal components analysis within 2D density estimation procedure proposed by Simonoff.

Dorota Pekasiewicz's paper on ***Application of Quantile Methods to Estimation of Cauchy Distribution Parameters*** focuses on using quantile methods to estimate population parameters when other methods such as the maximum likelihood method and the method of moments cannot be applied. The percentile method, the quantile least squares method and its two modifications are used for this purpose. The proposed methods allow estimators to be obtained with smaller bias and smaller mean squared error than estimators of the quantile least squares method. The proposed approach can be applied to estimation of the Cauchy distribution parameters. The theoretical considerations on the properties of the estimator are supported by results of the simulation analysis.

Margus Pihlak's presents an approach to ***Modelling of Skewness Measure Distribution***. After showing some results of matrix algebra useful in multivariate

statistical analyses, the central limit theorem on modelling of skewness measure distribution is presented. The paper concludes discussion of the idea of finding the confidence intervals of statistical model residuals' asymmetry measure. For example, by means of skewness confidence intervals it is possible to estimate the influence of outliers (which are typically present in forestry study).

Justyna Wilk and **Michał Bernard Pietrzak** discuss the issues involved in ***An Analysis of the Population Aging Phenomena in Poland from a Spatial Perspective***. The objective of this empirically-oriented demographic study is to characterize the degree of differentiation of the Polish population across subregions (66 in total) in terms of the proportion of senior citizens and its growth rate, and also determinants exerting impact on the demographic aging processes. Demographically the youngest and slowest aging population lives in south-eastern and central Poland. The most intensive population aging processes are seen in the selected subregions of south-western Poland. The latter also is characterized by extremely low fertility, old working-age population, and significant migration outflow of younger people.

The volume is concluded by **Jan Kordos'** remarks on the recently published textbook *Correlation and regression of economic qualitative features,* by J. W. Wiśniewski, which are presented in the 'book review' section.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 7

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://pts.stat.gov.pl/czasopisma/statistics-in-transition/

STATISTICS IN TRANSITION new series, Winter 2014
Vol.15, No. 1, pp. 9-22

TRIANGULAR METHOD OF SPATIAL SAMPLING

Tomasz Bąk[1]

ABSTRACT

In this paper a new adaptive method of spatial sampling - a triangular method of spatial sampling is presented. The theory of this method is developed. Benefits of decreased size of a sample, when this method is used, are discussed. Initial sampling of the first three elements is described and density of sampling at the initial stage is obtained by Monte Carlo method. The density is defined on the basis of the logarithm of inverse square of the Euclidean distance function. Simulation of the triangular method of spatial sampling is conducted. An example is research on a forest. The aim of this research is to approximate the ability of trees to absorb carbon dioxide. In this example the triangular method of spatial sampling is used at the strata sampling stage. Density of sampling in the simulated forest is obtained using Monte Carlo method.

Key words: adaptive sampling, spatial sampling, stratified sampling, research on a forest, minimizing costs of sampling.

1. Sampling based on triangles

In adaptive sampling probabilities of inclusion depend on values of the variable of interest. Adaptive sampling is primarily used in the studies of rare phenomena and agglomerations of the characteristics under study on a relatively large area. This method allows, in such a case, to increase the probability of inclusion of elements which are near to the element with given characteristics. This feature gives the researcher some control over sample composition.

Several authors have dealt with the problem of selection of the elements with specified characteristics. For instance, Thompson and Seber [1996] describe searching for the people infected by a rare disease. If the infected person is found using simple random sampling, then automatically all people

[1]University of Economics in Katowice. E-mail: tbak88@wp.pl.

close to the infected person are also selected to the sample. Similar methods are applied in various fields of science. In this paper a new method of spatial sampling is constructed. Theoretical construction of the sampling method based on triangles is put in the context of the simulation of the ability of forest to absorb the carbon dioxide.

2. The method of sampling the first triangle

At the first stage three elements are selected. Let us denote $X1$ and $X2$ by some characteristics of population under study. Let us consider the realization of spatial sampling from space $X1 \times X2$. Three circular areas with a fixed radius are used as a base of the estimation. However, the selection will be conducted by sampling points which are the centres of circles. The first element is taken uniformly from the space $X1 \times X2$. In this way, none of the fragments of the forest is preferred. Let us denote the first sampled element by $(x1_1, x2_1)$. This element becomes a central point for sampling the next two elements. In order to draw next two elements, let us define distance function from point $(x1_1, x2_1)$ on the space $X1 \times X2$:

$$g(x1, x2) = dist((x1, x2), (x1_1, x2_1)), \tag{1}$$

where $dist$ is a function which satisfies the definition of distance function (satisfies the conditions of non-negativity, identity of indiscernibles, symmetry and subadditivity) and is integrable. When the $dist$ is selected, the specificity of the space $X1 \times X2$ should be taken into account. The g function is used to define density function on the space $X1 \times X2$, which is as follows:

$$f(x1, x2) = \begin{cases} (\alpha g(x1, x2))^{-1}, \text{ when } (x1, x2) \neq (x1_1, x2_1), \\ 0, \text{ when } (x1, x2) = (x1_1, x2_1), \end{cases} \tag{2}$$

where $\alpha = \iint_{X1 \times X2} g(x1, x2)^{-1} dx1 dx2$. Sampling of X_1-coordinate is independent from sampling X_2-coordinate, thus the g function can be defined separately for each subspace:

$$\begin{aligned} g_1(x1) &= \text{dist}(x1, x1_1), \\ g_2(x2) &= \text{dist}(x2, x2_1). \end{aligned} \tag{3}$$

The density function can be also defined separately for each subspace:

$$f_1(x1) = \begin{cases} (\alpha_1 g_1(x1))^{-1}, \text{ when } x1 \neq x1_1, \\ 0, \text{ when } x1 = x1_1, \end{cases}$$
$$f_2(x2) = \begin{cases} (\alpha_2 g_2(x2))^{-1}, \text{ when } x2 \neq x2_1, \\ 0, \text{ when } x2 = x2_1, \end{cases} \tag{4}$$

where $\alpha_1 = \int_{X1} g_1(x1)^{-1} dx1$ and $\alpha_2 = \int_{X2} g_2(x2)^{-1} dx2$.

Finally, we define (using independence of sampling of each coordinate) the density function $f(x1, x2)$ on space $X1 \times X2$ as a product $f_1(x1) \cdot f_2(x2)$.

Next, two elements are sampled with probabilities defined by density function $f(x1, x2)$. The sampling plan of sample

$$s = \{K(x1_1, x2_1), K(x1_2, x2_2), K(x1_3, x2_3)\}, \tag{5}$$

where $K(x1_i, x2_i)$ denotes a circle centered at the point $(x1_i, x2_i)$, is as follows:

$$\begin{aligned} P(s) = c \Bigg[& \iint_{K(x1_2, x2_2)} f_1(x1, x2) dx1 dx2 \iint_{K(x1_3, x2_3)} f_1(x1, x2) dx1 dx2 \\ & + \iint_{K(x1_1, x2_1)} f_2(x1, x2) dx1 dx2 \iint_{K(x1_3, x2_3)} f_2(x1, x2) dx1 dx2 \\ & + \iint_{K(x1_1, x2_1)} f_3(x1, x2) dx1 dx2 \iint_{K(x1_2, x2_2)} f_3(x1, x2) dx1 dx2 \Bigg], \end{aligned} \tag{6}$$

where c is a ratio of the area of the circle to the area of entire space $X1 \times X2$, and f_i, $i = 1, 2$, denotes density functions defined by the equations (4) in the case when the first sampled element is a circle centred at the point $(x1_i, x2_i)$, $i = 1, 2, 3$.

Let us consider, for example, sampling of three elements in the way described above. Let $X1 \times X2$ be a plane $[0, 1] \times [0, 1]$. After determining on the plane $X1 \times X2$ the uniform distribution, the element with coordinates $(x1_1, x2_1)$ is sampled. The g function is defined as a product of g_1 and g_2 functions. These functions are as follows:

$$g_1(x1) = \begin{cases} \log((x1 - x1_1)^{-2}), \text{ when } x1 \neq x1_1, \\ 0, \text{ when } x1 = x1_1, \end{cases}$$
$$g_2(x2) = \begin{cases} \log((x2 - x2_1)^{-2}), \text{ when } x2 \neq x2_1, \\ 0, \text{ when } x2 = x2_1. \end{cases} \tag{7}$$

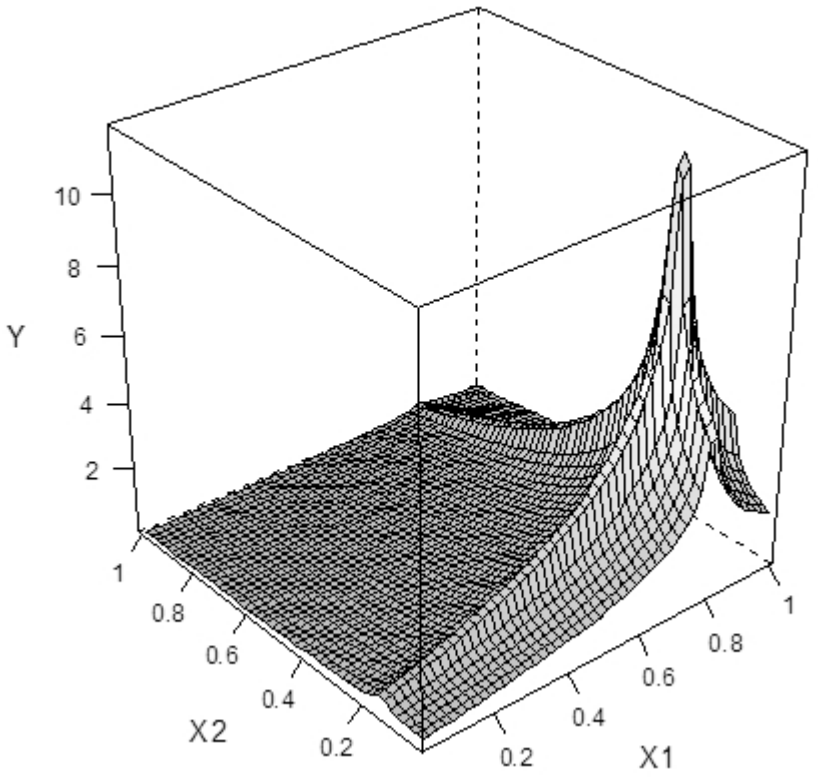


Figure 1. Density function defined on the basis of the logarithm of inverse square of the distance from the point (0.8026372,0.1422940).

Then, the probability of selecting the next element is given by the density function on $X_1 \times X_2$, which is a product of densities on X_1 and X_2. Density functions on spaces $X1$ and $X2$ are as follows:

$$\begin{aligned} f_1(x1) &= \begin{cases} (\alpha_1 \log((x1 - x1_1)^{-2}))^{-1}, \text{ when } x1 \neq x1_1, \\ 0, \text{ when } x1 = x1_1, \end{cases} \\ f_2(x2) &= \begin{cases} (\alpha_2 \log((x2 - x2_1)^{-2}))^{-1}, \text{ when } x2 \neq x2_1, \\ 0, \text{ when } x2 = x2_1, \end{cases} \end{aligned} \tag{8}$$

where
$\alpha_1 = \int_{X1} \log((x1 - x1_1)^{-2})^{-1} dx1$ and $\alpha_2 = \int_{X2} \log((x2 - x2_1)^{-2})^{-1} dx2$

Figure 1 presents a density function in a situation when the first sampled element has coordinates $(x1_1, x2_1) = (0.8026372, 0.1422940)$.

A more fundamental problem than determing the probability of sampling the second and the third element, when the first is already sampled, is to determine the probability of sampling without any assumption about the elements already sampled. Such analysis was performed for the example presented above. Let us denote density function which determines this probability by $f_0(x1, x2)$. In order to approximate this function, sampling of the first element, defining the density function $f(x1, x2)$ and sampling of the second and the third elements were repeated 10 000 times. The density function $f_0(x1, x2)$ was defined as:

$$f_0(x1, x2) = \frac{1}{10000} \sum_{i=1}^{10000} f_i(x1, x2), \tag{9}$$

where functions $f_i(x1, x2)$, $i = 1, \ldots, 10000$ are sequentially created densities defined as a product of the function (8). Approximation of density function $f_0(x1, x2)$ was obtained therefore by Monte Carlo method. The density function $f_0(x1, x2)$ could be used to determine the inclusion probabilities for initial sampling (sampling of the first triangle). The inclusion probabilities for initial sampling are

$$\pi_{K(x1_0,x2_0)} = \iint_{K(x1_0,x2_0)} f_0(x1, x2)dx1dx2,\ K(x1_0, x2_0) \in X_1 \times X_2. \tag{10}$$

Empirical inclusion probabilities defined in such way may be adopted in the Horvitz-Thompson statistics as well as in the variance estimators instead of their true counterparts [4]. Function $f_0(x1, x2)$ is presented in figure 2.

Figure 2 confirms that at the initial sampling stage (sampling of vertices of the first triangle) it is more likely to sample elements which are located more centrally. Of course, the shape of the density function $f_0(x1, x2)$ can be changed by another choice of distance function.

3. The method of sampling the subsequent triangles

Let us denote the characteristic under study by Z and the characteristic strongly correlated with the characteristic under study by Y. Let $\bar{y}$ be an average value of Y characteristic. The sampling method described below allows (of course with a certain lack of precision) the values of the elements which are included in the sample in next steps to be controlled. This sampling method is a kind of adaptive sampling. As in other adaptive sampling procedures, initial elements need to be chosen. The procedure of the triangular sampling method requires three initial elements. They form the vertices

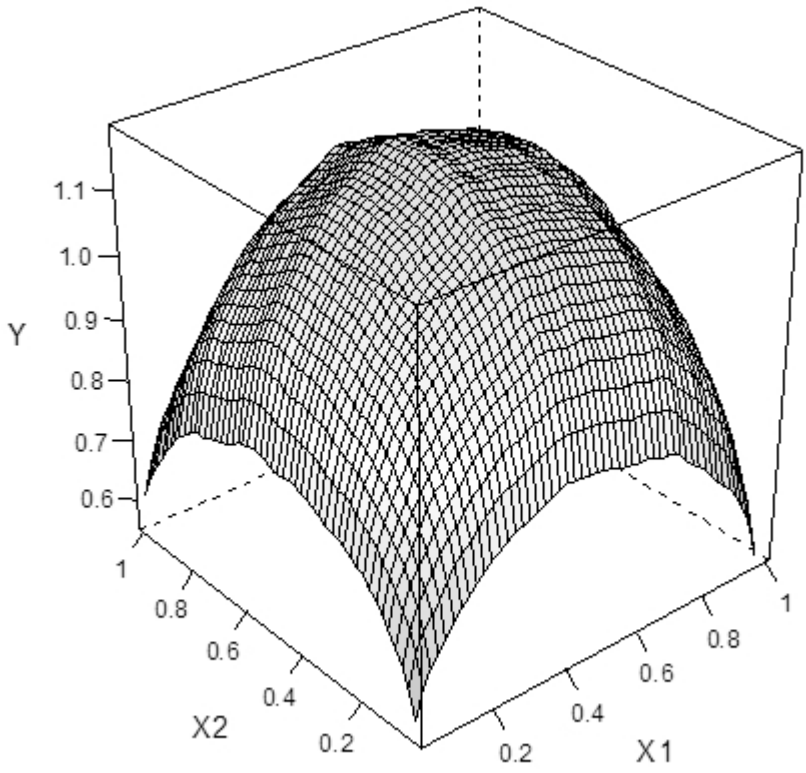


Figure 2. Density function $f_0(x1, x2)$.

of the first triangle. For the reasons of cost and ease of the implementation of the research, it is important to choose these 3 elements relatively close to each other.

Let us assume that one has information which suggests that values of Y change in the majority of the population monotonically (precisely, weakly monotonically, which means some that of the values of Y can be party constant). Let us denote Y-values of the first 3 sampled elements by $y_{1,1},\ y_{1,2},\ y_{1,3}$.

Consider 3 points $(x1_{1,1}, x2_{1,1}, y_{1,1})$, $(x1_{1,2}, x2_{1,2}, y_{1,2})$, $(x1_{1,3}, x2_{1,3}, y_{1,3})$, embedded in three-dimensional space $X1 \times X2 \times Y$. Selection of the next, fourth element is dependent on the values $y_{1,1}, y_{1,2}$ and $y_{1,3}$. The condition which defines sampling of the fourth element is as follows:

$$\exists_{i,j\in\{1,2,3\}, i\neq j} \quad y_{1,i} \leq \bar{y} \wedge y_{1,j} \geq \bar{y}. \tag{11}$$

There are, of course, two possible situations (compliance and non-compliance with condition (11)):

- compliance with condition (11). At least one of the elements $y_{1,1}$, $y_{1,2}$, $y_{1,3}$ has the value equal or lower than $\bar{y}$ and at least one of the elements $y_{1,1}$, $y_{1,2}$, $y_{1,3}$ has the value equal or higher than $\bar{y}$. Then, inside the triangle with vertices $(x1_{1,1}, x2_{1,1}, y_{1,1})$, $(x1_{1,2}, x2_{1,2}, y_{1,2})$, $(x1_{1,3}, x2_{1,3}, y_{1,3})$ we can find a point $(x1_{1,4}, x2_{1,4}, y_{1,4})$ such that $y_{1,4} = \bar{y}$. More precisely, it is a plane (segment should be treated as a degenerated plane) composed with points with given characteristics. Let us introduce the coefficient $d \in (0, 1)$. This coefficient defines the probability of sampling of the fourth element (point $(x1_{1,4}, x2_{1,4}, y_{1,4})$) from the plane inside a triangle $(x1_{1,1}, x2_{1,1}, y_{1,1})$, $(x1_{1,2}, x2_{1,2}, y_{1,2})$, $(x1_{1,3}, x2_{1,3}, y_{1,3})$. Therefore, the value of d coefficient should be set after the first triangle is built. When the value of d coefficient is set, the area (relative to population) of the first triangle should be taken into account. The smaller area of the first triangle is, the lower the value of d coefficient should be. However, other factors should also be considered, such as the diversity of population. The main feature of d coefficient is that the higher the value of d is, the greater 'adaptability' (ability to learn on already sampled elements) the triangular method of spatial sampling has.

  Naturally, coefficient $1 - d$ defines the probalitity of sampling of an element outside the triangle, but located relatively "close" to the triangle. In the second case, to sample the point it is suggested to use the sampling scheme which prefers elements situated in the neighborhood of the first selected element, such as sampling method presented in Chapter . In other words, the fourth element could be sampled in the same way in which the third and the second element were sampled.
- non-compliance with condition (11). With probability 1 we select an element located relatively "close" to the triangle $(x1_{1,1}, x2_{1,1}, y_{1,1})$, $(x1_{1,2}, x2_{1,2}, y_{1,2})$, $(x1_{1,3}, x2_{1,3}, y_{1,3})$ (cf. the sampling method presented in Chapter)

Then, having selected four points on the plane $X1 \times X2$, we create new triangles by choosing for each of the sampled points two other points which are closest to them (in the sense of Euclidean distance). Triangles with the same vertices are treated as one.

Let us consider the situation when, by using the method described above, we have created k triangles from m sampled elements (points in space). Let us denote these triangles by $\Delta_{1,1}, \Delta_{1,2}, \ldots, \Delta_{1,k}$. For each triangle we verify

the compliance with the following condition:

$$\exists_{i,j\in\{1,2,3\},\, i\neq j} \quad y_{1,l_i} \leq \bar{y} \wedge y_{1,l_j} \geq \bar{y}, \tag{12}$$

where $y_{1,l_1}, y_{1,l_2}, y_{1,l_3}$ are vertices of the triangle $\Delta_{1,l}, l = 1, \ldots, k$.

Let us denote the triangles fulfilling condition (12) by $\Delta_{1,p_1}, \Delta_{1,p_2}, \ldots, \Delta_{1,p_q}$, $q \leq k$. Then, the d coefficient denotes the probability of sampling the next element from interior of one of the triangles $\Delta_{1,p_1}, \Delta_{1,p_2}, \ldots, \Delta_{1,p_q}$. Especially, the probability of sampling the next element from the interior of the triangle $\Delta_{1,i}$ is

$$p_{1,m+1}(i) = d\,\frac{h(\Delta_{1,i})}{\sum_{j=1}^{q} h(\Delta_{1,p_j})}, \quad i \in \{p_1, p_2, \ldots, p_q\}, \tag{13}$$

where $h(\Delta_{1,i})$ is the area of the projection of the triangle $\Delta_{1,i}$ on the plane $X1 \times X2$. We can also simplify this scheme by taking $h(\Delta_{1,i}) = 1$, $i \in \{p_1, p_2, \ldots, p_q\}$.

After sampling the triangle we define uniform distribution on the plane inside the triangle, on which (according to conjecture) the value of the variable under study should be equal to $\bar{y}$. In the last step we sample the point from the plane according to the uniform distribution.

The sampling procedure is carried out until the sample size is equal to n_0. After sampling of n_0 elements from the area covered by the triangles we select a new element to the sample, "distant" from the previously sampled elements. Construction of the density for such sampling could be made by analogy to construction described in . The difference is that in 1 instead of the distance function the inverse of the distance function is used (the values of the function are higher for elements more distant from the hub). On the base of the $n_0 + 1$ element presented sampling scheme is repeated. In other words, we start for the second time the initial sampling of three elements - the vertices of the next first triangle. The main feature of n_0 coefficient is quite similarly to the main feature of d coefficient. However, it could be assumed that $n_0 = n$.

Generally, condition (12) adapted for sampling from k-th area (i.e. after sampling $(k-1)n_0$ elements) is as follows

$$\exists_{i,j\in\{1,2,3\},\, i\neq j} \quad y_{k,l_i} \leq \bar{y} \wedge y_{k,l_j} \geq \bar{y}. \tag{14}$$

Finally, using empirical inclusion of probabilities of Horvitz-Thompson statistic as well as variance estimators could be calculated [4]. Empirical density function and, therefore, this empirical inclusion probabilities for the triangular method of spatial sampling could be obtained using Monte Carlo method. Naturally, the density function is strictly dependent on the values of Y characteristic. In the example presented below such function is defined.

4. Example of the triangular method of spatial sampling

Let us consider a research on the forest which aim is to approximate the ability of the trees to absorb carbon dioxide. Diameter at breast height (DBH) is a standard method of expressing the diameter of a trunk of a standing tree. In continental Europe, Australia, the UK and Canada the diameter is measured at 1.3 metres above ground [2]. This characteristic is strongly correlated with the weight of a tree, thus it could be used to estimate carbon dioxide absorption [3]. In addition, DBH average in strata can be assessed by the average age of trees in strata. Therefore, if there is a possibility to sample only few trees in strata, then it is necessary to construct a sampling scheme in such a way that the trees with DBH values close to the average in strata are choosen. Then, the risk of sampling a tree with DBH value distant from average is reduced and the precision of estimation is improved.

Let us consider the sampling of points in a space (a land on which the forest grows) which are centres of circles with fixed radius, as the method of selecting a sample in this research. A point in a space for which a part of a circle is outside the forest could be sampled. In this case the radius of the circle is increased in such way that the area of the circle inside the forest is equal to the area of a 'normal' circle. Trees inside the circle will be later used to estimate the total carbon dioxide absorption. The sample is, therefore, taken from infinite (uncountable) population, but the estimation is made from finite population (trees). This method is a common approach in studying forests (cf. Fattorini et al. 2006). Because of differences between sampling elements and elements which are used in estimation, the assumption about monotonic changes of Y in the majority of the population refers to an average DBH value inside a circle. An average value of DBH is assigned to the point in $X_1 \times X_2$ space - centres of the circle (in practice, it is assigned to part of $X_1 \times X_2$ space - circles with centres close to each other can have the same contents). Sampling is conducted with replacement. The sampled trees are cut down and weighted in the next phase of the study. If the tree was only partly located in the sampled circle, the weight of the tree is multiplied by the share of the circle in the area of the trunk 1.3 metres above ground. By relying on these measurements, one can assess the amount of carbon dioxide absorbed by the forest.

The forest can be divided into strata, using the economic map of forest area. Those maps provide ready-to-use division of the forest, based on dominant species, its share in total afforestation of the area and the average age of dominant tree species. The sample is taken from infinite (uncountable)

population, thus each strata is a part of two-dimensional space, and consists of infinite (uncountable) elements - points in the space.

Referring to the theory, if the population can be divided into strata, then the division should maximize the differentiation between average values of the variable under study in strata (in other words, the aim is to maximize the intergroup variance). Thus, having strata which strongly differentiate capability to absorb the carbon dioxide (species and age are strongly correlated with this ability), it is important to select a "good" representative of each strata. Since the costs of weighting trees are very large, the perspective of sampling only a few elements in strata is very important.

Measurements of a tree diameter at breast height (DBH) on a simulated forest were used as a testing field of the triangular method of spatial sampling. In order to test this sampling method apprioprate simulations were performed. Also an apprioprate program was written in the R package. Simulations were conducted on the simulated matrix of DBH values. This matrix was equivalent to the real forest, the cells of matrix corresponded to the fragments of the forest and the values in the cells are mean DBH values in certain fragments of the forest. Matrix of DBH values was a square matrix, consisted of 10 000 cells. As a result of the simulation, a density function which determines the probabilities of inclusion of $X1 \times X2$ space fragments was obtained.

The triangular method of spatial sampling is based on selection of points from space - from an uncountable population. Therefore, the matrix of DBH values was equated to a space $[0, 1] \times [0, 1]$. The space $[0, 1] \times [0, 1]$ is a square forest area for which we can set two-dimensional coordinates. Each cell of DBH matrix is equivalent to the fragment of the space $[0, 1] \times [0, 1]$ of the size equal to $[0, 0.01] \times [0, 0.01]$. The value in each cell is the mean DBH value of the trees which grow on the area determined by coordinates of this cell in DBH matrix.

Further, as shown in the previous section, points are sampled from the space $[0, 1] \times [0, 1]$. Rather than choosing to sample circles centered in sampled points, as is described in Chapter , cells were selected from DBH matrix, within which sampled points were located. This way of proceeding was due to restrictions which result from algorithmization of mathematical models.

In the first step, the matrix of simulated average DBH values of the trees in certain forest was constructed. Spatial autoregressive model (SAR) was used in simulation (cf. Anselin [1980]). The form of this model was relatively simple, which facilitates further interpretation of the results of the simulation. The DBH value in the cell was influenced by the values of the adjacent cells which already have set the DBH values. In other words, cells $[i - 1, j]$, $[i, j - 1]$ and $[i - 1, j - 1]$ influenced on cell $[i, j]$, $i \in \{2, \ldots, 100\}$, with

following weights: 0.4, 0.4 and 0.2. For simulation of the DBH value in first the row/column of matrix, DBH value from the previous cell in the same row/column was used. The model was expanded by including a random element, which changed the value obtained from the basic model by 1% at the most. For the starting point (cell[1,1]) DBH value equal to 20cm was set. Formally, this model can be described as follows:

$$[i,j] = \begin{cases} \left(0.4\left([i-1,j]+[i,j-1]\right)+0.2[i-1,j-1]\right)\left(1+0.01\varepsilon_{i,j}\right), \\ \text{when } i,\ j \in \{2,\ldots,100\}, \\ [i-1,j]\left(1+0.01\varepsilon_{i,j}\right), \text{ when } i \in \{2,\ldots,100\},\ j=1, \\ [i,j-1]\left(1+0.01\varepsilon_{i,j}\right), \text{ when } i=1,\ j \in \{2,\ldots,100\}, \\ 20 \text{ cm, when } i=j=1, \end{cases} \tag{15}$$

where $\varepsilon_{i,j}$ $i,\ j \in \{1,\ldots,100\}$, have uniform distribution on segment $[-1,1]$.

In addition, limits on the maximum and minimum value of DBH average were imposed on the model. The average of diameter at breast height was no more than 30 cm and no less than 10 cm. Figure 3 shows the result of simulation of DBH values.

An average of DBH values on the forest created by simulation was 19.835 cm, with a standard devation equal to 0.746 cm.

On such a simulated forest the triangular method of spatial sampling was conducted. The sample consisted of 20 elements. At the first stage (initial sampling) three points from the space $[0,1] \times [0,1]$ were sampled and used to create a triangle. Further, 17 elements were sampled either from interior of one of the triangles, in compliance with condition (14) (with probability $d = 0.9$), or in accordance with density defined after drawing the first element (with probability $1-d = 0.1$). Probabilities of sampling each of the triangles are the same, that is they are not weighted by the areas of projection of triangles on space $X1 \times X2$. As the expected average DBH value (the constant $\bar{y}$) 20 cm was set.

Sampling of 20 elements was repeated 10 000 times. The necessary time to make all simulations was nearly 6 hours. 200 000 observations were obtained this way. Using Monte Carlo method the density function was obtained. This density function determines the probability of sampling of each part of the forest. This density function is, of course, strongly dependent on the average DBH value obtained by simulation. Density function is shown in Figure 4.

The average DBH from 200 000 sampled elements was 19.966 cm, with the standard devation equal to 0.593 cm. As expected, the sampled elements were close to the sought value, which was 20 cm. In addition, standard devation from elements in sample was less than standard deviation in the population.

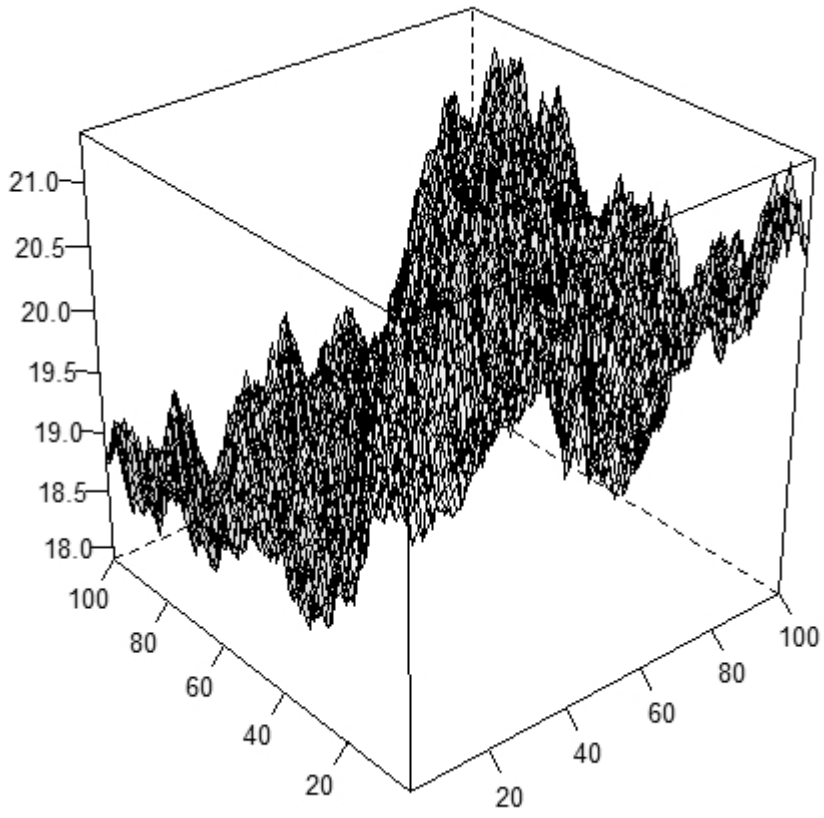


Figure 3. Simulated diameter at breast height.

5. Conclusion

The triangular method of spatial sampling can be an interesting alternative to classical methods of spatial sampling, especially in the case of stratified sampling, where this method allows the selected elements around a predetermined value to e stabilized. As a result, triangular method of spatial sampling increases the chance of achieving a high interstrata variance, which is a desirable feature in proportional stratified sampling.

It should be emphasized that an important advantage of the triangular method of spatial sampling could be the reduction of the cost of research. For instance, if $[0,1] \times [0,1]$ is a space which defines location of the element (for example length and width), then by choosing elements which are close to

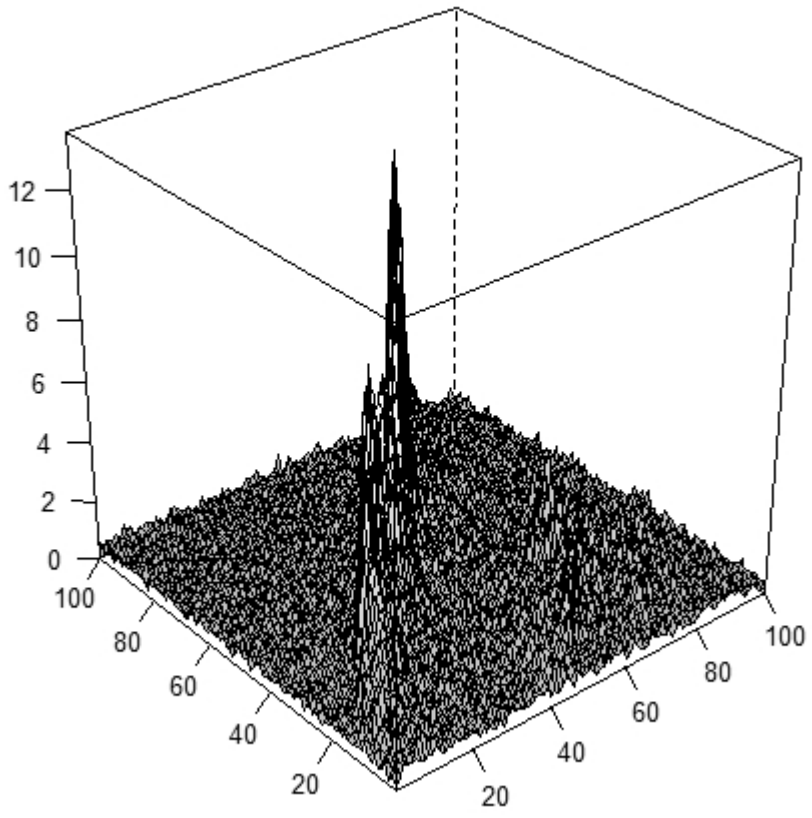


Figure 4. Density function obtained by Monte Carlo method.

each other, time and costs of research could be reduced. This aspect is particularly important in the case of spatial sampling, where the implementation is often a very expensive aspect of survey. It should be noted that for the efficiency of implementation of the triangular method of spatial sampling mobile electronic devices are necessary. One should not, however, consider this as a serious problem. Statistics is a branch of science more and more computerized, so one should try to begin using computers (especially mobile electronic devices) in the process of the research implementation. Then, a response to elements included in the sample could be made already during the implementation of the research. The response to elements included in the sample can be done by changing the probabilities of inclusion or transferring to/focusing on a certain fragment of the population.

In the end it should be emphasized that the presented method can be the subject of modifications, depending on the needs of the researcher.

The author thinks that the change of $\bar{y}$ – constant value into a variable could be an interesting problem, which gives new posibilities to use the triangular method of spatial sampling.

REFERENCES

ANSELIN, L., (1980). Estimation methods for spatial autoregressive structures, Regional Science Dissertation & Monograph Series, Program in Urban and Regional Studies, Cornell University 8, p. 273.

Agriculture, Fisheries and Conservation Department – Conservation Branch: Measurement of Diameter at Breast Height (DBH), Nature Conservation Practice Note No. 02, 2006.

BĄK, T., Udział lasów w procesie redukcji CO2 – aspekty ekonomiczne (Economic aspects of the share of forests in reduction of CO2), Zeszyty naukowe konferencji PITWIN 2/2012, p. 17–21.

FATTORINI, L., (2006). Applying the Horvitz_Thompson criterion in complex designs: a computerintensive perspective for estimating inclusion probabilities, Biometrika 93, p. 269–278.

FATTORINI, L., MARCHESELLI, M., PISANI, C., (2006). A three-phase sampling strategy for large-scale multiresource forest inventories, Journal of Agricultural, Biological, and Environmental Statistics, s. 296–316.

THOMPSON, S., SEBER, G., (1996). Adaptive Sampling, John Wiley & Sons, Inc.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 23–36

A MODIFIED TWO-PARAMETER ESTIMATOR IN LINEAR REGRESSION

Ashok V. Dorugade[1]

ABSTRACT

In this article, a modified two-parameter estimator is introduced for the vector of parameters in the linear regression model when data exists with multicollinearity. The properties of the proposed estimator are discussed and the performance in terms of the matrix mean square error criterion over the ordinary least squares (OLS) estimator, a new two-parameter estimator (NTP), an almost unbiased two-parameter estimator (AUTP) and other well known estimators reviewed in this article is investigated. A numerical example and simulation study are finally conducted to illustrate the superiority of the proposed estimator.

Key words: liu estimator, multicollinearity, two-parameter estimator, mean squared error matrix.

1. Introduction

In practice, there can be strong or near to strong linear relationships among the explanatory variables. In that case the independent assumptions are no longer valid, which causes the problem of multicollinearity. In the presence of multicollinearity, it is impossible to estimate the unique effects of individual variables in the regression equation. Also, the OLS estimator yields regression coefficients whose absolute values are too large and whose signs can actually reverse with negligible changes in the data (see Buonaccorsi, 1996). Therefore, multicollinearity becomes one of the serious problems in the linear regression analysis. The method of ridge regression, proposed by Hoerl and Kennard (1970a) is a popular technique for estimating the regression parameter for the ill-conditioned multiple linear regression models.

Much of the discussion on ridge regression concerns the problem of finding better alternative to the OLS estimator. Some popular numerical techniques to deal with multicollinearity are the ridge regression due to Stein estimator (Stein, 1956), contraction estimator (Mayer and Willke, 1973), modified ridge regression

[1] Y C Mahavidyalaya Halkarni, Tal-Chandgad, Kolhapur, Maharashtra, India - 416552.
E-mail: adorugade@rediffmail.com.

(MRR) estimator (Swindel, 1976), Kadiyala (1984), Ohtani (1986), Singh and Chaubey (1987), Nomura (1988), and Gruber (1998) Sing et al. (1988), Liu (1993), Akdeniz and Kaciranlar (1995), Crouse et al. (1995), Ozkale and Kaciranlar (2007), Batah et al. (2008), Sakallioglu and Kaciranlar (2008), Yang and Chang (2010), Wu and Yang (2011), Dorugade and Kashid (2011) and others.

In this paper we introduce a modified two-parameter estimator for the vector of parameters in the linear regression model when data exists with multicollinearity. The rest of the paper is organized as follows. The model and some well known estimators are reviewed in section 2. The modified two-parameter estimator is introduced in section 3. Performances of the proposed estimator with respect to the scalar MSE criterion are discussed in section 4. In section 5, we give methods to choose the biasing parameters. A simulation study to justify the superiority of the suggested estimator is given in section 6. Some concluding remarks are given in section 7.

2. Model and estimators

Consider a widely used linear regression model

$$Y = X\beta + \varepsilon, \tag{1}$$

where Y is an n×1 vector of observations on a response variable. β is a p×1 vector of unknown regression coefficients, X is a matrix of order (n × p) of observations on 'p' predictor (or regressor) variables and ε is an n × 1 vector of errors with E(ε) = 0 and V(ε) = $\sigma^2 I_n$. For the sake of convenience, we assume that the matrix X and the response variable Y are standardized in such a way that $X'X$ is a non-singular correlation matrix and $X'Y$ is the correlation between X and Y. The paper is concerned with data exhibited with multicollinearity leading to a high MSE for β meaning that $\hat{\beta}$ is an unreliable estimator of β.

Let Λ and T be the matrices of eigenvalues and eigenvectors of $X'X$, respectively, satisfying $T'X'XT = \Lambda$ = diagonal ($\lambda_1, \lambda_2, ..., \lambda_p$), where λ_i being the ith eigenvalue of $X'X$ and $T'T = TT' = I_p$. We obtain the equivalent model

$$Y = Z\alpha + \varepsilon, \tag{2}$$

where Z = XT. It implies that $Z'Z = \Lambda$, and α = $T'\beta$ (see Montgomery et al., 2006).

Then, the OLS estimator of α is given by

$$\hat{\alpha} = (Z'Z)^{-1}Z'Y = \Lambda^{-1}Z'Y \tag{3}$$

Therefore, the OLS estimator of β is given by

$$\hat{\beta} = T\hat{\alpha}$$

2.1. Ordinary ridge estimator (ORR)

A popular estimator for combating multicollinearity is the ridge estimator, originally introduced by Hoerl and Kennard (1970a) as

$$\hat{\beta}_R = \mathrm{T}\,\hat{\alpha}_R = \mathrm{T}\left[I - k\left(\Lambda + kI\right)^{-1}\right]\hat{\alpha} \tag{4}$$

where k is the ridge parameter (or biasing constant), and it normally lies between 0 and 1. $\hat{\alpha}_i$ is the i$^{\text{th}}$ element of $\hat{\alpha}$, $i = 1,2,...,p$ and $\hat{\sigma}^2$ is the OLS estimator of σ^2 i.e. $\hat{\sigma}^2 = (Y'Y - \hat{\alpha}'Z'Y)/(n-p-1)$.

The ridge regression method has been considered by various researchers. The drawback of the ridge regression method is that it is a complicated function of *k*. To overcome this problem Liu (1993) proposed an estimator which combines the benefit of both the estimators given by Hoerl and Kennard (1970a) and Stein (1956), respectively.
It is given as

$$\hat{\alpha}_{Liu} = \left(\Lambda + I\right)^{-1}\left(\Lambda + dI\right)\hat{\alpha} \qquad 0 < d < 1 \tag{5}$$

Liu estimator has been considered by several researchers several times for different perspectives. Following Liu many researchers propose two-parameter ridge estimators. Ozkale and Kaciranlar (2007) obtained the two-parameter (TP) estimator given as

$$\hat{\alpha}_{TP} = \left(\Lambda + kI\right)^{-1}\left(\Lambda + kdI\right)\hat{\alpha} \tag{6}$$

MSE of $\hat{\alpha}_{TP}$ is given as

$$MSE\left(\hat{\alpha}_{TP}\right) = \sigma^2 \sum_{i=1}^{p}\left[\frac{(\lambda_i + kd)^2}{\lambda_i(\lambda_i + k)^2}\right] + \sum_{i=1}^{p}\left[\frac{k^2(1-d)^2}{(\lambda_i + k)^2}\right]\alpha_i^2 \tag{7}$$

Sakallioglu and Kaciranlar (2008) suggested the following two-parameter estimator:

$$\hat{\alpha}_{LTE(3)} = \left(\Lambda + I\right)^{-1}\left(\Lambda + (d+k)I\right)^{-1}\left(\Lambda + kI\right)^{-1} Z'Y\,. \tag{8}$$

MSE of $\hat{\alpha}_{LTE(3)}$ is given as

$$MSE\left(\hat{\alpha}_{LTE(3)}\right) = \sigma^2 \sum_{i=1}^{p}\left[\frac{\lambda_i(\lambda_i + (k+d))^2}{(\lambda_i+1)^2(\lambda_i + k)^2}\right] + \sum_{i=1}^{p}\left[\frac{(\lambda_i(1-d)+k)^2}{(\lambda_i+1)^2(\lambda_i + k)^2}\right]\alpha_i^2 \tag{9}$$

since the ridge parameter $\hat{k} = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p}\hat{\alpha}_i^2$ given by Hoerl et al. (1975) performs fairly well and the well-known estimate of 'd' proposed by Liu (1993) is given as

$$\hat{d} = \sum_{i=1}^{p}(\hat{\alpha}_i^2 - \hat{\sigma}^2)/(\lambda_i+1)^2 \Bigg/ \sum_{i=1}^{p}(\hat{\sigma}^2 + \lambda_i\hat{\alpha}_i^2)/(\lambda_i+1)^2\,\lambda_i$$

The above calculated values of $\hat{k}$ and $\hat{d}$ are used in determination of estimators given in equations (6) and (8).

On the other hand, Yang and Chang (2010) introduce a new two-parameter (NTP) estimator given as

$$\hat{\alpha}_{NTP} = (\Lambda + I)^{-1}(\Lambda + dI)(\Lambda + kI)^{-1} Z'Y, \tag{10}$$

where $\hat{k} = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p} \hat{\alpha}_i^2$

and $\hat{d} = \dfrac{\sum_{i=1}^{p} \left\{ [(k+1)\lambda_i + k]\lambda_i \hat{\alpha}_i^2 - \lambda_i \hat{\sigma}^2 \right\} \Big/ \left[(\lambda_i + 1)^2 (\lambda_i + k)^2 \right]}{\sum_{i=1}^{p} (\hat{\sigma}^2 + \lambda_i \hat{\alpha}_i^2) \Big/ \left[(\lambda_i + 1)^2 (\lambda_i + k)^2 \right]}$.

It includes the OLS, RR, and Liu estimators as special cases and provides an alternative method to overcome multicollinearity in linear regression.

Also, MSE of $\hat{\alpha}_{NTP}$ is given as

$$MSE(\hat{\alpha}_{NTP}) = \sigma^2 \sum_{i=1}^{p} \left[\frac{\lambda_i (\lambda_i + d)^2}{(\lambda_i + 1)^2 (\lambda_i + k)^2} \right] + \sum_{i=1}^{p} \left\{ \frac{[(k + 1 - d)\lambda_i + k]^2}{(\lambda_i + 1)^2 (\lambda_i + k)^2} \right\} \alpha_i^2 \tag{11}$$

Recently Wu and Yang (2011) introduced an almost unbiased two-parameter (AUTP) estimator alternative to the OLS estimator in the presence of multicollinearity. These estimators are given as

$$\hat{\alpha}_{AUTP} = \hat{\alpha}_{TP} + k(1 - d)(\Lambda + kI)^{-1} \hat{\alpha}_{TP}, \tag{12}$$

where $\hat{d} < 1 - \min\left(\hat{\sigma} \Big/ \sqrt{\lambda_i \hat{\alpha}_i^2 + \hat{\sigma}^2} \right)$ and $\hat{k} = \lambda_i \hat{\sigma} \Big/ \left[(1 - d)\sqrt{\lambda_i \hat{\alpha}_i^2 + \hat{\sigma}^2} - \hat{\sigma} \right]$

Also, MSE of $\hat{\alpha}_{AUTP}$ is given as

$$MSE(\hat{\alpha}_{AUTP}) = \sigma^2 \sum_{i=1}^{p} \left\{ \frac{\left[\lambda_i (\lambda_i + 2k) + d\,k^2 (2 - d) \right]^2}{\lambda_i (\lambda_i + k)^4} \right\} + \sum_{i=1}^{p} \left[\frac{k^4 (1 - d)^4}{(\lambda_i + k)^4} \right] \alpha_i^2 \tag{13}$$

Estimators given in equations (6), (8), (10) and (12) used for estimating α are used in section 6.

3. Proposed ridge estimator

In this article we introduce a modified two-parameter estimator and it can be computed in two steps. Initially, following a similar method proposed by Liu (1993), Kaciranlar et al. (1999) and Yang and Chang (2010) we introduce two-parameter estimator as

$$\hat{\alpha}^* = (\Lambda + k\,d\,I)^{-1} Z'Y \tag{14}$$

Then, following Kadiyala (1984), Ohtani (1986) and Wu and Yang (2011) the estimator defined in equation (14) can be rewritten as

$$\hat{\alpha}_{MTP} = \hat{\alpha}^* + k(1-d)\left(\Lambda + kdI\right)^{-1}\hat{\alpha}^* \tag{15}$$

or

$$\hat{\alpha}_{MTP} = \left[I + k(1-d)(\Lambda + kdI)^{-1}\right]\left[I - kd\,(\Lambda + kdI)^{-1}\right]\hat{\alpha}\,.$$

It is termed as a modified two-parameter (MTP) estimator of α.

Thus, the coordinate wise estimators can be written as

$$\hat{\alpha}_{iMTP} = \left[\frac{\lambda_i(\lambda_i + k)}{(\lambda_i + kd)^2}\right]\hat{\alpha}_i \qquad i = 1,2,\ldots, p \tag{16}$$

where $\hat{\alpha}_i$ are the individual components of $\hat{\alpha}$.

We can see that it is a general estimator which includes the OLS and RR estimators as special cases:

at $\left(k = 0 \;\; or \;\; d = 0\right)$ $\hat{\alpha}_{MTP} = \Lambda^{-1} Z'Y$, the OLS estimator

at $d = 1$ $\hat{\alpha}_{MTP} = \left(\Lambda + k\,I\right)^{-1} Z'Y$, the RR estimator

Obviously,

$\hat{\alpha}_{iMTP} = (\hat{\alpha}_i)_{OLS}$ at $\left(k = 0 \;\; or \;\; d = 0\right)$

and $\hat{\alpha}_{iMTP} = (\hat{\alpha}_i)_R$ at $d = 1$

3.1. Bias, variance and MSE of MTP estimator

It is clear that $\hat{\alpha}_{MTP}$ is a biased estimator, with the bias of the MTP estimator is given by:

$$Bias\left(\hat{\alpha}_{MTP}\right) = E[\hat{\alpha}_{MTP}] - \alpha = [k(1-2d)(\Lambda + kdI)^{-1} - k^2 d\,(1-d)(\Lambda + kdI)^{-2}]\alpha$$

$$= \sum_{i=1}^{p}\left\{\frac{k\left[(1-2d)\lambda_i - kd^2\right]}{\left(\lambda_i + kd\right)^2}\right\}\alpha_i \tag{17}$$

$$V\left(\hat{\alpha}_{MTP}\right) = \sigma^2 V\Lambda^{-1}V'$$

where $V = \left[I + k(1-d)(\Lambda + kdI)^{-1}\right]\left[I - kd\,(\Lambda + kdI)^{-1}\right]$

$$= \sigma^2 \sum_{i=1}^{p}\left[\frac{\lambda_i\left(\lambda_i + k\right)^2}{\left(\lambda_i + kd\right)^4}\right] \tag{18}$$

The MSE of MTP estimator is

$$MSE\left(\hat{\alpha}_{MTP}\right) = V\left(\hat{\alpha}_{MTP}\right) + \left[Bias\left(\hat{\alpha}_{MTP}\right)\right]\left[Bias\left(\hat{\alpha}_{MTP}\right)\right]',$$

$$MSE\left(\hat{\alpha}_{MTP}\right) = \sigma^2 \sum_{i=1}^{p}\left[\frac{\lambda_i\left(\lambda_i + k\right)^2}{\left(\lambda_i + kd\right)^4}\right] + \sum_{i=1}^{p}\left\{\frac{k^2\left[(1-2d)\lambda_i - kd^2\right]^2}{\left(\lambda_i + kd\right)^4}\right\}\alpha_i^2 \tag{19}$$

Setting $k = 0$ or $d = 0$ in equation (19), we obtain

$$MSE(\hat{\alpha}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{20}$$

Also, setting $d = 1$ in equation (19), we obtain

$$MSE(\hat{\alpha}_R) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \tag{21}$$

4. Performance of proposed estimator

This section compares the performance of the $\hat{\alpha}_{MTP}$ with the $\hat{\alpha}$, $\hat{\alpha}_{AUTP}$ and $\hat{\alpha}_{NTP}$ using smaller MSE criteria.

4.1. Comparison between $\hat{\alpha}_{MTP}$ and $\hat{\alpha}$

In order to compare $\hat{\alpha}_{MTP}$ with $\hat{\alpha}$ in the MSE sense, using equations (19) and (20) we investigate the following difference:

$$MSE(\hat{\alpha}) - MSE(\hat{\alpha}_{MTP}) = \sigma^2 \sum_{i=1}^{p} \left[\frac{1}{\lambda_i} - \frac{\lambda_i (\lambda_i + k)^2}{(\lambda_i + kd)^4} \right] - k^2 \sum_{i=1}^{p} \left\{ \frac{\left[(1-2d)\lambda_i - k^2 d^2 \right]^2}{(\lambda_i + kd)^4} \right\} \alpha_i^2$$

$$= \sum_{i=1}^{p} \left\{ \frac{\sigma^2 \left[(\lambda_i + kd)^4 - \lambda_i^2 (\lambda_i + k) \right]^2 - \lambda_i \alpha_i^2 k^2 \left[(1-2d)\lambda_i - k^2 d^2 \right]^2}{\lambda_i (\lambda_i + kd)^4} \right\}$$

From above equation it can be seen that $MSE(\hat{\alpha}_{OLS}) \geq MSE(\hat{\alpha}_{MTP})$ if and only if

$$\sigma^2 \left[(\lambda_i + kd)^4 - \lambda_i^2 (\lambda_i + k)^2 \right] \geq \lambda_i \alpha_i^2 k^2 \left[(1-2d)\lambda_i - k^2 d^2 \right]^2$$

4.2. Comparison between $\hat{\alpha}_{MTP}$ and $\hat{\alpha}_{AUTP}$

Wu and Yang (2011) proposes the almost unbiased two-parameter estimator ($\hat{\alpha}_{AUTP}$) given in equation (12). Also, they compare performance of their estimator with the OLS estimator and the two-parameter estimator given in equation (6). To compare $\hat{\alpha}_{MTP}$ with $\hat{\alpha}_{AUTP}$ in the MSE sense, using equations (19) and (13) we investigate the following difference:

$$MSE(\hat{\alpha}_{AUTP}) - MSE(\hat{\alpha}_{MTP})$$

$$= \sigma^2 \sum_{i=1}^{p} \left\{ \frac{\left[\lambda_i (\lambda_i + 2k) + d\,k^2 (2-d) \right]^2}{\lambda_i (\lambda_i + k)^4} - \frac{\lambda_i (\lambda_i + k)^2}{(\lambda_i + kd)^4} \right\}$$

$$+ k^2 \sum_{i=1}^{p} \left\{ \frac{k^2 (1-d)^4}{(\lambda_i + k)^4} - \frac{\left[(1-2d)\lambda_i - kd^2 \right]^2}{(\lambda_i + kd)^4} \right\} \alpha_i^2$$

$$= \sigma^2 \sum_{i=1}^{p} \left\{ \frac{\left[\lambda_i(\lambda_i+2k)+d\,k^2(2-d)\right]^2(\lambda_i+kd)^4 - {\lambda_i}^2(\lambda_i+k)^6}{(\lambda_i+kd)^4\,\lambda_i(\lambda_i+k)^4} \right\}$$

$$+ k^2 \sum_{i=1}^{p} \left\{ \frac{k^2(1-d)^4(\lambda_i+kd)^4 - \left[(1-2d)\lambda_i - kd^2\right]^2(\lambda_i+k)^4}{(\lambda_i+kd)^4(\lambda_i+k)^4} \right\} \alpha_i^2$$

$$= \sum_{i=1}^{p} \left\{ \frac{\left\{\sigma^2\left[\lambda_i(\lambda_i+2k)+d\,k^2(2-d)\right]^2 + \lambda_i k^4(1-d)^4\alpha_i^2\right\}(\lambda_i+kd)^4 - \left\{\sigma^2{\lambda_i}^2(\lambda_i+k)^2 + \lambda_i\alpha_i^2\left[(1-2d)\lambda_i - kd^2\right]^2\right\}(\lambda_i+k)^4}{(\lambda_i+kd)^4\,\lambda_i(\lambda_i+k)^4} \right\}$$

From above equation it can be seen that $MSE(\hat{\alpha}_{AUTP}) \geq MSE(\hat{\alpha}_{MTP})$ if and only if

$$\left\{\sigma^2\left[\lambda_i(\lambda_i+2k)+d\,k^2(2-d)\right]^2 + \lambda_i k^4(1-d)^4\alpha_i^2\right\}(\lambda_i+kd)^4$$
$$\geq \left\{\sigma^2{\lambda_i}^2(\lambda_i+k)^2 + \lambda_i\alpha_i^2\left[(1-2d)\lambda_i - kd^2\right]^2\right\}(\lambda_i+k)$$

4.3. Comparison between $\hat{\alpha}_{MTP}$ and $\hat{\alpha}_{NTP}$

Yang and Chang (2010) introduced a new two-parameter (NTP) estimator and studied superiority of their estimator over the OLS estimator, Liu estimator and the two-parameter estimator. In order to compare $\hat{\alpha}_{MTP}$ with $\hat{\alpha}_{NTP}$ in the MSE sense, using equations (19) and (11) we investigate the following difference:

$$MSE(\hat{\alpha}_{NTP}) - MSE(\hat{\alpha}_{MTP})$$

$$= \sigma^2 \sum_{i=1}^{p} \left\{ \frac{\lambda_i(\lambda_i+d)^2}{(\lambda_i+1)^2(\lambda_i+k)^2} - \frac{\lambda_i(\lambda_i+k)^2}{(\lambda_i+kd)^4} \right\}$$

$$+ \sum_{i=1}^{p} \left\{ \frac{\left[(k+1-d)\lambda_i+k\right]^2}{(\lambda_i+1)^2(\lambda_i+k)^2} - \frac{\left[(1-2d)\lambda_i - kd^2\right]^2}{(\lambda_i+kd)^4} \right\} \alpha_i^2$$

$$= \sigma^2 \sum_{i=1}^{p} \left\{ \frac{\left[\lambda_i(\lambda_i+d)^2(\lambda_i+kd)^4\right] - \lambda_i(\lambda_i+1)^2(\lambda_i+k)^4}{(\lambda_i+kd)^4(\lambda_i+1)^2(\lambda_i+k)^2} \right\}$$

$$+ \sum_{i=1}^{p} \left\{ \frac{((k+1-d)\lambda_i+k)^2(\lambda_i+kd)^4 - k^2\left[(1-2d)\lambda_i - kd^2\right]^2(\lambda_i+k)^2(\lambda_i+1)^2}{(\lambda_i+kd)^4(\lambda_i+k)^2(\lambda_i+1)^2} \right\} \alpha_i^2$$

$$= \sum_{i=1}^{p} \left\{ \frac{\left\{\sigma^2\left[\lambda_i(\lambda_i+d)^2\right] + ((k+1-d)\lambda_i+k)^2\alpha_i^2\right\}(\lambda_i+kd)^4 - \left\{\sigma^2\lambda_i(\lambda_i+k)^2 + \alpha_i^2k^2\left[(1-2d)\lambda_i - kd^2\right]^2\right\}(\lambda_i+k)^2(\lambda_i+1)^2}{(\lambda_i+kd)^4(\lambda_i+k)^2(\lambda_i+1)^2} \right\}$$

From above equation it can be seen that $MSE(\hat{\alpha}_{NTP}) \geq MSE(\hat{\alpha}_{MTP})$ if and only if

$$\left\{\sigma^2\left[\lambda_i(\lambda_i+d)^2\right]+((k+1-d)\lambda_i+k)^2\alpha_i^2\right\}(\lambda_i+kd)^4$$
$$\geq\left\{\sigma^2\lambda_i(\lambda_i+k)^2+\alpha_i^2k^2\left[(1-2d)\lambda_i-kd^2\right]^2\right\}(\lambda_i+k)^2(\lambda_i+1)^2$$

5. Determination of ridge parameters k and d

In ridge regression the additional parameter, the ridge parameter k, plays a vital role to control the bias of the regression towards the mean of the response variable. Although these estimators result in a bias for certain value of k they yield minimum mean squared error (MSE) compared to the OLS estimator (see Hoerl and Kennard, 1970a). Similarly, d is another ridge parameter which serves the same role as k used in determination of two-parameter estimators (see Liu, 1993).

In order to determine and evaluate the performance of our proposed estimator $\hat{\alpha}_{MTP}$ as compared to the OLS estimator and others, we will find the optimal values of k and d. Let k be fixed and determined using one of the available methods for choosing the ridge parameter value. Some of the well known methods are listed below.

$$k_1 = p\hat{\sigma}^2\Big/\sum_{i=1}^{p}\hat{\alpha}_i^2 \qquad \text{(Hoerl et al., (1975))} \qquad (22)$$

$$k_2 = Median\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_i^{2'}}\right) \qquad i=1,2,...,p \qquad \text{(Kibria, (2003))} \qquad (23)$$

$$k_3 = p\hat{\sigma}^2\Big/\sum_{i=1}^{p}\left\{\hat{\alpha}_i^{\,2}\Big/\left[\left[\left(\hat{\alpha}_i^{\,4}\lambda_i^2/4\hat{\sigma}^2\right)+\left(6\hat{\alpha}_i^{\,4}\lambda_i/\hat{\sigma}^2\right)\right]^{1/2}-\left(\hat{\alpha}_i^{\,2}\lambda_i/2\hat{\sigma}^2\right)\right]\right\}$$

(Batah et al., (2008)) (24)

Then, the optimal value of d can be considered to be the d that minimize $MSE(\hat{\alpha}_{MTP})$.

Let $g(k,d) = MSE(\hat{\alpha}_{MTP}) = \sigma^2\sum_{i=1}^{p}\left[\frac{\lambda_i(\lambda_i+k)^2}{(\lambda_i+kd)^4}\right]+\sum_{i=1}^{p}\left\{\frac{k^2\left[(1-2d)\lambda_i-kd^2\right]^2}{(\lambda_i+kd)^4}\right\}\alpha_i^2$

Then, by differentiating $g(k,d)$ w.r.t. d and equating to 0, we have

$$d = \sum_{i=1}^{p}\left[\frac{(\lambda_i+k)\left(\sigma^2+\lambda_i\alpha_i^2\right)-\lambda_i}{k\alpha_i^2}\right] \qquad (25)$$

Unfortunately, d depends on the unknown σ^2 and α_i. For practical purposes we replace them with their unbiased estimator $\hat{\sigma}^2$ and $\hat{\alpha}_i$, and obtain

$$\hat{d} = \sum_{i=1}^{p} \left[\frac{(\lambda_i + k)(\hat{\sigma}^2 + \lambda_i \hat{\alpha}_i^2) - \lambda_i}{k\hat{\alpha}_i^2} \right] \tag{26}$$

6. Comparative study

6.1. Numerical illustration

In this section we demonstrate the performance of the proposed estimator by considering a numerical example; we use Hald cement data (see Montgomery et al., 2006). We use the ridge parameters given in equations (22) to (24) and d given in equation (26) to compute our modified two-parameter (MTP) estimator. Also, $\hat{\alpha}_{OLS}$, $\hat{\alpha}_{TP}$, $\hat{\alpha}_{LTE(3)}$, $\hat{\alpha}_{NTP}$, $\hat{\alpha}_{AUTP}$ estimators are computed and their estimated MSE values are obtained by replacing all unknown model parameters respectively with their OLS estimators in the corresponding expressions, and the values are reported in Table 1.

Table 1. Values of estimates and MSE

Estimator	$\hat{\alpha}_{OLS}$	$\hat{\alpha}_{TP}$	$\hat{\alpha}_{LTE(3)}$	$\hat{\alpha}_{NTP}$	$\hat{\alpha}_{AUTP}$	$\hat{\alpha}_{MTP}$ at k_1	$\hat{\alpha}_{MTP}$ at k_2	$\hat{\alpha}_{MTP}$ at k_3
MSE	1.3709	0.1485	1.2605	0.1484	1.3662	0.1492	0.1487	0.1490

From Table 1 we can see that the estimated MSE value of the modified two-parameter estimator is always smaller than the one of the OLS, AUTP and LTE(3) estimators. However, we also find that the estimated MSE value of the modified two-parameter estimator for each choice of the ridge parameter is approximately equal to those of the TP and NTP estimators. The results agree with our theoretical findings in section 4.

6.2. Simulation study

Here, we examine the performance of the modified two-parameter estimator $(\hat{\alpha}_{MTP})$ over different estimators $\hat{\alpha}_{OLS}$, $\hat{\alpha}_{TP}$, $\hat{\alpha}_{LTE(3)}$, $\hat{\alpha}_{NTP}$, $\hat{\alpha}_{AUTP}$. We examine the average MSE (AMSE) ratio of the $\hat{\alpha}_{MTP}$ and other estimators over the OLS estimator. We will discuss the simulation study that compares the performance of different estimators under several degrees of multicollinearity. We consider the true model as $Y = X\beta + \varepsilon$. Here, ε follows a normal distribution $N(0, \sigma^2 I_n)$ and the explanatory variables are generated (see Batah et al., 2008) from

$$x_{ij} = (1 - \rho^2)^{1/2} u_{ij} + \rho u_{ip}, \quad i = 1,2,...,n \qquad j = 1,2,...,p$$

where u_{ij} are independent standard normal random numbers and ρ^2 is the correlation between x_{ij} and $x_{ij'}$ for j, $j' < p$ and $j \neq j'$. j, $j' = 1, 2, \ldots, p$. When j or $j' = p$, the correlation is be 'ρ'. Here, we consider predictor variables p = 4 and $\rho = 0.9, 0.95$. These variables are standardized such that $X'X$ is in the correlation matrix form and it is used for the generation of Y with $\beta = (2, 1, 4, 5)'$. We have simulated the data with sample sizes n = 20, 50 and 100. The variance of the error terms is taken as $\sigma^2 = 1, 5, 10$ and 25. Estimators $\hat{\alpha}_{OLS}$, $\hat{\alpha}_{TP}$, $\hat{\alpha}_{LTE(3)}$, $\hat{\alpha}_{NTP}$, $\hat{\alpha}_{AUTP}$ are computed. The modified two-parameter estimator ($\hat{\alpha}_{MTP}$) is computed for different choices of ridge parameters given in equations (22) to (24) and d given in equation (26). The experiment is repeated 1500 times and obtains the average MSE (AMSE) of estimators using the following expression:

$$AMSE(\hat{\alpha}) = \frac{1}{1500} \sum_{i=1}^{4} \sum_{j=1}^{1500} (\hat{\alpha}_{ij} - \alpha_i)^2$$

where $\hat{\alpha}_{ij}$ denotes the estimator of the ith parameter in the jth replication and α_i, i=1,2,3, 4 are the true parameter values.

Firstly, we have computed the AMSE ratios ($AMSE(\hat{\alpha}_{OLS})/AMSE(\hat{\alpha})$) of the OLS estimator over different estimators for various values of triplet (ρ, n, σ^2) and reported them in Table 2. We consider the method that leads to the maximum AMSE ratio to be the best from the MSE point of view.

The same procedure for another choice of p = 3 and $\beta = (1, 2, 5)'$ is performed and AMSE ratios are computed and reported in Table 3.

Table 2. Ratio of AMSE of OLS over various two-parameter estimators (p = 4 and $\beta = (2, 1, 4, 5)'$)

ρ = 0.90													
n		20				50				100			
$\hat{\sigma}^2$		1	5	10	25	1	5	10	25	1	5	10	25
$\hat{\alpha}$	$\hat{\alpha}_{TP}$	1.185	1.2908	1.3352	1.472	1.228	1.2809	1.3539	1.478	1.235	1.2563	1.3283	1.484
	$\hat{\alpha}_{LTE(3)}$	1	1.0009	1.0023	1.004	1	1.0003	1.0007	1.001	1	1.0002	1.0008	1.001
	$\hat{\alpha}_{NTP}$	1.359	1.5054	1.5857	1.978	1.473	1.5372	1.6313	2.067	1.496	1.4705	1.576	2.089
	$\hat{\alpha}_{AUTP}$	0.997	0.9903	0.9758	1.002	0.999	0.9946	1.032	0.989	1.003	0.9959	0.9883	0.983
	$\hat{\alpha}_{MTP}$ at k_1	1.369	1.5169	1.6342	2.094	1.477	1.5402	1.6379	2.195	1.497	1.4848	1.607	2.164
	$\hat{\alpha}_{MTP}$ at k_2	1.368	1.5069	1.6059	2.044	1.471	1.5367	1.6325	2.110	1.498	1.4835	1.605	2.172
	$\hat{\alpha}_{MTP}$ at k_3	1.369	1.5167	1.6344	2.097	1.470	1.5401	1.638	2.141	1.497	1.4848	1.6071	2.165

Table 2. Ratio of AMSE of OLS over various two-parameter estimators (p = 4 and $\beta = (2, 1, 4, 5)'$) (cont.)

ρ = 0.95													
n		20				50				100			
$\hat{\sigma}^2$		1	5	10	25	1	5	10	25	1	5	10	25
$\hat{\alpha}$	$\hat{\alpha}_{TP}$	1.214	1.2792	1.4212	1.554	1.211	1.256	1.3406	1.495	1.222	1.2798	1.3456	1.502
	$\hat{\alpha}_{LTE(3)}$	1	1.0002	1.0006	1.001	1	1.0003	1.0006	1.001	1	1.0003	1.001	1.002
	$\hat{\alpha}_{NTP}$	1.448	1.5177	1.8593	2.434	1.438	1.4786	1.6337	2.128	1.46	1.5168	1.6237	2.12
	$\hat{\alpha}_{AUTP}$	1	0.9974	0.9929	1.005	0.998	1.001	1.002	0.987	0.999	0.9947	0.9954	0.983
	$\hat{\alpha}_{MTP}$ at k_1	1.447	1.5282	1.8693	2.485	1.442	1.5075	1.6492	2.172	1.468	1.5207	1.6325	2.182
	$\hat{\alpha}_{MTP}$ at k_2	1.447	1.5271	1.872	2.499	1.443	1.5053	1.6367	2.176	1.467	1.5186	1.6458	2.175
	$\hat{\alpha}_{MTP}$ at k_3	1.447	1.5281	1.8697	2.486	1.442	1.5074	1.6493	2.174	1.468	1.5206	1.6227	2.183
ρ = 0.99													
n		20				50				100			
$\hat{\sigma}^2$		1	5	10	25	1	5	10	25	1	5	10	25
$\hat{\alpha}$	$\hat{\alpha}_{TP}$	1.2819	1.3621	1.397	1.552	1.2745	1.3373	1.466	1.563	1.2615	1.37	1.457	1.565
	$\hat{\alpha}_{LTE(3)}$	1.0011	1.0021	1.003	1.004	1.0007	1.0023	1.004	1.004	1.0009	1.0026	1.004	1.005
	$\hat{\alpha}_{NTP}$	1.4962	1.6641	1.798	2.355	1.4737	1.5729	1.939	2.37	1.4394	1.6433	1.9	2.382
	$\hat{\alpha}_{AUTP}$	0.9847	1.0003	1.003	0.963	0.9923	0.9761	0.963	1.003	0.9929	0.9782	1.002	0.958
	$\hat{\alpha}_{MTP}$ at k_1	1.5321	1.7232	1.909	2.55	1.4708	1.6249	2.071	2.617	1.4374	1.6832	2.004	2.621
	$\hat{\alpha}_{MTP}$ at k_2	1.5293	1.7102	1.918	2.592	1.479	1.6116	2.081	2.544	1.4319	1.6666	2.030	2.671
	$\hat{\alpha}_{MTP}$ at k_3	1.532	1.7237	1.911	2.555	1.47	1.6253	2.073	2.622	1.4376	1.6835	2.006	2.627

Table 3. Ratio of AMSE of OLS over various two-parameter estimators (p = 3 and $\beta = (1, 2, 5)'$)

ρ = 0.90													
n		20				50				100			
$\hat{\sigma}^2$		1	5	10	25	1	5	10	25	1	5	10	25
$\hat{\alpha}$	$\hat{\alpha}_{TP}$	1.1772	1.3769	1.554	1.637	1.1193	1.3248	1.511	1.58	1.113	1.3425	1.539	1.671
	$\hat{\alpha}_{LTE(3)}$	1.0047	1.0063	1.01	1.01	1.0007	1.0017	1.002	1.002	1.0012	1.0016	1.003	1.002
	$\hat{\alpha}_{NTP}$	1.1245	1.5709	2.098	2.572	1.0138	1.4679	2.218	2.588	1.0059	1.493	2.247	2.905
	$\hat{\alpha}_{AUTP}$	1.003	0.9369	0.92	0.93	0.9882	1.0002	0.975	0.978	1.002	0.9788	0.975	1.003

Table 3. Ratio of AMSE of OLS over various two-parameter estimators (p = 3 and $\beta = (1,2,5)'$) (cont.)

ρ = 0.90													
n		20				50				100			
$\hat{\sigma}^2$		1	5	10	25	1	5	10	25	1	5	10	25
$\hat{\alpha}$	$\hat{\alpha}_{MTP}$ at k_1	1.1551	1.6875	2.429	3.004	1.0116	1.4822	2.347	2.712	1.0033	1.5099	2.343	3.081
	$\hat{\alpha}_{MTP}$ at k_2	1.1336	1.5861	2.52	3.12	1.0017	1.4733	2.326	2.754	1.003	1.5067	2.268	2.972
	$\hat{\alpha}_{MTP}$ at k_3	1.1545	1.6884	2.437	3.016	1.0112	1.4823	2.349	2.715	1.0029	1.51	2.346	3.085
ρ = 0.95													
n		20				50				100			
$\hat{\sigma}^2$		1	5	10	25	1	5	10	25	1	5	10	25
$\hat{\alpha}$	$\hat{\alpha}_{TP}$	1.217	1.2942	1.4217	1.543	1.147	1.214	1.2792	1.4212	1.209	1.299	1.4227	1.559
	$\hat{\alpha}_{LTE(3)}$	1	1.0001	1.0006	1.001	1	1	1.0002	1.0006	1	1.0003	1.0007	1.001
	$\hat{\alpha}_{NTP}$	1.457	1.5275	1.8817	2.38	1.314	1.448	1.5177	1.8593	1.434	1.5311	1.834	2.445
	$\hat{\alpha}_{AUTP}$	1	0.9971	0.9921	1.002	1	1	0.9974	0.9929	1.020	0.9968	1.003	0.99
	$\hat{\alpha}_{MTP}$ at k_1	1.454	1.5328	1.9085	2.391	1.31	1.447	1.5282	1.8693	1.435	1.5322	1.845	2.508
	$\hat{\alpha}_{MTP}$ at k_2	1.454	1.5322	1.9105	2.398	1.32	1.457	1.5271	1.872	1.439	1.5394	1.8461	2.45
	$\hat{\alpha}_{MTP}$ at k_3	1.454	1.5327	1.9089	2.392	1.31	1.457	1.5281	1.8697	1.441	1.5322	1.8454	2.509
ρ = 0.99													
n		20				50				100			
$\hat{\sigma}^2$		1	5	10	25	1	5	10	25	1	5	10	25
$\hat{\alpha}$	$\hat{\alpha}_{TP}$	1.3596	1.534	1.624	1.65	1.2018	1.3461	1.561	1.654	1.1787	1.3767	1.577	1.575
	$\hat{\alpha}_{LTE(3)}$	1.0053	1.006	1.008	1.01	1.0033	1.005	1.008	1.008	1.004	1.0065	1.008	1.009
	$\hat{\alpha}_{NTP}$	1.5565	2.09	2.489	2.66	1.1904	1.5099	2.133	2.555	1.1288	1.5669	2.173	2.31
	$\hat{\alpha}_{AUTP}$	1.001	1.002	0.933	0.93	0.9607	0.9514	1.002	0.933	0.9541	1.003	0.927	0.927
	$\hat{\alpha}_{MTP}$ at k_1	1.6313	2.376	2.972	3.24	1.2261	1.592	2.394	2.949	1.1592	1.638	2.499	2.843
	$\hat{\alpha}_{MTP}$ at k_2	1.6133	2.379	2.989	3.42	1.2051	1.5716	2.41	3.062	1.1312	1.6141	2.564	2.658
	$\hat{\alpha}_{MTP}$ at k_3	1.6317	2.381	2.984	3.26	1.2255	1.5926	2.399	2.957	1.1585	1.6385	2.506	2.854

From Tables 2 and 3 we observe that the performance of our proposed modified two-parameter estimator $\hat{\alpha}_{MTP}$ is better than $\hat{\alpha}_{OLS}$, $\hat{\alpha}_{TP}$, $\hat{\alpha}_{LTE(3)}$, and $\hat{\alpha}_{AUTP}$. At the same time $\hat{\alpha}_{MTP}$ perform equivalently and is slightly better than $\hat{\alpha}_{NTP}$ for all combinations of correlation between predictors (ρ), the numbers of explanatory variables (p), the sample size (n), the choice of the ridge parameter (k) and the variance of the error term (σ^2) used in this simulation study.

7. Conclusion

In this article a modified two-parameter estimator alternative to the OLS estimator is proposed for estimating the regression parameter in the presence of multicollinearity. The performance of the proposed estimator is evaluated in terms of scalar mean-squared error criterion. Through the simulation study the performance of the proposed estimator is evaluated, for different combinations of ρ, p, n, k and σ^2 over the OLS and other two-parameter estimators reviewed in this article. Finally, it is found that the performance of the proposed estimator is satisfactory over the other estimators in the presence of multicollinearity.

Acknowledgements

The author is very grateful to the reviewers and the editor for their valuable comments and constructive suggestions which certainly improved the quality of the paper in the present version.

REFERENCES

AKDENIZ, F., EROL, H., (2003). Mean squared error matrix comparisons of some biased estimators in linear regression. Commun. Statist. Theor. Meth. 32(23), 89–2413.

AKDENIZ, F., KACIRANLAR, S., (1995). On the almost unbiased generalized Liu estimator and unbiased estimation of the bias and MSE. Commun. Statist. Theor. Meth. 24, 1789–1797.

BATAH, F. S., RAMNATHAN, T., GORE, S. D., (2008). The efficiency of modified jackknife and ridge type regression estimators: a comparison. Surveys in Mathematics and its Applications 24(2), 157–174.

BUONACCORSI, J. P., (1996). A Modified estimating equation approach to correcting for measurement error in regression. Biometrika 83, 433–440.

CROUSE, R. H., JIN, C., HANUMARA, R. C., (1995). Unbiased ridge estimation with prior information and ridge trace. Commun. Statist. Theor. Meth. 24, 2341–2354.

DORUGADE, A. V., KASHID, D. N., (2011). Parameter estimation method in Ridge Regression. Interstat May 2011.

HOERL, A. E., KENNARD, R. W., (1970a). Ridge regression: biased estimation for nonorthogonal problems. Tech. 12, 55–67.

HOERL, A. E., KENNARD, R. W., BALDWIN, K. F., (1975). Ridge regression: Some Simulations. Commun. Statist. 4, 105–123.

KACIRANLAR, S., SAKALLIOGLU, S., AKDENIZ, F., STYAN, G. P. H., WERNER, H. J., (1999). A new biased estimator in linear regression and a detailed analysis of the widely-analysed dataset on Portland Cement. Sankhya Ind. J. Statist. 61, 443–459.

KADIYALA, K., (1984). A class almost unbiased and efficient estimators of regression coefficients. Econom. Lett. 16, 293–296.

KIBRIA, B. M., (2003). Performance of some new ridge regression estimators. Commun. Statist. –Simulation 32 (2), 419–435.

LIU, K., (1993). A new class of biased estimate in linear regression. Commun. Statist. Theor. Meth. 22, 393–402.

MASSY, W. F., (1965). Principal components regression in exploratory statistical research. JASA 60, 234–266.

NOMURA, M., (1988). On the almost unbiased ridge regression estimation. Commun. Statist. –Simulation 17(3), 729–743.

MAYER, L. S., WILLKE, T. A., (1973). On biased estimation in linear models. Technometrics 15, 497–508.

MONTGOMERY, D. C., PECK, E. A., VINING, G. G., (2006). Introduction to linear regression analysis. John Wiley and Sons, New York.

OZKALE, M. R., KACIRANLAR, S., (2007). The restricted and unrestricted two-parameter estimators. Commun. Statist. Theor. Meth. 36, 2707–2725.

OHTANI, K., (1986). On small sample properties of the almost unbiased generalized ridge estimator. Commun. Statist. Theor. Meth. 15, 1571–1578.

SAKALLIOGLU, S., KACIRANLAR, S., (2008). A new biased estimator based on ridge estimation. Stat Papers 49, 669–689.

STEIN, C., (1956). Inadmissibility of the usual estimator for mean of multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Probab. 1, 197–206.

SWINDEL, B. F., (1976). Good ridge estimators based on prior information. Commun. Statist. Theor. Meth. A5, 1065–1075.

SINGH, B., CHAUBEY, Y. P., (1987). On some improved ridge estimators. Stat Papers 28, 53–67.

YANG, H., CHANG, X., (2010). A New Two-Parameter Estimator in Linear Regression. Commun. Statist. Theor. Meth. 39, 923–934.

WU, J., YANG, H., (2011). Efficiency of an almost unbiased two-parameter estimator in linear regression model. Statistics 47(3), 535–545.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 37–58

THE USE OF NON-SAMPLE INFORMATION IN EXIT POLL SURVEYS IN POLAND

Arkadiusz Kozłowski[1]

ABSTRACT

Exit poll is a commonly used tool to predict election outcome in democratic countries. The aims of this survey, however, go beyond the standard prediction which usually loses its value after 1-2 days. Lasting benefits of exit poll result from the possibility of estimating vote distribution in socio-demographic groups, changes of political preferences, the motives for choosing a candidate, etc. No other survey is capable of providing such detailed data with satisfactory precision. Nonetheless, the exit poll accuracy, both in Poland and abroad, often leaves much to be desired. It seems that while conducting the research the non-sample information is not used sufficiently, which could improve the quality and the precision of the survey.

The sources of auxiliary variables, which can be used in exit poll, along with the analysis of technical aspects of their acquisition and combination are outlined in this paper. Statistical methods aiming at incorporating the information about those variables to the survey, both at the stage of selecting the sample of precincts and at the stage of forecasting election results are proposed. Developed solutions were subjected to the simulation testing on the parliamentary election to the Sejm 2011 data. The results confirm the possibility of a significant increase in the effectiveness of estimates by improving 'representativeness' of a sample and by applying complex estimation of parameters.

Key words: exit poll, auxiliary variables, balanced sampling, complex estimation.

1. Introduction

An exit poll, conducted on the election day in which respondents (voters) leaving the selected polling stations answer, i.a. on who they cast their votes, is a commonly used tool to predict the election outcome in democratic countries. Thanks to work of a few thousand pollsters within appropriately organized logistics and IT operation, the TV viewers can know the approximate election results on the same day, right after the last polling station has been closed. These

[1] University of Gdańsk. E-mail: arekozlowski@wzr.ug.edu.pl.

estimates allow first commentaries and live analysis to be presented on the election night (which usually guarantees wide audience).

The aim of exit poll is not only to predict the election result (the very same forecast quickly becomes useless). The survey gives the opportunity for in-depth analysis of voting results in such aspects as vote distribution in different socio-demographic groups, the changes of political preferences in relation to previous election, the motives of choosing a particular party or candidate, the motives of participating in election, etc. These analyses remain to be the most reliable source of citizens' political behaviours until the next election, due to the fact that current political surveys are unable to provide such detailed data with the necessary precision (Szreder, 2011).

In terms of statistics, exit poll is a unique survey because it is a sample survey research the general result of which is quickly confronted with the result of complete enumeration (the same cannot be said about pre-election surveys as the population of interest is much larger – apart from actual voters, the pre-election survey encompasses also people entitled to vote but not participating in the election). The degree of compatibility between the general forecast and official results announced by the National Electoral Commission (PKW) is the basis for validation of applied methodology, and also has influence on the quality of more detailed data sets.

The key elements of the exit poll methodology are the sampling design and the method of estimation. The sample is chosen in two stages. In the primary stage the precincts are sampled and in the secondary stage the voters leaving the polling station are chosen. As far as the selection of the respondents to the sample is concerned there is an agreement between theorists and practitioners that the best choice in this case is systematic sampling (in Poland it is usually every tenth person leaving polling station). This approach mainly results from the uneven distribution of particular party voters during the day, which was the object of study, for instance by Klorman (1976), Busch and Lieske (1985). It is especially important in countries where the election day falls on working day, like in the US (Tuesday) and the UK (Thursday). In Poland, as in the majority of countries, election takes place on holiday.

The more problematical stage is the choice of accurate sample of precincts which would be the most representative of the population. Barreto et al. (2006, p. 479) state, "In fact, this is *the* most important step in exit polling", suggesting that selecting the inaccurate sample of precincts was the reason of unsatisfactory exit poll results during U.S. presidential election 2004 (the survey conducted by Edison-Miofsky Research forecasted the victory of John Kerry over Georg W. Bush). In Poland around 25-26 thousand of precincts are created during the election (around 24 thousand are the so-called regular precincts). The conventional approach towards the issue of sampling the precincts is stratified sampling, in which strata are created based on geographical regions and the type of territorial unit (city/village). Such a solution increases the representativeness of a sample compared to unrestricted sampling, however, by increasing the degree of

the use of a prior knowledge about the population of interest, the representativeness can be further increased. The sources of knowledge about population can be, on the one hand, the official statistics of the Central Statistical Office (GUS), on the other hand, the official results of the past elections shared by PKW. PKW databases are particularly valuable because they contain detailed voting results for each of over 25 thousand precincts.

The same data sources which help selecting the right sample (supporting sampling process) can also support the process of estimating the election result. It is commonly known that electorates of particular political groups vary between themselves in respect of many demographic, economic and social variables. If this type of characteristics was known for every precinct, it would be, along with information about the previous elections vote structures, a rich source of auxiliary variables for complex estimations.

The aim of this paper is the empirical verification of the assumption that incorporating additional auxiliary variables to the exit poll strategy increases effectiveness of estimating the election result. The additional variables are non-sample and are not directly connected with the unit of research, i.e. the precinct, therefore, statistical analysis is preceded by the presentation of technical aspects of data acquisition and combination along with pointing out the advantages and limitations of a given source. At the stage of selecting the sample, an innovative method of balanced sampling, the so-called cube method, is applied. The object of the analysis is estimation of relative result, i.e. the fraction of votes cast on particular committees across the country. Proposed strategies are subjected to simulation testing on the parliamentary election to the Sejm 2011 data.

2. Characteristics of exit poll

The first exit poll was conducted in the United States in the 60s of the 20th century at the request of CBS (Levy, 1983). The creation and development of the survey methodology is ascribed to Warren Mitofsky (Moore, 2003). In Poland the first this type of research was conducted by Ośrodek Badania Opinii Publicznej (OBOP) during the first and second round of presidential election in 1990.

Exit poll differs from other political preferences surveys in many aspects. First of all, the population of interest is different. Apart from actual voters, the political surveys encompasses also people entitled to vote but not participating in the election, whereas the exit poll surveys only people taking part in the election (this is one of the main arguments of the opinion research centres refuting accusations of the discrepancy between pre-election surveys and the actual election results). Secondly, the exit poll questions refer to facts (the actual votes), not to the intentions which often are different from the actual voting decisions. The survey is conducted directly after leaving the polling station which minimizes the errors connected with ‘gaps in memory’ and the ‘bandwagon effect’ due to the

fact that the final result is still unknown. Another distinguishing feature is much higher percentage of the conducted interviews. In standard surveys conducted by applying the CATI method, the percentage goes up to several percents, in face-to-face interviews it goes up to 50-60% whereas in exit polls in Poland it remains at the level of 85-95% (Domański et al., 2010). It is worth emphasizing that this is a particularly high level, practically unparalleled in the Western countries, where the general trend for the decrease of the sample response rates has also affected exit poll (the examples of the response rate: Germany: 70-72% (Hofrichter, 1999), USA: 45-55% (Lenski, 2008), Great Britain: 86% (Moon 2008)). Furthermore, the survey scale is also noteworthy – the sample size measured by the number of individual respondents is usually several times higher than the sample size of a standard pre-election survey. For example, during the parliamentary election 2011, TNS OBOP conducted research on the sample of 900 polling stations, conducting around 100 000 interviews in total.

The above-mentioned characteristics raise the value of the exit poll information compared to other preferences and political behaviours surveys. The challenge to maintain this advantage is that in a few countries there is a possibility of voting indirectly, not in a polling station (the so-called *absentee ballot*), i.e. via mail, Internet or attorney. Additionally, people voting through mail can do this within a certain period of time before the official election day. This complicates the survey and forces the organizer to apply different techniques, e.g. telephone surveys, in order to supplement the interviews conducted in front of polling stations. However, for the time being, this is not a problem of Polish researchers.

The main focus of this paper is on reducing errors relating to the selection of precincts and estimation, nonetheless, it is worth mentioning other potential sources of errors. They mainly occur during the selection of voters and interaction between the respondent and interviewer. One of them is the faulty implementation of systematic sampling scheme. The threat is that the selection discipline of taking every n^{th} person will break down and interviewers will approach individuals who they think will respond. That would naturally introduce selection bias. Another problem is that of *co-location* of precincts, when two or more precincts are in the same physical facility. For someone who is leaving such a facility it is usually not clear which precinct he/she voted in, which makes it difficult to maintain the desired selection probabilities (Scheuren and Alvey, 2008, p.12).

Another potential source of errors are non-respondents. As stated above, the response rate in Polish exit polls is still quite high, but it will probably decline in the near future. There are two types of non-respondents in exit polls: refusals and misses. A refusal is when a sampled voter refuse to participate in the survey and a miss is when a sampled voter cannot be asked to fill out the questionnaire because the interviewer is too busy or the voter does not pass the interviewer. Merkle and Edelman (2002) estimate that about three-fourths of non-response in exit polls is attributable to refusals and about a quarter to misses. Refusals pose a greater threat to the survey outcome than misses because the voter's reluctance to participate in exit polls can result from specific political attitudes and

consequently cause certain bias. Merkle and Edelman (2002) studied factors correlated with non-response on the basis of various exit polls conducted from 1992 through 1998 in the USA. Conclusions from their investigation are as follows: of the voter's characteristics, age is the most strongly related factor to the response rate (older voters have lower response rates), of the Election Day factors, interviewing position at the polling place (closeness to the polling place) is the main factor that can have a dramatic effect on response rates (response rates decline as the interviewer moves further away from the voting room), and of the interviewer characteristics, again age is the most significant factor (the age of the interviewer is positively correlated with response rates). What is surprising is the authors found very little or no correlation between response rates and exit poll error measures. Neither refusal rates nor miss rates were significant predictors of errors.

Apart from unit non-response, when information is missing on all questionnaire variables, researchers conducting exit polls experience item non-response (when only some answers are missing) and false answers. The crucial factor for the scale of these occurrences seems to be the mode of data collection. One of the most popular solution here is the so-called secret ballot. In this mode respondents chosen to the sample are interviewed by the use of self-administered questionnaire, which is then put in the envelope or deposited in the specially prepared ballot box. Bishop and Fisher (1995) proved experimentally that this mode of data collection decreases item non-responses and gives socially desirable responses, compared to the face-to-face interview. Using secret ballot one assumes that voters can read and understand questions well enough to give a reasonable answer. This can not be the case in countries with low literacy level. Bautista et al. (2006) give the example of Mexico, where due to low educational level mixed-mode of data collection were used (face-to-face with secret ballot).

3. Sources of additional information

Exit poll does not always meet recipients' expectations as far as the compatibility between the forecast and the actual result is concerned, irrespective of the fact that the survey is conducted on a large sample size with a low rate of refusals and potentially low measurement error. As a result, the need to strengthen the survey with non-sample information arises. Two main sources of non-sample information are specified: data referring to past election results and GUS data not directly connected with the elections but strongly related to voters' decisions.

As far as the past general election data is concerned, the election results starting from the presidential election 2000 are available on each aggregation level (from voivodeship to precincts) on the PKW website. The key issue for using this data is the possibility of confronting the results for two or more elections between the corresponding precincts. This process, however, causes some problems. The main issue is that according to the Election code (2011) the

division into the election precincts is made by authority of municipality, however, the division is not permanent. Before each election a new division is made and both the number and the boarders of precincts within municipality can change. Further difficulties arise from typical demographic changes (reaching voting age, deaths, migrations), voting outside the voter's district. The above-mentioned reasons lead to a situation in which the voters from *i*-th precinct in the *X* municipality are not exactly the same voters who participated in the elections a few years ago. However, the differences are insignificant, thus the informative value of the past election results should remain high.

Another aspect of the use of information about past results is choosing the elections which will serve as a reference point to strengthen the estimates of the current survey. The most reasonable option seems to be choosing the chronologically nearest election as in such case, the changes on the political scene along with demographic and organizational changes are not so significant. In the case of the parliamentary election 2011 such a reference point can be the presidential election 2010 (some of the main candidates can be linked to political parties). However, the parliamentary election's character is different in respect of the division into electoral districts and different set of candidates in each district, hence the parliamentary election 2007 can be considered as a better reference point. The question arises as how much the informative value of the data has deteriorated due to the changes in precincts during 4 years. Another possibility is choosing the European Parliament election 2009, however, irrespective of being nearest in time, this choice has some drawbacks, i.e. low election turnout (24,53%), different division into electoral districts and generally speaking different attitude towards the European election among both the voters and the politicians than towards the national elections. In the conducted simulation analysis the use of the presidential election 2010 results (first round) and the parliamentary election to the Sejm 2007 results was studied.

Only the technical issue of matching the corresponding precincts in the mentioned elections needs to be resolved. Due to the fact that the division into precincts lies within the competence of municipality, there is no main key identifying the precincts between elections. By comparison of the precinct address, precinct number (numeration applied within municipality) and the number of registered voters, the corresponding precincts can be identified with the high credibility. The probably correctly linked precincts in which the difference in the number of people entitled to vote exceeded 200 were excluded (this operation also eliminates the precincts in tourist resorts, in which the vast majority of voters are out-of-towners). In further analysis only the regular precincts were taken into account, as only this type of precinct encompasses relatively unchanged voter groups. The number of linked precincts and the number of registered voters in comparison with the whole population are presented in Table 1.

Table 1. Data on populations subjected to simulation test

	Number of precincts (Number of registered voters)			
	all	regular	linked S11 – P10	linked S11 – P10 – S07
Population:	U	U1	U2	U3
Sejm 2011 (S11)	25 993 (30 762 931)	24 217 (30 387 730)	23 553 (29 382 340)	22 209 (27 305 152)
Presidential 2010 (P10)	25 774 (30 813 005)	24 144 (30 382 814)	23 553 (29 388 485)	22 209 (27 349 352)
Sejm 2007 (S07)	25 476 (30 615 471)	23 903 (30 188 868)		22 209 (27 332 149)

Source: Own calculation based on PKW data.

Another source of information that can increase the quality of exit poll are GUS official statistics for the units of territorial division of the country referring to social and economic characteristics. In this case, the main factors limiting the possible uses are: the level of data aggregation, the range of described population and the timeliness of data. As far as the aggregation level is concerned, the most helpful would be the data at the level of precincts, which of course does not exist. The lowest available level of aggregation is municipality and only for a limited range of variables. The second limitation is a problem due to the fact that GUS data refers naturally to the whole population and not only to the active voter groups which are analyzed in exit poll. As far as the timeliness of data is concerned, it depends on the type of variables, however, usually a few years' delay in relation to the election date has to be taken into account. In that case, the most reasonable approach is to use the features which do not change significantly in time.

Despite these limitations, it is believed that incorporating certain variables can improve both selecting a sample and the result estimations. After the analysis of available socio-economic data and their relation to the past election results, it was decided to incorporate to the study two variables at the level of municipality:

- economic entities registered in REGON per every 10 000 population (2010, *podm_gm*),
- the area of agricultural land (in ha) per every 1 000 population (2005, *uzyt_gm),*

and two variables at the level of powiat (the second-level unit of local government and administration in Poland):

- registered unemployment rate (2011, *bezr_pow),*
- the average monthly gross salary in comparison to national average salary (2010, *wyn_pow)*.

The very valuable source of the supportive information can be the distribution of voters in view of the features like sex, age and education. The character of election does not enable the official collection of such data, however, the opinion research centre conducting such survey in the past has own estimates at its disposal and can use them to correct the estimates at the level of a single unit. In view of the fact that presented analysis encompasses only official data in which the most detailed information is the general election result in the precinct, this possibility is not taken into consideration.

4. Sampling plan

The proposed sampling technique, which expands the conception of restricted sampling compared to typical stratified sampling, is balanced sampling. The sampling design is called balanced in relation to certain additional characteristics (*auxiliary variables)* if it generates samples from which the estimates of additional variables sums (by Horvitz–Thompson estimator, HT) match the known actual sums (Deville, 2004). In other words, in balanced sampling the auxiliary variables are estimated without an error. The above definition can be generalised for any samples, not necessarily chosen in random sampling.

The idea of balanced sampling is not new. It appeared along with the representative method and is connected with the very same term of representativeness. The first use of this conception in practice refers to famous sampling of precincts during the Italy census (Gini, Galvani, 1929, after Langel, Tille, 2011). 29 precincts were selected in such way that the averages from the sample for a few auxiliary variables would match the average from population. Both Nayman and Yates (Langel, Tille, 2011) condemned such behaviour as the sample was selected purposive. It was later observed that the balanced sample can be selected in a probabilistic way. A special example is stratified sampling, in which the sample is random and at the same time balanced on the specified indicator variables of the strata (such variables take on the value 1 for the units belonging to stratum and, otherwise, 0; the number of variables corresponds with the number of strata).

From a technical side, the probabilistic way of selecting a balanced sample is not evident. There is a number of methods enabling this choice, the majority of which is based on elimination process (the so-called *rejective sampling*), i.e. rejecting some of the sampled units (or the whole sample) if the condition of the balance is not satisfied (one of the variants of the method is the so-called tied sampling (Kozłowski, 2012). This requires conducting a number of interactions, which, depending on complexity of limiting conditions, are more or less time-consuming. The majority of methods also have some constraints resulting from the possible applicability only in a chosen sampling schemes, the lack of possibility of differentiating inclusion probabilities as well as limited number and type of auxiliary variables. The method which overcomes these difficulties and, in

this context, is the most general is the so-called cube method proposed by Deville (2004).

The starting point in the cube method is geometrical conceptualisation of all possible samples (in sampling without replacement) of N-element population as vertices of N-cube C, i.e. $C=[0,1]^N$. Any sample s is defined as a vector $(s_1,\ldots,s_k,\ldots s_N)'$, where s_k takes on the value 1 if the k^{th} unit is in the sample and, otherwise, 0. The number of all possible samples (of any size) equals to the number of the vertices of the cube C, i.e. 2^N. In the instance of 3-element population (N=3) the sample space can be presented as vertices of a cube (Figure 1). Starting from the point defined by the vector of the first order inclusion probabilities $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_k,\ldots,\pi_N)$, the selection of the sample can be illustrated as random 'reaching' to the one of vertices.

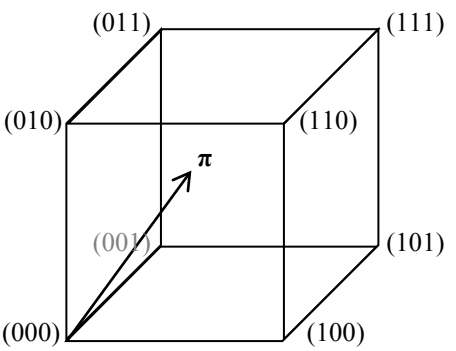


Figure 1. Geometrical representation of sample space for 3-element population

Source: Deville, 2004, p. 896.

The design is balanced on auxiliary variables only if the data is at the unit level, i.e. for every unit of the population the vector $\boldsymbol{x_k}=(x_{k1},\ldots,x_{kj},\ldots,x_{kp})'$ should be known, where p – the number of auxiliary variables. The totals of auxiliary variables $\boldsymbol{X}=\sum_{k\in U} x_k$ are estimated by $\widehat{\boldsymbol{X}}_{HT}=\sum_{k\in U}\frac{x_k S_k}{\pi_k}$, where U – population. The balanced sampling, as per definition, reassures:

$$\widehat{\boldsymbol{X}}_{HT}=\boldsymbol{X} \tag{1}$$

for every possible sample, in other words $\widehat{\boldsymbol{X}}_{HT}$ variation equals 0. In practice, this condition is usually fulfilled only approximately. The equation (1) is a set of limiting conditions, which defines an affine subspace (hyperplane) Q in $\mathbb{R}^N$ in dimension *N-p.* The idea of balanced sampling is to randomly 'reach' to such a vertex of cube C, which at the same time belongs to hyperplane Q (exactly balanced) or is located as close as possible (approximately balanced).

In the cube method the vector $\boldsymbol{\pi}$ is randomly transformed into a vector containing only values 0 and 1 (i.e. vector $\boldsymbol{s}$ defining a sample) in such a way that inclusion probabilities are exactly satisfied and balance condition (1) for every variable is satisfied to the furthest extent possible (Tillé, 2011). The method is divided into two phases: flight phase and landing phase. Flight phase is a random walk on in the intersection of the cube C and the constraint subspace, which starts from the point defined by the vector $\boldsymbol{\pi}$ and ends on the vertex of intersection ($\boldsymbol{\pi}^*$). If the reached point is at the same time the vertex of the cube C (i.e. all elements of $\boldsymbol{\pi}^*$ equals 0 or 1), then the balance is exactly satisfied and the process of sampling is finished. Otherwise, the landing phase begins in which (by applying linear programming) the vertex of cube C located as close as possible to the point reached in the flight phase and at the same time satisfying the inclusion probabilities is set.

In most cases the perfect balance cannot be achieved due to the so-called rounding problem. Nevertheless, it is proved that (Tillé 2006, p. 165):

$$\left|\widehat{\mathrm{X}}_{\mathrm{j\,HT}} - \mathrm{X_j}\right| \leq \mathrm{p} * \max_{\mathrm{k}\in\mathrm{U}} \left|\frac{\mathrm{x_{kj}}}{\pi_\mathrm{k}}\right| \tag{2}$$

The accuracy of the balance is decreased along with the increase in the amount of auxiliary variables, and is improved along with the increase in a sample size if it is set before the sampling.

5. Methods of estimation

The estimated parameter is the fraction of votes cast on *J* committee, which can be presented as a quotient of two sums:

$$\mathrm{P_J} = \frac{\mathrm{Y_J}}{\mathrm{Y}} = \frac{\sum_{\mathrm{k}\in\mathrm{U}} \mathrm{y_{Jk}}}{\sum_{\mathrm{k}\in\mathrm{U}} \mathrm{y_k}} \tag{3}$$

where:

Y_J – the sum of valid votes cast on the *J* committee across the country,

Y – the sum of valid votes in total across the country,

y_{Jk} – the number of valid votes cast on *J* committee in the k^{th} precinct,

y_k – the number of valid votes in total in k^{th} precinct.

The problem of estimation is to estimate the total number of valid votes and the total number of valid votes cast on *J* committee based on *n*-element sample of precincts. In the case of both sums, it was decided to test three types of estimators: Horvitz-Thompson estimator (HT), ratio estimator (q) and estimator using the log-linear model (P). The model is the so-called Poisson regression – the type of a generalized regression model, in which it is assumed that the response variable Y has a Poisson distribution. The function linking linear combination of explanatory variables with the response variable is a natural logarithm. This model was chosen because it is particularly useful in the analysis of count variables (taking on integer nonnegative values).

The Horvitz-Thompson estimators used for estimating the sum of votes in total and the sum of votes cast on *J* committee respectively are presented with the following formulas:

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{4}$$

$$\hat{Y}_{J,HT} = \sum_{i \in s} \frac{y_{J,i}}{\pi_i} \tag{5}$$

where:

y_i, $y_{J,i}$ – the number of valid votes, in total and on *J* committee respectively, cast in i^{th} precinct selected to the sample,

π_i - the first order inclusion probability.

The Horvitz-Thompson estimator does not use the auxiliary variables directly, however, it can use them indirectly if in the sampling design the additional variable is used to establish the inclusion probabilities.

The analogous set of ratio estimators is as follows:

$$\hat{Y}_q = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}} X \tag{6}$$

$$\hat{Y}_{J,q} = \frac{\hat{Y}_{J,HT}}{\hat{X}_{J,HT}} X_J \tag{7}$$

where:

$\hat{X}_{HT}$, $\hat{X}_{J,HT}$ - HT estimators of the sums for auxiliary variables,

X ,X_J – the known sums of auxiliary variables.

Ratio estimators use the information about one auxiliary variable for which the values from the sample and the actual sum in population are known. In the conducted analysis the auxiliary variable is usually the same variable but in the past (e.g. the result of the same committee in the previous election).

In the third case, the numbers of votes were modelled by the Poisson regression, by using the following formula (the same for both parameters):

$$\hat{y}_k = e^{x_k \hat{\beta}} \cdot X_k^* \tag{8}$$

where:

$\hat{y}_k$ - the theoretical number of votes in k^{th} precinct,

$\boldsymbol{x_k}$ – the vector of explanatory variables (independent variables),

$\widehat{\boldsymbol{\beta}}$ – the vector of regression factor estimated based on the sample *s*,

X_k^* - offset variable.

The aim of the offset variable, which is added to the basic model, is to eliminate the differences between number of votes resulting only from the precinct's size. In the case of modelling the number of votes in total, the offset variable was the number of registered voters in S11, whereas in the case of modelling the number of votes cast on committee *J*, the offset variable was a previously estimated total number of votes. Estimation of the sum of votes across the country is the simple prediction aggregation for all precincts in the population.

The estimators are presented as follows:

$$\widehat{Y}_P = \sum_{k \in U} \hat{y}_k \tag{9}$$

$$\widehat{Y}_{J,P} = \sum_{k \in U} \hat{y}_{Jk} \tag{10}$$

where:

$\hat{y}_k$, $\hat{y}_{Jk}$ - the theoretical number of votes in k^{th} precinct according to the model (8), in total and on committee *J*, respectively.

The general rule was adopted (with the exception of formula (12)) that the final estimator of fraction of votes cast on committee *J* would be the quotient of two sums estimated with the same type of estimator. Complex estimators can use supportive data from other sources and to different extents, which results in a high number of possible variants. Finally, it was decided to separate seven estimators, the effectiveness of which will be subjected to simulation testing later in this paper:

$$\widehat{P}_J^{(HT)} = \frac{\widehat{Y}_{J,HT}}{\widehat{Y}_{HT}} \tag{11}$$

$$\widehat{P}_J^{(Q-S11)} = \frac{\widehat{Y}_{J,HT}}{\widehat{Y}_q^{(S11)}} \tag{12}$$

where:

$\hat{Y}_q^{(S11)}$ - the ratio estimator according to formula (6) in which the auxiliary variable is the number of registered voters during the parliamentary election to the Sejm 2011;

$$\widehat{P}_J^{(Q-P10)} = \frac{\widehat{Y}_{J,q}^{(P10)}}{\widehat{Y}_q^{(P10)}} \tag{13}$$

where:

$\widehat{Y}_{J,q}^{(P10)}$ - the ratio estimator according to formula (7), in which the auxiliary variable is the number of votes cast on the candidate linked to the committee J during the presidential election 2010 (in the case of the Ruch Palikota (RuchP) committee, the auxiliary variable was the result of Bronisław Komorowski committee),

$\widehat{Y}_q^{(P10)}$ - ratio estimator according to formula (6), in which the auxiliary variable is the number of valid votes in total during the presidential election 2010;

$$\widehat{P}_J^{(Q-S07)} = \frac{\widehat{Y}_{J,q}^{(S07)}}{\widehat{Y}_q^{(S07)}} \tag{14}$$

where:

$\widehat{Y}_{J,q}^{(S07)}$ - the ratio estimator according to formula (7), in which the auxiliary variable is the number of votes cast on the same political party during the parliamentary election to the Sejm 2007 (in the case of Ruch Palikota committee, the auxiliary variable was the result of Platforma Obywatelska RP committee),

$\hat{Y}_q^{(S07)}$ - the ratio estimator according to formula (6), in which the auxiliary variable is the number of valid votes in total during parliamentary election to the Sejm 2007.

$$\hat{P}_J^{(Poiss-S11)} = \frac{\hat{Y}_{J,P}^{(S11)}}{\hat{Y}_P^{(S11)}} \tag{15}$$

where:

$\hat{Y}_{J,P}^{(S11)}$ - the estimator according to formula (10) based on the model (8) with the explanatory variables:

teren – the type of the area where the precinct is based (*large city* – above 80 thousand registered voters, *town, village*),

region – the group of voivodeships (*first group:* Małopolskie, Podkarpackie, Świętokrzyskie, Lubelskie, Łódzkie, Mazowieckie, Podlaskie; *second group:* the remaining voivodeships),

podm_gm, uzyt_gm.

$\hat{Y}_P^{(S11)}$ - the estimator according to formula (9) based on the model (8) with the explanatory variables:

podm_gm, uzyt_gm, bezr_pow, wyn_pow, terrain.

$$\hat{P}_J^{(Poiss-P10)} = \frac{\hat{Y}_{J,P}^{(P10)}}{\hat{Y}_P^{(P10)}} \tag{16}$$

where:

$\hat{Y}_{J,P}^{(P10)}$ - the estimator according to formula (10) based on the model (8) with explanatory variables:

P10_Komorowski – the number of votes cast on Bronisław Komorowski during the presidential election 2010,

P10_KaczynskiJ – the number of votes cast on Jarosław Kaczyński during the presidential election 2010,

P10_Napieralski – the number of votes cast on Grzegorz Napieralski during the presidential election 2010,

podm_gm, uzyt_gm, terrain, region.

$$\hat{P}_J^{(Poiss-S07)} = \frac{\hat{Y}_{J,P}^{(S07)}}{\hat{Y}_P^{(S07)}} \tag{17}$$

where:

$\hat{Y}_P^{(P10)}$ – the estimator according to formula (10), based on model (8) with the explanatory variables:

S07_PO – the number of votes cast on Platforma Obywatelska RP committee during the parliamentary to the Sejm 2007,

S07_PiS – the number of votes cast on Prawo i Sprawiedliwość (PiS) (Law and Justice) during the parliamentary election to the Sejm 2007,

S07_LiD – the number of votes cast on Lewica i Demokraci (LiD) (Left and Democrats) during the parliamentary election to the Sejm 2007,

terrain, region, podm_gm, uzyt_gm.

$\hat{Y}_P^{(S07)}$ - the estimator according to formula (9) based on model (8) with explanatory variables:

S07_votes – the number of valid votes in total during the parliamentary election to the Sejm 2007,

podm_gm, uzyt_gm, bezr_pow, wyn_pow, S07_PO, terrain.

6. Description of simulation

In the conducted simulation test the single-stage cluster sampling was applied instead of two-stage sampling typical for exit poll, which results from the character of available data. No sampling at the second stage was simulated as no unit information being capable to support estimation process was available, therefore, the result in the sampled precinct was taken as given without errors. The sampled units are precincts, i.e. the groups of voters participating in voting.

For reference purposes along with the balanced sampling, the simple random sampling without replacement (SRS) and stratified sampling were tested (STRAT). The division into strata was made based on the variation of the past election results in the section of following variables: *teren* and *region.* 6 strata were created as combination of 3 variants of a *teren* variable and 2 variants of a *region* variable. With regard to the large disproportion between the number of precincts and the number of votes cast in a stratum, the location was chosen proportionally to the number of valid votes cast in the parliamentary election to the Sejm 2007.

The balanced sampling was conducted in 3 variants depending on the type of auxiliary variables that were used:

- balance in reference to GUS variables (podm_gm, uzyt-gm, bezr_pow, wyn_pow) and the number of registered voters during S11 (BALS11),
- balance in reference to GUS variables (podm_gm, uzyt_gm, bezr_pow, wyn_pow) and the variables from the presidential election 2010 (P10_votes, P10_Komorowski, P10_KaczynskiJ, P10_Napieralski) (BALP10),

- balance in reference to GUS variables (podm_gm, uzyt_gm, bezr_pow, wyn_pow) and variables from the parliamentary election to the Sejm 2007. (S07_votes, S07_PO, S07_PiS, S07_*LiD)* (BALS07).

Additionally, each balanced sample was at the same time a stratified sample according to the above-described scheme. Within strata the precincts were sampled with the same probability of being selected, however, between strata the probabilities differed due to the allocation being disproportionate to the number of precincts in a stratum. In stratified sampling the estimators using auxiliary variables had a form of combined estimators, which means that the model is estimated for the whole sample altogether and not separately for each stratum like in the case of separate estimators. Due to small sizes of a sample in strata separate estimators would be in this case less stable.

The use of the past election results, due to the incomplete link of precincts between elections, implies the restriction of frame population to the set U2 or U3 (see Tab. 1). Even in the case of using only GUS variables, or if auxiliary variables are completely excluded, the frame population is restricted to regular precincts (set U1), which reflects the practical way of conducting the research. Nevertheless the aim of the survey is to estimate the actual fraction of the whole population (U). Thus, it seems appropriate to validate the estimates against non-included units. The correction is not necessary in the case of ratio estimators, which use the sum of additional features for the whole population, thus the estimates can be generalized to the entire population U. The estimates obtained by using estimators based on the log-linear model can be generalized only to the particular frame population (U1, U2 or U3). The same applies to the Horvitz-Thompson estimator, due to the restriction of frame population to the regular precincts. Therefore, the part of estimators was extended with the correction based on the past election results of the entire population in relation to the result of the particular frame population. General formula of the correction is as follows:

$$\widehat{\mathrm{P}}_{\mathrm{J}}^{*} = \widehat{\mathrm{P}}_{\mathrm{J}} \frac{\mathrm{P}_{\mathrm{J}\prime}^{(\mathrm{W})}}{\mathrm{P}_{\mathrm{J}\prime}^{(\mathrm{W,Ui})}} \tag{18}$$

where:

$P_{J\prime}^{(W)}$ – the actual fraction of votes cast on committee/candidate linked to the committee *J* in the election W∈{S07, P10}, in population U,

$P_{J\prime}^{(W,Ui)}$ – the actual fraction of votes cast on committee/candidate linked to the committee *J* in election W, in population Ui (Ui ∈ {U1, U2, U3}).

The correction applied only to irregular precincts (Ui=U1) is based on the assumption that the voters abroad, in prisons, on vessels, etc. are different from the rest of voters and the directions of those differences remain constant over at least a few years. The analysis of the past election results indicates the presence of some constant trends, i.a. the result of PO in irregular precincts was usually higher than the result in regular precincts (pkw.gov.pl). These trends, however, do

not have to sustain in the future, thus in case of Horwitz-Thompson estimators it was decided to test both estimators with correction or without it.

Juxtaposition of all sampling plans and methods of estimation, taking into account the fact that not all combinations are possible, gives 30 possible strategies of research. Some of the strategies use the same auxiliary variables, both at the stage of selecting the balanced sample and at the stage of estimation. Such a solution is not inconsistent due to the fact that the sample is almost never exactly balanced, thus using the complex estimators in the sample approximately balanced can bring additional benefits (Tille 2011, p. 223).

Due to the fact that in the majority of elections, three first parties usually get the vast majority of votes and estimating their results is of primary importance, the number of estimated parameters was limited to the results of three committees with the highest results. Besides, estimating separately very low fractions would artificially lower the mean absolute estimation error.

The sample size in each analysed scheme was set at the level of n=100 precincts, which (taking account of all voters in the sampled precinct) corresponds to the 50-70 thousand of elementary units. Every strategy was simulated M=1000 times. The effectiveness of a strategy was measured in two ways: separately for each of three committees and altogether. In the first case the Empirical Root Mean Squared Error (ERMSE), was used:

$$\mathrm{ERMSE}_{\mathrm{J}} = \sqrt{\frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}}\left(\widehat{\mathrm{P}}_{\mathrm{J,i}} - \mathrm{P}_{\mathrm{J}}\right)^{2}} \cdot 100 \tag{19}$$

where:

$\widehat{\mathrm{P}}_{\mathrm{J,i}}$ – the estimates of fraction of votes cast on *J* committee in i^{th} iteration.

In the second case, the estimates for the three main committees altogether were taken into account and for every iteration the Average Manhattan Distance (AMD) was calculated and subsequently the Mean AMD was computed (MAMD):

$$\mathrm{MAMD} = \frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}}\mathrm{AMD}_{\mathrm{i}} \cdot 100 \tag{20}$$

where:

$$\mathrm{AMD}_{\mathrm{i}} = \frac{1}{3}\sum_{\mathrm{J}=1}^{3}\left|\widehat{\mathrm{P}}_{\mathrm{J,i}} - \mathrm{P}_{\mathrm{J}}\right| \tag{21}$$

Both measures were multiplied by 100, thus the obtained values can be interpreted in categories of percentage points. The simulation analysis was conducted in the R environment.

7. Simulation results

In Table 2 the ERMSE for all strategies for the three subsequent committees with the highest final result are presented. The estimators marked with asterisk (*) were corrected according to the formula (18). It turned out that the best strategy in the case of all three committees was the strategy {BALP10, Q_P1t0}, in which the stratified, balanced against the chosen official statistics at the level of

municipality and powiat, and against the results of presidential election 2010 sample is drawn. This sample also uses the ratio estimator in which the auxiliary variable is the result of the candidate associated with a given party, also during the 2010 election. Distribution of the effectiveness of other strategies is similar in the case of PO and PiS, whereas it differs slightly in the case of Ruch Palikota. Nevertheless, the best sampling design in all cases, irrespective of the method of estimation, turned out to be BALP10.

Table 2. ERMSE for fraction of votes cast on three winning parties

Platforma Obywatelska (Civil Platform)		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	1.458	1.112	0.980	x	x
	HT*	1.468	1.130	1.000	0.601	0.637
	Q-S11	2.672	1.904	1.071	x	x
	Q-P10	0.589	0.624	0.560	0.557	x
	Q-S07	0.615	0.593	0.576	x	0.586
	Poiss-S11*	1.201	1.054	1.024	x	x
	Poiss-P10*	1.030	0.898	0.759	0.679	x
	Poiss-S07*	1.128	0.884	0.807	x	0.757
Prawo i Sprawiedliwość (Law and Justice)		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	1.152	1.020	0.963	x	x
	HT*	1.151	1.020	0.966	0.454	0.518
	Q-S11	1.604	1.721	1.000	x	x
	Q-P10	0.444	0.460	0.452	0.416	x
	Q-S07	0.447	0.504	0.478	x	0.468
	Poiss-S11*	0.921	0.943	0.947	x	x
	Poiss-P10*	0.676	0.686	0.629	0.560	x
	Poiss-S07*	0.722	0.740	0.627	x	0.590
Ruch Palikota (Palikot's Movement)		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	0.359	0.332	0.322	x	x
	HT*	0.355	0.323	0.314	0.267	0.323
	Q-S11	0.625	0.499	0.319	x	x
	Q-P10	0.367	0.348	0.354	0.265	x
	Q-S07	0.453	0.405	0.414	x	0.323
	Poiss-S11*	0.335	0.317	0.312	x	x
	Poiss-P10*	0.296	0.288	0.287	0.268	x
	Poiss-S07*	0.385	0.355	0.352	x	0.331

Source: Own calculation.

In the case of two main parties, the ratio estimators have proved greater efficiency as they correct the direct estimation only against the past election

results of the same party or the candidate associated with the party, which presumably results from the fact that the electorates of those parties remain almost unchanged. In the case of Ruch Palikota, which was the new party on the political scene, supporting the estimation only with the past PO or Bronisław Komorowski results did not work out; generally, the simple estimators or estimators based on log-linear model (the exception is the best strategy, which like in the case of two other parties was {BALP10, Q-P10}) would be a better choice. The mean square error is an absolute measure, thus the differences in values between three committees result mainly from the differences between the values of estimated parameters.

The simulation results with respect to the second criteria of evaluation of strategies are presented in Table 3. The table includes the mean of average absolute differences (for three first parties) between the actual result and estimations in each iteration. In the case of this criteria, the strategy {BALP10, Q-P10} again turned out to be the most effective. Taking into consideration only the sampling design, irrespective of the estimator, the best solution turned out to be BALP10 – the design using the information from the presidential election 2010. BALS07, i.e. the design using the information from the previous election 2007, turned out to be slightly worse. BALS11, which balanced the sample only on data referring to municipalities and powiats, showed similar effectiveness to BALS07 in case of complex estimators, however, in case of simple estimators the effectiveness was worse. The plan using the relatively little additional information, i.e. the stratified sampling performed poorly in terms of drawing the most representative sample. The simple random sampling turned out to be the least effective.

Table 3. MAMD for the three committees with the highest results

		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	0.794	0.653	0.600	x	x
	HT*	0.798	0.656	0.603	0.348	0.393
	Q-S11	1.301	1.088	0.632	x	x
	Q-P10	0.369	0.381	0.362	0.326	x
	Q-S07	0.401	0.400	0.389	x	0.365
	Poiss-S11*	0.652	0.612	0.604	x	x
	Poiss-P10*	0.516	0.498	0.445	0.401	x
	Poiss-S07*	0.582	0.522	0.478	x	0.445

Source: Own calculation.

The most favourable assessment of the effectiveness of the methods of estimations for the ratio estimators is, as in the first criteria, using the presidential election 2010 results (Q-P10) as the auxiliary variables. The same estimators using the parliamentary election to the Sejm 2007 results (Q-S07) as the auxiliary variables performed slightly worse. Subsequently, the estimators based on the

log-linear model, which irrespective of the higher number of auxiliary variables used do not surpass the ratio estimators in terms of effectiveness, are ranked. This results from the strong correlation between the estimated parameter and the same parameter in the previous election and not so strong link to the other auxiliary variables, and also from the relatively small sample size which leads to the less stable estimations of the model with many variables. The least effective estimator turned out to be Q-S11 estimator, in which the sum of votes in total was estimated by the ratio estimator and the sum of votes cast on *J* committee was estimated by the HT estimator.

The correction in Horwitz-Thompson estimator (HT*), calculated with respect to exclusion of irregular precincts from the frame population in plans SRS, STRAT and BALS11, led to minimal change in the estimate. The result of this correction for estimating the particular parties results (Table 2) is not unequivocal, however, as far as MAMD is concerned it is negative for every sampling plan. This confirms the above-mentioned assumptions that the electorate in irregular precincts can differ from the voters across country, however, the directions of those differences do change in time, thus they do not qualify for the correction of estimates from regular precincts. Consequently, the restriction of frame population to regular precincts should not systematically bias the results of research.

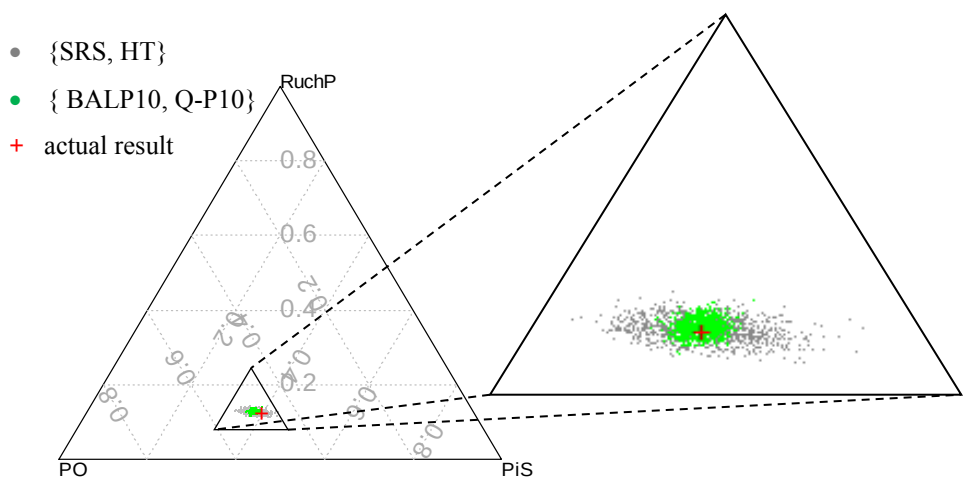


Figure 2. Ternary plot of the simulation results for strategies {SRS, HT} and {BALP10, QP-10}

Source: Own calculation.

To illustrate the difference between the results of the best strategy using auxiliary variables and the results of the classical strategy, i.e. without any auxiliary variables, the ternary plot was created, which is presented in the

Figure 2. The ternary plot is a type of scatter diagram for three variables adding up to a constant. In order to be able to present the estimates for three main parties in this way, the estimates were transformed. Thanks to the transformation, the sum of the estimates equalled 1, that is, as if only these three parties took part in the election. Each point represents the result of one out of M simulations. The location of the point indicates the distribution of votes over three parties – the closer to the vertex of a triangle, the larger part of votes is distributed to the committee described on the particular vertex. Smaller scatter of points for strategy {BALP10, Q-P10} compared to the strategy {SRS, HT} is the reflection of higher effectiveness of the first one.

7. Conclusions

The subject of this paper was the evaluation of the usefulness of available additional data to strengthen the process of estimating the distribution of votes cast during the election in exit poll survey. The additional data were taken from two sources: the Central Statistical Office and the National Electoral Commission. A priori information was included in the strategy of survey both at the stage of selecting a sample and at the stage of estimating parameters. The proposed strategies were tested on the detailed results of the parliamentary election to the Sejm 2011. The results of the conducted simulation indicate that drawing a sample balanced against the selected auxiliary variables as well as the use of those variables in the estimation process significantly improves the effectiveness of the survey. This conclusion was not obvious in the beginning, as the auxiliary features did not refer to the units of research directly; data from GUS refer to the higher aggregation level and data from PKW are not linked to the current research and somehow force the restriction of frame population. Out of two past elections tested as a reference point for the correction of current estimates, the chronologically nearest presidential election 2010 turned out to be best.

Acknowledgement

The research was supported by the grant number 538-2320-0842-12 from the Faculty of Management of University of Gdańsk. I am grateful to professor Mirosław Szreder for his valuable comments.

REFERENCES

BARRETO, M. A., GUERRA, F., MARKS, M., NUÑO, S. A., WOODS, N. D. (2006). Controversies in exit poll, Political Science and Politics, Vol. 39, No. 3, pp. 477–483.

BAUTISTA, R., CALLEGARO, M., VERA, J. A., ABUNDIS, F., (2006). Nonresponse in Exit Poll Methodology: A Case Study in Mexico, Paper presented at the annual meeting of the American Association For Public Opinion Association, Fontainebleau Resort, Miami Beach, FL. Available at: <http://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000612.pdf> [Accessed on 12 January 2014].

BISHOP, G. F., FISHER, B. S., (1995). 'Secret ballots' and self-reports in an exit-poll experiment, Public Opinion Quarterly, Vol. 59, No. 4, pp. 568–588.

BUSCH, R. J., LIESKE, J. A. (1985). Does time of voting affect exit poll results?, Public Opinion Quarterly, Vol. 49, No. 1, pp. 94–104.

DEVILLE, J.-C., TILLÉ, Y. (2004). Efficient Balanced Sampling: The Cube Method, Biometrika, Vol. 91, No. 4, pp. 893–912.

DOMAŃSKI, H., MARKOWSKI, R., SAWIŃSKI, Z., SZTABIŃSKI, P. B., (2010). Assessment of methodology and the results of research conducted before the first and second round of the presidential elections in 2010, Warsaw: OFBOR.

Election Code, Act of 5 January 2011, Journal of Laws 2011 No. 21, item 112, Article 12.

HILMER, R., (2008). Exit polls in Germany, Berlin: 3MC Conference Proceedings.

HOFRICHTER, J., (1999). Exit polls and elections campaigns. In B.I. Newman, ed. Handbook of political marketing, Thousand Oaks: Sage Publications.

KLORMAN, R., (1976). What Time Do People Vote?, Public Opinion Quarterly, Vol. 40, No. 2, pp. 182–193.

KOZŁOWSKI, A., (2012). The usefulness of past data in sampling design for exit poll surveys, Economic Studies, University of Economics in Katowice, Vol. 120, pp. 45–57.

LANGEL, M., TILLÉ, Y., (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory, METRON – International Journal of Statistics, Vol. LXIX, No. 1, pp. 45–65.

LENSKI, J., (2008). New methodological Issues in conducting exit polls, Berlin: 3MC Conference Proceedings.

LEVY, M. R., (1983). The methodology and performance of election day polls, Public Opinion Quarterly, Vol. 47, No. 1, pp. 54–67.

MERKLE, D. M., EDELMAN, M., (2002). Nonresponse in exit polls: a comprehensive analysis, in: Groves, R. M., Dillman, D. A., Eltinge, J. L., Little, R. J. A. (Eds.), Survey Nonresponse, New York: John Wiley & Sons.

MOON, N., (2008). Predicting the Election Result from an Exit Poll – the UK example, Berlin: 3MC Conference Proceedings.

MOORE, D. W., (2003). New Exit Poll Consortium Vindication for Exit Poll Inventor, Inside the polls, Gallup. Available at: <http://www.gallup.com/poll/9472/new-exit-poll-consortium-vindication-exit-poll-inventor.aspx> [Accessed on 24 January 2013].

Ośrodek Badania Opinii Publicznej, archival site, <www.obop.com.pl> [Accessed on 24 January 2013].

Państwowa Komisja Wyborcza, <pkw.gov.pl> [Accessed on 24 January 2013]

SCHEUREN, F., ALVEY, W., (2008). Elections and Exit Polling, New Jersey: John Wiley & Sons.

SZREDER, M., (2010). Methods and techniques of opinion polls surveys, Warsaw: PWE.

SZREDER, M., (2011). Emotions and the truth of election night, Rzeczpospolita, 28.09.2011, No. 227.

The Election code, Act of 5 January 2011, Journal of Laws 2011, No. 21, item 112, Article 12.

TILLÉ, Y., (2006). Sampling Algorithms, New York: Springer.

TILLÉ, Y., (2011). Ten years of balanced sampling with the cube method: An appraisal, Survey Methodology, Vol. 37, No. 2, pp. 215–226.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 59–66

AN IMPROVED ESTIMATOR FOR POPULATION MEAN USING AUXILIARY INFORMATION IN STRATIFIED RANDOM SAMPLING

Sachin Malik, Viplav K. Singh, Rajesh Singh[1]

ABSTRACT

In the present study we propose a new estimator for population mean $\overline{Y}$ of the study variable y in the case of stratified random sampling using the information based on auxiliary variable x. An expression for the mean squared error (MSE) of the proposed estimator is derived up to the first order of approximation. The theoretical conditions have also been verified by a numerical example. An empirical study demonstrates the efficiency of the suggested estimator over sample mean estimator, usual separate ratio, separate product estimator and other proposed estimators.

Key words: study variable, auxiliary variable, stratified random sampling, separate ratio estimator, bias and mean squared error.

1. Introduction

The problem of estimating the population mean in the presence of an auxiliary variable has been widely discussed in the finite population sampling literature. Many ratio, product and regression methods of estimation are good examples in this context. Diana (1993) suggested a class of estimators of the population mean using one auxiliary variable in the stratified random sampling and examined the MSE of the estimators up to the k^{th} order of approximation. Kadilar and Cingi (2003), Singh et al. (2007), Singh and Vishwakarma (2008) as well as Koyuncu and Kadilar (2009) proposed estimators in stratified random sampling. Bahl and Tuteja (1991) and Singh et al. (2007) suggested some exponential ratio type estimators.

[1] Department of Statistics, Banaras Hindu University, Varanasi-221005, India. E-mails: sachinkurava999@gmail.com, viplavkumarsingh0802@gmail.com, rsinghstat@gmail.com. Corresponding author.

Consider a finite population of size N is divided into L strata such that $\sum_{h=1}^{L} N_h = N$, where N_h is the size of h^{th} stratum (h=1,2,...,L). We select a sample of size n_h from each stratum by simple random sampling without replacement (SRSWOR), such that $\sum_{h=1}^{L} n_h = n$, where n_h is the stratum sample size. Let (y_{hi}, x_{hi}, z_{hi}) denote the observed values of y, x, and z on the i^{th} unit of the h^{th} stratum, where i=1, 2, 3…N_h.

We use the following notations:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \,,\ \bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} \bar{y}_{hi} \,,\ \bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{n_h} \bar{Y}_{hi},$$

$$Y = \bar{Y}_{st} = \sum_{h=1}^{L} W_h \bar{Y}_h, \qquad w_h = \frac{N_h}{N}.$$

Let

$$S_{yh}^2 = \sum_{i=1}^{N_h} \frac{\left(\bar{y}_h - \bar{Y}_h\right)^2}{N_h - 1}, \quad S_{xh}^2 = \sum_{i=1}^{N_h} \frac{\left(\bar{x}_h - \bar{X}_h\right)^2}{N_h - 1}$$

$$S_{yxh} = \sum_{i=1}^{N_h} \frac{\left(\bar{x}_h - \bar{X}_h\right)\left(\bar{y}_h - \bar{Y}_h\right)}{N_h - 1} \text{ and } f_h = \frac{1}{n_h} - \frac{1}{N_h}$$

2. Established estimators

When the population mean $\bar{X}_h$ of the stratum h of the auxiliary variable x is known then the usual separate ratio and product estimators for the population mean $\bar{Y}$ are respectively given as

$$t_1 = \sum_{h=1}^{L} W_h \bar{y}_h \frac{\bar{X}_h}{\bar{x}_h} \tag{2.1}$$

$$t_2 = \sum_{h=1}^{L} W_h \bar{y}_h \frac{\bar{x}_h}{\bar{X}_h} \tag{2.2}$$

Following Bahl and Tuteja (1991), we propose the following ratio and product exponential estimators

$$t_3 = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\bar{X}_h - \bar{x}_h}{\bar{X}_h + \bar{x}_h}\right) \tag{2.3}$$

$$t_4 = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\bar{x}_h - \bar{X}_h}{\bar{x}_h + \bar{X}_h}\right) \tag{2.4}$$

The MSEs of these estimators are respectively given by

$$\mathrm{MSE}(t_1) = \sum_{h=1}^{L} W_h^2 f_h \left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yxh}\right] \tag{2.5}$$

$$\mathrm{MSE}(t_2) = \sum_{h=1}^{L} W_h^2 f_h \left[S_{yh}^2 + R_h^2 S_{xh}^2 + 2R_h S_{yxh}\right] \tag{2.6}$$

$$\mathrm{MSE}(t_3) = \sum_{h=1}^{L} W_h^2 f_h \left[S_{yh}^2 + \frac{R_h^2}{4} S_{xh}^2 - R_h S_{yxh}\right] \tag{2.7}$$

$$\mathrm{MSE}(t_4) = \sum_{h=1}^{L} W_h^2 f_h \left[S_{yh}^2 + \frac{R_h^2}{4} S_{xh}^2 + R_h S_{yxh}\right] \tag{2.8}$$

The usual regression estimator of the population mean $\bar{Y}$ is

$$t_{lr} = \sum_{h=1}^{L} w_h \left[\bar{y}_h + b_h\left(\bar{X}_h - \bar{x}_h\right)\right] \tag{2.9}$$

The MSE of the regression estimator is given by

$$\mathrm{var}(t_{lr}) = \sum_{h=1}^{L} W_h^2 f_h S_{yh}^2 \left(1 - \rho_h^2\right) \tag{2.10}$$

The variance of the usual sample mean estimator $\bar{y}_h$ is given as

$$\mathrm{var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 f_h S_{yh}^2 \tag{2.11}$$

Yadav et al. (2011) proposed an exponential ratio-type estimator for estimating $\bar{Y}$ as

$$t_R = \sum_{h=1}^{L} w_h \bar{y}_h \exp\left(\frac{\bar{X}_h - \bar{x}_h}{\bar{X}_h + (a_h - 1)\bar{x}_h}\right) \tag{2.12}$$

The MSE of the estimator t_R is given by

$$\mathrm{MSE}(t_R) = \sum_{h=1}^{L} W_h^2 f_h \left[S_{yh}^2 + \frac{R_h^2}{a_h^2} S_{xh}^2 - 2\frac{R_h}{a_h} S_{yxh}\right] \tag{2.13}$$

At the optimum value of a_h the MSE of the estimator t_R is equal to the MSE of the regression estimator t_{lr} given in equation (2.9).

3. The proposed estimator

Motivated by Singh and Solanki (2012), we propose an estimator of population mean $\overline{Y}$ of the study variable y as

$$t_P = \sum_{h=1}^{L} w_h \left[\lambda_1 \bar{y}_h + \lambda_2 \left(\overline{X}_h - \bar{x}_h\right)\right]\left\{2 - \left(\frac{\overline{X}_h}{\bar{x}_h}\right)\exp\left(\frac{\overline{X}_h - \bar{x}_h}{\overline{X}_h + \bar{x}_h}\right)\right\} \tag{3.1}$$

To obtain the bias and MSE of t_P, we write

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h = \overline{Y}(1+e_0), \quad \bar{x}_{st} = \sum_{h=1}^{L} w_h \bar{x}_h = \overline{X}(1+e_1)$$

such that

$$E(e_{0h}) = E(e_{1h}) = 0,$$

and $$E(e_0^2) = \frac{\sum_{h=1}^{L} W_h^2 f_h S_{yh}^2}{\overline{Y}^2}, \; E(e_1^2) = \frac{\sum_{h=1}^{L} W_h^2 f_h S_{xh}^2}{\overline{X}^2}, \; E(e_0 e_1) = \frac{\sum_{h=1}^{L} W_h^2 f_h S_{yxh}^2}{\overline{Y}\overline{X}}.$$

Expressing equation (3.1) in terms of es, we have

$$\begin{aligned} t_P &= \sum_{h=1}^{L} w_h \left\{\left[\lambda_1 \overline{Y}_h (1+e_0) - \lambda_2 \overline{X}_h\right]\left[2 - (1+e_1)^{-1}\exp\left(-\frac{e_1}{2} + \frac{e_1^2}{4}\right)\right]\right\} \\ &= \sum_{h=1}^{L} w_h \left\{\left[\lambda_1 \overline{Y}_h (1+e_0) - \lambda_2 \overline{X}_h\right]\left[1 + \frac{3e_1}{2} - \frac{15}{8}e_1^2\right]\right\} \end{aligned} \tag{3.2}$$

By neglecting the terms of e's power greater than two in expression (3.2), we obtain

$$\begin{aligned} t_P - \overline{Y} = \sum_{h=1}^{L} w_h \Bigg[& \lambda_1 \overline{Y}_h (1+e_0) - \lambda_2 \overline{X}_h e_1 + \frac{3}{2}\lambda_1 \overline{Y}_h e_1 + \frac{3}{2}\lambda_1 \overline{Y}_h e_0 e_1 \\ & - \frac{3}{2}\lambda_2 \overline{X}_h e_1^2 - \frac{15}{8}\lambda_1 \overline{Y}_h e_1^2 - \overline{Y}_h \Bigg] \end{aligned} \tag{3.3}$$

Taking expectations on both sides of (3.3), we have the bias of the estimator t_P up to the first order of approximation as

$$B(t_P) = \sum_{h=1}^{L} w_h \left\{\overline{Y}_h(\lambda_1 - 1) + \frac{3}{2}\lambda_1 f_h \frac{S_{yxh}}{\overline{X}_h} - \frac{3}{2}\lambda_2 f_h \frac{S_{xh}^2}{\overline{X}_h} - \frac{15}{8}\lambda_1 \overline{R}_h \frac{S_{xh}^2}{\overline{X}_h}\right\} \tag{3.4}$$

Squaring both sides of (3.3) and neglecting the terms with power greater than two, we have

$$\left(t_P - \overline{Y}\right)^2 = \sum_{h=1}^{L} w_h{}^2 \left[\lambda_1 \overline{Y}_h \left(1 + e_0\right) - \lambda_2 \overline{X}_h e_1 + \frac{3}{2}\lambda_1 \overline{Y}_h e_1 - \overline{Y}_h \right]^2$$

$$\left(t_P - \overline{Y}\right)^2 = \sum_{h=1}^{L} w_h{}^2 \left[\lambda_1^2 \left(\overline{Y}_h^2 e_0 + \overline{Y}_h^2 + \frac{9}{4}\overline{Y}_h^2 e_1^2 + 3\overline{Y}_h^2 e_0 e_1 \right) + \lambda_2^2 \overline{X}_h^2 \right.$$
$$\left. + \overline{Y}_h^2 - 2\lambda_1 \overline{Y}_h^2 - 2\lambda_1\lambda_2 \overline{Y}_h \overline{X}_h e_0 e - 3\lambda_1\lambda_2 \overline{Y}_h \overline{X}_h e_1^2 \right] \tag{3.5}$$

Taking expectations of both sides of (3.5), we have the mean squared error of the estimator t_P up to the first order of approximation as

$$\mathrm{MSE}\left(t_P\right) = \lambda_1^2 P_1 + \lambda_2^2 P_2 - 2\lambda_1\lambda_2 P_3 - 3\lambda_1\lambda_2 P_4 - 2\lambda_1 \sum_{h=1}^{L} w_h^2 \overline{Y}_h^2 + \sum_{h=1}^{L} w_h^2 \overline{Y}_h^2 \tag{3.6}$$

where,

$$\left.\begin{aligned} P_1 &= \sum_{h=1}^{L} W_h^2 f_h S_{yh}^2 + \frac{9}{4}\sum_{h=1}^{L} W_h^2 f_h R_h^2 S_{xh}^2 + \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 + 3\sum_{h=1}^{L} W_h^2 f_h R_h S_{yxh} \\ P_2 &= \sum_{h=1}^{L} W_h^2 f_h S_{xh}^2 \\ P_3 &= \sum_{h=1}^{L} W_h^2 f_h S_{yxh} \\ P_4 &= \sum_{h=1}^{L} W_h^2 f_h R_h S_{xh}^2 \end{aligned}\right\} \tag{3.7}$$

Partially differentiating expression (3.6) with respect to λ_1 and λ_2, we get the optimum values of λ_1 and λ_2 as

$$\lambda_1(\mathrm{opt}) = \frac{4P_2 \sum_{h=1}^{L} w_h^2 \overline{Y}_h^2}{4P_1P_2 - \left[2P_3 + 3P_4\right]^2} \quad \text{and} \quad \lambda_2(\mathrm{opt}) = \frac{2\left[2P_3 + 3P_4\right]\sum_{h=1}^{L} w_h^2 \overline{Y}_h^2}{4P_1P_2 - \left[2P_3 + 3P_4\right]^2}$$

Substituting these values of λ_1 and λ_2 in expression (3.7), we get the minimum value of the MSE(t_P).

4. Numerical study

For numerical study we use the data set used earlier by Kadilar and Cingi (2003). 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 in Table 4.1. The indices 1,2,…,6 indicate the strata.

Table 4.1. Population data

N=854	n=140				
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
$\overline{X}_1 = 24375$	$\overline{X}_2 = 27421$	$\overline{X}_3 = 72409$	$\overline{X}_4 = 74365$	$\overline{X}_5 = 26441$	$\overline{X}_6 = 9844$
$\overline{Y}_1 = i536$	$\overline{Y}_2 = 2212$	$\overline{Y}_3 = 9384$	$\overline{Y}_4 = 5588$	$\overline{Y}_5 = 967$	$\overline{Y}_6 = 404$
$\beta_{x1} = 25.71$	$\beta_{x2} = 34.57$	$\beta_{x3} = 26.14$	$\beta_{x4} = 97.60$	$\beta_{x5} = 27.47$	$\beta_{x6} = 28.10$
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
$\rho_1 = 0.82$	$\rho_2 = 0.86$	$\rho_3 = 0.90$	$\rho_4 = 0.99$	$\rho_5 = 0.71$	$\rho_6 = 0.89$
$f_1 = 0.102$	$f_2 = 0.049$	$f_3 = 0.016$	$f_4 = 0.009$	$f_5 = 0.138$	$f_6 = 0.006$
$w_1^2 = 0.015$	$w_2^2 = 0.015$	$w_3^2 = 0.012$	$w_4^2 = 0.04$	$w_5^2 = 0.057$	$w_6^2 = 0.041$

To compare the efficiency 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:

$$PRE\left(t,\bar{y}_{st}\right)=\frac{MSE\left(\bar{y}_{st}\right)}{MSE(t)}*100\text{, where } t=(t_1,t_2,t_3,t_{lr},t_P)$$

The findings are given in the Table 4.2.

Table 4.2. Percent relative efficiencies (PREs) of estimators

S. No.	ESTIMATORS	PREs
1	$\bar{y}_{st}$	100
2	t_1	423.20
3	t_2	37.60
2	t_3	199.14
3	t_4	12.83
4	t_{lr}	629.03
5	t_R	629.03
6	t_P	789.87

5. Conclusion

In this paper we have proposed a new estimator of the population mean of the study variable using auxiliary variables. Expressions for bias and MSE of the estimator are derived up to first order of approximation. The proposed estimator is compared with the usual mean estimator and other considered estimators. A numerical study is carried out to support the theoretical results. From Table 4.2. it is clear that the proposed estimator t_P is more efficient than the unbiased sample mean estimator $\bar{y}_{st}$, the usual ratio and product estimators t_1 and t_2, the usual exponential ratio and product type estimators t_3 and t_4, and Yadav et al. (2011) estimator t_R.

Acknowledgement

The authors are grateful to the learned referees for their valuable comments and suggestions regarding the improvement of the paper.

REFERENCES

BAHL, S., TUTEJA, R. K., (1991). Ratio and product type exponential estimator. Infrm. and Optim. Sci., XII, I, 159–163.

DIANA, G., (1993). A class of estimators of the population mean in stratified random sampling. Statistica 53 (1): 59–66.

KADILAR,C., CINGI,H., (2003). Ratio Estimators in Stratified Random Sampling. Biometrical Journal 45 (2003) 2, 218–225.

KOYUNCU, N., KADILAR, C., (2009). Family of estimators of population mean using two auxiliary variables in stratified random sampling. Comm. In Stat. – Theory and Meth., 38: 14, 2398–2417.

SOLANKI, R. S., SINGH, H. P., RATHOUR, A., (2012). An alternative estimator for estimating the finite population mean using auxiliary information in sample surveys. ISRN Prob. and Stat. doi: 10.5402/2012/657682.

SINGH, H., P., VISHWAKARMA, G. K., (2008). A family of estimators of population mean using auxiliary information in stratified sampling. Communication in Statistics Theory and Methods, 37 (7), 1038–1050.

SINGH, R., CHAUHAN, P., SAWAN, N., SMARANDACHE, F., (2007). Auxiliary information and a priori values in construction of improved estimators. Renaissance High Press. Zip publishing Columbas, Ohio, USA.

YADAV R., UPADHYAYA L. N., SINGH H. P., CHATTERJEE S., (2011). Improved separate exponential estimator for population mean using auxiliary information. Statistics in Transition new series, 12 (2), pp. 401–412.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 67–82

A MODIFIED MIXED RANDOMIZED RESPONSE MODEL

Housila P. Singh, Tanveer A. Tarray[1]

ABSTRACT

Socio-economic investigations often relate to certain personal features that people wish to hide from others in comprehensive inquiries, detailed questionnaires include numerous items. Data on most of them are frequently easy to procure merely by asking, but a few others can be on sensitive issues for which people are not inclined to state honest responses. For example, most people prefer to conceal the truth regarding their savings, the extent of their accumulated wealth, their history of intentional tax evasion and other illegal and or unethical practices leading to earnings from clandestine sources, crimes, trade in contraband goods, susceptibility to intoxication, expenditures on addictions of various forms, homosexuality, and similar issues which are customarily disapproved of by society. Open or direct queries often fail to yield reliable data on such confidential aspects of human life. Warner (1965) developed an alternative survey technique that is known as randomized response (RR) technique. Greenberg et al. (1971) presented a revised version of Warner's (1965) technique for qualitative variables. Later various modifications were given by several researchers [see Chaudhuri (2011)]. Kim and Warde (2005) and Nazuk and Shabir (2010) presented mixed randomized response models using simple random sampling with replacement sampling scheme which improves the privacy of respondents. In this paper we have suggested a modified mixed randomized response model to estimate the proportion of a qualitative sensitive variable. Properties of the proposed randomized response model have been studied along with recommendations. It has been shown that the suggested randomized response model is always better than Kim and Warde's (2005) model while it is better than Nazuk and Shabbir's (2010) model under some realistic conditions. Numerical illustrations and graphs are also given in support of the present study.

Key words: randomized response technique, simple random sampling, dichotomous population, estimation of proportion, privacy of respondents, sensitive characteristics.

[1] School of Studies in Statistics, Vikram University, Ujjain – 456010 – India.
E-mail: tanveerstat@gmail.com.

1. Introduction

In situations where potentially embarrassing or incriminating responses are sought, the randomized response (RR) technique is effective in reducing non-sampling errors in sample surveys.

Refusal to respond and lying in surveys are two main sources of such non-sampling errors, as the stigma attached to certain practices (e.g. sexual behaviours and the use of illegal drugs) often leads to discrimination. Warner (1965) was first to introduce a randomized response (RR) model to estimate the proportion for sensitive attributes including homosexuality, drug addiction or abortion. Greenberg et al. (1969) proposed the unrelated question RR model that is a variation of Warner's (1965) RR model. Since the work by Warner (1965) a huge literature has emerged on the use and formulation of different randomization devices to estimate the population proportion of a sensitive attribute in survey sampling. Mention may be made of the work of Tracy and Mangat (1996), Cochran (1977), Singh and Mangat (1996), Chaudhuri and Mukherjee (1988), Ryu et al. (1993), Fox and Tracy (1986), Singh (2003), Singh and Tarray (2012, 2013 a,b,c,d) and the references cited therein.

Mangat et al. (1997) and Singh et al. (2000) pointed out the privacy problem with Moors' (1971) model. To implement the privacy problem with the Moors' (1971) model, Mangat et al. (1997) and Singh et al. (2000) presented several strategies as an alternative to Moors' model, but their models can lose a large portion of data information and require a high cost to obtain confidentiality of the respondents. Kim and Warde (2005) suggested a mixed randomized response model using simple random sampling which rectifies the privacy problem. Amitava (2005) and Hussain and Shabbir (2007) suggested improvements over Kim and Warde's (2005) mixed randomized response technique in complex surveys situations and illustrated the superiority of their models over Kim and Warde's (2005) procedure. Later, Nazuk and Shabbir (2010) presented a modification of Kim and Warde's (2005) model to estimate the proportion of a qualitative sensitive variable using simple random sampling with replacement (SRSWR), which reduces the variance of the estimator and improves the privacy protection of respondents.

In this paper we have suggested a modified mixed randomized response model and its properties are studied. We have shown that the suggested mixed randomized response model is always better than Kim and Warde's (2005) model and it is more efficient than the one recently proposed by Nazuk and Shabbir's (2010) estimator under some realistic conditions.

2. Kim and Warde (2005) and Nazuk and Shabbir's (2010) models

2.1. Kim and Warde's (2005) mixed randomized response model

Kim and Warde (2005) introduced a mixed randomized response procedure for estimating the proportion π_S of a population possessing a sensitive attribute using simple random sampling with replacement (SRSWR) which rectifies the privacy problem. Following them, a single sample with the size n is selected by SRSWR from the population. Each respondent selected in the sample is instructed to answer the direct question "I am a member of the innocuous trait group". If a respondent answers "Yes" to the direction question, then she or he is instructed to go to the randomization device R_1 consisting of the statements (i) "I am a member of the sensitive trait group" and (ii) "I am a member of the innocuous trait group" with pre-assigned probability of selection P_1 and $1\text{-}P_1$, respectively. If a respondent answers "No" to the direct question, then the respondent is instructed to use the randomization device R_2 consisting of the statement (i) "I am a member of the sensitive trait group" and (ii) "I am not a member of the sensitive trait group" with pre-assigned probability P and 1-P, respectively. The survey procedures are performed under the assumption that both the sensitive and innocuous questions are unrelated and independent in the randomization device R_1. To protect the respondents' privacy, the respondents should not disclose to the interviewer the question they answered from either R_1 or R_2. Let n be the sample size confronted with a direct question and n_1 and n_2 (= n- n_1) denote the number of "Yes" and "No" answers from the sample. Since all respondents using the randomization device R_1 already responded "Yes" from the initial direct innocuous question, the proportion Y of getting "Yes" answers from the respondents using the randomization device R_1 should be

$$Y = P_1\pi_S + (1-P_1)\pi_1 = P_1\pi_S + (1-P_1), \tag{2.1}$$

where π_S is the proportion of "Yes" answers from the sensitive trait and π_1 is the proportion of "Yes" answers from the innocuous question [see Kim and Warde (2005,p.212)].

An unbiased estimator of π_S is given by

$$\hat{\pi}_a = \frac{\hat{Y} - (1-P_1)}{P_1} \tag{2.2}$$

where $\hat{Y}$ is the sample proportion of "Yes" responses.

The proportion of "Yes" answers from the respondents using the randomization device R_2 is given by

$$X = [P\pi_S + (1-P)(1-\pi_S)] = [(2P-1)\pi_S + (1-P)] \tag{2.3}$$

Thus, an unbiased estimator of π_S, in terms of the sample proportion of "Yes" responses $\hat{X}$, is

$$\hat{\pi}_b = \frac{\hat{X} - (1-P)}{(2P-1)} \tag{2.4}$$

Pooling the two unbiased estimators $\hat{\pi}_a$ and $\hat{\pi}_b$ using weights, Kim and Warde (2005) suggested an unbiased estimator $\hat{\pi}_a$ and $\hat{\pi}_b$ for π_S as

$$\hat{\pi}_{kw} = \frac{n_1}{n}\hat{\pi}_a + \frac{(n-n_1)}{n}\hat{\pi}_b \text{ , for } 0 < \frac{n_1}{n} < 1 \tag{2.5}$$

Applying Lanke's (1976) arguments, Kim and Warde (2005) derived

$$P = \frac{1}{2-P_1} \tag{2.6}$$

and hence obtained the variance of the estimator $\hat{\pi}_{kw}$ as

$$V(\hat{\pi}_{kw}) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-P_1)\,[\,\lambda P_1(1-\pi_S) + (1-\lambda)]}{nP_1^2} \tag{2.7}$$

for $n = n_1 + n_2$ and $\lambda = \frac{n_1}{n}$.

2.2. Nazuk and Shabbir's (2010) model

Nazuk and Shabbir (2010) presented a modified version of Kim and Warde's (2005) model which differs from Kim and Warde's (2005) procedure only in the formation of the randomization device R_2. The description of Nazuk and Shabbir's (2010) model is given below.

Let a random sample of size n be selected using SRSWR. Each respondent in the sample is instructed to answer an innocuous question "I possess the innocuous character Y". If the answer to the initial direct question is "Yes" then the respondent is instructed to go the randomization device R_1, otherwise R_2, where R_1 consists of two statements (i) "I belong to the sensitive group" and (ii) "I belong to the innocuous group", with respective probability P_1 and $(1\text{-}P_1)$, while R_2 consists of the same pair of statements as in R_1 but with respective probability P_2 and $(1\text{-}P_2)$. In order to offer privacy to the respondents they are not required to say that which randomization device they have used. Let n_1 and n_2 be the number of respondents using R_1 and R_2 respectively such that $(n_1 + n_2) = n$. Note that the respondents coming to R_1 have reported "Yes" to the initial direct question, therefore $\pi_1 = 1$ in R_1 [see Nazuk and Shabbir (2010, pp.186-187)].

The probability of "Yes" answers is (the same as given in (2.1))

$$Y = [P_1\pi_S + (1-P_1)\pi_1] = P_1\pi_S + (1-P_1), \tag{2.8}$$

An unbiased estimator of π_S is (the same as given in (2.2))

$$\hat{\pi}_a = \frac{\hat{Y} - (1-P_1)}{P_1} \tag{2.9}$$

where $\hat{Y}$ is the same as defined earlier.

Note that the respondents using R_2 have reported a "No" to the initial direct question, therefore $\pi_1 = 0$ in R_2. Denote by X_2 the probability of "Yes" answers from the respondents using R_2 which is given by

$$X_2 = [P_2\pi_S + (1-P_2)\pi_1] = P_2\pi_S \tag{2.10}$$

Let $\hat{X}_2$ be the sample proportion of "Yes" response from the randomization device R_2, then an unbiased estimator of π_S is given by

$$\hat{\pi}_c = \frac{\hat{X}_2 - (1-P_2)}{P_2} \tag{2.11}$$

Pooling the two unbiased estimators $\hat{\pi}_a$ and $\hat{\pi}_c$, Nazuk and Shabbir (2010) suggested an unbiased estimator for π_S as

$$\hat{\pi}_{ns} = \frac{n_1}{n}\hat{\pi}_a + \frac{(n-n_1)}{n}\hat{\pi}_c \text{ , for } 0 < \frac{n_1}{n} < 1 \tag{2.12}$$

With $P_2 = \frac{1}{2-P_1}$ Nazuk and Shabbir (2010) obtained the variance of $\hat{\pi}_{ns}$ as

$$V(\hat{\pi}_{ns}) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-P_1)\,[\,\lambda(1-\pi_S) + (1-\lambda)\pi_S P_1\,]}{nP_1} \tag{2.13}$$

3. The suggested model

The suggested procedure differs from Kim and Warde (2005) and Nazuk and Shabbir's (2010) procedures only in the contribution of the randomization device R_2. Let a random sample of size n be selected using simple random sampling with replacement (SRSWR). Each respondent from the sample is instructed to answer the direct question "I am a member of the innocuous group". If a respondent

answers "Yes" to the direct question, then she or he is instructed to go to the randomization device R_1 consisting of the statements (i) "I am the member of the sensitive trait group" and (ii) "I am a member of the innocuous trait group" with respective probabilities P_1 and $(1\text{-}P_1)$. If a respondent answers "No" to the direct question, then the respondent is instructed to use the randomization device R_2 using three statements: (i) "I possess the sensitive attribute "A" ", (ii) "Yes" and (iii) "No" with known probabilities P, (1-P)w and (1-P)w respectively, where $w \in [0,1]$. It is to be mentioned that the randomization device R_2 is due to Singh et al. (1995). The survey procedures are performed under the assumption that both the sensitive and innocuous questions are unrelated and independent in the randomization device R_1. To protect the respondent's privacy, the respondents should not disclose to the interviewer the question they answered from either R_1 or R_2.

We explain the suggested procedure with the help of an example earlier considered by Hussain and Shabbir (2007). Consider that we are interested in the estimation of the proportion π_S of carriers of HIV in a particular county/ locality/ town or district. Each survey respondent is asked a direct innocuous (non-sensitive) question "Were you born in the first three months of a calendar year?". On receiving a "Yes" response he/she is requested to use the randomization device R_1 consisting of the two statements, (i) "I do carry HIV" and (ii) "My birthday falls in the first three months of a calendar year" presented with predetermined probabilities P_1 and $(1\text{-}P_1)$. If the respondent says "No" to the direct question he/she is requested to use the randomization device R_2. Now, from this random device, if the statement (i) is chosen, the respondent will reply according to his actual status with respect to carriers of HIV. In the case the statement (ii) or (iii) is selected, one will report "Yes" or "No" as observed on the outcome of the random device R_2 presented with predetermined probabilities P, (1-P)w and (1-P)w respectively, where $w \in [0,1]$.

Let n be the sample size confronted with a direct question and n_1 and n_2 (= $n - n_1$) denote the number of "Yes" and "No" answers from the sample. Note that the respondents coming to R_1 have reported a "Yes" to the initial direct question, therefore $\pi_1 = 1$ in R_1, where π_1 is the proportion of "Yes" answers from the innocuous question.

Denote by 'Y' the probability of "Yes" from the respondents using R_1. Then

$$Y = P_1\pi_S + (1 - P_1)\pi_1 = P_1\pi_S + (1 - P_1), \tag{3.1}$$

where π_S is the proportion of "Yes" answers from the sensitive trait.

An unbiased estimator of π_S, in terms of the sample proportion of "Yes" responses $\hat{Y}$, becomes

$$\hat{\pi}_a = \frac{\hat{Y} - (1 - P_1)}{P_1} \quad . \tag{3.2}$$

The variance of $\hat{\pi}_a$ is

$$V(\hat{\pi}_a) = \frac{Y(1-Y)}{n_1 P_1^2} = \frac{(1-\pi_S)[\,P_1\pi_S + (1-P_1)]}{n_1 P_1}$$
$$= \frac{1}{n_1}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P_1)}{P_1}\right]. \tag{3.3}$$

The proportion of "Yes" answers from the respondents using randomization device R_2 follows:

$$X_3 = P\pi_S + (1-P)w \tag{3.4}$$

An unbiased estimator of π_S, in terms of the sample proportion of "Yes" responses $\hat{X}_3$, becomes

$$\hat{\pi}_d = \frac{\hat{X}_3 - (1-P)w}{P} \quad . \tag{3.5}$$

The variance of $\hat{\pi}_d$ is given by

$$V(\hat{\pi}_d) = \frac{X_3(1-X_3)}{n_2 P^2} = \left[\frac{\pi_S(1-\pi_S)}{n_2} + \frac{(1-\pi_S)(1-P)w}{n_2 P^2}\right] \tag{3.6}$$

The estimator of π_S, in terms of the sample proportions of "Yes" responses $\hat{Y}$ and $\hat{X}_3$, is

$$\hat{\pi}_t = \frac{n_1}{n}\hat{\pi}_a + \frac{n_2}{n}\hat{\pi}_d$$
$$= \frac{n_1}{n}\hat{\pi}_a + \frac{(n-n_1)}{n}\hat{\pi}_d \text{ , for } 0 < \frac{n_1}{n} < 1 \tag{3.7}$$

As both $\hat{\pi}_a$ and $\hat{\pi}_d$ are unbiased estimators, the expected value of $\hat{\pi}_t$ is

$$E(\hat{\pi}_t) = E\left[\frac{n_1}{n}\hat{\pi}_a + \frac{n_2}{n}\hat{\pi}_d\right]$$
$$= \frac{n_1}{n}\pi_S + \frac{(n-n_1)}{n}\pi_S = \pi_S$$

Thus, the proposed estimator $\hat{\pi}_t$ is an unbiased estimator π_S.

Now, the variance of $\hat{\pi}_t$ is given by

$$V(\hat{\pi}_t) = \left(\frac{n_1}{n}\right)^2 V(\hat{\pi}_a) + \left(\frac{n_2}{n}\right)^2 V(\hat{\pi}_d)$$

$$= \left(\frac{n_1}{n}\right)^2 \frac{1}{n_1}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P_1)}{P_1}\right]$$

$$+ \left(\frac{n_2}{n}\right)^2 \frac{1}{n_2}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P)w}{P^2}\right]$$

$$= \frac{n_1}{n^2}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P_1)}{P_1}\right] + \frac{n_2}{n^2}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P)w}{P^2}\right] \tag{3.8}$$

Since our mixed RR model also uses Horvitz's et al. (1967) method when $\pi_1 = 1$, we can apply Lanke's (1976) idea to our suggested model. Thus, using Lanke's (1976) result for P with $\pi_1 = 1$, we get

$$P = \frac{1}{2 - P_1}. \tag{3.9}$$

Putting $P = (2 - P_1)^{-1}$ in (3.6), we get

$$V(\hat{\pi}_d) = \frac{\pi_S(1-\pi_S)}{(n-n_1)} + \frac{(1-\pi_S)(1-P_1)w}{(n-n_1)}$$

$$= \frac{1}{(n-n_1)}\left[\pi_S(1-\pi_S) + (1-\pi_S)(1-P_1)w\right] \tag{3.10}$$

Thus, we have established the following theorem.

Theorem 3.1. The variance of $\hat{\pi}_t$ is given by

$$V(\hat{\pi}_t) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-P_1)\left[\lambda(1-\pi_S) + (1-\lambda)P_1(1-\pi_S)w\right]}{nP_1} \tag{3.11}$$

for $n = n_1 + n_2$ and $\lambda = \frac{n_1}{n}$.

Remark 3.1. Following Chaudhuri (2001, 2004), Amitava (2005) and Hussain and Shabbir (2007), the present study can be extended for complex surveys.

4. Efficiency comparisons

In this section we have made a comparison of the suggested model under a completely truthful reporting case with Kim and Warde (2005) and Nazuk and Shabbir's (2010) models.

From (2.7) and (3.11) we have

$$V(\hat{\pi}_t) < V(\hat{\pi}_{kw}) \text{ if}$$

$$[\lambda(1-\pi_S)+(1-\lambda)P_1(1-\pi_S)w] < \frac{[\lambda P_1(1-\pi_S)+(1-\lambda)]}{P_1}$$

$$\text{i.e. if } \quad P_1^2(1-\pi_S)w < 1$$

which is always true.

Thus, the proposed model is always better than Kim and Warde's (2005) model.

Further, from (2.13) and (3.11) we have

$$V(\hat{\pi}_{ns}) - V(\hat{\pi}_t) = \frac{(1-P_1)(1-\lambda)}{n}\{\pi_S(1+w)-w\}$$

which is positive if

$$\pi_S > \frac{w}{(1+w)}. \tag{4.1}$$

It follows from (4.1) that f for $\pi_S \geq 1/2$, the proposed randomized response model is always superior to Nazuk and Shabbir's (2010) model. Further, for $\pi_S = 2/5, 3/5, 1/5, 1/10$, the proposed model is better than Nazuk and Shabbir's (2010) model in the respective ranges of w:

$$w \in (0, 2/3),\ w \in (0, 3/7),\ w \in (0, 1/4) \text{ and } w \in (0, 1/9)$$

It is observed from the above that when the value of $\pi(<1/2)$ decreases the ranges of w decrease.

To have a tangible idea about the performance of the proposed estimator $\hat{\pi}_t$ over Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$ and Nazuk and Shabbir's (2010) estimator $\hat{\pi}_{ns}$, we have computed the percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$ and Nazuk and Shabbir's (2010) estimator $\hat{\pi}_{ns}$ by using the formulae:

$$PRE\ (\hat{\pi}_t, \hat{\pi}_{kw}) = \frac{V(\hat{\pi}_{kw})}{V(\hat{\pi}_t)} \times 100$$

$$= \frac{\left[\pi_S(1-\pi_S)P_1^2 + (1-P_1)\{\lambda P_1(1-\pi_S) + (1-\lambda)\}\right]}{P_1\left[\pi_S(1-\pi_S)P_1 + (1-P_1)\{\lambda(1-\pi_S) + (1-\lambda)P_1(1-\pi_S)w\}\right]} \times 100 \quad (4.2)$$

and

$$\text{PRE}\,(\hat{\pi}_t, \hat{\pi}_{ns}) = \frac{V(\hat{\pi}_{ns})}{V(\hat{\pi}_t)} \times 100$$

$$= \frac{\left[\pi_S(1-\pi_S)P_1 + (1-P_1)\{\lambda(1-\pi_S) + (1-\lambda)\pi_S P_1\}\right]}{\left[\pi_S(1-\pi_S)P_1 + (1-P_1)\{\lambda(1-\pi_S) + (1-\lambda)P_1(1-\pi_S)w\}\right]} \times 100 \quad (4.3)$$

for different values of , P_1,w, n and n_1.

We have obtained the values of the percent relative efficiencies $\text{PRE}\,(\hat{\pi}_t, \hat{\pi}_{kw})$ for λ= (0.7, 0.5, 0.3), n = 1000 and for different cases of of π_S , w, n_1 and P_1. Findings are shown in Table 1. Diagrammatic representation is also given in Fig. 1.

It is observed from Fig. 1 and Table 1that:

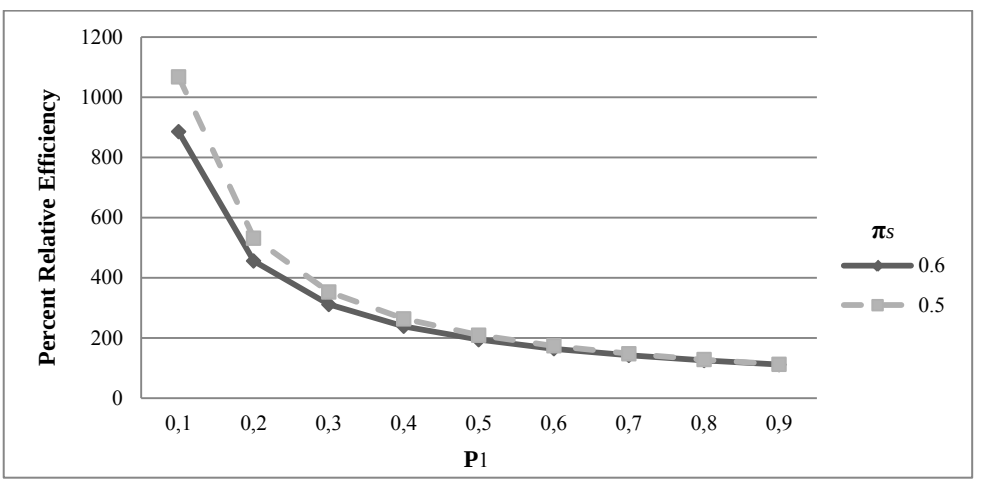


Figure 1. Percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$ when $\lambda = 0.7$ and $w = 0.25$

Table 1. Percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$

π_S	n = 1000		λ	w	P_1								
	n_1	n_2			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.50	700.00	300.00	0.70	0.25	885.33	455.36	311.09	238.24	193.94	163.88	141.93	125.00	111.37
0.50	700.00	300.00	0.70	0.50	876.71	447.37	303.80	231.71	188.24	159.09	138.15	122.34	109.97
0.50	700.00	300.00	0.70	0.75	868.26	439.66	296.84	225.52	182.86	154.57	134.56	119.79	108.59
0.50	500.00	500.00	0.50	0.25	1858.19	865.38	538.40	377.36	282.35	220.13	176.45	144.23	119.53
0.50	500.00	500.00	0.50	0.50	1818.18	833.33	512.82	357.14	266.67	208.33	168.07	138.89	116.96
0.50	500.00	500.00	0.50	0.75	1779.86	803.57	489.56	338.98	252.63	197.74	160.44	133.93	114.49
0.50	300.00	700.00	0.30	0.25	3848.10	1614.13	914.09	587.68	405.63	292.93	218.14	165.98	128.20
0.50	300.00	700.00	0.30	0.50	3675.68	1500.00	836.60	534.48	369.23	268.52	202.53	156.98	124.25
0.50	300.00	700.00	0.30	0.75	3518.04	1400.94	771.22	490.12	338.82	247.86	189.00	148.90	120.54
0.60	700.00	300.00	0.70	0.25	1067.81	531.79	352.90	263.27	209.35	173.25	147.34	127.77	112.42
0.60	700.00	300.00	0.70	0.50	1057.57	522.73	344.98	256.47	203.64	168.64	143.83	125.39	111.20
0.60	700.00	300.00	0.70	0.75	1047.52	513.97	337.40	250.00	198.23	164.27	140.48	123.09	110.01
0.60	500.00	500.00	0.50	0.25	2256.12	1022.22	619.63	423.68	309.68	236.16	185.44	148.75	121.23
0.60	500.00	500.00	0.50	0.50	2208.45	985.71	591.70	402.50	293.88	224.73	177.62	143.95	119.00
0.60	500.00	500.00	0.50	0.75	2162.76	951.72	566.19	383.33	279.61	214.36	170.44	139.45	116.86
0.60	300.00	700.00	0.30	0.25	4650.76	1896.91	1048.23	659.09	445.57	315.45	230.45	172.10	130.51
0.60	300.00	700.00	0.30	0.50	4448.13	1769.23	965.12	604.17	409.30	291.96	215.94	164.01	127.09
0.60	300.00	700.00	0.30	0.75	4262.43	1657.66	894.22	557.69	378.49	271.73	203.14	156.65	123.85
0.70	700.00	300.00	0.70	0.25	1372.48	660.11	423.80	306.41	236.52	190.34	157.68	133.43	114.77
0.70	700.00	300.00	0.70	0.50	1359.50	649.17	414.67	298.91	230.51	185.69	154.28	131.22	113.68
0.70	700.00	300.00	0.70	0.75	1346.75	638.59	405.93	291.78	224.79	181.26	151.03	129.08	112.62
0.70	500.00	500.00	0.50	0.25	2921.41	1286.90	758.87	504.92	359.08	266.32	203.27	158.33	125.16
0.70	500.00	500.00	0.50	0.50	2860.83	1242.53	726.35	481.25	342.14	254.58	195.56	153.81	123.16
0.70	500.00	500.00	0.50	0.75	2802.71	1201.11	696.50	459.70	326.73	243.82	188.42	149.54	121.22
0.70	300.00	700.00	0.30	0.25	5998.59	2380.72	1283.95	788.84	521.07	360.06	256.18	185.70	136.04
0.70	300.00	700.00	0.30	0.50	5744.28	2227.83	1188.34	727.94	482.27	335.83	241.78	178.01	132.94
0.70	300.00	700.00	0.30	0.75	5510.65	2093.39	1105.98	675.77	448.84	314.65	228.90	170.93	129.98
0.90	700.00	300.00	0.70	0.25	3814.24	1694.15	1000.32	661.65	464.12	336.52	248.40	184.63	136.83
0.90	700.00	300.00	0.70	0.50	3779.14	1667.54	980.42	647.06	453.73	329.44	243.91	182.13	135.80
0.90	700.00	300.00	0.70	0.75	3744.68	1641.75	961.29	633.09	443.80	322.65	239.58	179.69	134.78
0.90	500.00	500.00	0.50	0.25	8261.22	3430.00	1901.23	1182.61	779.49	529.00	362.52	246.43	162.47
0.90	500.00	500.00	0.50	0.50	8096.00	3319.35	1827.01	1133.33	747.54	509.17	351.09	240.70	160.40
0.90	500.00	500.00	0.50	0.75	7937.25	3215.63	1758.38	1088.00	718.11	490.76	340.36	235.23	158.38
0.90	300.00	700.00	0.30	0.25	16862.28	6343.75	3253.67	1896.91	1180.58	758.78	491.58	313.12	189.05
0.90	300.00	700.00	0.30	0.50	16183.91	5970.59	3037.64	1769.23	1105.45	715.95	468.81	302.63	185.63
0.90	300.00	700.00	0.30	0.75	15558.01	5638.89	2848.51	1657.66	1039.32	677.69	448.06	292.82	182.34

The values of percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t, \hat{\pi}_{kw})$ is more than 100. We can say that the envisaged estimator $\hat{\pi}_t$ is more efficient than Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$. Fig. 1 shows the results for $\lambda = 0.7$, w = 0.25 and different values of P_1 and π_S.

We note from Table 1 that the values of the percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t, \hat{\pi}_{kw})$ decrease as the value of P_1 increases. Also, the values of the percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t, \hat{\pi}_{kw})$ increase as the value of λ decreases for fixed values of P_1.

We further note from the results of Fig. 1 that there is a large gain in efficiency by using the suggested estimator $\hat{\pi}_t$ over Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$ when the proportion of the stigmatizing attribute is moderately large.

Fig. 2 and Table 2exhibit that:

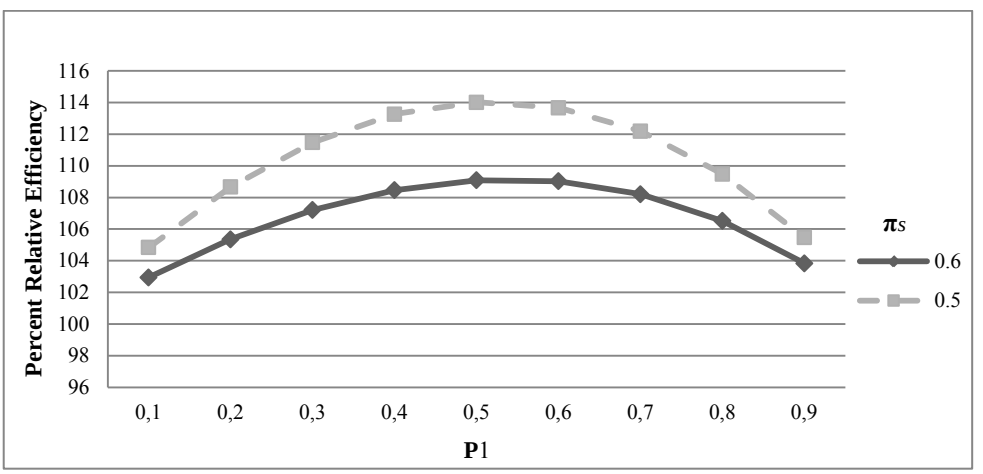


Figure 2. Percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Nazuk and Shabbir's (2010) estimator $\hat{\pi}_{ns}$ when $\lambda = 0.7$ and $w = 0.25$

Table 2. Percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Nazuk and Shabbir's (2010) estimator $\hat{\pi}_{ns}$

π_S	n = 1000		λ	w	P_1								
	n_1	n_2			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.50	700.00	300.00	0.70	0.25	102.95	105.36	107.21	108.46	109.09	109.03	108.21	106.52	103.84
0.50	700.00	300.00	0.70	0.50	101.95	103.51	104.69	105.49	105.88	105.84	105.33	104.26	102.53
0.50	700.00	300.00	0.70	0.75	100.96	101.72	102.29	102.67	102.86	102.84	102.59	102.08	101.25
0.50	500.00	500.00	0.50	0.25	106.60	111.54	114.96	116.98	117.65	116.98	114.96	111.54	106.60
0.50	500.00	500.00	0.50	0.50	104.31	107.41	109.50	110.71	111.11	110.71	109.50	107.41	104.31
0.50	500.00	500.00	0.50	0.75	102.11	103.57	104.54	105.08	105.26	105.08	104.54	103.57	102.11
0.50	300.00	700.00	0.30	0.25	114.07	122.83	127.79	129.86	129.58	127.27	123.13	117.21	109.53
0.50	300.00	700.00	0.30	0.50	108.96	114.14	116.96	118.10	117.95	116.67	114.31	110.85	106.16
0.50	300.00	700.00	0.30	0.75	104.29	106.60	107.81	108.30	108.24	107.69	106.68	105.15	102.99
0.60	700.00	300.00	0.70	0.25	104.84	108.67	111.48	113.27	114.02	113.68	112.20	109.49	105.47
0.60	700.00	300.00	0.70	0.50	103.84	106.82	108.98	110.34	110.91	110.65	109.52	107.45	104.33
0.60	700.00	300.00	0.70	0.75	102.85	105.03	106.59	107.56	107.96	107.78	106.98	105.49	103.21
0.60	500.00	500.00	0.50	0.25	110.79	118.52	123.60	126.32	126.88	125.42	122.01	116.67	109.36
0.60	500.00	500.00	0.50	0.50	108.45	114.29	118.03	120.00	120.41	119.35	116.87	112.90	107.35
0.60	500.00	500.00	0.50	0.75	106.21	110.34	112.94	114.29	114.56	113.85	112.14	109.38	105.41
0.60	300.00	700.00	0.30	0.25	122.78	136.08	143.06	145.45	144.30	140.23	133.61	124.65	113.44
0.60	300.00	700.00	0.30	0.50	117.43	126.92	131.72	133.33	132.56	129.79	125.19	118.79	110.47
0.60	300.00	700.00	0.30	0.75	112.52	118.92	122.04	123.08	122.58	120.79	117.78	113.46	107.65
0.70	700.00	300.00	0.70	0.25	107.96	114.04	118.34	120.89	121.74	120.89	118.34	114.04	107.96
0.70	700.00	300.00	0.70	0.50	106.94	112.15	115.79	117.93	118.64	117.93	115.79	112.15	106.94
0.70	700.00	300.00	0.70	0.75	105.94	110.33	113.35	115.12	115.70	115.12	113.35	110.33	105.94
0.70	500.00	500.00	0.50	0.25	117.65	129.76	137.31	140.98	141.25	138.46	132.83	124.51	113.56
0.70	500.00	500.00	0.50	0.50	115.21	125.29	131.43	134.38	134.59	132.35	127.80	120.95	111.74
0.70	500.00	500.00	0.50	0.75	112.87	121.11	126.03	128.36	128.53	126.76	123.13	117.59	109.98
0.70	300.00	700.00	0.30	0.25	136.89	157.19	167.05	169.72	167.05	160.14	149.66	136.01	119.42
0.70	300.00	700.00	0.30	0.50	131.09	147.09	154.61	156.62	154.61	149.36	141.24	130.37	116.70
0.70	300.00	700.00	0.30	0.75	125.76	138.22	143.89	145.39	143.89	139.94	133.72	125.19	114.10
0.90	700.00	300.00	0.70	0.25	132.51	155.85	171.06	178.95	180.15	175.18	164.42	148.17	126.64
0.90	700.00	300.00	0.70	0.50	131.29	153.40	167.66	175.00	176.12	171.50	161.45	146.15	125.69
0.90	700.00	300.00	0.70	0.75	130.09	151.03	164.39	171.22	172.26	167.96	158.58	144.20	124.74
0.90	500.00	500.00	0.50	0.25	171.43	216.67	242.17	252.17	249.57	236.36	213.95	183.33	145.19
0.90	500.00	500.00	0.50	0.50	168.00	209.68	232.71	241.67	239.34	227.50	207.21	179.07	143.34
0.90	500.00	500.00	0.50	0.75	164.71	203.13	223.97	232.00	229.92	219.28	200.87	175.00	141.54
0.90	300.00	700.00	0.30	0.25	246.71	318.75	348.91	352.58	337.86	309.40	269.97	221.29	164.42
0.90	300.00	700.00	0.30	0.50	236.78	300.00	325.75	328.85	316.36	291.94	257.47	213.88	161.45
0.90	300.00	700.00	0.30	0.75	227.62	283.33	305.46	308.11	297.44	276.34	246.07	206.94	158.58

The values of percent relative efficiencies $PRE(\hat{\pi}_t, \hat{\pi}_{ns})$ is more than 100. We can say that the envisaged estimator $\hat{\pi}_t$ is more efficient than Nazuk and Shabbir's (2010) estimator$_c$. Fig. 2 shows the results for $\lambda = 0.7$, w = 0.25 and different values of P_1 and π_S.

We note from Table 2 that the values of the percent relative efficiencies $PRE(\hat{\pi}_t, \hat{\pi}_{ns})$ increase as the value of P_1 increases up to $P_1 \leq 0.5$ and decreases $P_1 > 0.5$ onwards. Also, the values of the percent relative efficiencies $PRE(\hat{\pi}_t, \hat{\pi}_{ns})$ increase as the value of λ decreases for fixed value of P_1.

We further note from the results of Fig. 2 that there is a large gain in efficiency by using the suggested estimator $\hat{\pi}_t$ over Nazuk and Shabbir's (2010) estimator $\hat{\pi}_{ns}$ when the proportion π_S of the stigmatizing attribute and P_1 are moderately large.

It is observed from Table 1 and Table 2 that a larger gain in efficiency is obtained by using the proposed estimator $\hat{\pi}_t$ over Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$ as compared to Nazuk and Shabbir's (2010) estimator $\hat{\pi}_{ns}$.

5. Conclusions

In this paper we have proposed a mixed randomized response model to estimate the proportion of qualitative sensitive character. It has been shown that the proposed mixed randomized response model is more efficient than Kim and Warde (2005) and Nazuk and Shabbir's (2010) mixed randomize response models with a larger gain in efficiency. Thus, this paper attempts to extend the methodology of the mixed randomized response techniques.

Acknowledgements

The authors are grateful to the referee for fruitful comments to bring the original manuscript in the present form.

REFERENCES

AMITAVA, S., (2005). Kim and Warde's mixed randomized response technique for complex surveys. Jour Mod. Appl. Statist. Meth. , 4(2), 538–544.

CHAUDHURI, A., MUKERJEE, R., (1988): Randomized Response: Theory and Techniques. Marcel-Dekker, New York, USA.

CHAUDHURI, A., (2011). Randomized response and indirect questioning techniques in surveys. CRC Press, Taylor and Frances group, USA.

CHAUDHURI, A., (2004). Christofides' randomized response technique in complex sample surveys. Metrika, 60(3), 23–228.

CHAUDHURI, A., (2002). Estimating sensitive proportions from randomized responses in unequal probability sampling. Cal. Statist. Assoc. Bull., 52, 315–322.

COCHRAN, W. G., (1977). Sampling Technique, 3rd Edition. New York: John Wiley and Sons, USA.

FOX, J. A., TRACY, P. E., (1986). Randomized Response: A method of Sensitive Surveys. Newbury Park, CA: SEGE Publications.

GREENBERG, B., ABUL-ELA, A., SIMMONS, W. R., HORVITZ, D. G., (1969). The unreleased question randomized response: Theoretical framework. Jour. Amer. Statist. Assoc., 64, 529–539.

HORVITZ, D. G., SHAH, B. V., SIMMONS, W. R., (1967). The unrelated question randomized response model. Proc. Soc. Statist. Sec. Amer. Statistical Assoc. 65–72.

HUSSAIN, Z., SHABBIR, J., (2007). Improvement of Kim and Warde's mixed randomized response technique for complex surveys. InterStat, July # 003.

KIM, J. M., TEBBS, J. M., AN, S. W., (2006). Extensions of Mangat's randomized response model. Jour. Statist. Plan. Inference, 136, 1554–1567.

KIM, J. M., WARDE, W. D., (2005). A mixed randomized response model. Jour. Statist. Plan. Inference, 133, 211–221.

LANKE, J., (1976). On the degree of protection in randomized interview Internet. Statist. Rev. 44, 80–83.

MANGAT, N. S., SINGH, R., (1990). An alternative randomized procedure. Biometrika, 77, 439–442.

MANGAT, N. S., SINGH, R., SINGH, S., (1997). Violation of respondent's privacy in Moors model – its rectification through a random group strategy response model. Comm. Statist. Theo. Meth., (3), 243–255.

MOORS, J. A., (1971). Optimization of the unrelated question randomized response model. Jour. Amer. Statist. Assoc., 66, 627–629.

NAZUK, A., SHABBIR, J., (2010). A new mixed randomized response model. Inter. Jour Buss. Soc. Sci. 1, 186–190.

RYU, J. B., HONG, K. H., LEE, G. S., (1993). Randomized response model, Freedom Academy, Seoul, Korea.

SINGH, H. P., TARRAY, T. A., (2012). A Stratified Unknown repeated trials in randomized response sampling. Comm. Korean Statist. Soc., 19, (6), 751–759.

SINGH, H. P., TARRAY, T. A., (2013a). An alternative to Kim and Warde's mixed randomized response model. Statist. Oper. Res. Trans., 37 (2), 189–210.

SINGH, H. P., TARRAY, T. A., (2013b). An alternative to stratified Kim and Warde's randomized response model using optimal (Neyman) allocation, Model Assist. Statist. Appl., 9, 37–62.

SINGH, H. P., TARRAY, T. A., (2013c). An improved mixed randomized response model. Model Assist. Statist. Appl., 9, 73–87.

SINGH, H. P., TARRAY, T. A., (2013d). An alternative to Kim and Warde's mixed randomized response technique. Accepted in Statistica.

SINGH, R., MANGAT, N. S., (1996). Elements of Survey Sampling, Kluwer Academic Publishers, Dordrecht, The Netherlands.

SINGH, S., SINGH, R., MANGAT, N. S., TRACY, D. S., (1995). An improved two-stage randomized response strategy. Statistical Papers, 36, 265–271.

SINGH, S., (2003). Advanced sampling theory with applications. Kluwer Academic Publishers, Dordrecht.

SING, S., SINGH, R., MANGAT, N. S., (2000). Some alternative strategies to Moor's model in randomized response model. Jour. Statist. Plan. Inference, 83, 243–255.

TRACY, D. S., MANGAT, N. S., (1996). Some developments in randomized response sampling during the last decade – A follow up of review by Chaudhuri and Mukherjee. Jour. Applied. Statist. Sci., 4 (2/3), 147–158.

WARNER, S. L., (1965). Randomized response: A survey technique for eliminating evasive answer bias. Jour. Amer. Statist. Assoc., 60, 63–69.

STATISTICS IN TRANSITION-new series, Winter 2014
Vol. 15, No. 1, pp. 83–96

APPLICATION OF THE ORIGINAL PRICE INDEX FORMULA TO MEASURING THE CPI'S COMMODITY SUBSTITUTION BIAS

Jacek Białek[1]

ABSTRACT

This paper examines a possibility to apply the original price index formula to measuring the commodity substitution bias associated with the Consumer Price Index (CPI). Through simulation study the CPI bias values - calculated by using the original price index formula – is compared with those calculated on the basis of some known, superlative price indices.

Key words: CPI, COLI, superlative index, Laspeyres index, Fisher index.
JEL Classification: E17, E21, E30
AMS Classification: 62P20

1. Introduction

The Consumer Price Index (CPI) is used as a basic measure of inflation. The index approximates changes in the costs of household consumption that provide the constant utility (COLI, Cost of Living Index). In practice, the Laspeyres price index is used to measure the CPI (see White (1999), Clements and Izan (1987)). The Lapeyres formula does not take into account changes in the structure of consumption, which occur as a result of price changes in the given time interval. It means that the Laspeyres index can be biased due to the commodity substitution. Many economists consider the superlative indices (like the Fisher index or the Törnqvist index) to be the best approximation of COLI. Thus, the difference between the Laspeyres index and the superlative index should approximate the value of the commodity substitution bias. In this paper we propose the application of the original price index formula (see Białek (2012a), Białek (2013)) in measuring the commodity substitution bias associated with the Consumer Price Index (CPI). In our simulation study we compare the CPI bias values calculated by using the original price index formula with those calculated

[1] University of Lodz, Chair of Statistical Methods. E-mail: jbialek@uni.lodz.pl.

on the basis of some known, superlative price indices. It should be emphasized, that we do not consider other sources of the CPI biases, presented by White (1999).

2. Superlative price indices in the CPI bias measurement

Any discussion of consumer price index bias must first address the important issue of the target measure with respect to which the bias is measured. The final report of the Boskin Commission begins with a recommendation that "the Bureau of Labor Statistics (BLS) should establish a cost of living index (COLI) as its objective in measuring consumer prices" (see Boskin *et al.* (1996), page 2). Further discussions on the theory of the COLI can be found in the following papers: Diewert (1983), Jorgenson and Slesnick (1983), Pollak (1989).

Let $E(P,\overline{u}) = \min_{Q}\{P^T Q \,|\, U(Q) \geq \overline{u}\}$ be the expenditure function of a representative consumer which is dual to the utility function $U(Q)$. In other words it is the minimum expenditure necessary to achieve a reference level of utility $\overline{u}$ at vector of prices P. Then the Konüs cost of living price index is defined as

$$I_K = \frac{E(P^t,\overline{u})}{E(P^s,\overline{u})}, \tag{1}$$

where t denotes the current period, s denotes the base period, and in general, the vector of N considered prices at any moment τ is given by $P^\tau = [p_1^\tau, p_2^\tau, ..., p_N^\tau]^T$. I_K is a true cost of living index in which the commodity Q changes as the vector of prices facing the consumer changes. The CPI, in contrast, measures the change in the cost of purchasing a fixed basket of goods at a fixed sample of outlets over a time interval, *i.e.* $Q^s = [q_1^s, q_2^s, ..., q_N^s]^T = Q^t$.

The CPI is a Laspeyres-type index defined by

$$I_{La} = \frac{\sum_{i=1}^{N} q_i^s p_i^t}{\sum_{i=1}^{N} q_i^s p_i^s}, \tag{2}$$

so we assume here the constant consumption vector on the base period level. It can be shown (see Diewert (1993)) that under the assumption that the consumption vector Q^t solves the base period t expenditure minimization problem, that

$$I_K = \frac{E(P^t,U(Q^s))}{E(P^s,U(Q^s))} \leq I_{La}, \tag{3}$$

so $I_{La} - I_K$ is the extent of the commodity substitution bias, where I_K plays the role of the reference benchmark. In the so-called economic price index approach many authors use superlative price indices to approximate the I_K index (see White (1999)).

First we define a price index I to be *exact* for a linearly homogeneous aggregator function f (here a utility function), which has a dual unit cost function $c(P)$ and it holds

$$I = \frac{c(P^t)}{c(P^s)}. \tag{4}$$

In other words, an *exact* price index is one whose functional form is *exactly* equal to the ratio of cost functions for some underlying functional form representing preferences. The Fisher price index I_F (defined below as a geometric mean of the Laspeyres and Paasche indices) is exact for the linearly homogeneous quadratic aggregator function $f(x) = (x^T Ax)^{0,5}$, where A is a symmetric matrix of constants (see Diewert (1976)). The quadratic function above is an example of a *flexible functional form* (i.e. a function that provides a second order approximation to an arbitrary twice continuously differentiable function). Since I_F is exact for a flexible functional form, it is said to be a *superlative* index number (see Diewert (1976)). In the papers of Afriat (1972), Pollak (1971) and Samuelson-Swamy (1974) we can find other examples of exact index numbers and also superlative index numbers (see Diewert (1976), von der Lippe (2007)). Under all above remarks, an estimate of the commodity substitution bias B_{csub} can be given by (see White (1999))

$$B_{csub} \approx I_{La} - I_F. \tag{5}$$

In general, we can use any other superlative index $I_{\sup}$ to calculate the above mentioned CPI bias, namely then (see Hałka, Leszczyńska (2011))

$$B_{csub} \approx I_{La} - I_{\sup}. \tag{6}$$

In this paper we compare the CPI biases calculated by using some known, superlative price indices. In particular, we use the Fisher price index, the Törnqvist price index (I_T), the Walsh price index (I_W) and some original price index formula, defined in the next part of the paper (I_B). These indices are as follows (see von der Lippe (2007))

$$I_F = \sqrt{I_{La} I_{Pa}}, \tag{7}$$

$$I_W = \frac{\sum_{i=1}^{N} p_i^t \cdot \sqrt{q_i^s q_i^t}}{\sum_{i=1}^{N} p_i^s \cdot \sqrt{q_i^s q_i^t}}, \tag{8}$$

$$I_T = \prod_{i=1}^{N} \left(\frac{p_i^t}{p_i^s}\right)^{\overline{w}_i}, \tag{9}$$

where

$$w_i^s = \frac{p_i^s q_i^s}{\sum_{k=1}^{N} p_k^s q_k^s}, \quad w_i^t = \frac{p_i^t q_i^t}{\sum_{k=1}^{N} p_k^t q_k^t}, \quad \overline{w}_i = \frac{1}{2}(w_i^s + w_i^t). \tag{10}$$

3. The original price index formula

Białek (2012a) proposes the following price index

$$I_B = \sqrt{I_L \cdot I_U}, \tag{11}$$

where the *lower* (I_L) and *upper index* (I_U) we define as follows

$$I_L = \frac{\sum_{i=1}^{N} \min(q_i^s, q_i^t) p_i^t}{\sum_{i=1}^{N} \max(q_i^s, q_i^t) p_i^s}, \tag{12}$$

$$I_U = \frac{\sum_{i=1}^{N} \max(q_i^s, q_i^t) p_i^t}{\sum_{i=1}^{N} \min(q_i^s, q_i^t) p_i^s}. \tag{13}$$

In the above-mentioned paper it is proved that the index I_B^P satisfies price dimensionality, commensurability, identity, the mean value test, the time reversal test and linear homogeneity (see von der Lippe (2007)). Moreover, there are some interesting relations between this index and other formulas. For example, in the paper of Białek (2013) it is also proved that

$$\frac{I_F}{I_B} = \left[\frac{I_U}{I_L}\right]^{\frac{1}{2} - \frac{Log(I_U / I_F)}{Log(I_U / I_L)}}, \tag{14}$$

and thus

$$\sqrt{\frac{I_L}{I_U}} \le \frac{I_F}{I_B} \le \sqrt{\frac{I_U}{I_L}}. \tag{15}$$

It leads to the following conclusion

$$\forall i \in \{1,2,...,N\} \;\; q_i^s \approx q_i^t \;\; \Rightarrow \; I_L \approx I_U \; \Rightarrow \; I_F \approx I_B. \tag{16}$$

In the paper of von der Lippe (2012) it is proved that the Marshall-Edgeworth price index I_{ME} can be written as a weighted arithmetic mean of I_L and I_U, namely

$$I_{ME} = \frac{\sum_{i=1}^{N} p_i^s \max(q_i^s, q_i^t)}{\sum_{i=1}^{N} p_i^s \max(q_i^s, q_i^t) + \sum_{i=1}^{N} p_i^s \min(q_i^s, q_i^t)} I_L + \frac{\sum_{i=1}^{N} p_i^s \min(q_i^s, q_i^t)}{\sum_{i=1}^{N} p_i^s \max(q_i^s, q_i^t) + \sum_{i=1}^{N} p_i^s \min(q_i^s, q_i^t)} I_U \tag{17}$$

In fact, we can make a much more general observation – it can be proved (see Białek (2013)) that each of the above-mentioned indices (Fisher, Laspeyres, Paasche, Marhall-Edgeworth, Walsh formulas) have values between the *lower* and *upper index*. Thus, the formula I_B, as a geometric mean of I_L and I_U, seems to be well formed. In our simulation study (see Białek (2013)) it is shown, that the Fisher index and the Białek's price formula approximate each other. In the Section 4 we use the mentioned superlative price indices and the I_B index for calculating the CPI substitution bias in simulation studies.

4. Simulation study

Simulation 1

Let us take into consideration a group of $N = 50$ `commodities`, where random vectors of prices and quantities are as follows[1] (we present below only the first five commodities):

[1] The specification of prices and quantities does not mean that $p_i^t / p_i^s = a = const$ and $q_i^t / q_i^s = h(a) = const$, because random components of vectors P^t and Q^t are generated after the generating vectors P^s and Q^s. In other words, firstly we generate values of vectors of prices and quantities twice obtaining two pairs: (P^s, Q^s) and (P'^s, Q'^s) and after that we assume $P^t = aP'^s$ and $Q^t = h(a)Q'^s$.

$$P^s = [U(400,700), U(1000,6000), U(3,9), U(3000,7000), U(100,500),...]',$$

$$Q^s = [U(30000,70000), U(100,500), U(300,900), U(20000,50000), U(300,900),...]',$$

$$P^t = a \cdot [U(400,700), U(1000,6000), U(3,9), U(3000,7000), U(100,500),...]',$$

$$Q^t = h(a) \cdot [U(30000,70000), U(100,500), U(300,900), U(20000,50000), U(300,900),...]'$$

where $U(m,n)$ denotes a random variable with the uniform distribution which has values in an $[m,n]$ interval and $h(a)$ is some positive function of the parameter $a > 0$. We consider three cases:

Case 1: $h(a) = 1/a$ which means that prices and quantities are negatively correlated;

Case 2: $h(a) = 1$ which means that prices and quantities are uncorrelated and consumption is on the constant level;

Case 3: $h(a) = a$ which means that prices and quantities are positively correlated.

We consider these three cases although only the first one is the most common in practice. However, sometimes consumers stock up on commodities although they observe the rise in prices and in this case prices and quantities are positively correlated. We generate values of these vectors in $n = 100000$ repetitions. We get the results[1] presented in Tab. 1, Tab.2 and Tab. 3 and on Fig. 1.

Case 1

Table 1. Basic characteristics of the discussed CPI bias measures for the given values of parameter a

Characteristics	$I_{La} - I_F$	$I_{La} - I_T$	$I_{La} - I_W$	$I_{La} - I_B$
	$a = 0.2$			
Mean	0.003191	0.002693	0.003537	0.003323
Standard deviation	0.003921	0.003502	0.004200	0.004192
Volatility coefficient	1.228871	1.300712	1.187441	1.261510
	$a = 0.5$			
Mean	0.001521	0.001672	0.001381	0.001404
Standard deviation	0.005902	0.007801	0.005902	0.005898
Volatility coefficient	3.878860	4.665669	4.273692	4.199560

[1] To read more about mean value estimation and the bias of this estimation see Żądło (2006), Gamrot (2007), Małecka (2011) or Papież, Śmiech (2013).

Table 1. Basic characteristics of the discussed CPI bias measures for the given values of parameter a (cont.)

Characteristics	$I_{La} - I_F$	$I_{La} - I_T$	$I_{La} - I_W$	$I_{La} - I_B$
	$a = 1$			
Mean	0.002513	0.001069	0.002877	0.002756
Standard deviation	0.01165	0.01108	0.01179	0.01173
Volatility coefficient	4.62112	10.35642	4.09681	4.25761
	$a = 1.5$			
Mean	0.001744	0.002202	0.001910	0.001775
Standard deviation	0.017100	0.021628	0.017262	0.017180
Volatility coefficient	9.802840	9.821980	9.034320	9.676200
	$a = 2$			
Mean	0.006229	0.003213	0.006427	0.006033
Standard deviation	0.023712	0.022446	0.023920	0.023774
Volatility coefficient	3.806440	6.984651	3.721341	3.940150

Source: Own calculations using Mathematica 6.0.

Case 2

Table 2. Basic characteristics of the discussed CPI bias measures for the given values of parameter a

Characteristics	$I_{La} - I_F$	$I_{La} - I_T$	$I_{La} - I_W$	$I_{La} - I_B$
	$a = 0.2$			
Mean	-0.002382	-0.001877	-0.002532	-0.002517
Standard deviation	0.003318	0.002939	0.003414	0.003405
Volatility coefficient	1.392530	1.565551	1.348520	1.353140
	$a = 0.5$			
Mean	-0.002344	-0.002338	-0.002498	-0.002472
Standard deviation	0.006125	0.006040	0.006212	0.006156
Volatility coefficient	2.612710	2.582451	2.486670	2.490291
	$a = 1.5$			
Mean	-0.006760	-0.007030	-0.005637	-0.005642
Standard deviation	0.018090	0.017812	0.017742	0.017716
Volatility coefficient	2.675760	2.533680	3.147060	3.139941
	$a = 2$			
Mean	-0.004522	-0.004402	-0.005210	-0.005148
Standard deviation	0.023198	0.022387	0.023437	0.023359
Volatility coefficient	5.12930	5.084760	4.497961	4.537181

Source: Own calculations using Mathematica 6.0.

Case 3

Table 3. Basic characteristics of the discussed CPI bias measures for the given values of parameter a

Characteristics	$I_{La} - I_F$	$I_{La} - I_T$	$I_{La} - I_W$	$I_{La} - I_B$
	$a = 0.2$			
Mean	-0.001148	-0.001988	-0.001118	-0.001148
Standard deviation	0.002542	0.002984	0.002537	0.002542
Volatility coefficient	2.213850	1.500073	2.269740	2.213850
	$a = 0.5$			
Mean	-0.001893	-0.002050	-0.002111	-0.001893
Standard deviation	0.005976	0.005851	0.0061144	0.005984
Volatility coefficient	3.155621	2.853650	2.896052	3.160200
	$a = 1.5$			
Mean	-0.002294	-0.003171	-0.0024332	-0.002294
Standard deviation	0.0029232	0.003653	0.003022	0.002903
Volatility coefficient	1.274201	1.151940	1.242360	1.265571
	$a = 2$			
Mean	-0.003097	-0.003630	-0.002794	-0.002916
Standard deviation	0.003777	0.004212	0.003534	0.003627
Volatility coefficient	1.219540	1.160181	1.264651	1.243970

Source: Own calculations using Mathematica 6.0.

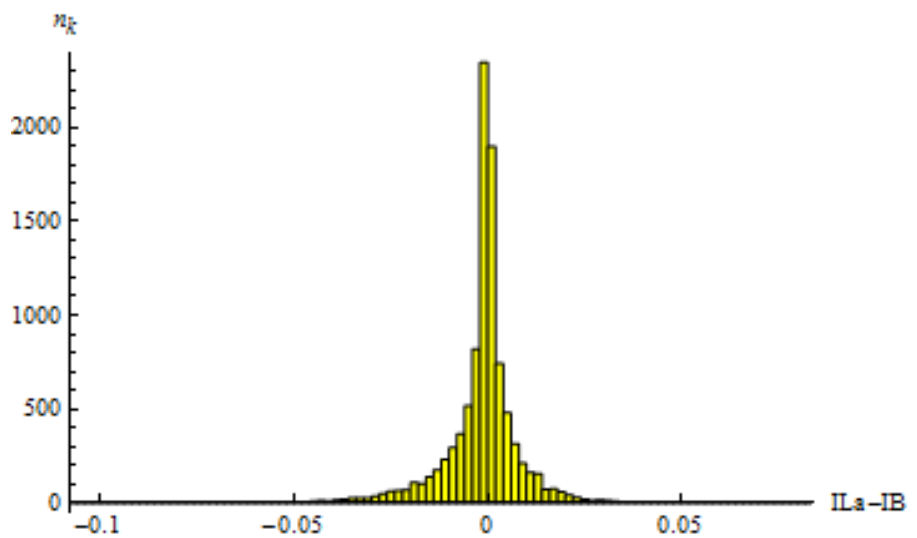


Figure 1. Histogram of $I_{La} - I_B$ in the case of $a = 1$

Source: Own calculations using Mathematica 6.0.

Simulation 2

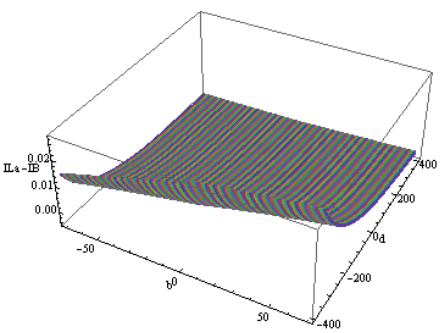
Let us take into consideration a group of only $N = 4$ commodities, where vectors of prices and quantities are as follows:

$P^s = [50,200,500,2500]'$, $P^t = [90,300,400,2000]'$,

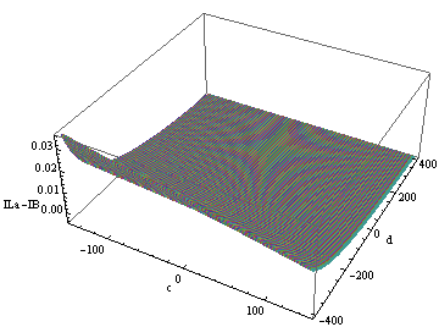
$Q^s = [300 + a, 90 + b, 200 + c, 500 + d]'$ and

$Q^t = [300,90,200,500]'$, for some real parameters a, b, c, d. The difference $I_{La} - I_B$ depending on these parameters is presented by Fig.2.

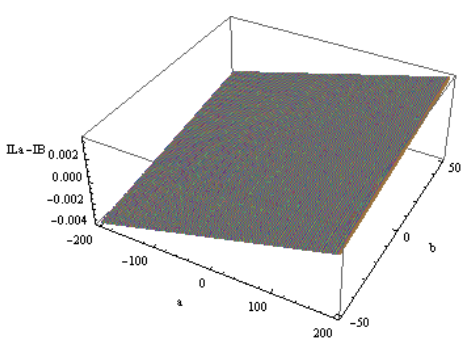
2.1. $c = d = 0$



2.2. $b = d = 0$



2.3. $b = c = 0$



2.4. $a = d = 0$

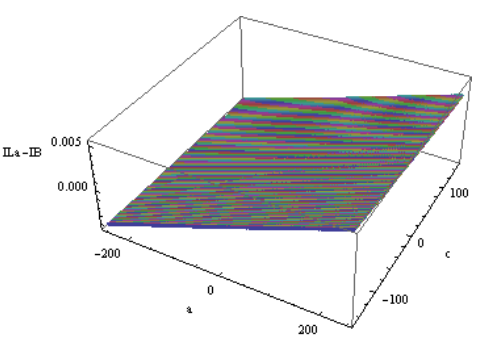


Figure 2. The CPI bias calculated as $I_{La} - I_B$ depending on parameters a, b, c and d

2.5. $a = c = 0$ 2.6. $a = b = 0$

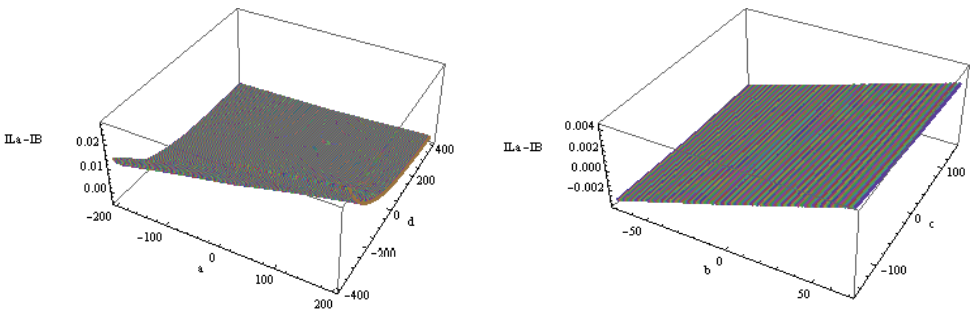


Figure 2. The CPI bias calculated as $I_{La} - I_B$ depending on parameters a, b, c and d (cont.)

Source: Own calculations using Mathematica 6.0.

5. Conclusions

In the simulation 1, case 1 (see Tab.1) we observe a positive expected value of the commodity substitution bias. This observation corresponds to the results coming from the report of the Boskin Commission, where we can find the estimated value of the commodity substitution bias on the level of 0,004 (see Boskin *et al.* (1996)). The similar conclusion can be also found in Cunnigham[1] (1996) or Crawford[2] (1998). We can notice that only in this case the estimator of the CPI bias calculated as $I_{La} - I_T$ is different in its expected value and has the highest volatility coefficient (the rest of estimators have similar values of this coefficient). Taking into consideration only the expected value of the commodity substitution bias we can find high similarity between the measures based on the Fisher and Białek formulas. Moreover, the scale of the commodity substitution bias does not seem to depend on the parameter a, which describes the changes in prices and quantities. In cases 2 and 3 (see Tab. 2 and Tab. 3 and Fig. 1) we observe negative expected value of the commodity substitution bias. Such situation can appear when the consumption does not depend on prices and is constant in time or quantities and prices are positively correlated. Then the real CPI is higher than the value obtained by Laspeyres formula (see researches by

[1] In this report the commodity substitution bias is in the interval 0 - 0.001.
[2] In this report the commodity substitution bias equals 0,001 and CPI is overestimated by 0.007.

Hałka, Leszczyńska[1] (2011), Ngasamiaku, Mkenda[2] (2009)). In the case 2 we can notice that bias measured by using the Walsh price index and the Białek index approximate one another. But the case 2 is not a real-life situation because it means that consumption remains on the constant level and is independent of changes in prices. It is worth noting that in case 3 we can also observe small differences between values of the CPI bias measures based on the Fisher and Białek formulas (as in the case 1). Taking into consideration good properties of the Białek formula (see also Białek (2012a), (2012b), (2013)) and the presented results we conclude that the I_B index can be a good alternative for the Fisher index in the CPI bias measurement.

In the simulation 2 we can notice that when parameters a, b, c and d increase then the Euklidean distance between quantities $d_e(Q^s, Q^t) = \sqrt{a^s + b^2 + c^2 + d^2}$ also increases and consequently the value of the difference $\left| I_{La}(Q^s, Q^t, P^s, P^t) - I_B(Q^s, Q^t, P^s, P^t) \right|$ becomes higher (see Fig. 2).

For $a = b = c = d = 0$ we have $I_{La}(Q^s, Q^t, P^s, P^t) - I_B(Q^s, Q^t, P^s, P^t) = 0$.

Let us also notice (see Fig. 2) that the CPI substitution bias is an increasing function of parameters a and b but a decreasing function of parameters c and d. When the value of any parameter increases, we observe that the consumption decreases. Thus, the above observation confirms the conclusion from the simulation 1 if we notice that the first two products have higher prices and the last two products have lower prices at time t compared with time s.

Note

This article is based on the paper presented at the 7th Scientific Conference on Modelling and Forecasting of Socio-Economic Phenomena, May 7-10, 2013, Zakopane, Poland.

[1] In this paper the commodity substitution bias equals - 0.1 or 0.
[2] In this report the commodity substitution bias equals about - 0.0027.

REFERENCES

AFRIAT, S. N., (1972). The Theory of International Comparisons of Real Income and Prices, 13–69, In D. J. Daly (ed.), International Comparisons of Prices and Outputs, New York: Columbia University Press.

BIAŁEK, J., (2012a). Propozycja indeksu cen, Statistical News 7, 13–24, Central Statistical Office, Warsaw.

BIAŁEK, J., (2012b). Proposition of the general formula for price indices, Communications in Statistics: Theory and Methods, vol. 41:5, s. 943–952.

BIAŁEK, J., (2013). Simulation study of an original price index formula, Communications in Statistics: Simulation and Computation (in press, DOI: 10.1080/03610918.2012.700367).

BOSKIN, M. J., DULBERGER E. R., GORDON R. J., GRILICHES Z., JORGENSON D., (1996). Toward a More Accurate Measure of the Cost of Living, Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index.

CLEMENTS, K., IZAN H. Y., (1987). The Measurement of Inflation: A Stochastic Approach, Journal of Business and Economic Statistics, 5(3), 339–350.

DIEWERT, W. E., (1974). Functional forms for revenue and factor requirement functions, International Economic Review, 15, 119–130.

DIEWERT, W. E., (1976). Exact and superlative index numbers, Journal of Econometrics, 4, 114–145, North-Holland Publishing Company.

DIEWERT, W. E., (1983). The theory of the cost-of-living index and the measurement of welfare change, Price level measurement: proceedings of a conference sponsored by Statistics Canada, Eds. W. E. Diewert and C. Montmarquette, 163–233, Ottawa: Statistics Canada.

CRAWFORD, A., (1998). Measurement Biases in the Canadian CPI: An Update, Bank of Canada Review, Spring, 38–56.

CUNNINGHAM, A. W., (1996). Measurement biases in price indexes: an application to the UK's RPI, Bank of England, Working Paper Series 47.

DIEWERT, W. E., (1993). The economic theory of index numbers: a survey, Essays in index number theory, vol. 1, Eds. W. E. Diewert and A. O. Nakamura, 177–221, Amsterdam.

FRISCH, R., (1936). Annual survey of general economic theory: The problem of index numbers, Econometrica, 4, 1–39.

GRAMOT, W., (2007). Mean Value Estimation Using Two-Phase Samples with Missing Data in Both Phases, Acta Appl Math 96, 215–220.

HAŁKA, A., LESZCZYŃSKA, A., (2011). Wady i zalety wskaźnika cen towarów i usług konsumpcyjnych – szacunki obciążenia dla Polski, Gospodarka Narodowa 9, 51–75.

JORGENSON, D. W., SLESNICK, D. T., (1983). Individual and social cost of living indexes, Price level measurement: proceedings of a conference sponsored by Statistics Canada, Eds. W.E. Diewert and C. Montmarquette, 241–336, Ottawa: Statistics Canada.

KONYUS, A., BYUSHGENS, S., (1926). K probleme pokupatelnoi cili deneg, Voprosi Konyunkturi II: 1, 151–172.

MAŁECKA, M., (2011). Prognozowanie zmienności indeksów giełdowych przy wykorzystaniu modelu klasy GARCH, Ekonomista 6, 843–860.

NGASAMIAKU, A., MKENDA, W., (2009). An Analysis of Alternative Weighting System on the National Price Index in Tanzania: The Imlication to Poverty Analysis, Botswana Journal of Economics, 6, no. 10, 50–70.

PAPIEŻ, M., ŚMIECH, S., (2013). Causality in mean and variance within the international steam coal market, Energy Economics 36, 594–604.

POLLAK, R. A., (1971). The theory of the cost of living index, Research Discussion Paper no. 11, Office of Prices and Living Conditions, U. S. Bureau of Labor Statistics, Washington.

POLLAK, R. A., (1989). The theory of the cost-of-living index, Oxford: Oxford University Press.

SAMUELSON, P. A., SWAMY, S., (1974). Invariant economic index numbers and canonical duality: Survey and synthesis, American Economic Review 64, 566–593.

WHITE, A. G., (1999). Measurement Biases in Consumer Price Indexes, International Statistical Review, 67, 3, 301–325.

ŻĄDŁO, T. (2006). On prediction of total value in incompletely specified domains, Australian & New Zealand Journal of Statistics 48(3), 269–283.

VON DER LIPPE, P., (2007). Index Theory and Price Statistics, Peter Lang, Frankfurt, Germany.

VON DER LIPPE, P., (2012). Some short notes on a price index of Jacek Białek, Ekonometria (Econometrics), 35, 76–83.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 97–110

WINSORIZATION METHODS IN POLISH BUSINESS SURVEY

Grażyna Dehnel[1]

ABSTRACT

One of the major problems involved in estimating information about economic activity across small domains is too small sample size and incompleteness of data sources. For instance, the distribution of enterprises by target variables tends to be considerably right-skewed, with high variation, high kurtosis and outliers. Therefore, it is not obvious that the implementation of traditional estimation methods meets the desired requirements, such as being free from bias or having competitive variance. Furthermore, the pressure to produce accurate estimates at a low level of aggregation or needs to substantially reduce sample size have increased the importance of exploring the possibilities of applying new, more sophisticated methods of estimation. The aim of the study was to test the usefulness of winsorization methods to estimate economic statistics from the DG1 survey.

Key words: domain estimation, business statistics, winsorized estimator.

1. Introduction

Nowadays the growing demand for business information at a low level of aggregation has called for estimation methods that could meet the requirements specified by the user's needs. In practice, business surveys often pose a variety of data problems. For example, target variables tend to be highly skewed and populations can contain a number of extreme values, the so-called outliers. Although outliers are extreme, they need not necessarily be incorrect but are an integral part of each survey population and cannot be dismissed in the analysis. Since outliers usually have a huge impact on estimates, outlier detection and their treatment are important elements of statistical analysis. This is true especially when estimation is carried out for small domains. In the case of small sample size, outliers can result in estimates greatly diverging from the real value for the population. Even if the sample size is large, the influence of an outlier can significantly increase the variance resulting in a decreased efficiency of estimation.

[1] Poznan University of Economics, Department of Statistics. E-mail: g.dehnel@ue.poznan.pl.

Dealing with outliers has two aspects: the first one involves identifying outlying observations in an objective way, while the second focuses on ways of handling them to reduce their effect on survey estimates.

There are three main methods of dealing with outliers in a finite population (Cox, 1995): reducing the weights of outliers (trimming weight), changing the values of outliers (winsorization, trimming), using robust estimation techniques such as M-estimation.

The paper describes implementation of winsorization - one frequently applied estimation method, used to reduce the impact of outlying units. The general idea of winsorization is that if an observation exceeds a preset cutoff value, then the observation is replaced by that cutoff value or by a modified value closer to the cutoff value.

The objective of the referred study was to assess the performance of four various methods use to estimate robust regression parameters, and hence estimate the cutoff values used in the winsorized estimator. The paper presents attempts to estimate basic economic information about small, medium-sized and large businesses at a low level of aggregation (in the joint cross-section of economic activity classification and the territorial division by province).

2. Estimation method

Winsorization is often used for data cleaning in statistical practice. Since outliers are a serious problem in many sample surveys (especially business surveys), an appropriate way of handling them is required. Winsorization involves identifying cutoff values. Sample observations whose values lie outside certain preset cutoff values are transformed in order to make them closer to the cutoff value.

Cutoff values are derived in a way that approximately minimizes the MSE of estimates. All sampled units are divided into two groups. One group contains typical observations which are left unmodified, the other one contains observations regarded as outliers. The classification is made on the basis of two preset cutoff values. Then, values of the study variable outside the cutoff values are transformed so that they are no longer regarded as outliers. It should be stressed, however, that the modified values are artificial and may sometimes be unacceptable. As a result of the winsorized estimation, we obtain a „new" sample, in which untypical observations have been replaced with typical ones. Further calculations are conducted for the modified sample. Any kind of estimation can be used at this stage. Here, GREG estimation is illustrated.

The winsorized estimator, with GREG estimation, can be expressed as:

$$\hat{Y}_{win} = \sum_{i \in s_d} \tilde{w}_i y_i^* = \sum_{i \in s_d} w_i g_i y_i^* \tag{1}$$

where, in the presence of outliers, modified values of the study variable y_i^* are calculated in the following manner (Gross, Bode Taylor, Lloyd-Smith, 1986):

$$y_i^* = \begin{cases} \left(\frac{1}{\widetilde{w}_i}\right) y_i + \left(1 - \frac{1}{\widetilde{w}_i}\right) K_{Ui} & \text{if} \quad y_i > K_{Ui} \\ y_i & \text{if} \quad K_{Li} \le y_i \le K_{Ui} \\ \left(\frac{1}{\widetilde{w}_i}\right) y_i + \left(1 - \frac{1}{\widetilde{w}_i}\right) K_{Li} & \text{if} \quad y_i < K_{Li} \end{cases} \tag{2}$$

$$g_i = \left(1 + x_i' \left(\sum_{i \in s_d} w_i x_i x_i'\right)^{-1} \left(t_x - \sum_{i \in s_d} w_i x_i\right)'\right) \tag{3}$$

where:

s_d - population parameter for domain *d*

$U = \{1,.....i,.....N\}$ - general population of size *N*

$s(s \subseteq N)$ - sample

$\widetilde{w}_i = w_i g_i$

$w_i = 1/\pi_i$ - sampling weights

g_i - weights dependent on the value of a vector of auxiliary variables for sampled units

$x_i = (x_{1i},..., x_{ki},..., x_{Ki})'$ - vector of auxiliary variables

$t_x = \sum_{i \notin U} x_i$ - population total

K_{Ui} - upper cutoff value

K_{Li} - lower cutoff value

The cutoff values are calculated to minimize MSE of the winsorized estimator under the model (Preston, Mackin, 2002):

$$K_{Ui} = \mu_i^* - \frac{B_U}{(\widetilde{w}_i - 1)} \tag{4}$$

$$K_{Li} = \mu_i^* - \frac{B_L}{(\widetilde{w}_i - 1)} \tag{5}$$

where:

$\mu_i^* = E(Y_i^*)$ - expectation under the assumed model

$B_U = E\left[\hat{Y}_{winU} - \hat{Y}_{DIR}\right]$ - bias of $\hat{Y}_{winU}$

$B_L = E\left[\hat{Y}_{winL} - \hat{Y}_{DIR}\right]$ - bias of $\hat{Y}_{winL}$

$\hat{Y}_{winU}$ - the winsorized estimator of the population total when only upper winsorization is performed

$\hat{Y}_{winL}$ - the winsorized estimator of the population total when only lower winsorization is performed.

When winsorization is mild and reasonably symmetric, being μ_i^* difficult to estimate, we can replace μ_i^* with μ_i . Then, the approximately optimal cutoffs are (Preston, Mackin, 2002):

$$K_{Ui} = \mu_i - \frac{B_U}{(\tilde{w}_i - 1)} = \mu_i + \frac{G}{(\tilde{w}_i - 1)} \tag{6}$$

$$K_{Li} = \mu_i - \frac{B_L}{(\tilde{w}_i - 1)} = \mu_i + \frac{H}{(\tilde{w}_i - 1)} \tag{7}$$

Under the assumption $\mu_i = \hat{\mu}_i = \hat{\beta} x_i$ (Preston, Mackin, 2002) the cutoff values are estimated based on the following formulas:

$$\hat{K}_{Ui} = \hat{\mu}_i - \frac{B_U}{(\tilde{w}_i - 1)} = \hat{\mu}_i + \frac{G}{(\tilde{w}_i - 1)} \text{ where } G = -B_U \tag{8}$$

$$\hat{K}_{Li} = \hat{\mu}_i - \frac{B_L}{(\tilde{w}_i - 1)} = \hat{\mu}_i + \frac{H}{(\tilde{w}_i - 1)} \quad \text{where } H = -B_L \tag{9}$$

where $\hat{\mu}_i = \hat{\beta} x_i$ - a robust estimate of regression parameter μ_i (see below).

In order to estimate the bias parameter B_U under winsorization we can use the Kokic and Bell approach (1994). According to that approach, the value of B_U can be calculated by solving the equation:

$$G - E\left[\sum_{i \in s} \max\{D_i - G, 0\}\right] = 0 \tag{10}$$

where $D_i = (Y_i - \mu_i^*)(\tilde{w}_i - 1)$ are weighted residuals. Assuming $\hat{\mu}_i$ is a robust estimate of parameter μ_i , we obtain $\hat{D}_i = (Y_i - \hat{\mu}_i)(\tilde{w}_i - 1)$.

We can write the function $\psi_U\left(\hat{D}_{(k)}\right)$ (Kokic, Bell, 1994).

$$\psi_U\left(\hat{D}_{(k)}\right) = \hat{D}_{(k)} - \sum_{i \in s} \max\left\{\hat{D}_i - \hat{D}_{(k)}, 0\right\} = (k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)} \tag{11}$$

where:

(k) - a number assigned to the unit drawn into the sample after ordering all units in the sample according to non-ascending estimated residuals $\hat{D}_i$: $\hat{D}_{(1)} \geq \hat{D}_{(2)} \geq ... \geq 0 \geq ....$ By solving $\psi_U(G) = 0$ one can obtain the value of G.

In order to estimate the cutoff values $\hat{K}_{Ui}$ and $\hat{K}_{Li}$, in addition to the above bias parameters $G = -B_U$ and $H = -B_L$, it is necessary to compute $\hat{\mu}_i = \hat{\beta}x_i$ which is an estimate of μ_i^*. For this purpose, robust regression methods can be used. Those recommended in the literature (Preston, Mackin, 2002) include: *Trimmed least squares (TLS), Trimmed least absolute value (ABS), Sample Splitting (HALF), Least median of squares (LMS).*

The method of ***Trimmed least squares (TLS)*** involves first fitting an Ordinary Least Squares (OLS) regression model to minimise the function:

$$F = \sum_{i \in s} \left(y_i - \beta^T x_i \right)^2 \tag{12}$$

Then fitted values are calculated, and then residuals. In the second step, units with the largest positive and negative residuals are removed. Finally, a new regression model is fitted to the reduced sample in order to estimate the value of μ_i^*.

Another method used in robust regression is ***Trimmed least absolute value (ABS)***. It consists in fitting a regression model to minimise the function:

$$F = \sum_{i \in s} \left| y_i - \beta^T x_i \right| \tag{13}$$

After evaluating fitted values and residuals, as is the case in the TLS method, units with the largest positive and negative residuals are removed. A new regression model is fitted to the reduced sample. It is expected that the *ABS* method is a more robust regression model than the *TLS* technique because large residuals which are not squared have less influence on the regression parameters.

Another example of robust regression is ***Sample Splitting Technique (HALF)*** based on Ordinary Least Squares (OLS). It is applied to data that has been randomly split into two halves. A regression model is fitted to each half of the data while the residuals are calculated using the model applied to the half of the data that was not used to fit the model. Then, after merging the data, units with the largest positive and negative residuals are removed. The process is repeated until a certain percentage of data has been deleted. The *HALF* technique is expected to be more robust than TLS because the residuals used to remove the 'outlier' units are not calculated from the regression model that has been generated using these 'outlier' units.

The list of robust regression techniques cannot be complete without the ***Least median of squares (LMS)*** technique. It was described by Rousseeuw and Leroy [2003]. It resembles the bootstrap method. It involves drawing subsamples of size $n - 1$ from a sample of size n using simple random sampling with replacement. For each subsample trial regression model parameters are calculated and then their squared residuals, which are used to calculate the median. The model with the smallest median of squared residuals is selected. The *LMS* technique should be more robust than TLS because an OLS regression model is fitted in the absence of ”outlier” units, without totally removing these “outlier” units (Preston, Mackin, 2002).

3. Data source

Information for the study came from the DG1 survey conducted by the Statistical Office in Poznan. The survey is conducted in the form of monthly reports submitted by all large and medium-sized enterprises and a 10% sample of small enterprises. Its objective is to collect up-to-date information about basic indicators of economic activity of enterprises, such as *revenue from sales (of products and services), number of employees, gross wages, volume of wholesale trade and retail sales, excise tax, specific subsidies.* The sample frame includes 98,000 units, of which 19,000 are medium-sized and large enterprises (with over 49 employees), 80,000 are small enterprises (from 10 to 49 employees). In effect, about 30,000 units participate in the survey every month.

4. Description of the study

The study was limited to enterprises that were active in August of 2012. *Gross wages* were the target variable, while *revenue from sales of products (goods and services)* was the auxiliary variable.

The general population included all enterprises that participated in the DG1 survey. This choice enabled access to detailed information about the target and auxiliary variables. With the general population defined in this way, it was possible to conduct a simulation study, which was then used to evaluate estimation precision.

The level of aggregation adopted for the study was a combination of economic activity classification (NACE Rev.2) and the territorial division by province.

5. Precision assessment methods

The precision of estimators analysed in the study was evaluated using the bootstrap method. 1000 iterations of drawing 20% samples were made, which

were then used to calculate:

- Relative estimation error (REE)

$$CV\left(\hat{Y}_d\right)=\frac{\sqrt{Var\left(\hat{Y}_d\right)}}{E\left(\hat{Y}_d\right)} \tag{14}$$

where:
$$Var\left(\hat{Y}_d\right)=\frac{1}{999}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d}-\hat{Y}_d\right)^2 \tag{15}$$

- Mean absolute relative bias (ARB)

$$ARB\left(\hat{Y}_d\right)=\frac{1}{1000}\left|\sum_{b=1}^{1000}\frac{\hat{Y}_{b,d}-Y_d}{Y_d}\right| \tag{16}$$

- Relative root mean square error (RMSE)

$$RMSE\left(\hat{Y}_d\right)=\frac{\sqrt{\frac{1}{1000}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d}-Y_d\right)^2}}{Y_d} \tag{17}$$

- To describe the general precision of combined estimates, for all small areas, mean values of Relative root mean square error applied to particular domains were calculated. The mean values were calculated as arithmetic means used in empirical studies and as weighted means

$$\overline{RMSE}=\frac{1}{D}\sum_{d=1}^{D}RMSE(\hat{Y}_d)$$

Owing to the large volume of estimation results, the following presentation is limited to estimates for the variable of *gross wages* for two PKD categories: *manufacturing*, *construction* and *trade*.

6. Estimation results and assessment of their precision

The effect of different winsorized estimation techniques on the value of the study variable is shown on a scatterplot (see Fig. 1). To illustrate the shift in values as a result of modification, only domains for *manufacturing* have been selected. Empirical values of the study variable in domains are marked by a black cross. Each domain is represented by five points: the real value and values modified as a result of each of the four robust regression techniques. The degree of modification depends on the type of robust regression technique. It is also worth noting that in nearly all the cases the HT estimates were significantly different from the winsorized estimates.

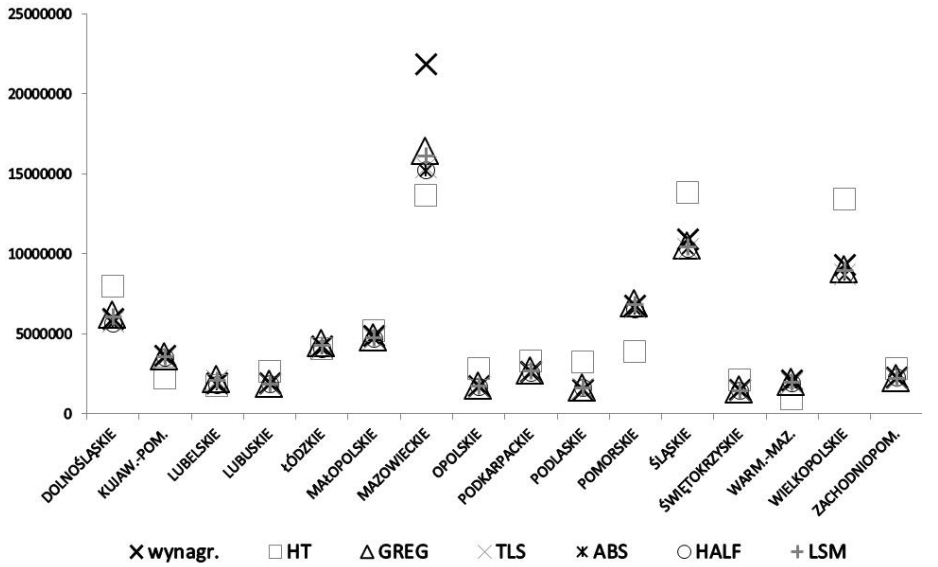


Figure 1. Real values (*Y*- Gross Wage) and values estimated by winsorization
Source: Own calculations based on DG1 survey, data from August 2012.

The scatterplot shows both the direction and the degree of modification of the study variable. In the case of units classified as *x-outliers*, namely for small values of wages paid by businesses with high revenue, the modification involved increasing the value of the study variable. The study variable was decreased in the case of outliers corresponding to businesses paying high wages but reporting low revenue.

Figures 2-7 present the distribution of three performance criteria: relative estimation error, mean absolute relative bias and relative root mean square error for two analyzed sections: *construction* and *trade*. From the results in Fig. 2 and 3 we can see that in most cases the winsorized estimator has considerably less REE than the HT and GREG estimators.

The amount of bias induced by winsorizing is for most cases almost insignificant except in the case of province characterised by high variation of the auxiliary variable (see Fig. 4 and 5). In terms of RMSE, the performance of the winsorized estimators is considerably better than the HT and GREG estimator (see Fig. 6 and 7).

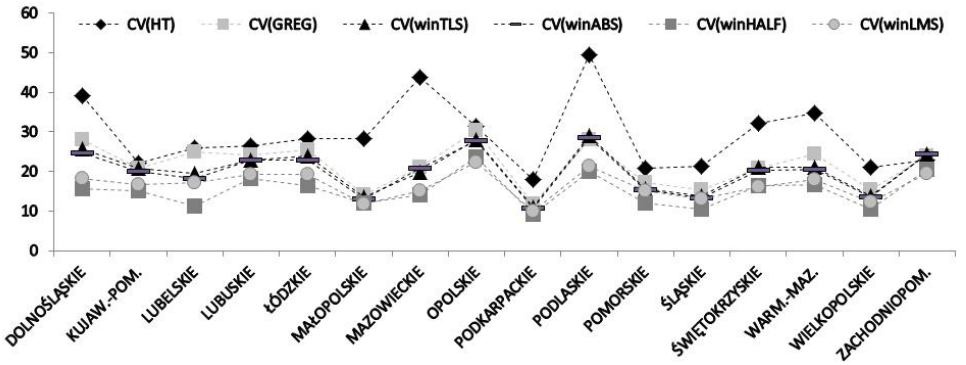


Figure 2. Relative estimation error for *construction*

Source: Own calculations based on DG1 survey, data from August 2012.

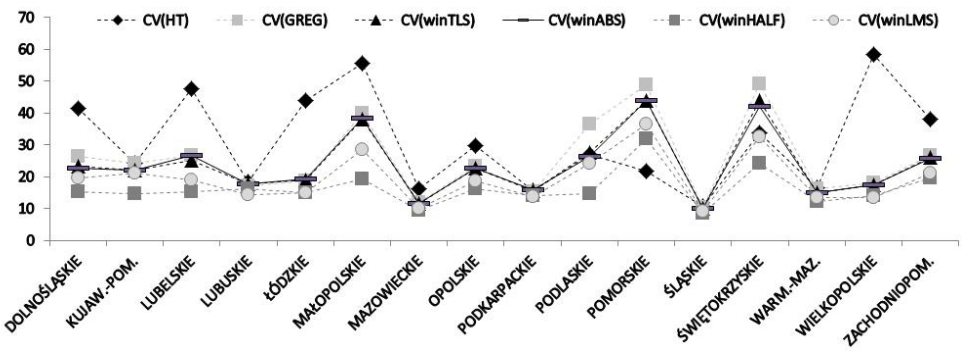


Figure 3. Relative estimation error for *trade*

Source: Own calculations based on DG1 survey, data from August 2012.

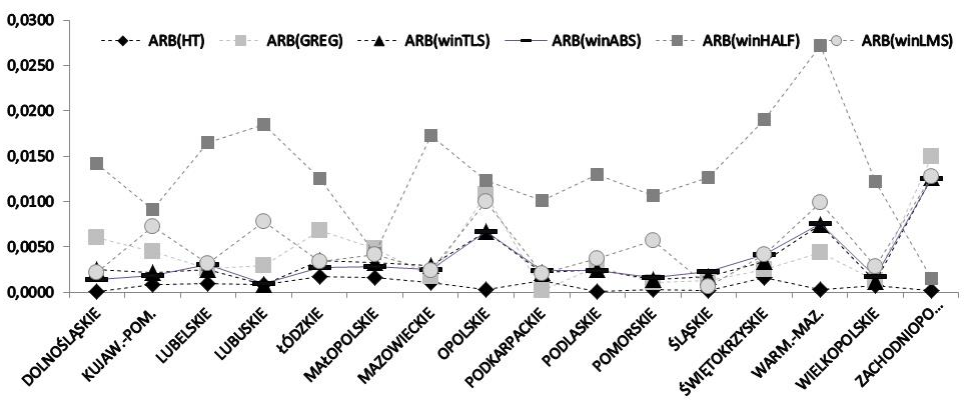


Figure 4. Mean absolute relative bias for *construction*

Source: Own calculations based on DG1 survey, data from August 2012.

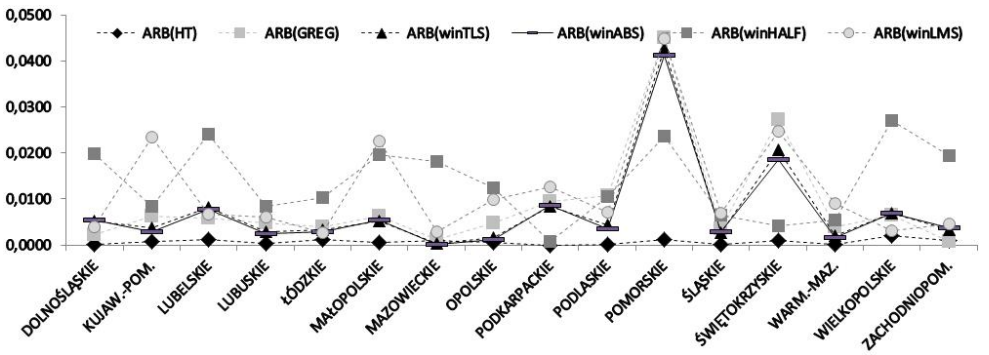


Figure 5. Mean absolute relative bias for *trade*

Source: Own calculations based on DG1 survey, data from August 2012.

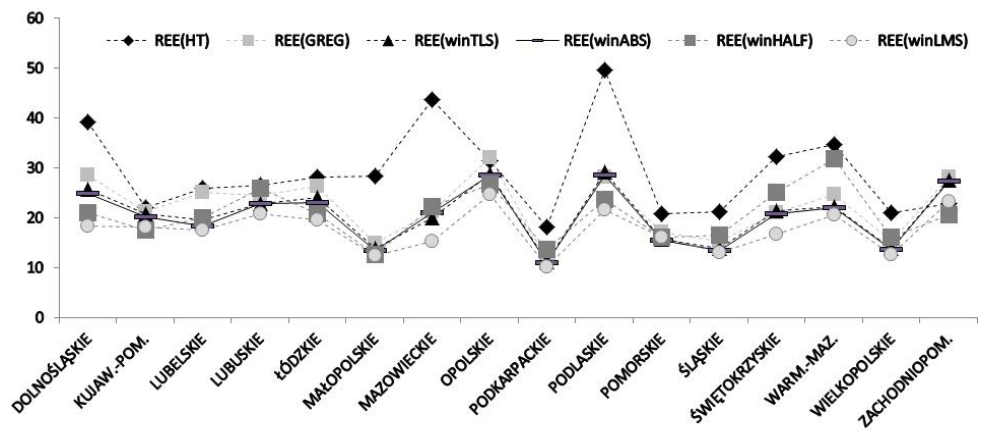


Figure 6. Relative root mean square error for *construction*

Source: Own calculations based on DG1 survey, data from August 2012.

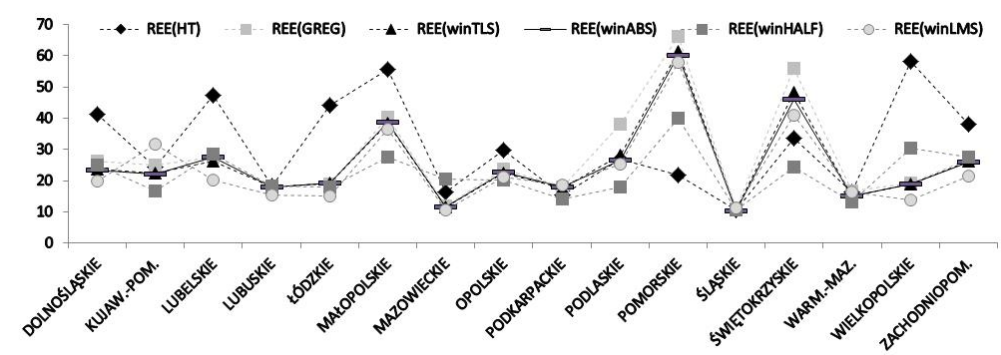


Figure 7. Relative root mean square error for *trade*

Source: Own calculations based on DG1 survey, data from August 2012.

For most values of the estimated cutoffs (calculated according to the four various methods of estimate robust regression parameters), the winsorized estimator significantly outperformed the expansion estimators (see Tab. 1). There are very few cases when the HT and GREG estimation is better than the winsorized estimator. The winsorized estimator nearly always had considerably smaller RMSE than the expansion estimators.

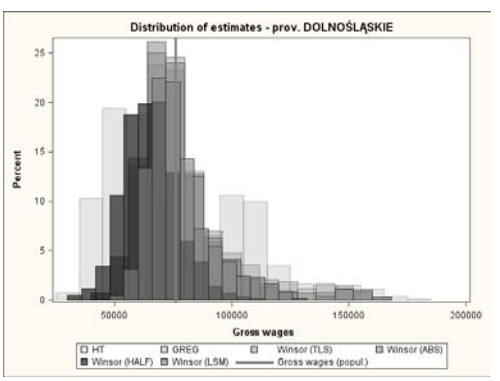
The results described above indicate that winsorizing optimize trade-off between variance and bias. The improvement in the general performance of the estimator that is obtained against extremely large errors from winsorizing is usually at the price of introducing a small amount of bias in estimation.

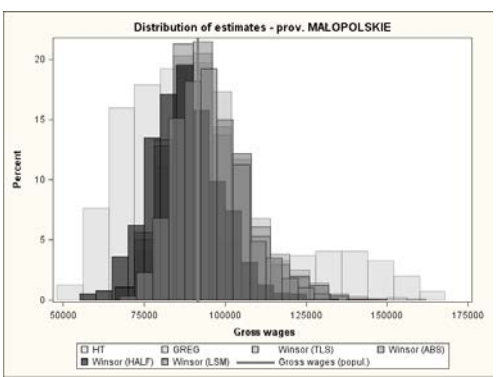
Table 1. Relative root mean square error for *construction* and *trade*

$\overline{RMSE}$	**HT**	**GREG**	**winTLS**	**winABS**	**winHALF**	**winLMS**
Construction	29.1	22.2	20.7	20.3	20.6	17.6
	$RMSE<RMSE_{HT}$ (%)	88	94	94	100	94
Trade	31.1	27.7	25.5	25.2	22.1	23.5
	$RMSE<RMSE_{HT}$ (%)	56	75	81	88	69

Source: Own calculations based on DG1 survey, data from August 2012.

Figures 8-9 present the distribution of estimates for selected provinces for *construction* and *trade*. The use of the winsorized estimation reduces estimator variance compared to direct estimation. The distribution of the winsorized estimates is significantly more leptokurtic than DIRECT or GREG estimates. In many cases it follows the normal distribution while the distribution of DIRECT or GREG estimators is sometimes multimodal or highly skewed. It is very difficult to point out which type of the winsorized estimators has better properties based on the presented figures.





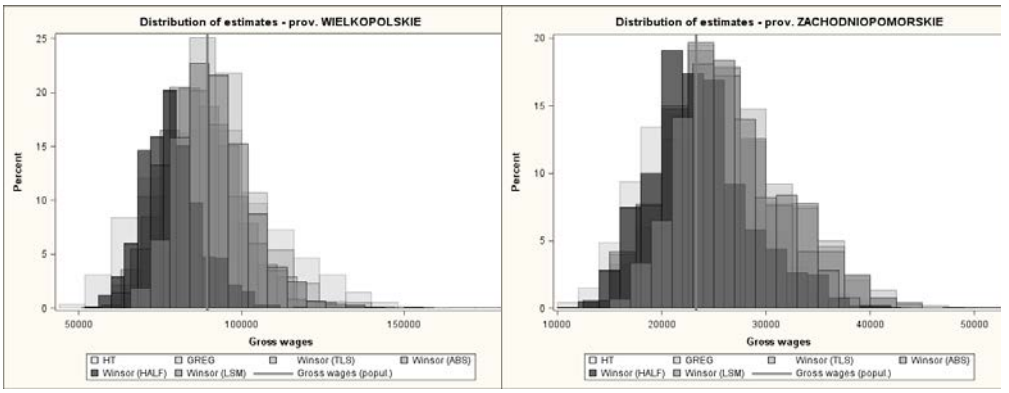


Figure 8. Distribution of estimates for selected provinces for *construction*
Source: Own calculations based on DG1 survey, data from August 2012.

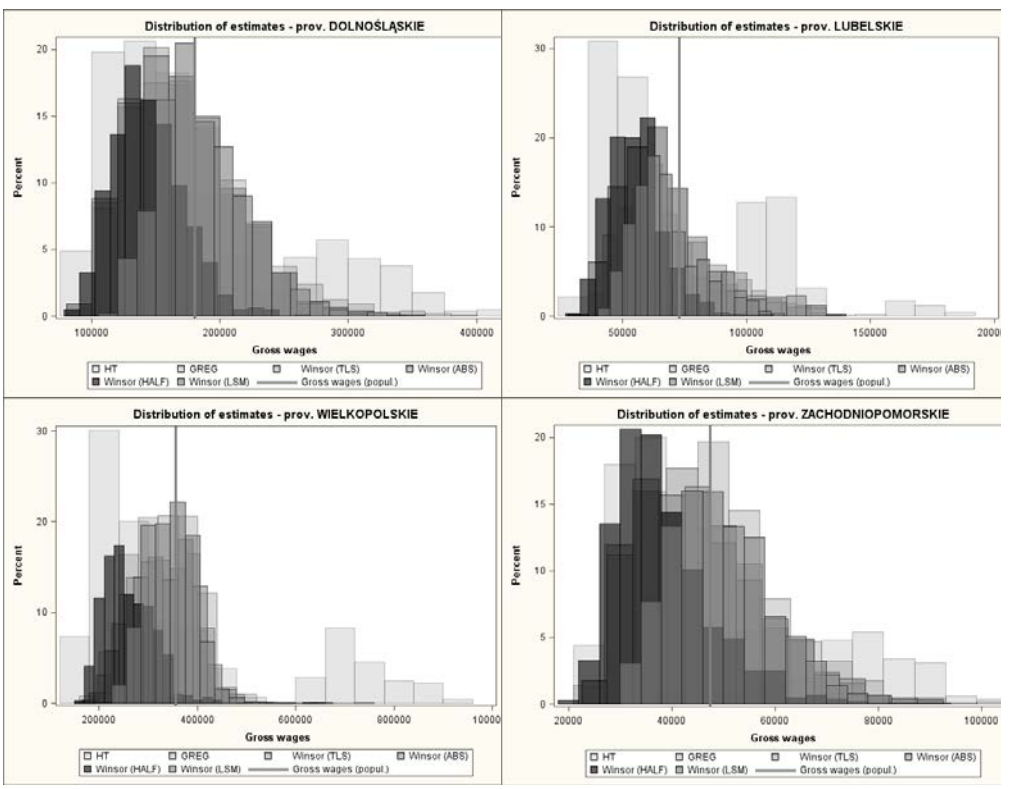


Figure 9. Distribution of estimates for selected provinces for *trade*
Source: Own calculations based on DG1 survey, data from August 2012.

7. Conclusion

- Simulation research demonstrated the relation between efficiency of estimation and a type of robust regression technique used.
- The effectiveness of the winsorized estimator in terms of its resistance to unusually large residuals depends on the choice of cutoff values - in other words, on methods of estimating bias parameters and regression parameters. The more robust regression technique was applied, the more efficient estimates were produced.
- The use of the winsorized estimation reduces estimator variance.
- Winsorization reduces outliers values, producing an insignificant estimated bias in the characteristic estimates.
- If cutoff values are chosen appropriately, the decline in variance is big enough to offset the bias of MSE. The winsorized estimator nearly always outperforms the expansion estimator in terms of MSE.

REFERENCES

CHAMBERS, R., KOKIC, P., SMITH, P., CRUDDAS, M., (2000). Winsorization for Identifying and Treating Outliers in Business Surveys, Proceedings of the Second International Conference on Establishment Surveys (ICES II), 687–696.

COX, B. G., BINDER, A., CHINNAPPA, N. B., CHRISTIANSON, A., COLLEDGE, M. J., KOTT, P. S., (1995). Business Survey Methods, John Wiley & Sons.

GROSS, W. F., BODE, G., TAYLOR, J. M., LLOYD–SMITH, C. W., (1986). Some finite population estimators which reduce the contribution of outliers, [in:] Proceedings of the Pacific Statistical Conference, 20–24 May 1985, Auckland, New Zealand.

KOKIC, P. N., BELL, P. A., (1994). Optimal winsorizing cutoffs for a stratified finite population estimator, Journal of Official Statistics, 10, 419–435.

PRESTON, J., MACKIN, C., (2002). Winsorization for Generalised Regression Estimation, Australian Bureau of Statistics.

PRESTON, J., MACKIN, C., (2002). Winsorization for Generalised Regression Estimation, Paper for the Methodological Advisory Committee, November 2002, Australian Bureau of Statistics.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 111–132

SPARSE METHODS FOR ANALYSIS OF SPARSE MULTIVARIATE DATA FROM BIG ECONOMIC DATABASES

Daniel Kosiorowski[1], **Dominik Mielczarek**[2], **Jerzy Rydlewski**[2], **Małgorzata Snarska**[3]

ABSTRACT

In this paper we present a novel perspective dedicated for *sparse* high-dimensional *data* sets, i.e. data which contain many zeros among coordinates of observations. Using jointly, selected *sparse methods* recently proposed in multivariate statistics, and kernel density framework for discrete data, we outline a general perspective for bringing out useful information from big economic databases. As a framework for our considerations we take the so-called functional data analysis, which originates from Ramsay and Silverman works. In particular we use functional principal components analysis within 2D density estimation procedure proposed by Simonoff.

Key words: sparse data, sparse methods, robust methods, categorical data, big data.

1. Introduction

In recent years several authors have investigated the use of smoothing methods for sparse multinomial data. In his excellent paper Simonoff (1983) considered probabilities in a large one-dimensional sparse contingency table estimated by maximizing the likelihood modified by a roughness penalty. It was shown in his paper that if certain smoothness criteria on the underlying probability vector are fulfilled, the maximum penalty estimator is consistent in a one-dimensional table under a sparse asymptotic framework. However, a proof of sparse asymptotic consistency for multidimensional tables was not found. It was shown that the bias of kernel estimates of probabilities for cells near the

[1] Department of Statistics, Faculty of Management, Cracow University of Economics. E-mail: daniel.kosiorowski@uek.krakow.pl.
[2] Department of Differential Equations, Faculty of Applied Mathematics, AGH University of Science and Technology. E-mail: dmielcza@wms.mat.agh.edu.pl.
[3] Department of Capital Markets, Faculty of Finance, Cracow University of Economics. E-mail: snarskam@uek.krakow.pl.

boundaries of the multinomial vector often dominates the mean sum of the squared error of the estimator. However, boundary kernels contrived to correct boundary effects for kernel regression estimators can achieve the same result for these estimators. Dong and Simonoff (1994) investigated the properties of estimators based on boundary kernels and compared them to unmodified kernel estimates and maximum penalized kernel likelihood estimates. They showed that the boundary-corrected estimates usually outperform uncorrected kernel estimates and are quite competitive with penalized likelihood estimates. Shane and Simonoff (2001) considered categorical data analysis using maximum likelihood. The problem with maximum likelihood estimates is their sensitivity to outlier cells. For this reason robust alternatives to maximum likelihood estimation were proposed in Shane and Simonoff (2001). The methods include the least median of chi-squared residuals, the least median of weighted squared residuals, and methods using the least trimmed functions. They also considered equivariance and breakdown properties of the estimators. They showed that the maximum likelihood estimates break down in the presence of outlying cells, while robust estimators do not as long as the contamination point does not exceed the breakdown point. Simonoff (1998) focused on nonparametric estimation of smooth functions. He considered categorical data smoothing and constructed effective categorical likelihood smoothing estimates. He also used an appropriate likelihood function yielding cell probability estimates with many desirable properties. Such estimates can be used to construct well-behaved density estimates using local or penalized likelihood estimation. Simonoff (1998) showed advantage of the local polynomial likelihood density estimate over the penalized likelihood density estimate. Namely, it is the structure which can be manipulated to allow local variation in the amount of smoothing.

In this paper we consider the estimator of the bivariate density function proposed in Simonoff (1988) and its modifications in the context of data mining in huge economic databases which may contain outliers.

2. Estimator of two-dimensional density function

Models using categorical data usually assume that there is no relation between adjacent cells. This is not the case for continuous distributions, where many estimation procedures are based on the fact that observations falling near the approximation site do give some information about the function we are trying to estimate, whether this is a density or a regression function. This information by proximity is at the base of the modifications that have been proposed to the histogram. The classical kernel or local polynomial estimators are, in fact, clever ways to use this idea to improve upon rough estimates. This idea has been used to smooth over discrete distributions, with increased interest when few observations are available when compared with the number of cells of the underlying distribution, or when the observations tend to concentrate too much in a few cells

of the support, indicating that the underlying distribution is quite peaked. Smoothing over adjacent cells does contribute to improve estimators in the similar cases. For one-dimensional distributions Simonoff (1983), Hall and Titterington (1987) smoothed the histogram with a uniform-like distribution, and Burman (1987) discretized the kernel estimator. More recently Simonoff (1995, 1996), Dong and Simonoff (1995) or Aerts et al. (1997) studied discrete versions of local polynomial estimators for higher dimensional data. Jacob and Oliveira (2011) used the local polynomial approach but with respect to a relativized L_2 - error, showing good performance for one-dimensional data. The extension of these methods to higher dimensional data introduces some difficulties.

Assume we consider objects with respect to (w.r.t.) two variables X_1 and X_2, and our aim is to estimate their joint probability density function. Our starting point is the estimator proposed in Simonoff (1995), which is based on binning the data and dedicated to sparse continuous data. Simonoff proposes to divide the range of X_1 into n_1 bins, the *i*-th bin being called I_{1i}, and to divide the range of X_2 into n_2, the *j*-th bin being called I_{2j}.

Table 1. Illustration for binning 2D continuous data

X_1 / X_2	I_{21}	I_{22}	…	I_{1k_2}	**total**
I_{11}	n_1	n_{k_1+1}	…		
I_{12}	n_2	n_{k_1+2}	…		
⋮					
I_{1k_1}	n_{k_1}	n_{2k_1}	…	$n_{k_2 \cdot k_1}$	
total					

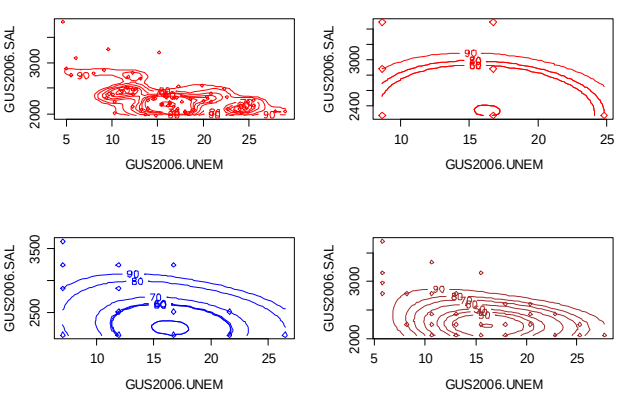


Figure 1. 2D kernel density estimates for binned data, unemployment vs. mean salary in Polish subregions in 2006

Next, let us consider $f_{2|1}(x_2 \mid x_1 \in I_{1i})$, the conditional density of x_2 given $x_1 \in I_{1i}$, $f_{1|2}(x_1 \mid x_2 \in I_{2j})$, the conditional density of x_1 given $x_2 \in I_{2j}$, the marginal densities of x_1 and x_2 to be $f_1(x_1)$ and $f_2(x_2)$. Integrating the conditional densities over the appropriate bins gives conditional probabilities:

$$P(x_1 \in I_{1i} \mid x_2 \in I_{2j}) = \int_{I_{1i}} f_{1|2}(u \mid x_2 \in I_{2j})du\,, \tag{2.1}$$

$$P(x_2 \in I_{2j} \mid x_1 \in I_{1i}) = \int_{I_{2j}} f_{2|1}(v \mid x_1 \in I_{1i})dv\,. \tag{2.2}$$

Simonoff proposes to estimate the conditional probabilities by treating each row and each column as **one-dimensional multinomial vector,** and then smooth them using **the penalized likelihood method proposed** by Simonoff (1983). The marginal probabilities were estimated using the marginal frequency estimates. He shows that when the number of rows $n_1 \to \infty$, and the number of columns $n_2 \to \infty$, then his estimator is a sparse asymptotic consistent one. For estimating the continuous density $f(x_1, x_2)$ we use an analogous technique.

Substituting into

$$f(x_1, x_2) = \left[f_{2|1}(x_2 \mid x_1) f_1(x_1) f_{1|2}(x_1 \mid x_2) f_2(x_2) \right]^{1/2}, \tag{2.3}$$

the kernel estimates of the conditional and marginal densities we obtain the 2D density estimate.

It is possible to generalize the estimator proposed by Simonoff for the multidimensional case. The main advantages of this estimator are relative computational simplicity in comparison to direct estimation of the multidimensional density, the effect of avoiding outlying cell propagation on the whole density estimate and its elasticity related to marginal and conditional density estimation method.

Further, we use a kernel density estimator for discrete data. Let us revise some basic notions related to this idea. Consider the estimation of a probability function defined for $X_i \in \mathrm{S} = \{0,1,...,c-1\}$.

The kernel estimator of $p(x)$

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} l(X_i, x)\ , \tag{2.4}$$

where $l(\cdot)$ is a kernel function defined by, say,

$$l(X_i, x) = \begin{cases} 1-\lambda & X_i = x \\ \lambda/(c-1) & otherwise \end{cases}, \tag{2.5}$$

and where $\lambda \in [0,(c-1)/c]$ is a "smoothing parameter" or "bandwidth". It is easy to show

$$E\hat{p}(x) = p(x) + \lambda\left\{\frac{1-cp(x)}{c-1}\right\}, \text{ var } \hat{p}(x) = \frac{p(x)(1-p(x))}{n}\left(1-\lambda\frac{c}{(c-1)}\right)^2.$$

This estimator was proposed by Aitchinson and Aitken (1976).

Theoretical results related to the Simonoff estimator (2.3) applied to binned data can be found in Simonoff (1995). Further, we use the estimator of (2.3) of the form

$$\hat{f}(x_1,x_2) = \left(\hat{f}_{2|1}(x_2 \mid x_1 \in I_{1i})\,\hat{f}_1(x_1)\,\hat{f}_{1|2}(x_1 \mid x_2 \in I_{2j})\,\hat{f}_2(x_2)\right)^{1/2}. \qquad (2.6)$$

The rate of the Mean Squared Error for this estimator equals to $O(n^{-4/7})$, and is worse than the rate of the common univariate kernel estimator $O(n^{-2/3})$. This inferiority has been called the **quantitative effectiveness of smoothing**. However, it is balanced by the adaptive nature of the proposed estimator in the sense of mode determination.

It is worth noting how important is the correct choice of bins for multimodality detection of the underlying distribution. Figure 1 presents the effects of kernel density estimation of the unemployment rate and the average salary in Polish subregions in 2006 for various number of bins. Obviously, the number of bins should increase as the sample size increases. As it has been shown, it should increase with a rate $n^{2/7}$, the best rate with respect to squared error.

3. Robustness in the case of sparse contingency table

Effective analysis of high-dimensional discrete sparse data requires a special attention especially in the context of robustness of the procedure and its computational complexity. Issues related to robustness of the procedure dedicated to analysis of discrete data are not so highly developed as in the case of continuous data analysis. In the predominant part, good multivariate robust procedures are computationally very intensive. This in particular affects methods of nonparametric estimation of probability density function for high-dimensional data. As a starting point for our considerations and proposals we take pioneering works of J. Simonoff related to automatic and adaptive estimation of bivariate density function (see Simonoff, 1985, 1988, 1995), developed now by Jacob and Oliveira (see Jacob & Oliveira,2011).

Categorical data analysis is typically performed by fitting models to the observed counts in a contingency table using maximum likelihood. An inherent problem with maximum likelihood fits is their sensitivity to outlier cells, the ones whose counts are not consistent with the assumed model. Maximum likelihood estimates break down in the presence of outlying cells. It is worth noting that in

categorical data analysis an outlier is a cell, i.e. a set of observations rather than a single observation, which deviates greatly from the expected count associated with the parametric model appropriate for the majority of cells.

Following Shane and Simonoff (2001), let us consider a D dimensional contingency table with d cells written as $d \times 1$ vector $n = (n_1, ..., n_d)$. Let $\mathbf{e} = (e_1, ..., e_d)$ be the vector of expected cell counts under a hypothesized model. The expected counts are $e_k = N \cdot p_k$, where N is the total sample size $\sum_{k=1}^{d} N_k$, where $\mathbf{p} = (p_1, ..., p_d)$ are theoretical cell probabilities. Assuming multinomial model for the cells we can understand robustness of the estimator in terms of goodness-of-fit statistics:

$$X^2 = \sum_{k=1}^{d} \chi_k^2(n_k, \hat{e}_k) = \sum_{k=1}^{d} \frac{(n_k - \hat{e}_k)^2}{\hat{e}_k} = \sum_{k=1}^{d} \frac{(n_k - p_k N)^2}{p_k N} \tag{3.1}$$

or equivalently the likelihood ratio goodness-of-fit statistics

$$G^2 = 2\sum_{k=1}^{d} n_k \log(n_k / \hat{e}_k) = 2\sum_{k=1}^{d} n_k \log(n_k / N \cdot p_k) \tag{3.2}$$

Let $X_{(l)}^2$ denote the $l-$ order statistics of X_k^2 . Shane and Simonoff (2001) define a robust Pearson estimate of a contingency table model as minimizing the criterion

$$\sum_{k=1}^{d} c_k X_{(k)}^2(n_k, e_k) \ , \tag{3.3}$$

where $\mathbf{c} = (c_1, ..., c_d)$ is an appropriate vector of weights.

The robust estimate according to Simonoff means a fit that is appropriate for the majority of cells and which is determined by the vectors of weights $\mathbf{c} = (c_1, ..., c_d)$. For continuous data this idea depends on the binning, the vector of weights and the measure used to assess the overall goodness of fit.

In the context of the analysis of sparse high-dimensional data for robustness of the procedure evaluation we propose to follow ideas presented in Mizera (2001). According to the ideas it is possible to define **halfspace depth** and **maximum depth based estimators** for the contingency tables. **General halfspace depth** can be defined as a measure data-analytic **admissibility of a fit with respect to the data**. **Depth** of $\mathbf{p}$ can be expressed as **the proportion of the data points** whose **omission causes** $\mathbf{p}$ to **become a nonfit**, a fit that can be uniformly dominated by another one.

For a contingency table with bins $\{I_{1i}\} \times \{I_{2j}\}$, $i = 1, ..., k_1$, $j = 1, ..., k_2$, we define the **depth of a fit** $\mathbf{p} = (p_1, ..., p_d)$ **as a minimal fraction of observations in the contingency table, whose replacement with other observations from the table will effect in taking the overall goodness-of-fit measure**

unacceptable value. As the overall goodness-of-fit measure we take Pearson statistics calculated for nonzero cells (we can use many other criteria functions instead, however):

$$F_{PEAR} = \sum_{n_k \neq 0} \frac{(n_k - Np_k)^2}{Np_k} \ . \tag{3.4}$$

As the **robust estimator** of the model we take the **maximum depth estimator**.

In Mizera (2002) it is shown how to reformulate the general criteria (3.4) into the first order optimization. Mizera introduces the tangent depth - the depth of the fit takes a form

$$d(\mathbf{p}) = \inf_{\mathbf{u}\neq\mathbf{0}} \#\left\{ n : \mathbf{u}^T \nabla_{\mathbf{p}} F_{PEAR}(\mathbf{p}) \geq 0 \right\}. \tag{3.5}$$

where $\acute{\mathrm{N}}_p f$ denotes gradient of a function f in a point p .

Attractive breakdown point robustness of the maximum depth estimator follows from Mizera (2002).

4. Our proposals

Sparse methods could be described as methods which make interpretation of the statistical analysis easier by forcing the statistical procedure to produce sparser output that is, for example, a sparser vector of regression coefficients. As a prototype for the sparse methods one can take the ridge regression, the LASSO regression, or the ELASTIC NET. Considering regression data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\} \subset \square^{p+1}$, in ridge and LASSO regression correspondingly, as regression parameters estimates we take vectors

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 , \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t , \tag{4.1}$$

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 , \text{ subject to } \sum_{j=1}^{p} \left|\beta_j\right| \leq t \ . \tag{4.2}$$

In the case of sparse PCA, taking into account the fact that an interpretation of the PCA components is conducted by examining the direction vectors known as **loadings** – we force the estimation procedure to produce sparser set of the loadings. Constraints encourages some loadings to be zero (for further details see Hastie et al. (2009)). The SCOTLASS procedure of Joliffe et al. (2003) focuses on maximum variance property of principal components by solving

$$\max \mathbf{v}^T \left(\mathbf{X}^T \mathbf{X} \right) \mathbf{v} \ , \text{ subject to } \sum_{j=1}^{p} \left|v_j\right| \leq t \ , \ \mathbf{v}^T \mathbf{v} = 1 . \tag{4.3}$$

Sparse and robust methods are relatively new and appeared during last 5 years (see Croux and Filzmoser, 2010).

Below we propose a general idea of producing a sparse and robust estimator of 2D density appealing **to functional data analysis.** The Simonoff estimator enables us to decompose 2D density estimation procedure (computationally a more complicated problem) into blocks which are estimated using 1D marginal densities and 1D conditional densities (computationally a less complicated problem).

Assuming a certain **sample of contingency tables** – for each of its cells we dispose of a certain number of marginal and conditional density estimates. We can successfully apply Functional Data Analysis (FDA) machinery to them. In particular we can use functional PCA of the estimated densities. Squared functional principal components fulfil density function postulates. We can decompose the overall density by means of them.

Let us consider functional data $x_1(t),...,x_s(t)$. Assuming we have chosen a basis $\phi_1,...,\phi_L$ (we advocate here on using basis consisted of splines), we consider representations of the data

$$x_r(t) = \sum_{j=1}^{L} c_{rj}\phi_j(t) \ , \tag{4.4}$$

where $c_{r1},...,c_{rL}$ are coefficients for *r*-th objects in this basis.

Coefficients $c_{r1},...,c_{rL}$ are chosen separately for every function $x_r(t)$. Assume we fixed L basis functions and then our data set consists of s functions $x_1(t),...,x_s(t)$. In the FDA we perform basic operations using $L \times s$ matrix containing object coefficients in the fixed basis (see Krzyśko et al., 2012). Introducing a quantity

$$\rho_{\xi}(x(t)) = \int \xi(t)x(t)dt \ , \tag{4.5}$$

our aim is to find a function $\xi(t)$ which in a best way underlines a variability of the data, i.e. for which $\rho_{\xi}(x(t))$ takes the maximal value.

$$\textbf{FPCA GOAL: } \mu = \max_{\xi}\left\{\sum_{i=1}^{s}\rho_{\xi}^2(x(t))\right\} \text{, under the condition } \int \xi^2(t)dt = 1 \, . \tag{4.6}$$

It is common to use a restriction on weight function ξ, $\int \xi^2(t)dt = 1$. In a similar manner as in the case of classical PCA a non-decreasing sequence of eigenvalues $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$ is developed recursively: $\int \xi_j(t)\xi_l(t)dt = 0$, $j = 1,...,l-1$, $\int \xi_l^2(t) = 1$. For further details see Ramsey et al. (2010) and Krzyśko et al. (2012).

5. Empirical examples

In order to illustrate the presented approach we used the Central Statistical Office (CSO) data concerning traceability of crimes and unemployment in Polish subregions in 2004 – 2010. We have analysed eight 5x5 contingency tables, each consisting of 66 observations. Figures 2 – 6 present kernel density estimates for marginal, conditional and joint probability distribution of the unemployment rate and traceability of crimes in Polish subregions in 2004 – 2010. Estimates were obtained using binned data presented in Table 1. Figures 7 – 18 present results of the functional PCA performed on the basis of 8 contingency tables consisting of data on traceability of crimes and the unemployment rate in Polish subregions. For simplicity of the presentation we focused only on one cell placed on the crossing of the shaded row and column in Table 2. We have performed similar analysis for the rest of the cells. It is easy to see that we can estimate the joint density of the variables using the idea of the Simonoff estimator (2.3) and using only the first or the second weight function (Fig. 9, Fig. 12, Fig. 15, Fig. 18). The output obtained in this way is much easier to interpret – the joint density function is decomposed into more evident layers. Although it is well known that the classical PCAs are not robust for outliers, several simulation studies we have performed using mixtures of various 2D discrete distributions show that our proposal seems to be robust to replacement of a small fraction of observations in the contingency table and in the spirit of Mizera (2002) ideas. It is possible, however, to directly the use robust PCA (see Croux et al., 2012) instead of classical PCA calculations during functional PCA. Our approach is computationally less intensive.

Table 2. A contingency table – traceability of crimes in Polish sub-regions in 2010

2010	X_{11}= 50.4	X_{12}= 58.4	X_{13}= 66.4	X_{14}= 74.4	X_{15}= 82.4	TOTAL
X_{21}=5.6	3	2	1	2	0	8
X_{22}=9.8	1	4	5	1	2	13
X_{23}=14.0	0	1	3	14	8	26
X_{24}=18.2	0	0	1	9	3	13
X_{25}=22.4	0	0	1	5	0	6
TOTAL	4	7	11	31	13	66

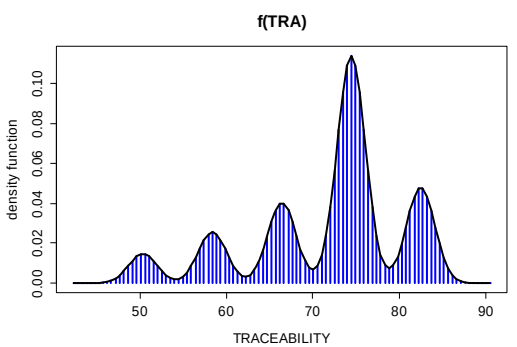


Figure 2. Kernel estimate of marginal density – traceability of crimes in Polish sub-regions in 2010

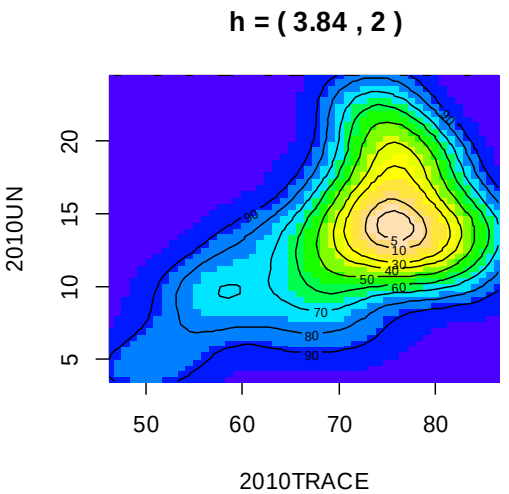


Figure 3. 2D kernel density estimate of unemployment rate vs. traceability of crimes in Polish sub-regions in 2010

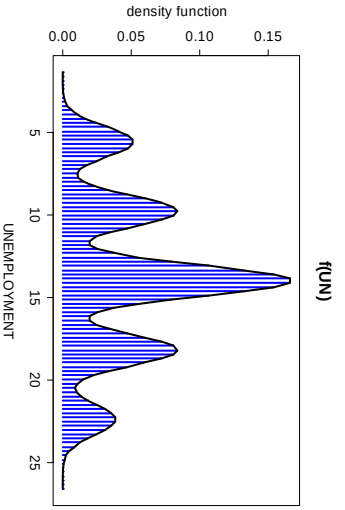


Figure 4. Kernel estimate of marginal density – unemployment rate in Polish sub-regions in 2010

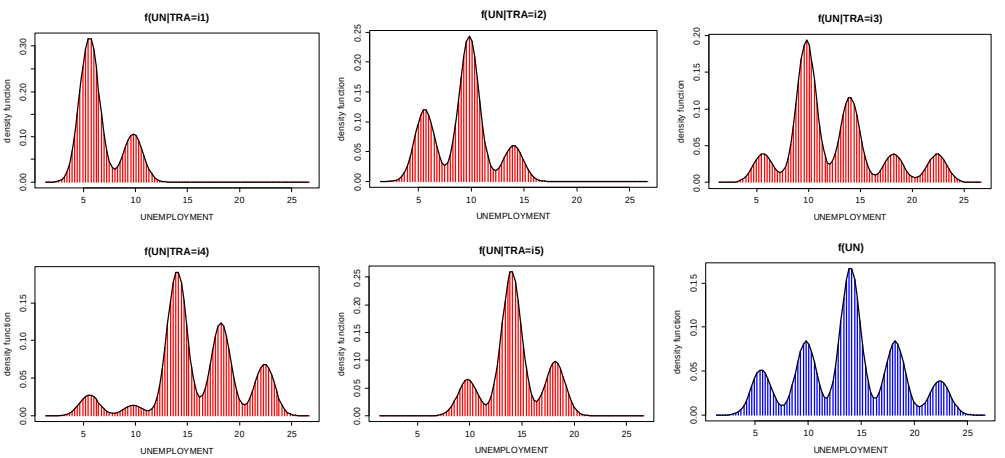


Figure 5. Conditional density estimate of unemployment rate under the condition that traceability of crimes takes value i1,…, i5. Last graph represents the unconditional density estimate of unemployment

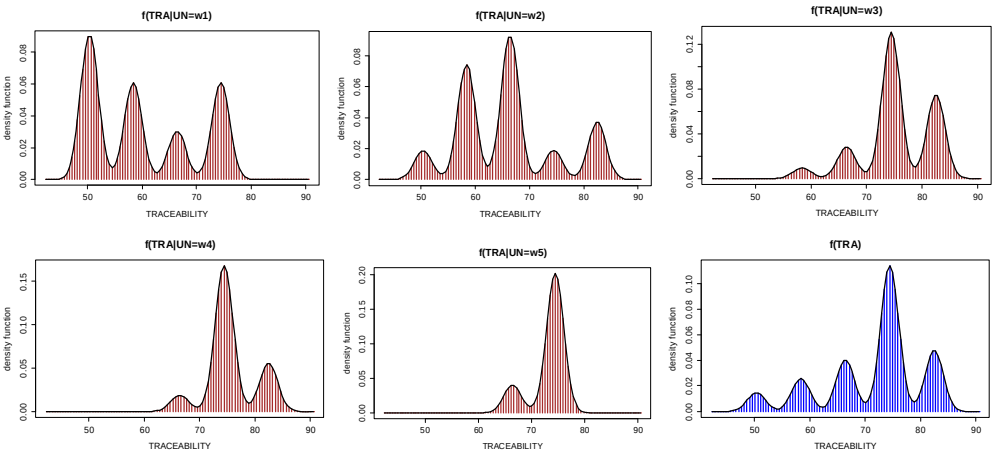


Figure 6. Conditional density estimate of traceability of crimes under the condition that unemployment rate takes value w1,…,w5. Last graph represents the unconditional density estimate of traceability

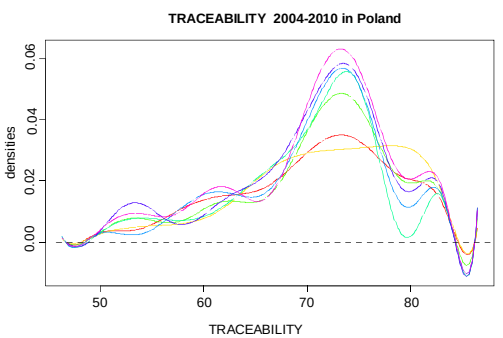


Figure 7. Density estimates for traceability of crimes in Polish subregions in 2004–2010

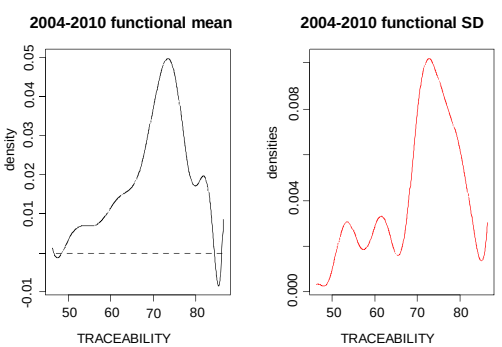


Figure 8. Functional mean (left) and functional SD (right) for density estimates for traceability of crimes in Polish subregions in 2004–2010

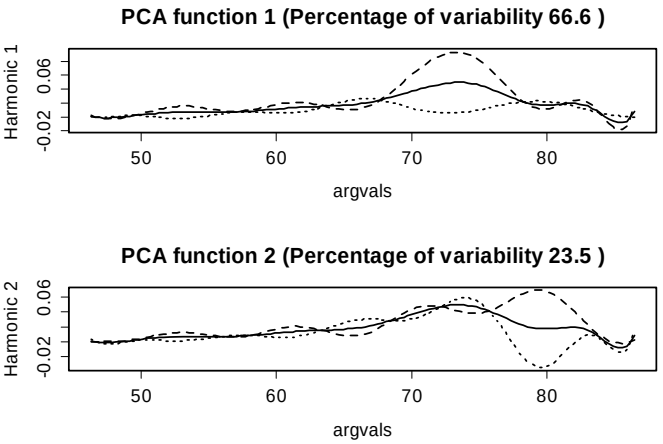


Figure 9. First and second weight functions (analogues of the eigenvectors) for density estimates for traceability of crimes in Polish subregions in 2004 – 2010

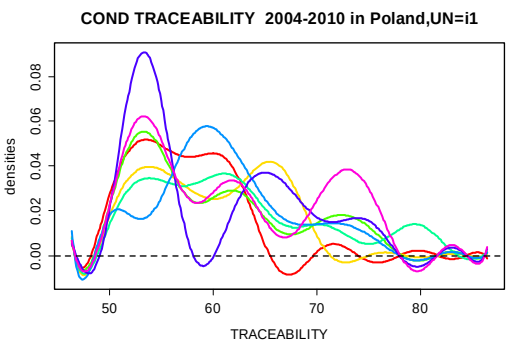


Figure 10. Density estimates for conditional traceability of crimes in Polish subregions in 2004–2010, condition unemployment rate = i1

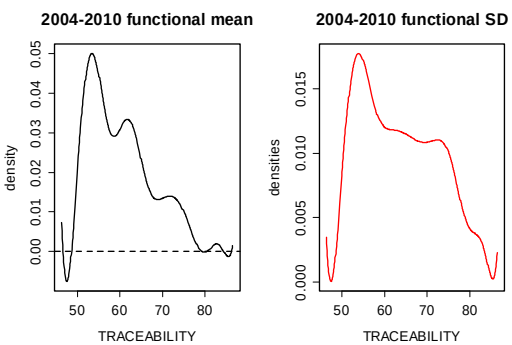


Figure 11. Functional mean (left) and functional SD (right) for conditional density estimates for traceability of crimes in Polish subregions

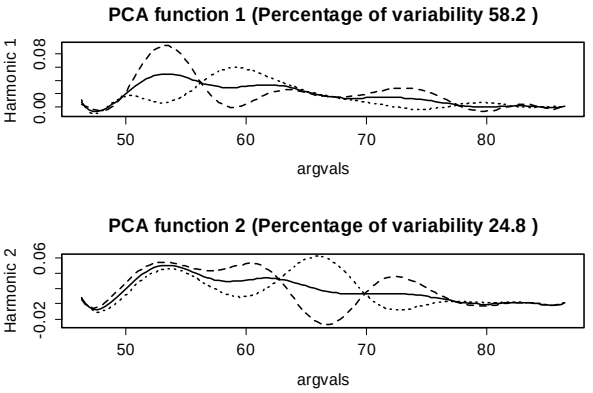


Figure 12. First and second weight functions (analogues of the eigenvectors) for conditional density estimates for traceability of crimes in Polish subregions in 2004–2010

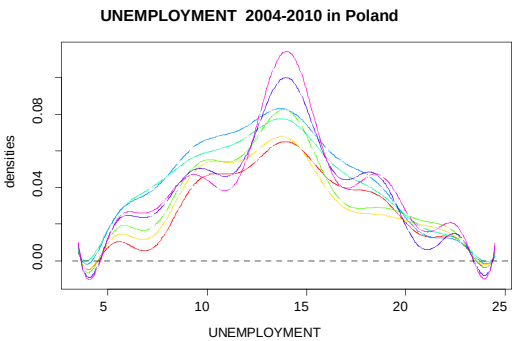


Figure 13. Density estimates for unemployment rate in Polish subregions in 2004–2010

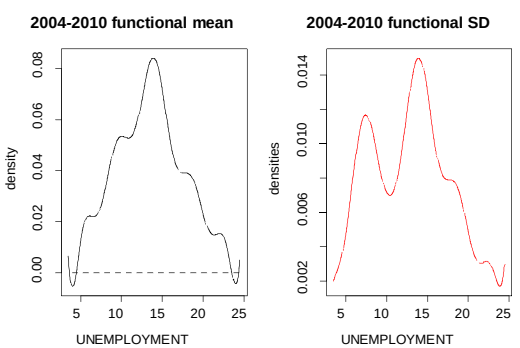


Figure 14. Functional mean (left) and functional SD (right) for density estimates for unemployment in Polish subregions in 2004–2010

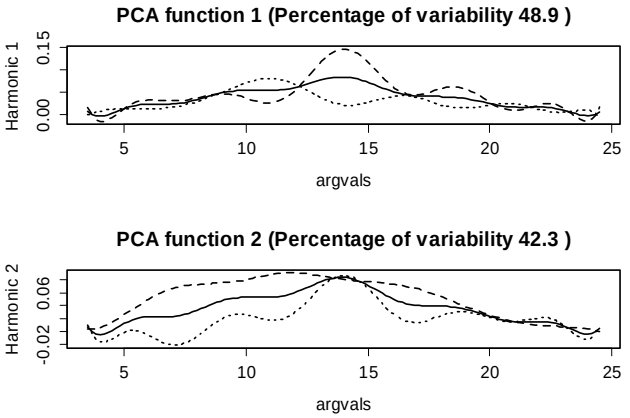


Figure 15. First and second weight functions (analogues of the eigenvectors) for density estimates for unemployment rate in Polish subregions in 2004–2010

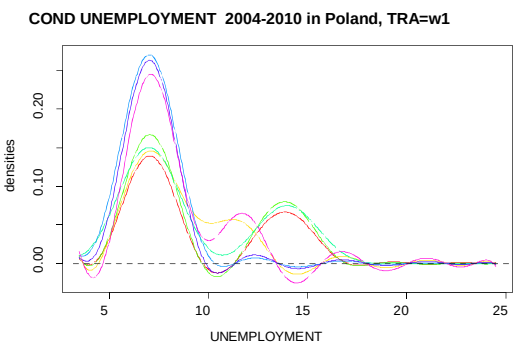


Figure 16. Density estimates for conditional traceability of crimes in Polish subregions in 2004–2010, condition unemployment rate = i1

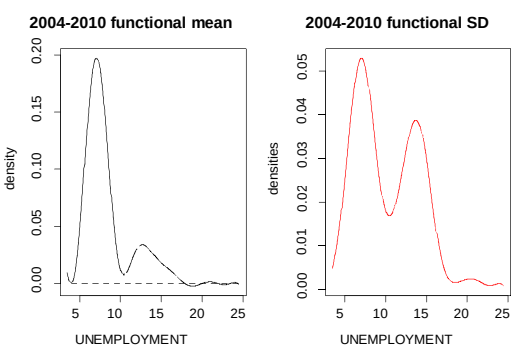


Figure 17. Functional mean (left) and functional SD (right) for conditional density estimates for traceability of crimes in Polish subregions

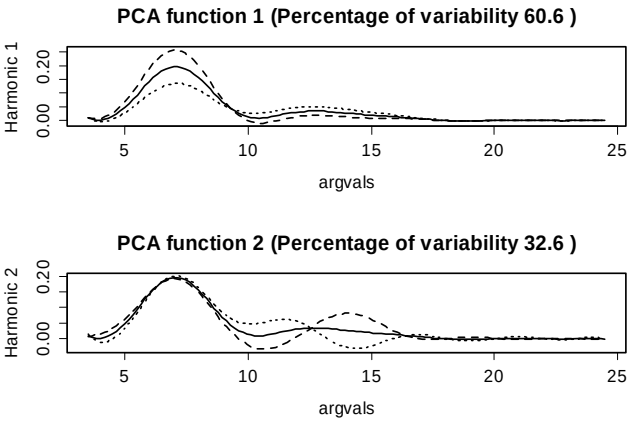


Figure 18. First and second weight functions (analogues of the eigenvectors) for density estimates for unemployment rate in Polish subregions in 2004–2010

6. The random matrix theory for detecting dependency between variables in a huge contingency table

Consider now that a contingency table, i.e. a data frame of p_1 input factors and p_2 output factors is observed continuously at n consecutive time moments. Let Y_{ia} be the value of the i-th ($i = 1,\ldots,p_1$) random variable at the a-th time moment ($a = 1,\ldots,n$); together, they make up a rectangular $p_1 \times n$ matrix $\mathbf{Y}$. Analogously, let X_{jb} be the value of the j-th ($j = 1,\ldots,p_2$) random variable at the b-th time moment ($b = 1,\ldots,n$); together, they make up a rectangular $p_2 \times n$ matrix $\mathbf{X}$. **In general** p_1, p_2, n **can be very large**. Further, we will assume that $p_1, p_2, n \to \infty$ but $p_1 / n = c_1$ and $p_2 / n = c_2$ are fixed. Under null hypothesis, each Y_{ia} and X_{jb} is supposed to be drawn from a Gaussian probability distribution, and that they have mean values zero. Specifically, the aim is to test the hypothesis:

H_0 **: x and y are independent; against** H_1 **: x and y are not independent,**

where $\mathbf{x} = (x_1,\ldots,x_{p_1})^T$ and $\mathbf{y} = (y_1,\ldots,y_{p_2})^T$. Without loss of generality, suppose that $p_1 \le p_2$.

It is well known that the canonical correlation analysis (CCA) deals with the correlation structure between two random vectors. Draw n independent and identically distributed (i.i.d.) observations from these two random vectors **x** and

$\mathbf{y}$ respectively, and group them into $p_1 \times n$ random matrix $\mathbf{X} = (x_1, \ldots, x_n) = (X_{ij})_{p_1 \times n}$ and $p_2 \times n$ random matrix $\mathbf{Y} = (y_1, \ldots, y_n) = (Y_{ij})_{p_2 \times n}$, respectively. The CCA seeks the linear combinations $\mathbf{a}^{\mathrm{T}}\,\mathbf{x}$ and $\mathbf{b}^{\mathrm{T}}\,\mathbf{y}$ that are most highly correlated, that is to maximize

$$\gamma = Corr(a^T x, b^T y) = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}} \tag{6.1}$$

where Σ_{XX} and Σ_{YY} are the population covariance matrices for x and y respectively, and Σ_{XY} is the population covariance matrix between x and y.

After finding the maximal correlation r_1 and associated vectors a_1 and b_1, CCA continues to seek a second linear combination $a_2^T x$ and $b_2^T y$ that has the maximal correlation among all linear combinations uncorrelated with $a_1^T x$ and $b_1^T y$. This procedure can be iterated and successive canonical correlation coefficients $\gamma_1, \ldots, \gamma_{p_1}$ can be found. It turns out that the population canonical correlation coefficients $\gamma_1, \ldots, \gamma_{p_1}$ can be recast as the roots of the determinant equation

$$\det(\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY'} - \gamma^2 \sigma_{XX}) = 0 \tag{6.2}$$

This equation can be replaced by:

$$\det(G_{XY}D_{YY}^{-1}G_{XY'} - r^2 D_{XX}) = 0 \tag{6.3}$$

$$D_{XX} = \frac{1}{n}XX^T \qquad D_{YY} = \frac{1}{n}YY^T \qquad G_{XY} = \frac{1}{n}XY^T$$

We also think of D_{XX}, D_{YY} and G_{XY} as sample covariance matrices. However, due to dimensionality curse these are not consistent estimators of population covariance matrices, when the dimensions p_1 and p_2 are both comparable to the sample size n. As a consequence, it is conceivable that the classical likelihood ratio statistics do not work well in the high dimensional case.

Moreover, $r_1^2, r_2^2, \ldots, r_{p_1}^2$ are the eigenvalues of the matrix

$$S_{XX} = D_{XX}^{-1}G_{XY}D_{YY}^{-1}G_{XY^T} \tag{6.4}$$

Evidently, D_{XX}^{-1} and D_{YY}^{-1} do not exist when $p_1 > n$ and $p_2 > n$. For this reason we also consider the eigenvalues of the **regularized matrix**

$$T_{XY} = D_{tX}^{-1}G_{XY}D_{XY}^{-1}G_{XY^T}, \tag{6.5}$$

where $D_{tX}^{-1} = (\frac{1}{n} XX' + tI_{P_1})^{-1}$, t is a positive constant number and I_{p_1} is a $p_1 \times p_1$ identity matrix.

In addition to proposing statistics for testing we will also establish the limit of the ESD of regularized sample canonical correlation coefficients and central limit theorems (CLT) of linear functionals of the classical and regularized sample canonical correlation coefficients $r_1, r_2, \ldots, r_{p_1}$, respectively. To derive the CLT for linear spectral statistics of classical and regularized sample canonical correlation coefficients, the strategy is to first establish the CLT under the Gaussian case, the entries of X are Gaussian distributed. In the Gaussian case, the CLT for linear spectral statistics of the matrix S_{XY} can be linked to that of an F-matrix, which was investigated in Bai and Silverstein (1995).

We make the following assumptions:

1. $p_1 = p_1(n)$ and $p_2 = p_2(n)$ with $p_1 \to c_1$ and $p_2 \to c_2$, $c_1, c_2 \in (0,1)$ as $n \to \infty$
2. $X = (X_{ij})_{i,j=1}^{p_1,n}$ and $Y = (Y_{ij})_{i,j=1}^{p_2,n}$ satisfy $X = \Sigma_{XX}^{1/2} W$ and $Y = \Sigma_{YY}^{1/2} V$, where $W = (w_1, \ldots, w_n) = (W_{ij})_{i,j=1}^{p_1,n}$ consists of i.i.d. real random variables $\{W_{ij}\}$ with $EW_{11} = 0$ and $E|W_{11}|^2 = 1$; $V = (v_1, \ldots, v_n) = (V_{ij})_{i,j=1}^{p_2,n}$ consists of i.i.d. real random variables $\{V_{ij}\}$ with $EV_{11} = 0$ and $E|V_{11}|^2 = 1$; $\Sigma_{XX}^{1/2}$, $\Sigma_{YY}^{1/2}$ are Hermitian square roots of positive definite matrices Σ_{XX} and Σ_{YY}.
3. $F^{\Sigma_{XX}} \to^D H$ a proper cumulative distribution function.

By the definition of the matrix S_{XY}, the classical canonical correlation coefficients between x and y are the same as those between w and v when w, v are i.i.d.

We now introduce some results from random matrix theory and free probability theory as presented by Voiculescu (1991).

Definition 6.1: Denote the ESD of any $n \times n$ matrix A with real eigenvalues $\mu_1 \le \mu_2 \le \ldots \le \mu_n$

$$F^A(x) = \frac{1}{n} \#\{i : \mu_i \le x\}, \tag{6.6}$$

where $\#\{\ldots\}$ denotes the cardinality of the set $\{\ldots\}$.

Theorem 6.2: When the two random vectors x and y are independent and each of them consists of i.i.d Gaussian random variables, under Assumptions 1 and 2, the empirical measure of the classical sample canonical correlation

coefficients $r_1, r_2, \ldots, r_{p_1}$ converges in probability to a fixed distribution whose density is given by

$$\rho(x) = \frac{\sqrt{(x-L_1)(x+L_1)(L_2-x)(L_2+x)}}{\pi c_1 x(1-x)(1+x)} \quad , \qquad (6.7)$$

$x \in [L_1, L_2]$, and atom size of $\max(0,(1-c_2)/c_1)$ at zero and size $\max(0,1-(1-c_2)/c_1)$ at unity, where $L_1 = |\sqrt{c_2 - c_2 c_1} - \sqrt{c_1 - c_1 c_2}|$ and $L_2 = |\sqrt{c_2 - c_2 c_1} + \sqrt{c_1 - c_1 c_2}|$.

Here, the empirical measure of $r_1, r_2, \ldots, r_{p_1}$ is defined as in the ESD with μ_i replaced by r_i.

Let us now introduce the test statistics. Under Assumption 1 and Assumption 3, if $Y = \sigma_1 W$ and $X = \sigma_2 W$ with $p_1 = p_2$ and both Σ_1 and Σ_2 being invertible, then $S_{XY} = 1$, which implies that the limit of $F^{S_{XY}}(x)$ is a degenerate distribution. Thus, we consider the following statistics

$$S_n = \int x dF^{S_{XY}}(x) = \frac{1}{p_1} \sum_{i=1}^{p_1} r_i^2 . \qquad (6.8)$$

In the classical CCA, the maximum likelihood ratio test statistics with fixed dimensions is

$$MLR_n = \sum_{i=1}^{p_1} \log(1-r_i^2) . \qquad (6.9)$$

Note that the density $\rho(x)$ has atom size of $\max(0,1-(1-c_2)/c_1)$ at unity. Thus, the normalized statistics MLR_n is not well defined when $c_1 + c_2 > 1$ (because $\int \log(1-x^2)dx$ is not meaningful). In addition, even when $c_1 + c_2 \le 1$, the right end point of $\rho(x)$, L_2, can be equal to one so that some sample correlation coefficients r_i are close to one. For example, $L_2 = 1$ when $c_1 = c_2 = 1$. This in turn causes a big value of the corresponding $\log(1-r_i^2)$. **Therefore, MLR_n is not stable.**

Here we would like to point out that the idea of testing independence between two random vectors x and y by the CCA is based on the fact that the lack of correlation between x and y is equivalent to independence between them when the random vector of size (p1 +p2) consisting of the components of x and y is a Gaussian random vector.

In addition, it can be proved that

$$\mathrm{Tr}(G_{XY}^{H_1} - G_{XY}^{H_0}) = O_p(n) \qquad (6.10)$$

ALGORITHM FOR THE PROCEDURE – "DOUBLE SPARSITY ALGORITHM"

STEP 1. Preparation of the dataset

Now we will extend our consideration to the case of n consecutive observations. First, let us divide all variables into two subsets, i.e. focus on p_1 input factors X_a $(a=1,\ldots,p_1)$ and p_2 output factors Y_α $(\alpha=1,\ldots,p_2)$ with the total number of observations being n. All series of observations are standardized to have zero mean and unit variance. The data can be completely different or can be the same variables but observed at different times. First, one has to remove potential correlations inside each subset, otherwise it may interfere with the out-of-sample signal. To remove the correlations inside each sample we form two correlation matrices which contain information about in-the-sample correlations:

$$\mathbf{D_{XX}} = \frac{1}{n} XX^T, \qquad \mathbf{D_{YY}} = \frac{1}{n} YY^T$$

STEP 2. Diagonalization

The matrices are then diagonalized, provided $n > p_1, p_2$, and the empirical spectrum is compared to the theoretical Bai, Silverstein (1995) result

$$\rho(x) = \frac{1}{2\pi x} \mathbf{Re}\sqrt{(x-L_1)(L_2-x)} \quad L_1 = (1\pm\sqrt{c_1})^2 \qquad L_1 = (1\pm\sqrt{c_1})^2$$

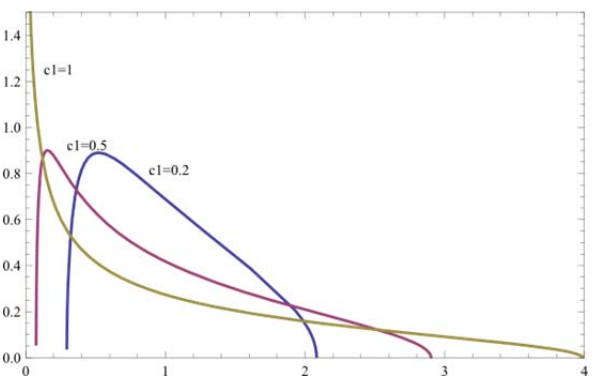


Figure 19. The spectrum of the single sparse matrices D_{XX} and D_{YY} when null hypothesis holds (i.e., there are no internal temporal correlations. The eigenvalues of ESD, which lie much below the lower edge of the spectrum, represent the redundant factors inconsistent with the null hypothesis)

STEP 3. Reconstruction

One can then construct a set of uncorrelated unit variance input variables $\hat{X}$ and output variables $\hat{Y}$

$$\hat{X}_{w_i} = \frac{1}{\sqrt{nw_i}} W^T X_i \qquad \hat{Y}_{v_j} = \frac{1}{\sqrt{nv_j}} V^T Y_j$$

where V,U, λ_a, λ_α are the corresponding eigenvectors and eigenvalues of D_{XX}, D_{YY}.

Finally, we can reproduce the asymmetric $p_1 \times p_2$ cross-correlation matrix G between the $\hat{Y}$ and $\hat{X}$:

$$G = \hat{X}\hat{Y}^T.$$

Under the null hypothesis of independence between X and Y, the ESD should follow the distribution with density (see, Snarska 2012)

$$\rho_G(x) = \max(1-c_1, 1-c_2)\delta(x) + \max(c_1+c_2-1, 0)\delta(x-1) + \frac{\mathrm{Re}\sqrt{(x^2 - s_-)(x_+ - s^2)}}{\pi x(1-x^2)},$$

where $x_\pm = c_1 + c_2 - 2c_1c_2 \pm 2\sqrt{c_1c_2(1-c_1)(1-c_2)}$ are the two positive roots of the quadratic expression under the square root. It is easy to see the fact that in the limit $n \to \infty$ at fixed p_1, p_2 all singular values collapse to zero as they should since there are no true correlations between X and Y; the allowed band in the limit $c_1, c_2 \to 0$ becomes: $x \in \left[|\sqrt{c_1} - \sqrt{c_2}|, \sqrt{c_1} + \sqrt{c_2} \right]$. When $c_1 \to c_2$, the support becomes $x \in [0, 2\sqrt{c_1(1-c_1)}]$ (plus a δ function at $x = 1$ when $c_1 + c_2 > 1$), while when $c_1 = 1$, the whole band collapses to a δ function at $x = \sqrt{1-n}$. For $c_1 + c_2 \to 1^-$ there is an initial singularity of $\rho(x)$ $x = 1$ diverging as $(1-x)^{-1/2}$. Ultimately, $c_1 \to 0$ at fixed c_2, one finds that the whole band collapses again to a δ function at $x = \sqrt{c_2}$.

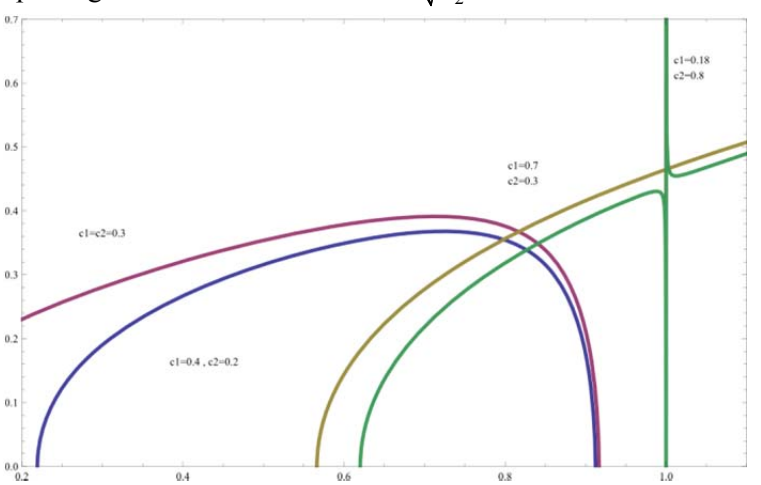


Figure 20. Theoretical distribution of singular values for G_{XY} under validity of null hypothesis. The eigenvalues of ESD, which lie much below the lower edge of the spectrum, represent the redundant factors inconsistent with the null hypothesis

7. Conclusions

A common application of the statistical procedures has changed business and the economy. Statistics have changed the ways we reason in a public debate, form our opinions, manage banking systems, perform interventions in a certain market, allocate energy stored in the capital between competing investments.

The innovative nature of the outlined approach to big economic databases analysis is manifested in formation of a complete methodology for a robust analysis of sparse high-dimensional discrete data in the economy. Our approach is still being developed and we hope to obtain interesting results in the near future. We are convinced that our proposal could find several applications in the on-line economy and exploration of the official statistics databases.

Acknowledgements

Daniel Kosiorowski thanks for financial support from Polish National Science Center grant UMO-2011/03/B/HS4/01138.

REFERENCES

CROUX, C., FILZMOSER, P., FRITZ, H., (2012). Robust Sparse Principal Component Analysis, Technometrics

DONG, J., SIMONOFF, J. S., (1994). The Construction and Properties of Boundary Kernels for Smoothing Sparse Multinomials. Journal of Computational and Graphical Statistics. Vol. 3, No. 1, 57–66.

HASTIE, T., TIBSHIRIANI, R., FRIEDMAN, J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer.

JACOB, P., OLIVEIRA, P. E., (2011). Local smoothing with given marginals, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2011.561436

JOLIFFE, I. T., TRENDAFILOV, N. T., UDDIN, M., (2003). A modified principal component technique based on the lasso, Journal of Computational and Graphical Statistics 12: 531–547.

KRZYŚKO, M., GÓRECKI, T., DERĘGOWSKI, K., (2012). Jądrowa i Funkcjonalna Analiza Składowych Głównych – spotkanie PTS o. w Poznaniu (referat dostępny na stronach PTS o. w Poznaniu http://www.stat.gov.pl/pts/)

MIZERA, I., (2002). On Depth and Depth Points: a Calculus. *The Annals of Statistics* (30), 1681–1736.

RAMSAY, J. O., HOOKER, G., GRAVES, S., (2010). *Functional Data Analysis with R and Matlab*, Springer, New York.

SHANE, K. V., SIMONOFF, J. S., (2001). A robust approach to categorical data analysis, Journal of Computational and Graphical Statistics, Vol. 10, No. 1, 135–157.

SILVERSTEIN, J., BAI, Z., (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. Journal of Multivariate Analysis 54, (2), 175–192.

SIMONOFF, J. S., (1985). An improved goodness-of-fit statistic for sparse multinomials, Journal of the American Statistical Association, Vol. 80, No. 391, 671–677.

SIMONOFF, J. S., (1988). Detecting outlying cells in two-way contingency tables via backward-stepping, Technometrics, Vol. 30, No. 3, 339–345.

SIMONOFF, J. S., (1995). A simple, automatic and adaptive bivariate density estimator based on conditional densities, Statistics and Computing, Vol. 5, 245–252.

SIMONOFF, J. S., (1983). A penalty function approach to smoothing large sparse contingency tables. The Annals of Statistics. Vol. 11, No. 1, 208–218.

SIMONOFF, J. S., (1998). Three sides of smoothing: categorical data smoothing, nonparametric regression, and density estimation. International Statistical Review, Vol. 66, No. 2, 137–156.

SNARSKA, M., (2012). A random matrix approach to dynamic factors in macroeconomic data, Acta Phys. Pol A, 121 (2B), 110–120**.**

VOICULESCU, D. V., (1991). Limit laws for random matrices and free products, Invent. Math. 104, 201.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 133–144

APPLICATION OF QUANTILE METHODS TO ESTIMATION OF CAUCHY DISTRIBUTION PARAMETERS

Dorota Pekasiewicz[1]

ABSTRACT

Quantile methods are used for estimation of population parameters when other methods such as the maximum likelihood method and the method of moments cannot be applied. In the paper the percentile method, the quantile least squares method and its two modifications are considered. The proposed methods allow estimators to be obtained with smaller bias and smaller mean squared error than estimators of the quantile least squares method. The considered methods can be applied to estimation of the Cauchy distribution parameters. The results of the simulation analysis of the estimator properties have allowed conclusions to be drawn as concerning the application of the considered methods.

Key words: quantile , percentile method, quantile least square method, Cauchy distribution.

1. Introduction

Quantile estimation methods can be used for estimating parameters of different distributions, particularly in the cases when we cannot use the maximum likelihood method and the method of moments, for example, for heavy tailed distributions. We focus on the percentile method, the quantile least squares method and its modifications.

Since in the percentile method distribution quantiles are compared with sample quantiles, its application requires the formula for the quantile function. The number of required quantiles depends on the number of distribution parameters. The orders of selected quantile s have impact on properties of estimators. Quantiles that give the estimators with small mean squared errors can be different for different types of distribution estimates. In Aitchison and Brown (1975) the orders of quantiles which should be selected in estimating lognormal distribution parameters are given and in Pekasiewicz (2012) they are computed for the Pareto distribution.

[1] University of Łódź, Department of Statistical Methods. E-mail: pekasiewicz@uni.lodz.pl.

The quantile least squares method has the advantage over the percentile method that there is no need to determine the ranks of used order statistics. However, in the case of the Cauchy distribution, the application of minimum or maximum statistics leads to very large mean squared errors of the parameter estimators because extreme statistics have infinite variances. Rejecting extreme order statistics significantly improves the properties of the estimators. Hence, we suggest the truncated quantile least squares method. In the case of the Cauchy distribution, we reject a fixed number of the largest and the smallest order statistics. Rejecting the same number of quantiles on both sides of the distribution appears to be justified in view of the symmetry of the distribution. The use of this method requires determination of the number of rejected quantiles. The second of the proposed methods does not require assumptions about the number of truncated quantiles. In this method all possible estimators are calculated by the truncated least squares and the median of them is chosen.

The properties of the Cauchy distribution parameter estimators are analysed by the Monte Carlo method. The received results allow some conclusions to be drawn regarding the choice of ranks of the order statistics in the percentile method or the number of rejected order statistics.

2. The percentile method

The percentile method (PM) allows for estimation of unknown parameters $\theta_1, \theta_2, ..., \theta_s$ of the continuous random variable X distribution with cumulative distribution function $F(\cdot, \theta_1, \theta_2, ..., \theta_s)$ by comparing theoretical quantiles and empirical quantiles (Wywiał, 2004, Castillo et al., 2004).

Let $X_1, X_2, ..., X_n$ be an i.i.d. sample with a cdf F. Let us denote by $X_{p_i;n}$ the sample quantile of order p_i, $i = 1, ..., s$. Estimators of the parameters $\theta_1, \theta_2, ..., \theta_s$ are the statistics $\hat{\theta}_1^{pm}, \hat{\theta}_2^{pm}, ..., \hat{\theta}_s^{pm}$ that are solutions of the equations:

$$\begin{cases} X_{p_1;n} = F^{-1}(p_1, \theta_1, \theta_2, ..., \theta_s), \\ X_{p_2;n} = F^{-1}(p_2, \theta_1, \theta_2, ..., \theta_s), \\ ..... \\ X_{p_s;n} = F^{-1}(p_s, \theta_1, \theta_2, ..., \theta_s), \end{cases} \tag{1}$$

where F^{-1} is the inverse of F.

When estimating parameters θ_1, θ_2 for the random variable X with cdf $F(\cdot, \theta_1, \theta_2)$, frequently the quantiles of orders p_1, p_2 are chosen, such that $p_1 + p_2 = 1$.

For the Cauchy distribution with cdf $F(x)=\frac{1}{2}+\frac{1}{\pi}arctg\left(\frac{x-m}{\lambda}\right)$ and orders of quantile s p and $1-p$, equations (1) take the following form:

$$\begin{cases} X_{p;n} = m+\lambda \operatorname{tg}(\pi(p-0{,}5)), \\ X_{(1-p);n} = m+\lambda \operatorname{tg}(\pi(0{,}5-p)), \end{cases} \tag{2}$$

and the estimators are defined by the formulas:

$$\hat{m}^{pm} = \frac{X_{p;n}+X_{(1-p);n}}{2}, \tag{3}$$

$$\hat{\lambda}^{pm} = \frac{X_{(1-p);n}-X_{p;n}}{2\operatorname{ctg}(\pi p)}. \tag{4}$$

3. The Quantile least squares method and its modifications

The quantile least squares method (QLSM) estimates the unknown parameters $\theta_1,\theta_2,...,\theta_s$ of random variable X with cdf F by minimizing the sum of squares of the differences between theoretical and empirical quantile s (Gilchrist, 2000; Castillo et al., 2004). Then, the function for which we calculate the global minimum has the following form:

$$G(\theta_1,\theta_2,...,\theta_s)=\sum_{i=1}^{n}\left(X_{i/n;n}-Q_{i/n}\right)^2, \tag{5}$$

where $X_{i/n;n}$ is the sample quantile of order $p_i=\frac{i}{n}$ from the i.i.d. sample $X_1,X_2,...,X_n$ and $Q_{i/n}=F^{-1}\left(\frac{i}{n},\theta_1,...,\theta_n\right)$.

The estimators of parameters $\theta_1,\theta_2,...,\theta_s$ obtained by QLSM are denoted by $\hat{\theta}_1^{qls},\hat{\theta}_2^{qls},...,\hat{\theta}_s^{qls}$.

Using all available quantile orders can, however, result in unsatisfactory estimate properties or in some cases cannot be feasible. For the Cauchy distribution extreme statistics have infinite variance, which means that the mean squared errors of estimators based on them are very large. Therefore, the minimum and maximum statistics must be rejected for estimation of the Cauchy distribution parameters.

The first suggested modification of the quantile least squares method is rejecting a fixed number of quantile s, which we call the truncated quantile least squares method (TQLMS). In this case the estimators of distribution parameters $\theta_1,\theta_2,...,\theta_s$ of the random variable X with distribution function $F(\cdot,\theta_1,\theta_2,...,\theta_s)$

are statistics $\hat{\theta}_1^{tqls}, \hat{\theta}_2^{tqls}, ..., \hat{\theta}_s^{tqls}$, for which the following expression reaches a global minimum:

$$G(\theta_1, \theta_2, ..., \theta_s) = \sum_{i \in I_n} \left(X_{p_i;n} - Q_{p_i} \right)^2, \tag{6}$$

where $p_i = \frac{i}{n}$ and I_n is the subset of $\{1, 2, ..., n\}$.

For symmetric or close to symmetric distributions we suggest skipping k quantiles, where k is the even number, that is $\frac{k}{2}$ the smallest and $\frac{k}{2}$ the largest quantiles. Then, the function (6) takes the form:

$$G(\theta_1, \theta_2, ..., \theta_s) = \sum_{i=1+k/2}^{n-k/2} \left(X_{p_i;n} - Q_{p_i} \right)^2. \tag{7}$$

For asymmetric distributions with right or left heavy tail we suggest skipping k the largest or the smallest order statistics, respectively. Then, one of the functions expressed by the formula:

$$G(\theta_1, \theta_2, ..., \theta_s) = \sum_{i=1}^{n-k} \left(X_{p_i;n} - Q_{p_i} \right)^2 \tag{8}$$

or

$$G(\theta_1, \theta_2, ..., \theta_s) = \sum_{i=k}^{n} \left(X_{p_i;n} - Q_{p_i} \right)^2, \tag{9}$$

is minimized, respectively for right and left heavy tailed distribution.

The second modification of the quantile least squares method is the median-quantile least squares method (MQLSM). The estimators of $\theta_1, \theta_2, ..., \theta_s$ parameters of the random variable X from the random sample $X_1, X_2, ..., X_n$ are statistics $\hat{\theta}_1^{mq}, \hat{\theta}_2^{mq}, ..., \hat{\theta}_s^{mq}$ of the following form:

$$\hat{\theta}_i^{mq} = Me\left(\hat{\theta}_{i;\,k=2}^{tqls}, \hat{\theta}_{i;\,k=4}^{tqls}, ..., \hat{\theta}_{i;\,k=r}^{tqls} \right) \qquad \text{for } i = 1, ..., s, \tag{10}$$

where $\hat{\theta}_{i;\,k=2}^{tqls}, \hat{\theta}_{i;\,k=4}^{tgls}, ..., \hat{\theta}_{i;\,k=r}^{tqls}$ are estimators of the parameter θ_i obtained through the truncated quantile least squares method with k quantiles left out and $r = n - 2$, when n is even and $r = n - 3$, when n is odd.

The proposed modifications can be used to estimate the Cauchy distribution parameters.

The application of the truncated quantile least squares method for the Cauchy distribution is related to the minimization of the function:

$$G(m,\lambda)=\sum_{i=1+k/2}^{n-k/2}\left(X_{p_i;n}-m+\lambda\operatorname{ctg}(\pi p_i)\right)^2 . \tag{11}$$

Therefore, it requires solving the system of equations:

$$\begin{cases}\sum\limits_{i=1+k/2}^{n-k/2}\left(X_{p_i;n}-m+\lambda\operatorname{ctg}(\pi p_i)\right)=0,\\ \sum\limits_{i=1+k/2}^{n-k/2}\left(X_{p_i;n}-m+\lambda\operatorname{ctg}(\pi p_i)\right)\operatorname{ctg}(\pi p_i)=0.\end{cases} \tag{12}$$

The estimators for parameters m and λ, received by the truncated quantile least squares method (TQLSM) are defined by the formulas:

$$\hat{m}^{tqls}=\frac{\sum\limits_{i=1+k/2}^{n-k/2}X_{p_i;n}}{n-k}-\frac{\sum\limits_{i=1+k/2}^{n-k/2}X_{p_i;n}\sum\limits_{i=1+k/2}^{n-k/2}\operatorname{ctg}(\pi p_i)-(n-k)\sum\limits_{i=1+k/2}^{n-k/2}X_{p_i;n}\operatorname{ctg}(\pi p_i)}{(n-k)\left(\sum\limits_{i=1+k/2}^{n-k/2}\operatorname{ctg}(\pi p_i)\right)^2-(n-k)^2\sum\limits_{i=1+k/2}^{n-k/2}\operatorname{ctg}^2(\pi p_i)}\sum_{i=1+k/2}^{n-k/2}\operatorname{ctg}(\pi p_i), \tag{13}$$

$$\hat{\lambda}^{tqls}=\frac{\sum\limits_{i=1+k/2}^{n-k/2}X_{p_i;n}\sum\limits_{i=1+k/2}^{n-k/2}\operatorname{ctg}(\pi p_i)-(n-k)\sum\limits_{i=1+k/2}^{n-k/2}X_{p_i;n}\operatorname{ctg}(\pi p_i)}{(n-k)\sum\limits_{i=1+k/2}^{n-k/2}\operatorname{ctg}^2(\pi p_i)-\left(\sum\limits_{i=1+k/2}^{n-k/2}\operatorname{ctg}(\pi p_i)\right)^2}, \tag{14}$$

where k is a fixed even number, $X_{p_i;n}$ is the quantile from the i.i.d. sample $X_1, X_2, ..., X_n$ and $p_i=\dfrac{i}{n}$ for $i=1+\dfrac{k}{2},...,n-\dfrac{k}{2}$.

The estimators of the Cauchy distribution parameters received by median-quantile least squares method (MQLSM) have the following form:

$$\hat{m}^{mq}=Me(\hat{m}^{tqls}_{k=2},\hat{m}^{tqls}_{k=4},...,\hat{m}^{tqls}_{k=r}), \tag{15}$$

$$\hat{\lambda}^{mq}=Me(\hat{\lambda}^{tqls}_{k=2},\hat{\lambda}^{tqls}_{k=4},...,\hat{\lambda}^{tqls}_{k=r}), \tag{16}$$

where $r=n-2$, when n is even and $r=n-3$, when n is odd.

4. Simulation analysis of Cauchy distribution parameter estimators

The properties of the Cauchy distribution parameter estimators were studied using the Monte Carlo methods. The following methods were considered:

- the percentile methods with different orders of selected quantiles,
- the truncated quantile least squares methods with different number of rejected quantile s from tails of the distribution,
- the median-quantile least squares method.

In each case the bias and the mean squared error were estimated over 20000 repetitions. For the percentile method the dependences of the estimator bias and the mean squared error on the quantile order are presented in Figures 1 and 2. The size of the sample n was set to 100.

From the obtained results it can be concluded that, for the considered distribution, if p increases both the bias and the mean squared error of the estimators decrease, but only to a certain point. When this point is exceeded the precision can deteriorate. For $p \approx 0.45$ the estimator $\hat{m}^{pm}$ of parameter m has the best of properties, and for $p \approx 0.25$ the estimator $\hat{\lambda}^{pm}$ has the smallest mean squared error.

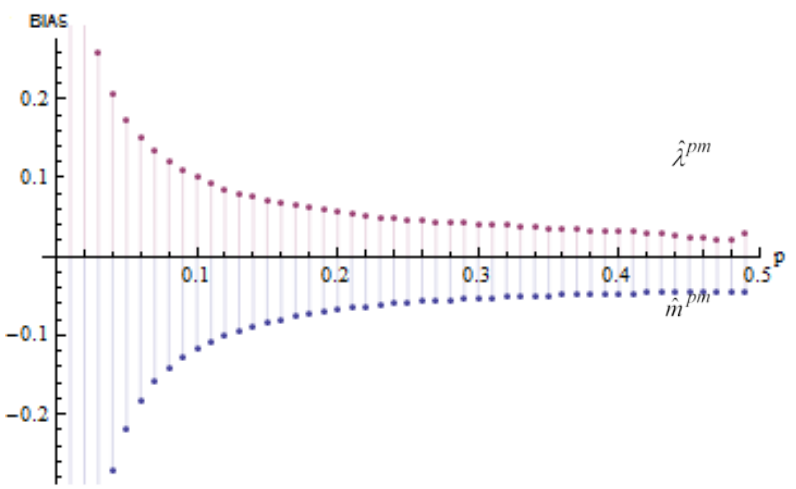


Figure 1. Dependence of the bias of $Ca(0,3)$ parameter estimators obtained by the percentile methods on the quantile order

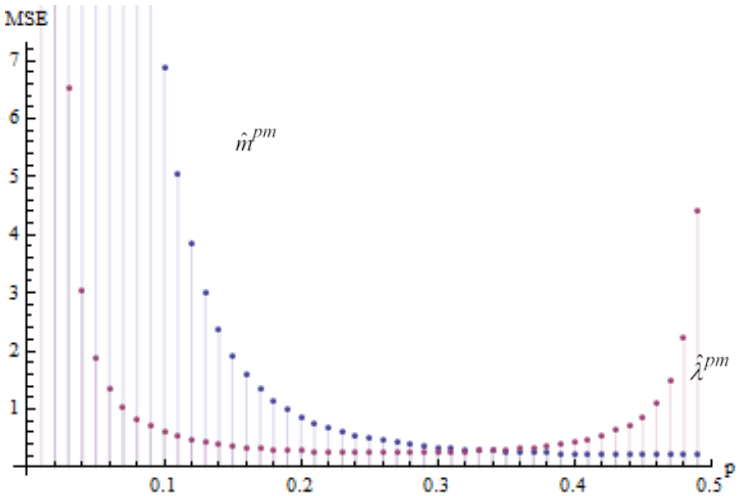


Figure 2. Dependence of the mean squared error of *Ca*(0,3) parameter estimators obtained by the percentile methods on the quantile order

The estimated values of the bias and the mean squared error for selected orders of quantile , for chosen parameters of the Cauchy distribution are shown in Table 1. The study included sample sizes $n = 60$ and $n = 100$. Pekasiewicz (2012) gives results for the Cauchy distribution with other parameters.

Table 1. The estimated bias and the mean square errors for the Cauchy distribution parameter estimators obtained by the percentile method

Distribution random variable	n	p	$\hat{BIAS}(\hat{m}^{pm})$	$\hat{BIAS}(\hat{\lambda}^{pm})$	$\hat{MSE}(\hat{m}^{pm})$	$\hat{MSE}(\hat{\lambda}^{pm})$
Ca(0, 3)	60	0.05	- 4.4690	0.7057	241.5300	5.7885
		0.10	- 0.9033	0.2788	14.1888	1.2504
		0.15	- 0.4011	0.1824	3.7038	0.7143
		0.20	- 0.2385	0.1400	1.5905	0.5356
		0.25	- 0.1614	0.1155	0.8972	0.4656
	100	0,05	- 2.4252	0.3735	81.5948	1.8908
		0,10	- 0.5821	0.1630	6.6861	0.6084
		0,20	- 0.1521	0.0817	0.8736	0.2842
		0,30	- 0.0818	0.0567	0.3424	0.2618
		0,40	- 0.0575	0.0500	0.2312	0.4381
		0,45	- 0.0525	0.0397	0.2233	0.8682

Table 1. The estimated bias and the mean square errors for the Cauchy distribution parameter estimators obtained by the percentile method (cont.)

Distribution random variable	n	p	$\hat{BIAS}\left(\hat{m}^{pm}\right)$	$\hat{BIAS}\left(\hat{\lambda}^{pm}\right)$	$\hat{MSE}\left(\hat{m}^{pm}\right)$	$\hat{MSE}\left(\hat{\lambda}^{pm}\right)$
$Ca(3, 2)$	60	0,05	- 3.1883	0.4949	143.4070	3.4363
		0,10	- 0.6597	0.2021	6.8335	0.6129
		0,15	- 0.2971	0.1285	1.7430	0.3321
		0,20	- 0.1729	0.0958	0.7408	0.2427
		0,25	- 0.1153	0.0759	0.4051	0.2056
	100	0,05	- 1.6606	0.2460	36.0447	0.8338
		0,10	- 0.3889	0.1115	3.1781	0.2740
		0,20	- 0.1032	0.0511	0.3826	0.1239
		0,30	- 0.0552	0.0372	0.1515	0.1136
		0,40	- 0.0402	0.0317	0.1022	0.1907
		0,45	- 0.0369	0.0361	0.0973	0.3835

The use of the truncated quantile least squares method for large k allowed the bias and the mean squared error to be significantly reduced, although rejecting too many quantiles gives poorer results. Dependences of the bias and the mean squared error on k are shown in Figures 3 and 4, respectively. The results for the selected parameters are given in Table 2.

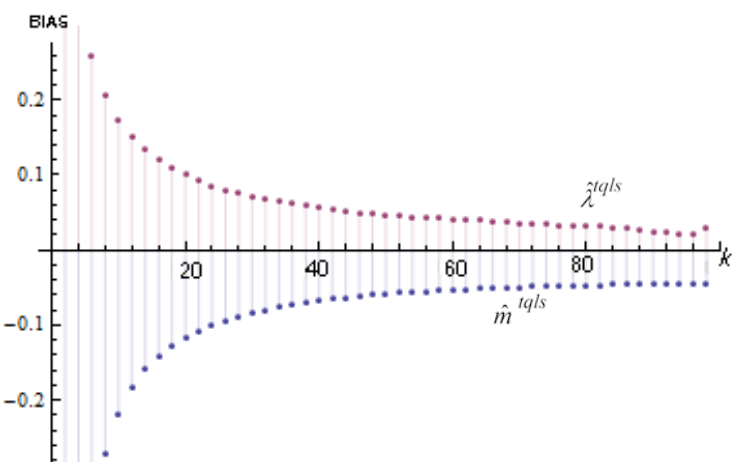


Figure 3. Dependence of the bias of $Ca(0,3)$ parameter estimators obtained by TQLSM on the number of rejected quantiles

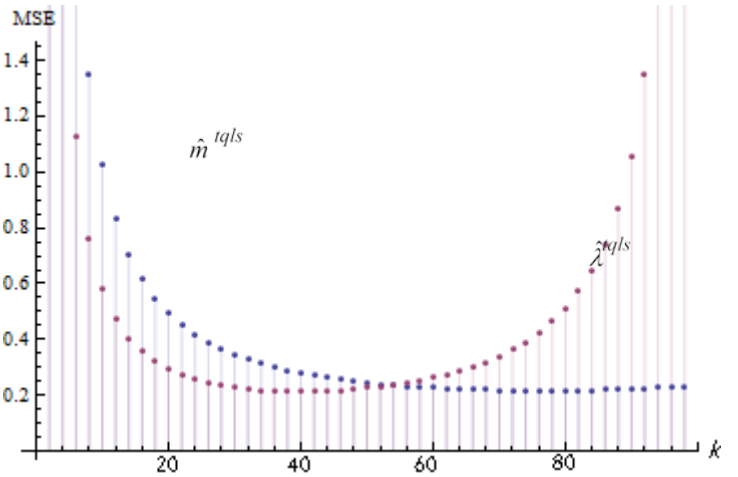


Figure 4. Dependence of the mean squared error of $Ca(0,3)$ parameter estimators obtained by TQLSM on the number of rejected quantiles

Table 2. The estimated bias and the mean squared errors for the Cauchy distribution parameter estimators obtained by the truncated quantile least squares method

Distribution random variable	n	k	$\hat{BIAS}\left(\hat{m}^{tqls}\right)$	$\hat{BIAS}\left(\hat{\lambda}^{tqls}\right)$	$\hat{MSE}\left(\hat{m}^{tqls}\right)$	$\hat{MSE}\left(\hat{\lambda}^{tqls}\right)$
$Ca(0, 3)$	60	2	- 1.1615	0.4884	14.5741	9.2934
		10	- 0.3844	0.2749	2.0822	1.2637
		20	- 0.2024	0.1687	0.9150	0.5702
		30	- 0.1460	0.1296	0.6253	0.4364
		40	- 0.1182	0.1068	0.4968	0.3990
		50	- 0.1016	0.0943	0.4303	0.4137
	100	2	- 1.1474	0.4970	15.3895	12.9429
		10	- 0,2189	0.1736	1.0284	0.5808
		30	- 0.0840	0.0716	0.3446	0.2286
		50	- 0.0585	0.0471	0.2446	0.2264
		70	- 0.0492	0.0362	0.2184	0.3392
		90	- 0.0451	0.0240	0.2222	1.0580

Table 2. The estimated bias and the mean squared errors for the Cauchy distribution parameter estimators obtained by the truncated quantile least squares method (cont.)

Distribution random variable	n	k	$\hat{BIAS}\left(\hat{m}^{tqls}\right)$	$\hat{BIAS}\left(\hat{\lambda}^{tqls}\right)$	$\hat{MSE}\left(\hat{m}^{tqls}\right)$	$\hat{MSE}\left(\hat{\lambda}^{tqls}\right)$
$Ca(3, 2)$	60	2	- 0.8459	0.3919	3.3330	1.3605
		10	- 0.3745	0.1965	0.7766	0.4571
		20	- 0.2330	0.0424	0.4033	0.1809
		30	- 0.2575	0.0192	0.3496	0.1500
		40	- 0.2631	- 0.0129	0.3209	0.1615
		50	- 0.2553	-0.0232	0.2996	0.1710
	100	2	- 0.7546	0.3381	7.3597	4.4003
		10	- 0.1461	0.1177	0.4638	0.2525
		30	- 0.0562	0.0522	0.1570	0.1014
		50	- 0.0389	0.0405	0.1111	0.1014
		70	- 0.0317	0.0363	0.0979	0.1532
		90	- 0.0292	0.0332	0.0996	0.4758

The median-quantile least squares method was also used for estimation of the Cauchy distribution parameters. The properties of the estimators are presented in Table 3 for selected parameters. This method offers a more convenient estimation algorithm in comparison to the percentile method and the truncated quantile least squares method because it does not require additional assumptions about the quantile orders or the number of rejected quantiles.

Table 3. The estimated bias and the mean squared errors for the Cauchy distribution parameter estimators obtained by the median-quantile least squares method

Distribution random variable	n	$\hat{BIAS}\left(\hat{m}^{mq}\right)$	$\hat{BIAS}\left(\hat{\lambda}^{mq}\right)$	$\hat{MSE}\left(\hat{m}^{mq}\right)$	$\hat{MSE}\left(\hat{\lambda}^{mq}\right)$
$Ca(0, 3)$	60	- 0.1136	0.0846	0.4383	0.3861
	100	- 0.0659	0.0455	0.2468	0.2181
$Ca(3, 2)$	60	- 0.0780	0.0535	0.2005	0.1685
	100	- 0.0462	0.0340	0.1119	0.0983

In Table 4 the results of simulation analysis about the considered methods: the percentile method (PM) for selected p, the truncated quantile least squares method (TQLSM) and the median-quantile least squares method (MQLSM) for selected k are presented. The estimators obtained from the percentile methods have the smallest mean squared errors for the value of p given in the table. In the case of the truncated quantile least squares estimators for k=78 the estimator $\hat{m}$ has the smallest mean squared errors and the estimator $\hat{\lambda}$ for k =40. The number of k =52 was chosen as an intermediate value for comparison of the properties of the estimators.

Table 4. The estimated mean squared errors for the Cauchy distribution parameter estimators obtained by the considered quantile methods for n = 100

Distribution random variable	Estimators	Estimation method					
		PM	*TQLSM*				MQLSM
			k=2	k=40	k=52	k=78	
Ca(0,3)	$\hat{m}$	0.2227 (p= 0.44)	15.3895	0.2787	0.2399	0.2167	0.2468
	$\hat{\lambda}$	0.2536 (p= 0.27)	12.9429	0.2135	0.2318	0.4228	0.2181
Ca(3,2)	$\hat{m}$	0.0973 (p= 0,45)	7.3597	0.1273	0.1089	0.0972	0.1118
	$\hat{\lambda}$	0.1098 (p= 0,27)	4.4003	0.0957	0.1014	0.1910	0.0983

The results of the analysis indicate that rejecting a number of the smallest and the largest quantiles significantly improved properties of the Cauchy distribution parameter estimators as compared to the quantile least squares method, which rejects only extreme statistics. In both methods estimation of each parameter requires choosing different orders and different number of rejected quantiles, which ensures the smallest mean squared errors.

The application of the median-quantile least squares method gives results which are similar to the truncated quantile least squares method on condition that the appropriate quantile order is chosen.

5. Conclusions

Quantile methods are used for estimation of the Cauchy distribution parameters because the method of moments and the maximum likelihood method cannot be used. Practical conclusions as to their application result from the simulation analysis of the estimator properties. The appropriate value of the quantile orders in the percentile method and the number of rejected quantiles in the truncated quantile least squares method lead to estimators with small bias and mean squared error.

In the case of the Cauchy distribution, which is a heavy tailed distribution, rejecting a fixed number of the smallest and the largest quantiles significantly improves the properties of the parameter estimators. In order to minimize the mean squared errors of estimators, it is possible to use different number of rejected quantiles for each estimator.

The second suggested modification of the quantile least squares method is more convenient in applications, as it does not require any additional assumptions and allows estimators with good properties to be obtained.

Both methods can be applied to estimation of the Cauchy distribution parameters. The application of these methods to estimation of other distribution parameters requires simulation analysis of the quantile orders in the percentile methods and of the number of truncated quantiles in the truncated quantile least squares method.

REFERENCES

AITCHISON, J., BROWN, J. A. C., (1975). The lognormal distribution. Cambridge University Press, Cambridge.

CASTILLO, E., HADI, A. S., BALAKRISHNAN, N., SARABIA, J. M., (2004). Extreme value and related models with application in engineering and science. Wiley Interscience, A. John Wiley & Sons, Inc. New Jersey.

GILCHRIST, W. G., (2000). Statistical modelling with quantile functions. Chapman & Hall/CRT, Boca Raton.

PEKASIEWICZ, D., (2012). The use of simulation methods to study the properties of estimators obtained by quantile method, [in:] collective work edited by Z. E. Zielinski: „The role of information technology in economic and social sciences. Innovations and interdisciplinary implications". Publishing House of Higher School of Commerce, 236–244.

WYWIAŁ, J., (2004). Introduction to statistical inference. Publishing House of University of Economics in Katowice, Katowice.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 145–152

MODELLING OF SKEWNESS MEASURE DISTRIBUTION

Margus Pihlak[1]

ABSTRACT

In this paper the distribution of random variable skewness measure is modelled. Firstly, we present some results of matrix algebra useful in multivariate statistical analyses. Then, we apply the central limit theorem on modelling of skewness measure distribution. Finally, we give an idea for finding the confidence intervals of statistical model residuals' asymmetry measure.

Key words: central limit theorem, multivariate skewness measure, skewness measure distribution, statistical model residuals.

1. Introduction and basic notations

Firstly, we introduce some notations used in the paper. The zero vector is denoted as $\mathbf{0}$. The transposed matrix $\mathbf{A}$ is denoted as $\mathbf{A}^{\mathrm{T}}$.

Let us have random vectors $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ik})^{\mathrm{T}}$ where index $i = 1,2,\ldots,n$ is for observations and k denotes the number of variables. These random vectors are independent and identically distributed copies (observations) of a random k-vector $\mathbf{X}$. Let

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

and

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{x}})(\mathbf{X}_i - \overline{\mathbf{x}})^{\mathrm{T}}$$

be the estimators of the sample mean $E(\mathbf{X}) = \boldsymbol{\mu}$ and the covariance matrix $D(\mathbf{X}) = \boldsymbol{\Sigma}$, respectively.

[1] Tallinn University of Technology (Estonia), Department of Mathematics, Ehitajate tee 5 19086 Tallinn. E-mail: margus.pihlak@ttu.ee.

Now, we present matrix operations used in this paper. One of the widely used matrix operation in multivariate statistics is Kronecker product (or tensor product) $\mathbf{A} \otimes \mathbf{B}$ of $\mathbf{A} : m \times n$ and $\mathbf{B} : p \times q$ which is defined as a partitioned matrix

$$\mathbf{A} \otimes \mathbf{B} = \left\lfloor a_{ij}\mathbf{B} \right\rfloor, \quad i = 1,2,\ldots,m;\ j = 1,2,\ldots,n.$$

By means of Kronecker product we can present the third and the fourth order moments of vector $\mathbf{X}$:

$$m_3(\mathbf{X}) = E(\mathbf{X} \otimes \mathbf{X}^{\mathrm{T}} \otimes \mathbf{X})$$

and

$$m_4(\mathbf{X}) = E(\mathbf{X} \otimes \mathbf{X}^{\mathrm{T}} \otimes \mathbf{X} \otimes \mathbf{X}^{\mathrm{T}}).$$

The corresponding central moments

$$\overline{m}_3(\mathbf{X}) = E\left\{(\mathbf{X} - \boldsymbol{\mu}) \otimes (\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} \otimes (\mathbf{X} - \boldsymbol{\mu})\right\}$$

and

$$\overline{m}_4(\mathbf{X}) = E\left\{((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}) \otimes ((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}})\right\}.$$

The operation $\mathrm{vec}(\mathbf{A})$ denotes a mn-vector obtained from $m \times n$-matrix by stacking its columns one under another in the natural order. For the properties of Kronecker product and vec-operator the interested reader is referred to Harville (1997), Kollo (1991) or Kollo and von Rosen (2005). In the next section skewness measure will be defined be means of the star-product of the matrices. The star-product was introduced in (MacRae, 1974) where some basic properties of the operation were presented and proved.

Definition 1. *Let us have a matrix* $\mathbf{A} : m \times n$ *and a partitioned matrix* $\mathbf{B} : mr \times ns$ *consisting of* $r \times s$*-blocks* B_{ij}, $i = 1,2,\ldots,m;\ j = 1,2,\ldots,n$. *Then, the star-product* $\mathbf{A} * \mathbf{B}$ *is a* $r \times s$*-matrix*

$$\mathbf{A} * \mathbf{B} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}\mathbf{B}_{ij}.$$

The star-product is an inverse operation of Kronecker product in a sense of increasing and decreasing the matrix dimensions. One of the star-product applications is presented in the paper (Pihlak, 2004).

We also use the matrix derivative defined following Neudecker (1969).

Definition 2. *Let the elements of a matrix* $\mathbf{Y} : r \times s$ *be functions of a matrix* $\mathbf{X} : p \times q$. *Assume that for all* $i = 1,2,\ldots,p$, $j = 1,2,\ldots,q$, $k = 1,2,\ldots,r$ *and* $l = 1,2,\ldots,s$ *partial derivatives* $\dfrac{\partial Y_{kl}}{\partial X_{ij}}$ *exist and are continuous in an open set A. Then, the matrix* $\dfrac{d\mathbf{Y}}{d\mathbf{X}}$ *is called matrix derivative of the matrix* $\mathbf{Y} : r \times s$ *by the matrix* $\mathbf{X} : p \times q$ *in a set A, if*

$$\frac{d\mathbf{Y}}{d\mathbf{X}} = \frac{d}{d\text{vec}^T(\mathbf{X})} \otimes \text{vec}(\mathbf{Y})$$

where

$$\frac{d}{d\text{vec}^{\mathrm{T}}(\mathbf{X})} = \left(\frac{\partial}{\partial x_{11}}, \ldots, \frac{\partial}{\partial x_{p1}}, \ldots, \frac{\partial}{\partial x_{1q}}, \ldots, \frac{\partial}{\partial x_{pq}} \right).$$

The matrix derivative defined by Definition 2 is called Neudecker matrix derivative. This matrix derivative has been in the last 40 years a useful tool in multivariate statistics.

2. Multivariate measures of skewness

In this section we present a multivariate skewness measure by means of the matrix operation described above. A skewness measure in multivariate case was introduced in Mardia (1970). Mori et al. (1993) introduced a skewness measure as a vector. B. Klar (2002) gave a thorough overview of the skewness problem. In this paper asymptotic distribution of different skewness characteristics is also examined. In Kollo (2008) a skewness measure vector is introduced and applied in Independent Component Analyses (ICA). In this paper we give an idea for the application of a skewness measure to residuals of statistical models. Our aim is to estimate the distribution of skewness measure and to find confidence intervals of the asymmetry characteristics.

The skewness measure in the multivariate case is presented through the third order moments:

$$\mathbf{s}(\mathbf{X}) = E(\mathbf{Y} \otimes \mathbf{Y}' \otimes \mathbf{Y}) \tag{2.1}$$

where

$$\mathbf{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}).$$

In Kollo (2008) the skewness measure based on (2.1) is introduced by means of the star product:

$$\mathbf{b}(\mathbf{X}) = \mathbf{1}_{k \times k} * \mathbf{s}(\mathbf{X}) \tag{2.2}$$

where $k \times k$ -matrix

$$\mathbf{1}_{k \times k} = \begin{pmatrix} 1 & \cdots & 1 \\ \cdots & \ddots & \cdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

In Kollo and Srivastava (2004) the Mardia's skewness measure is presented through the third order moment:

$$\beta = \text{tr}(m_3^{\mathrm{T}}(\mathbf{Y}) m_3(\mathbf{Y}))$$

where operation tr denotes the trace of matrix.

The equality (2.2) generalizes the univarite ($k = 1$) skewness measure

$$b(X) = \frac{E(X-\mu)^3}{\sigma^3}$$

where σ denotes standard deviation of the random variable X. Thus, we can express the estimator of the univariate skewness measure as

$$\hat{b}(X) = \frac{E(X_i - \bar{x})^3}{s^3} \tag{2.3}$$

where s denotes unbiased estimator of standard deviation σ and $\bar{x}$ is the sample mean estimator.

3. Modelling distribution of the univariate skewness measure

In this section we model the distribution of the random variable $\hat{b}(X)$ defined by the equation (2.3). Let us have independent and identically distributed random variables $X_1, X_2, \ldots, X_n$.

Let $E(X) = \mu$ and $D(X) = \sigma^2$. Then, according to the central limit theorem the distribution of the random variable $\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}$ converges to the normal distribution $N(0,1)$. In the multivariate case the distribution of the random vector $\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})$ converges to normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$.

Let us have $k + k^2$-vector

$$\mathbf{Z}_n = \begin{pmatrix} \bar{\mathbf{x}} \\ \mathrm{vec}(\mathbf{S}) \end{pmatrix}.$$

Then

$$\sqrt{n}(\mathbf{Z}_n - E(\mathbf{Z}_n)) \mapsto N(\mathbf{0}, \boldsymbol{\Pi})$$

in distribution. Here, $(k^2 + k) \times (k^2 + k)$-dimensional partitioned matrix

$$\boldsymbol{\Pi} = \begin{pmatrix} \boldsymbol{\Sigma} & \bar{m}_3^{\mathrm{T}}(\mathbf{X}) \\ \bar{m}_3(\mathbf{X}) & \boldsymbol{\Pi}_4 \end{pmatrix}$$

where $k^2 \times k^2$ -matrix $\boldsymbol{\Pi}_4 = \bar{m}_4(\mathbf{X}) - \mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}^{\mathrm{T}}(\boldsymbol{\Sigma})$ (Parring, 1979). This convergence can be generalized by means of the following theorem.

Theorem 1. Let $\{\mathbf{Z}_n\}$ be a sequence of $k+k^2$-component random vectors and $\boldsymbol{\nu}$ be a fixed vector such that $\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\nu})$ has the limiting distribution $N(\mathbf{0}, \boldsymbol{\Pi})$ as $n \to \infty$. Let the function $g: R^{k^2+k} \to R^k$ have continuous partial derivatives at $\mathbf{z}_n = \boldsymbol{\nu}$. Then, the distribution of random variable $\sqrt{n}\{g(\mathbf{Z}_n) - g(\boldsymbol{\nu})\}$ converges to the normal distribution $N(\mathbf{0}, g_{\mathbf{z}_n}^{\mathrm{T}} \boldsymbol{\Pi} g_{\mathbf{z}_n})$ where $(k^2+k)\times k$ -matrix

$$g_{\mathbf{z}_n} = \frac{dg(\mathbf{z}_n)}{d\mathbf{z}_n}\Big|_{\mathbf{z}_n = \boldsymbol{\nu}}$$

is Neudecker matrix derivative at $\mathbf{z}_n = \boldsymbol{\nu}$.

The proof of the theorem can be found in the book of T. W. Anderson (2003, page 132). In Theorem 1 vector $\boldsymbol{\nu} = E(\mathbf{Z}_n)$.

In the univariate case

$$\mathbf{Z}_n = \begin{pmatrix} \overline{X} \\ s^2 \end{pmatrix},$$

$E(\mathbf{Z}_n) = \begin{pmatrix} \mu & \sigma^2 \end{pmatrix}^{\mathrm{T}}$ and the function $g(\mathbf{z}_n)$ is defined by equality (2.3).

In this case 2×2 -matrix

$$\boldsymbol{\Pi} = \begin{pmatrix} \sigma^2 & \overline{m}_3(X) \\ \overline{m}_3(X) & \Pi_4 \end{pmatrix}$$

where

$$\Pi_4 = \overline{m}_4(X) - \sigma^4.$$

Let us take

$$g(\mathbf{z}_n) = g(\overline{x}, s^2) = \hat{b}(X) = \frac{E(X-\overline{x})^3}{(s^2)^{\frac{3}{2}}}.$$

We get

$$g_{\mathbf{z}_n} = \left(\frac{\partial}{\partial \overline{x}} \frac{E(X-\overline{x})^3}{(s^2)^{\frac{3}{2}}} \quad \frac{\partial}{\partial s^2} \frac{E(X-\overline{x})^3}{(s^2)^{\frac{3}{2}}} \right)^{\mathrm{T}} \Bigg|_{\mathbf{z}_n = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}} =$$

$$= \left(\frac{-3E(X-\overline{x})^2}{s^3} \quad \frac{-3E(X-\overline{x})^3}{2s^5} \right)^{\mathrm{T}} \Bigg|_{\mathbf{z}_n = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}} =$$

$$= -\left(\frac{3}{\sigma} \quad \frac{3\overline{m}_3(X)}{2\sigma^5} \right)^{\mathrm{T}}.$$

According to the Theorem 1 we can say that the random variable $\sqrt{n}\left\{\hat{b}(X)-b(X)\right\}$ has approximately normal distribution $N(0,\sigma_b^2)$. Variance σ_b^2 can be found in the following way:

$$\sigma_b^2 = g_{\mathbf{z}_n}^{\mathrm{T}}\mathbf{\Pi} g_{\mathbf{z}_n} =$$

$$= \begin{pmatrix} \frac{-3}{\sigma} & \frac{-3\overline{m}_3(X)}{2\sigma^5} \end{pmatrix} \begin{pmatrix} \sigma^2 & \overline{m}_3(X) \\ \overline{m}_3(X) & \Pi_4 \end{pmatrix} \begin{pmatrix} \frac{-3}{\sigma} \\ \frac{-3\overline{m}_3(X)}{2\sigma^5} \end{pmatrix} =$$

$$= 9\frac{\overline{m}_3^2(X)}{\sigma^6}\left(\frac{\Pi_4}{4\sigma^4}+1\right)+9.$$

Thus, we have

$$\sigma_b^2 = 9\frac{\overline{m}_3^2(X)}{\sigma^6}\left(\frac{\Pi_4}{4\sigma^4}+1\right)+9. \tag{3.1}$$

Example. Let us generate *m* times random variable *X* with a sample size *n*. Let random variable *X* have exponential distribution with parameter $\lambda > 0$. Then, the *i*-th order moment $E(X^i)=\frac{i!}{\lambda^i}$. Using these moments we get that $\overline{m}_3(X)=\frac{2}{\lambda^3}$ and $\Pi_4=\frac{8}{\lambda^4}$. According to the formula (3.1) variance $\sigma_b^2=117$. Thus, we get the following approximate 0.95-confidence interval for the skewness measure $b(X)$:

$$\hat{b}(X)\pm\frac{1.96\sqrt{117}}{\sqrt{nm}}.$$

4. Summary: skewness confidence intervals for statistical models

The problem concerns the estimation of statistical models. This is the problem of skewness or lack of symmetry, which means the distribution of statistical model residuals is frequently non-gaussian, as Kolmogorov-Smirnov test shows. In this case the skewness has to be estimated for testing the goodness of models. The confidence intervals of that parameter have to be found. This enables us to improve the diagnosis of statistical models. By means of skewness confidence intervals we can estimate the influence of outliers. These outliers are typical in

forestry. The main question is: does the zero value belong to the estimated confidence interval? To answer this question we can estimate the variance of residuals by means of equality (3.1). This variance depends on standard deviation, skewness and kurtosis of residuals.

Acknowledgements

This paper is supported by Estonian Ministry of Education and Science target financed theme No. SF0140011s09.

REFERENCES

ANDERSON, T. W., (2003). An Introduction to Multivariate Statistical Analysis. Wiley Interscience.

HARVILLE, A., (1997). Matrix Algebra from a Statistician's Perspective. Springer, New York.

KLAR, B., (2002). A Treatment of Multivariate Skewness, Kurtosis, and Related Statistics. Journal of Multivariate Analysis, 83, 141–165.

KOLLO, T., (1991). Matrix Derivative in Multivariate Statistics. Tartu University Press, Tartu (in Russian).

KOLLO, T., SRIVASTAVA, M. S., (2004). Estimation and testing of parameters in multivariate Laplace distribution. Comm. Statist., 33, 2363–2687.

KOLLO, T., VON ROSEN, D., (2005). Advanced Multivariate Statistics with Matrices. Springer, Dordrecht.

KOLLO, T., (2008). Multivariate skewness and kurtosis measures with an application in ICA. Journal of Multivariate Analyses, 99, 2328–2338.

MacRAE, E. C., (1974). Matrix derivatives with an application to an adaptive linear decision problem. Ann. Statist., 2, 337–346.

MARDIA, K. V., (1970). Measures of multivariate skewness and kurtosis measures with applications. Biometrika, 57, 519–530.

MORI, T. F., ROHATGI, V. K., SZÉKELY., (1993). On multivariate skewness and kurtosis. Theory Probab. Appl, 38, 547–551.

NEUDECKER, H., (1969). Some theorems on matrix differentiations with special reference to Kronecker matrix products. Journal of the American Statistical Association, 64, 953–963.

PARRING, A-M., (1979). Estimation asymptotic characteristic function of sample (in Russian). Acta et Commetationes Universitatis Tartuensis de Mathematica, 492, 86–90.

PIHLAK, M., (2004). Matrix integral. Linear Algebra and its Applications, 388, 315–325.

STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 153–170

AN ANALYSIS OF THE POPULATION AGING PHENOMENA IN POLAND FROM A SPATIAL PERSPECTIVE

Justyna Wilk[1], **Michał Bernard Pietrzak**[2]

ABSTRACT

The processes of socio-economic development are continuously accompanied by the process of population aging. It is seen as a growing percentage share of people aged 65 and over in the general population. It covers the majority of the European Union countries and also refers to Poland. The objective of the paper is to analyse the population aging phenomenon from a spatial perspective. The study has been carried out for 66 subregions (NUTS 3) and covered the period 1995-2012.

Poland is characterized by strong spatial diversification regarding the proportion of senior citizens and its growth rate, and also determinants exerting impact on the demographic aging processes. Demographically the youngest and slowest aging population lives in south-eastern and also central Poland. The most intensive population aging processes are seen in the selected subregions of south-western Poland. Here, we observe extremely low fertility, demographically old working-age population and also significant migration outflow of younger people.

Key words: population aging, socio-economic development, spatial approach, taxonomic analysis, regression analysis.

1. Introduction

The paper discusses the problem of demographic aging of the Polish population. Aging processes are defined as changes in the age structure of population where the percentage of older population compared to the total population number is increasing (Rosset, 1959; Frątczak, 1984; Uhlenberg, 2009). They are accompanied by social, cultural and economic factors; among other things, intentional delaying of procreation time and changing life priorities,

[1] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. E-mail: Justyna.Wilk@ue.wroc.pl.
[2] Nicolaus Copernicus University in Toruń, Department of Econometrics and Statistics, Toruń. E-mail: Michal.Pietrzak@umk.pl.

leading more and more towards a healthy life style, progress in medicine and wider access to medical services, etc.

These processes are present in almost all European Union countries and also refer to Poland (see Kurkiewicz et. al., 2006, 2012; Muenz, 2007; Giannakouris, 2008; Kurek, 2008; Strzelecki, 2009; Lindh and Malmberg, 2009; Kurkiewicz, 2010; Dragan, 2011; Hrynkiewicz, 2012; Pociecha, 2003). The processes of demographic aging are affected by a slow, although often irreversible, change in time. From the perspective of the state policy it is observed as a major problem which strongly determines the situation in the country in terms of its financial, social and economic issues.

In the long-term prospect, without taking appropriate actions in the socio-economic sphere, the population aging processes can lead to a gap in the labour market, disturb the retirement system, decrease the efficiency of social systems (e.g. health care, welfare). The growing proportion of the older population also imposes the need to adapt adequate social policy within the framework of which indispensable care will be offered to people included in this age group (see Golinowska, 2008; Jurek, 2012; Magnus, 2008; Prskawetz and Lindh, 2011; Wilk and Bartłomowicz, 2012).

There are significant regional disparities in socio-economic development processes in Poland. This affects the conditions of the population aging processes. Therefore, we can assume that the dynamics of the population aging processes are also spatially diversified and their conditions are peculiar to particular regions of the country. The objective of the paper is to examine the level and rate of the population aging processes in Poland, and also to reveal their demographic conditions from a spatial perspective.

The analysis of the population aging phenomenon was carried out for 66 subregions (NUTS 3) and covered the period 1995-2012. In the first part of the paper the degrees and rates of the population aging processes in subregions will be discussed. Therefore, the subregions demonstrating significant progress in the population aging processes will be revealed.

In the next part of the paper the econometric model will be constructed. The investigation will be carried out based on the synthetic measure values specified considering the proportion of senior citizens, as well as the growth rate of this proportion. Next, for the selected demographic factors (such as fertility, migrations, etc.), the identification of their relations with the population aging processes will be carried out. In the third part, the demographic conditions of population aging processes in Polish subregions will be presented.

2. The population aging phenomena in Poland

2.1. The background

The overall trends in changing the age structure are shown by the shape of the age pyramid, which is a graphical illustration of the distribution of various age

groups by sex in a population, and its changes in time. It forms the shape of a triangle when the young population is growing, whereas a diagram in the shape of an upside-down vase is characteristic of a demographically old population.

In 1975 the Polish population was relatively young demographically. There was a higher percentage proportion of young people than old people in the total population. The age pyramid formed a shape similar to a pyramid. In subsequent years the shape of the diagram changed; the middle of the pyramid extended while the lower part narrowed. The percentage share of young population significantly decreased in the period before 2012 while the number of older people increased (see Figure 1).

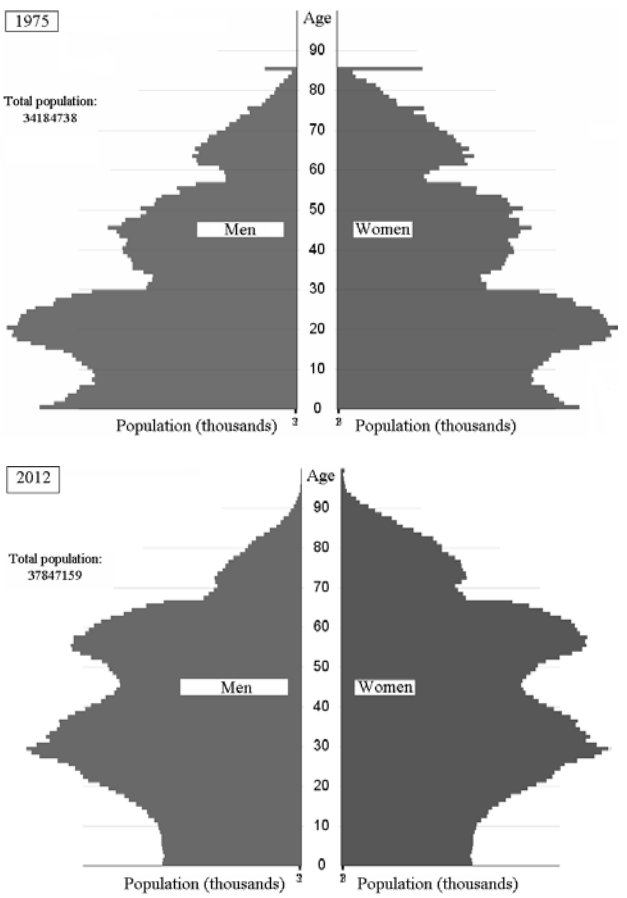


Figure 1. Age pyramids for Poland in 1975 and 2012

Source: See Central Statistical Office of Poland, http://www.stat.gov.pl/PI_gus/ludnosc_piramida/struktura_ludnosci.svg.

2.2. The proportion of senior citizens and its growth rate in Poland 1995-2012, by subregions

Population aging is demonstrated by a high proportion of senior citizens (people aged 65 and over) and also by the dynamics of this proportion. The proportion indicates the current state of the population aging processes in a region (saturation), while the growth rate indicates the direction of changes and their dynamics. Both variables function as stimulants of the population aging processes. High values of these variables confirm the advancement (intensity) of the population aging processes in a region.

In 2012 the proportion of senior citizens in the total population of Poland reached 14.24%, which was approximately 3.0 percentage points more than in 1995. This implies 1.41% yearly increases (see Table 1). Polish subregions demonstrated diversified proportion of senior citizens, from approximately 10.0% in Gdański subregion to approximately 19.0% in the city of Łódź. Positive increase rates of the proportion, to a greater or lesser degree, were observed in all subregions in the period 1995-2012. The values of the growth rate near to zero were observed in Krakowski (0.23%) and Sandomiersko-Jędrzejowski (0,26%) subregions, while very high dynamics were demonstrated by Rybnicki (3.44%) and Gliwicki (3.19%) subregions.

Table 1. Basic statistics for the proportion of senior citizens and its growth rate

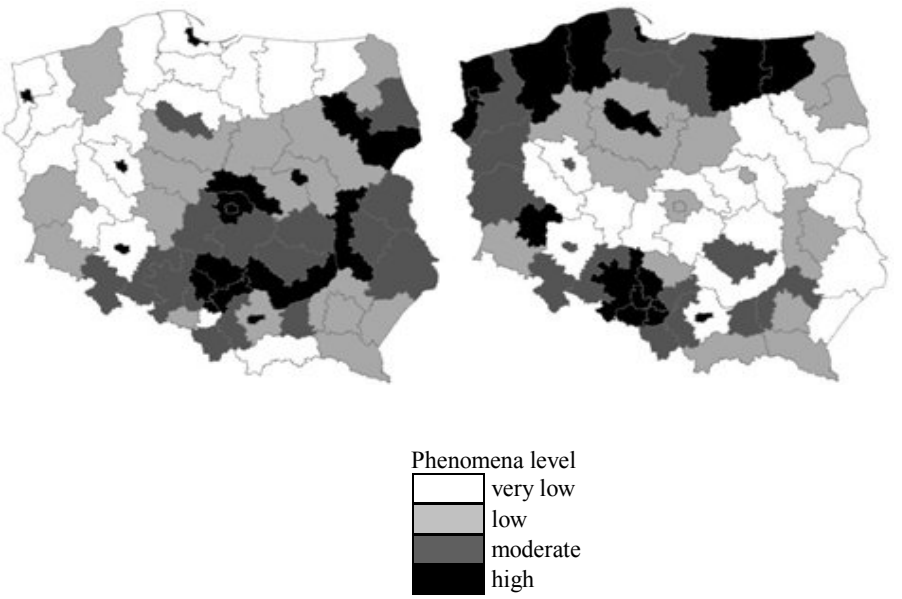
Specification	The proportion of senior citizens (population aged 65 and over) in 2012 (%)	The growth rate of the proportion of senior citizens in the period 1995-2012 (%)
POLAND	14.24	1.41
Minimum	9.9	0.23
Maximum	18.6	3.44
Coefficient of variation (%)	12.2	46.6
Pearson's correlation [-1, 1]	0.08	

Source: own estimation based on data provided by the Local Data Bank of the Central Statistical Office of Poland.

The subregions were divided, on the basis of quartiles, into four classes representing very low, low, moderate and high proportion of senior citizens. The same procedure was conducted in respect to the growth rate of the proportion of senior citizens. Figure 2 presents the results of this classification. Relatively high proportions of senior citizens were observed for subregions located in eastern, central and also in a part of south Poland. In particular, this situation occurred in the biggest cities of Poland such as the cities of Warsaw, Łódź, Poznań, Wrocław, Szczecin, Tricity and Cracow.

Relatively high and moderate levels were recorded within Łódzkie, Śląskie, Lubelskie, Opolskie, Świętokrzyskie and Podlaskie voivodeships (NUTS 2). Conversely, very low proportions of senior citizens were observed within

Zachodniopomorskie, Wielkopolskie, Pomorskie, Lubuskie and Warmińsko-Mazurskie voivodeships, and low proportions within Mazowieckie and Podkarpackie voivodeships.



a) the proportion of senior citizens in the total population in 2012

b) the growth rate of the proportion of senior citizens in the period 1995-2012

Figure 2. Spatial diversification of the proportion of senior citizens and its growth rate Source: own compilation based on data provided by the Local Data Bank of the Central Statistical Office of Poland

The situation is quite different regarding the growth rate of the proportion. Relatively high or moderate dynamics were characteristic of the subregions of Zachodniopomorskie, Pomorskie, Warmińsko-Mazurskie, Lubuskie, Opolskie and Śląskie voivodeships. A very low or low growth rate was true for Wielkopolskie, Łódzkie, Mazowieckie, Podlaskie, Lubelskie and Podkarpackie voivodeships. The highest dynamics were recorded in northern, western and south-western Poland, and also in selected bigger cities.

It is also interesting that discussed indicators are not statistically correlated (see Table 1). Therefore, we cannot conclude that the higher the growth rate, the bigger (or lower) the proportion of senior citizens in Polish subregions. The exception is Śląskie voivodeship, for which the values of both variables represent a relatively high percentage share. This is manifested in the intense advancement of the population aging processes occurring in this subregion.

3. Modelling the population aging processes in Poland

3.1. The dependent variable

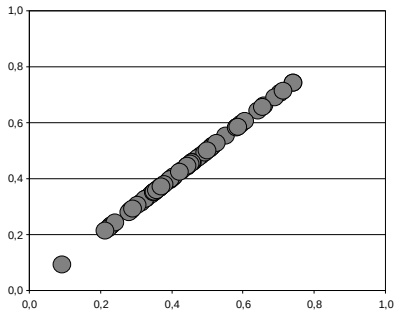
In order to examine the advancement of the population aging processes in Poland and their conditions the regression analysis was applied. The synthetic measure of the intensity of the population aging processes served as the dependent variable while several significant demographic indicators formed a set of explanatory variables. The application of a synthetic measure in cause-effect models was proposed in Hellwig, Siedlecka and Siedlecki, 1995.

The approach using a taxonomic measure of development (TMD) was applied to assess the spatial diversification of the population aging processes in Polish subregions. It allows us to cover a set of indicators at the same time and to provide the synthetic description of the situation regarding the analyzed phenomenon (see Grabiński, Wydymus and Zeliaś, 1989; Nowak, 1990; Młodak, 2006).

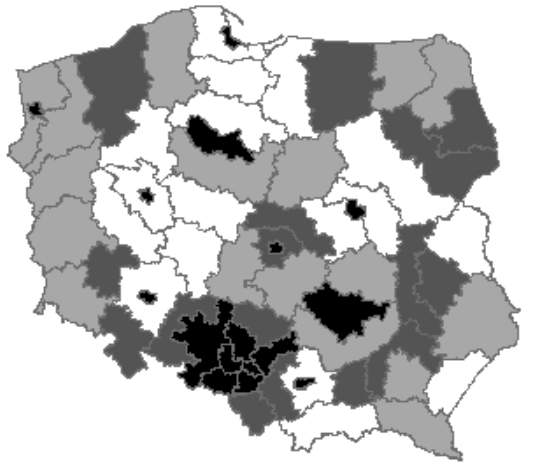
The construction of TMD was based on two diagnostic variables presented above, such as the proportion of senior citizens in 2012 and also the growth rate of the senior citizens proportion in the period 1995-2012. Both variables function as stimulants of the population aging processes. Therefore, the highest values of variables reached by the Polish subregions served as the features of the pattern object, while the lowest values of variables determined the coordinates of the anti-pattern object. These two ideal objects (the pattern object and anti-pattern object) constituted referring points in the comparative analysis.

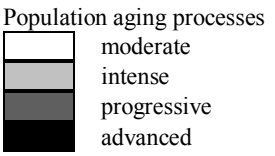
The values of the variables were normalized, using the unitization with zero as the minimum, to standardize their implementation intervals. For each subregion the distances to the pattern object and anti-pattern object were calculated using Euclidean distance. Application of the TOPSIS formula (see Hwang and Yoon, 1981) resulted in determination of the TMD value for each subregion. High TMD values confirm the advancement of the population aging process in a subregion while its low values are seen as a moderate intensity of the phenomenon.

TMD takes values of [0, 1] while the Polish subregions recorded values from approximately 0.20 to approximately 0.70, apart from Gdański subregion which took the value of 0.91 (see Figure 3a). The situation in Poland is relatively highly diversified from a spatial perspective; the classical coefficient of variation took the value of 32%. On the basis of the calculated TMD values the subregions were divided into four classes using quartiles. The obtained classes illustrate advanced (TMD took values above 0.52), progressive (between 0.42 and 0.51), intense (between 0.36 and 0.41) or moderate (up to 0.35) levels of the population aging processes advancement (see Figure 3b).



a) the dispersion of TMD values in Polish subregions





b) classes of Polish subregions according to TMD values

Figure 3. Spatial diversification of the population aging processes in Poland
Source: Own compilation.

Advanced population aging processes are observed in south-western Poland. In particular the advanced and progressive situation is typical of Śląskie and Opolskie voivodeships. Moderate or intense progresses relate to Mazowieckie, Wielkopolskie, Lubuskie, Pomorskie and selected subregions of Małopolskie voivodeship.

3.2. Explanatory variables

High territorial disparities of population aging processes in Poland are seen in the specificities of Polish subregions which differ in the social and cultural conditions: the dynamics of economic growth and socio-economic development, and also in respect of the resources of the natural environment and its condition (see, e.g. Müller-Frączek and Pietrzak, 2009a, 2009b; Bal-Domańska and Wilk, 2011; Pietrzak, 2012; Wilk and Bartłomowicz, 2012; Wilk, Pietrzak and Matusik, 2013). This affects the economic and financial situation of households, the lifestyle led and priorities held, health condition of society, access to medical services, etc.

All these factors have a significant influence on demographic processes occurring in subregions; this is reflected in statistical data. Propensity to procreate and the duration of life are regarded as the basic determinants of the population aging processes from the perspective of the state. The migration outflow of the young (working-age) population to other regions or abroad can also be a significant factor, as seen from the regional perspective.

The population aging processes are also intensified by the aging of the working-age population (population aged between 18 years and retirement age). The higher the proportion of people aged 45 and over in this group, the bigger the "portion" of people who will supply the population of senior citizens in the future. The population aging processes and their economic consequences are also determined by the proportion of the oldest-old population (population aged 80 and over) in the population of senior citizens, and also its growth rate. Table 2 presents a set of accepted explanatory variables.

Table 2. The set of explanatory variables

No.	Variable name	Definition
1	Total fertility rate (person)	The mean number of children that would be born alive to a woman during her lifetime if she were to pass through her childbearing years conforming to the fertility rates by age of a given year
2	Life expectancy of men at the age of 65 (year)	The mean number of years still to be lived by a man who has reached the age of 65, if subjected throughout the rest of his life to the current mortality conditions (age-specific probabilities of dying)
3	Net migration rate of population aged 20-59 (person)	Net migration of people aged between 20 and 59, expressed per 10 000 inhabitants in this age group. Data covers registered migration inflows and outflows for permanent residence between subregions and abroad
4	Working-age population aging rate (%)	The percentage share of immobile working-age people in the working-age population. The working-age population is defined as people aged 18 to 59 (women) and 18 to 64 (men), while the group of people aged 45 and over is seen as immobile working-age population
5	Oldest-old-age population rate (%)	The percentage share of people aged 80 and over in the total population aged 65 and over

Source: Own compilation according to the definitions of Central Statistical Office of Poland.

Extremely low fertility has been typical of Poland for many years and does not provide the so-called simple generation substitutability which is seen as the level 2.1 of total fertility rate. Although a slight increase of the indicator value was observed within the last 10 years, in 2012 the value reached only 1.3 in Poland (see Table 3). However, in the majority of Polish subregions (apart from Gdański subregion) there were only 1.5 children per woman. In respect to the current socio-economic processes occurring in Poland, we cannot expect rapid changes to this situation even in the long-term perspective. On the other hand, we notice a decrease in regional disparities regarding the fertility level within Poland.

Low fertility is accompanied by extending life duration in Poland. According to statistical data, a man aged 65 is expected to live a further 15.4 years. It means that in average terms he would be up to a little more than 80 years old. The situation differs for Polish women who, on average, live 4.3 years longer than men. Statistical correlation between values of these two indicators (for men and women) is relatively high; Pearson's correlation took a value above 0.7, while territorial disparities are slightly higher for men.

In the period 2007-2012 yearly increases in life expectancy of men aged 65 were recorded by all subregions. However, their dynamics is territorially diversified; the difference between subregions reaches 2.7 years. The shortest expected life duration in 2012 was typical of men living in Skierniewicki (79.2 years), Elbląski and Łódzki (79.3 years) and also Starogardzki (79.4 years) subregions. The longest life expectancy of men at the age of 65 is observed in big agglomerations, such as the cities of Warsaw and Wrocław, where an average man is expected to live for 82 years.

In some regions of Poland the population aging processes are significantly affected by migration outflows of young people. Approximately, three in four subregions showed a negative balance of migration flows in 2012. The highest intensity of the phenomenon is typical in Łomżyński subregion, where the net migration coefficient reached -55.2. A similar situation (negative net migration rate under 50.0) was in Puławski subregion, which is, as Łomżyński subregion, located in eastern Poland, and also adjoins Mazowieckie voivodeship.

Very high positive values (above 50.0) of net migration coefficient were presented by 7 subregions while extremely high value (100.2) was observed in Poznański subregion, which surrounds the city of Poznań. This means that many more young people settle down in this subregion than leave it. This results from the suburbanization processes (see e.g. Matusik, Pietrzak and Wilk, 2012; Pietrzak et al., 2012; Pietrzak, Drzewoszewska and Wilk, 2012; Pietrzak, Wilk and Matusik, 2013a, 2013b; Pietrzak and Wilk, 2013; Wilk and Pietrzak, 2013; Pietrzak, Wilk and Siekaniec, 2013).

Table 3. Basic statistics for explanatory variables

No	Variable name	Year*	POLAND	Minimum	Maximum	Coefficient of variation (%)
1	Total fertility rate (person)	2002	1.249	0.893 (Wrocław)	1.614 (Nowosądecki)	13.3
		2012	1.299	1.091 (Kraków)	1.632 (Gdański)	8.4
2	Life expectancy of men at the age of 65 (years)	2007	14.6	13.6 (Włocławski)	16.4 (Warszawa)	4.1
		2012	15.4	14.2 (Skierniewicki)	16.9 (Warszawa)	4.0
3	Net migration rate of population aged 20-59 (person)	1995	x	x	x	x
		2012	x	-55.2 (Łomżyński)	100.2 (Poznański)	x
4	Working-age population aging rate (%)	2009	37.8	33.7 (Nowosądecki)	41.7 (Łódź)	4.4
		2012	37.4	34.4 (Nowosądecki, Rzeszowski)	41.1 (Sosnowiecki)	4.1
5	Oldest-old-age population rate (%)	2005	20.3	16.2 (Rybnicki)	24.2 (Sandomiersko-jędrzejowski)	8.3
		2012	26.3	19.5 (Rybnicki)	31.3 (Łomżyński)	8.4

* according to the availability of statistical data.

Explanations: "x" – not applicable.

Source: Own estimation based on data provided by Local Data Bank of Central Statistical Office of Poland.

In Poland we can observe a relatively high percentage share of old people within the working-age population. In 2009-2012 more than one in three people at the working age was 45 years old or over. The youngest working-age populations live in Nowosądecki and Rzeszowski subregions (34.4% immobile working-age people) while the oldest population lives in Sosnowiecki subregion (41.1%).

Oldest-old population is also increasing in Poland; the growth rate of the oldest-old-age population coefficient is relatively high. The proportion of this group in the population of senior citizens increased by 6 percentage points within the last 7 years. Values of this indicator are spatially diversified; the differences reach over 10%. The highest proportion of the oldest-old-age people is typical of Łomżyński subregion, where one in three senior citizens represents an oldest-old-age person. Relatively high values of this indicator are also observed for the Sandomiersko-Jędrzejowski subregion (30.1%), the city of Warsaw (29.9%), Bialski subregions (29.8%) and Suwalski subregion (29.7%).

4. Demographic conditions of the population aging processes in Poland

The estimated values of the structural parameters of the regression model, representing explanatory variables impacts, were estimated using the least squares method. In this way the significance and the impact direction of the adopted variables on the phenomenon of population aging in the Polish subregions was analysed.

Table 4 presents the results of the estimation. They turned out statistically significant for of all explanatory variables. This means that each of the examined factors exerts a significant impact on the population aging processes in Poland. The resulting value of the determination coefficient confirms high adjustment of the model to the empirical data.

Table 4. Results of the estimation of the parameters of the regression model

No	Variable name	Estimate	*p*-value*
1	Total fertility rate	-0.4013	0.0034
2	Life expectancy of men at the age of 65	0.1539	0.0001
3	Net migration rate of population aged 20-59	-0.0071	0.0169
4	Working-age population aging rate	0.0492	0.0001
5	Oldest-old-age population rate	0.0259	0.0001
Determination coefficient		0.7356	

* at 5% degree of significance.

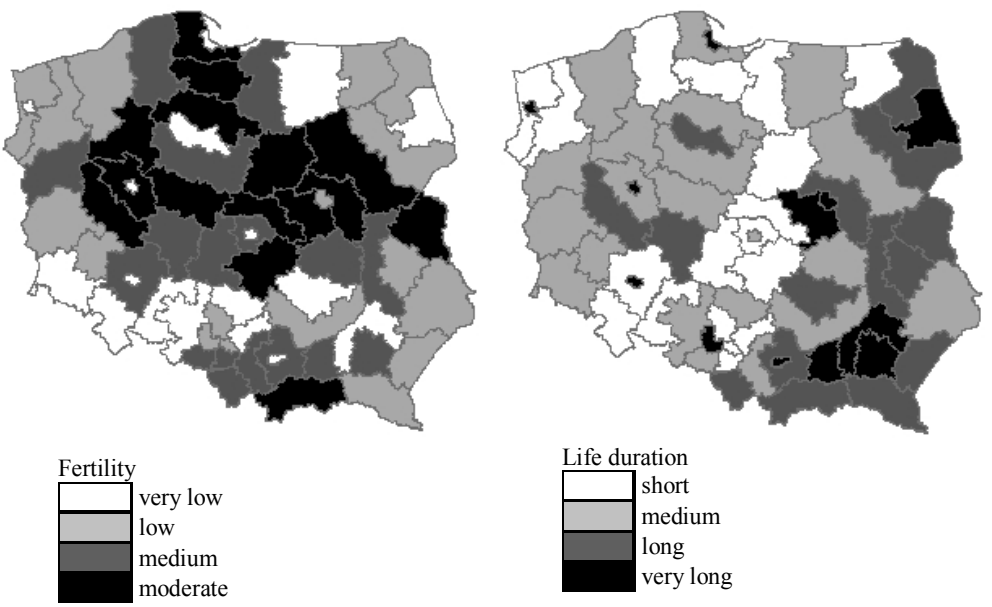
Source: Own estimation based on data provided by Local Data Bank of Central Statistical Office of Poland in R-CRAN.

The highest estimate (considering absolute values) was determined for the total fertility rate. A relatively high value of the structural parameter was also showed by the indicator of life expectancy. The negative estimate was determined for the total fertility rate, while the positive value is typical of the life expectancy. This indicates inversely proportional influence of fertility and directly proportional relations between life expectancy and the synthetic measure. Therefore, the progress in the population aging processes in the Polish subregions is most significantly affected by low fertility and extending the duration of life.

The subregions were divided into four classes (according to quartiles), representing very low (the indicator takes values of [1.00, 1.22)), low [1.22, 1.30), medium [1.30, 1.37) and moderate [1.37, 1.65] level of fertility (see Figure 4a). It can be seen that moderate or medium level is present in the subregions of Pomorskie, Wielkopolskie, Mazowieckie, Łódzkie and Małopolskie voivodeships. Additionally, moderate intensity of births occurs for the following subregions: Grudziądzki (Kujawsko-pomorskie voivodeship), Skierniewicki and Piotrowski (Łódzkie voivodeship), Nowosądecki (Małopolskie voivodeship) and Bialski (Lubelskie voivodeship).

On the other hand, it is quite characteristic to observe a very low level of fertility in the case of the largest Polish cities. Extremely low fertility was also recorded in Opolski region, as well as in Jeleniogórski and Wałbrzyski (Dolnośląskie voivodeship), Częstochowski and Sosnowiecki (Śląskie voivodeship), Kielecki (Świętokrzyskie voivodeship), Tarnobrzeski (Podkarpackie voivodeship), Białostocki (Podlaskie voivodeship) and Olsztyński (Warmińsko-mazurskie voivodeship) subregions.

The estimate of the regression parameter relating to life expectancy of men aged 65 was 0.1539, which is nearly three times less than in the case of the fertility rate. According to the values of the variables we can distinguish subregions with short (the indicator takes values of [14.1, 14.8)), medium [14.8, 15.2), long [15.2, 15.6) and very long [15.6, 16.9) expected duration of life (see Figure 4b). In this case, very low life expectancy is characteristic of the subregions of Łódzkie voivodeship and some subregions of Zachodniopomorskie, Pomorskie, Warmińsko-Mazurskie, Dolnośląskie and Śląskie voivodeships. Relatively high values of the indicator were recorded in the subregions related to regional capital cities.



a) total fertility rate in 2012 b) life expectancy of men at the age of 65 in 2012

Figure 4. Spatial diversification of the most significant factors affecting the population aging processes in Polish subregions

Source: Own compilation based on data provided by Local Data Bank of Central Statistical Office of Poland.

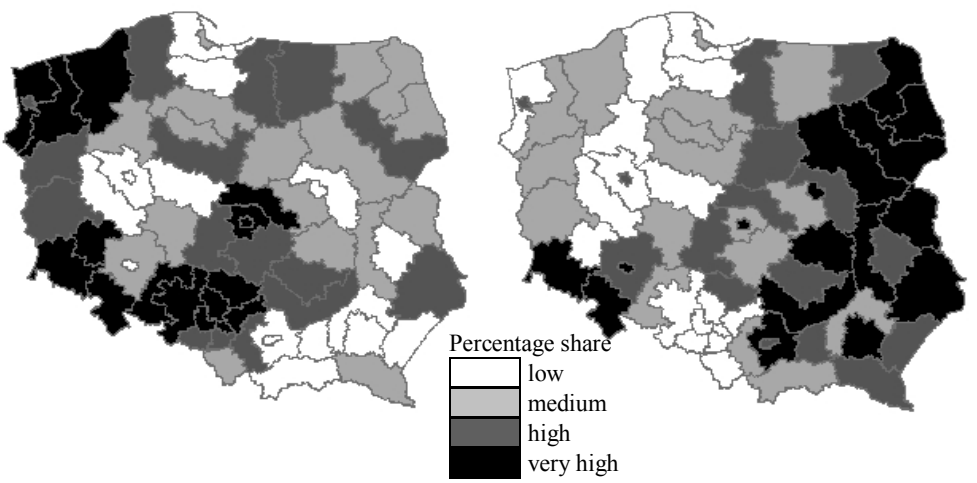
Very low fertility and simultaneously long life expectancy are observed in Białostocki (Podlaskie voivodeship), Tarnobrzeski (Podkarpackie voivodeship) and Gliwicki (Śląskie voivodeship) subregions. A similar situation, resulting from

different reasons (probably due to suburbanization processes) is seen in the following cities: Szczecin, Poznań, Tricity, Wrocław and Cracow.

The conducted regression analysis also allowed observation of the substantial relations between the working-age population aging rate and the oldest-old-age population rate, as well as the intensity of the population aging processes. Slightly higher impact is presented by the working-age population aging rate. The interval of the indicator values was divided into four subintervals representing very high (the indicator takes values of [38.60, 41.10]), high [37.75, 38.60), medium [36.61, 37.75) and low [34.31, 36.61) percentage share of the immobile working-age population in the total working-age population (see Figure 5a).

A relatively high or very high level of this indicator is true for the Zachodniopomorskie, Opolskie and Śląskie voivodeships, and some subregions from Dolnośląskie and Łódzkie voivodeships. A relatively young working-age population lives in the biggest Polish cities (such as Poznań, Warsaw, Wrocław, Cracow and Lublin), as well as in the subregions of Podkarpackie voivodeship.

Figure 5b presents the classes of subregions in relation to the values of the oldest-old-age population rate which can be defined as very high (the indicator takes values of [26,97; 31,30]), high [26,33; 26,97), medium [25,04; 26,33) and low [19,51; 25,04) percentage share of the oldest-old population. Except for a few subregions, relatively high or very high values of the oldest-old-age population rate were recorded in eastern voivodeships, such as: Warmińsko-Mazurskie, Podlaskie, Mazowieckie, Lubelskie, Świętokrzyskie, Małopolskie and Podkarpackie voivodeships, while a low or medium percentage share of the oldest-old seniors relates to the remaining part of the state, in particular western Poland.



a) the working-age population aging rate b) the oldest-old-age population rate

Figure 5. Spatial diversification of the working-age population aging rate and the oldest-old-age population rate in 2012

Source: Own compilation based on data provided by Local Data Bank of Central Statistical Office of Poland.

Statistically, there are no significant relations between the working-age population aging rate and the oldest-old-age population rate; the Pearson's correlation coefficient took the value near to zero (-0.142). However, there are subregions in which both indicators take relatively high values: subregions of Świętokrzyskie voivodeship, Jeleniogórski and Wałbrzyski subregions (Dolnośląskie voivodeship), Sieradzki and the city of Łódź (Łódzkie voivodeship), Częstochowski (Śląskie voivodeship), Elbląski (Warmińsko-Mazurskie voivodeship), Łomżyński (Podlaskie voivodeship) and Chełmsko-Zamojski subregions (Lubelskie voivodeship).

The process of population aging in many Polish subregions is also deepened by migration outflow of young people (negative estimate). The higher the negative migration rate coefficient, the stronger the processes of population aging in a region. The values of quartiles served in defining four classes of subregions (see Figure 6):

- high outflow (the indicator took values of [-55,30; -29,40)), resulting from a much higher migration outflow than inflow of young people,
- medium outflow (the indicator took values of [-29,40; -20,62)),
- low outflow or slight inflow; the coefficient took relatively low absolute values within the interval [-20,62; 4,21),
- high inflow (the indicator took values of [4,21; 100,30]), resulting from a much higher migration inflow than outflow of young people.

High positive balance of migration flows occurs in subregions related to big agglomerations and their neighbouring subregions (the city of Szczecin and Szczeciński subregion, the city of Wrocław and Wrocławski subregion, the city of Warsaw and Warszawski Wschodni and Warszawski Zachodni subregions, the city of Cracow and Krakowski subregion). A similar situation is observed in subregions which contain a bigger city inside, such as Bielski (Śląskie voivodeship), Rzeszowski (Podkarpackie voivodeship), Białostocki (Podlaskie voivodeship), Bydgosko-Toruński (Kujawsko-pomorskie voivodeship), and also Szczeciński subregions.

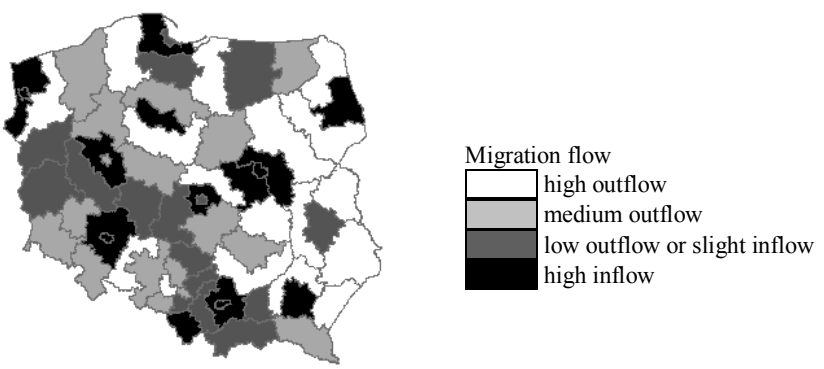


Figure 6. Spatial diversification of the net migration rate values referring to population aged 20-59 in 2012

Source: Own compilation based on data provided by Local Data Bank of Central Statistical Office of Poland.

This indicates a progressing processes of concentrating population in the economically well developed areas and their surroundings, and also depopulation of economically poor regions. An alarmingly high outflow of young people is noticeable for the voivodeships of eastern Poland, such as Podlaskie, Lubelskie and Podkarpackie voivodeships, as well as selected subregions of Mazowieckie and Świętokrzyskie voivodeships. A similar situation is also seen in subregions (e.g. Stargardzki subregion in Zachodniopomorskie voivodeship, Słupski subregion in Pomorskie voivodeship, etc.) which are located close to bigger cities and their nearest neighbours.

5. Conclusions

The population aging processes result in long-term consequences in the socio-economic sphere relating to public finance, regional labour markets, increasing demand on selected goods and services, organizational problems, etc. Therefore, they affect the socio-economic aspects of the functioning of territorial units and formation of their policies of regional development. The results of the conducted empirical research allow drawing the following conclusions which can provide significant support in these cases.

Population aging is usually a consequence of a few simultaneously occurring phenomena. The most significant factor in progressing the population aging processes is extremely low fertility which is also supported by extension of life expectancy in the Polish subregions. A significant role is also played by aging processes of the working-age population and increasing number of oldest-old people within the population of senior citizens. The majority of subregions also struggle with a significant migration outflow of younger people, which deepens the population aging processes.

High territorial diversification of the population aging processes, according to the senior citizens proportion and its growth rate, and also demographic conditions of these processes, occur in Poland. The most advanced population aging processes are observed in south-western Polish subregions, in which the most significant problems are very low fertility, relatively old working-age population and also a significant migration outflow of younger people. On the other hand, the youngest and slowest aging population lives in south-eastern and central Poland. It is characterized by higher fertility but shorter life expectancy, and also a younger working-age population.

REFERENCES

BAL-DOMAŃSKA, B., WILK, J., (2011). Gospodarcze aspekty zrównoważonego rozwoju województw – wielowymiarowa analiza porównawcza [Economic aspects of sustainable development of provinces – multivariate comparative analysis], Statistical Review, no. 3–4, pp. 300–322.

DRAGAN, A., (2011). Starzenie się społeczeństwa i jego skutki [Population aging and its effects], Warsaw: The Polish Senate. The Office of Analyses and Documentation, Thematic Studies, No. 601.

FRĄTCZAK, E., (1984). Proces starzenia się ludności Polski a proces urbanizacji [The population aging process of Poland's population and urbanization process], Warsaw: Main School of Planning and Statistics.

GIANNAKOURIS, K., (2008). Ageing characterises the demographic perspectives of the European societies, Eurostat Statistics in Focus, No. 72.

GOLINOWSKA, S., (2008). Społeczno-ekonomiczne konsekwencje starzenia się populacji [Socio-economic consequences of population aging], [in: J. Kleer (Ed.), Konsekwencje ekonomiczne i społeczne starzenia się społeczeństwa], Warsaw: Polish Academy of Science Publishing House.

GRABIŃSKI, T., WYDYMUS, S., ZELIAŚ, A., (1989). Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych [Numerical taxonomy methods in socio-economic phenomena modelling], Warsaw: PWN Publishers.

HELLWIG, Z., SIEDLECKA, U., SIEDLECKI, J., (1995). Taksonometryczne modele zmian struktury gospodarczej Polski [Taxonometric models of changes in Poland's economic structure], Warsaw: Institute of Development and Strategic Studies.

HRYNKIEWICZ, Z. (Ed.), (2012). O sytuacji ludzi starszych [On seniors' situation], Warsaw: Governmental Population Council.

HWANG, C. L., YOON, K., (1981). Multiple Attribute Decision Making Methods and Applications, Berlin Heidelberg: Springer.

JUREK, Ł., (2012). Ekonomia starzejącego się społeczeństwa [The economics of population aging], Warsaw: Difin Publishers.

KUREK, S., (2008). Typologia starzenia się ludności Polski w ujęciu przestrzennym, Cracow: Pedagogical Academy of Science Publishing House.

KURKIEWICZ, J. (Ed.), (2006). Ludzie starsi w rodzinie i w społeczeństwie [Seniors within family and society], Cracow: Cracow University of Economics Publishing House.

KURKIEWICZ, J. (Ed.), (2010). Procesy demograficzne i metody ich analizy [Demographic processes and methods for their analysis], Cracow: Cracow University of Economics Publishing House.

KURKIEWICZ, J. (Ed.), (2012). Demograficzne uwarunkowania i wybrane społeczno-ekonomiczne konsekwencje starzenia się ludności w krajach europejskich [Demographic conditions and selected socio-economic consequences of population aging in European countries], Cracow: Cracow University of Economics Publishing House.

LINDH, T., MALMBERG, B., (2009). European Union Economic Growth and the Age Structure of the Population, Economic Change and Restructuring, Vol. 42, No. 3, pp. 159–187.

MAGNUS, G., (2008). The age of aging: how demographics are changing the global economy and our world, John Wiley & Sons.

MATUSIK, S., PIETRZAK, M., WILK, J., (2012). Ekonomiczne-społeczne uwarunkowania migracji wewnętrznych w Polsce w świetle metody drzew klasyfikacyjnych [Economic-social conditions of internal migrations in Poland in the light of classification tree method], Demographic Studies, No. 2(162), pp. 3–28.

MŁODAK, A., (2006). Analiza taksonomiczna w statystyce regionalnej [Taxonomic analysis in regional statistics], Warsaw: Difin Publishers.

MUENZ, R., (2007). Aging and Demographic Change in European Societies: Main Trends and Alternative Policy Options, SP Discussion Paper, No. 0703.

MÜLLER-FRĄCZEK, I., PIETRZAK, M. B., (2009a). Analiza porównawcza rozwoju ekonomicznego województwa kujawsko-pomorskiego w latach 2003 i 2007 z wykorzystaniem narzędzi statystyki przestrzennej[Comparative analysis of the economic development of the Kujawsko-Pomorskie province in the years 2003 and 2007 with the use of tools of spatial statistics], Acta Universitatis Nicolai Copernici. Ekonomia, z. 39, pp. 135–145.

MÜLLER-FRĄCZEK, I., PIETRZAK, M. B., (2009b). Potencjał ekonomiczny jako miara społeczno-ekonomicznego rozwoju regionu na przykładzie województwa kujawsko-pomorskiego [The economic potential as a measure of the economic development of the Kujawsko-Pomorskie province], Acta Universitatis Nicolai Copernici. Ekonomia, z. 40, pp. 87–100.

NOWAK, E., (1990). Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych [Taxonomic methods in the classification of socio-economic objects], Warsaw: PWE Publishers.

PIETRZAK, M. B., (2012). Wykorzystanie przestrzennego modelu regresji przełącznikowej w analizie regionalnej konwergencji w Polsce , Ekonomia i Prawo, T XI, nr 4/2012, pp. 167–186.

PIETRZAK, M. B., DRZEWOSZEWSKA, N., WILK, J., (2012). The analysis of interregional migrations in Poland in the period of 2004-2010 using panel gravity model, Dynamic Econometric Models, vol. 12, pp. 111–122.

PIETRZAK, M. B., ŻUREK, M., MATUSIK, S., WILK, J., (2012). Application of Structural Equation Modeling for analysing internal migration phenomena in Poland, Statistical Review, nr 4, R. LIX, pp. 487–503.

PIETRZAK, M. B., WILK, J., SIEKANIEC, M., (2013). The impact of metropolitan areas on internal migrations in Poland. The case of southern regions, [in: M. Papież, S. Śmiech (ed.), Proceedings of the 7th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena], Cracow: Foundation of the Cracow University of Economics, pp. 124–132.

PIETRZAK, M. B., WILK, J., MATUSIK, S., (2013a). Gravity model as a tool for internal migration analysis in Poland in 2004-2010, [in: J. Pociecha (ed.), Quantitative Methods for Modelling and Forecasting Economic Processes], Cracow: Foundation of the Cracow University of Economics, pp. 108–120.

PIETRZAK, M. B., WILK, J., MATUSIK, S., (2013b). Analiza migracji wewnętrznych w Polsce z wykorzystaniem modelu grawitacji, Acta Universitatis Lodziensis, Folia Oeconomia, 293, pp. 27–36.

PIETRZAK, M. B., WILK, J., (2013). Obszary metropolitalne Polski południowej a ruch migracyjny ludności [Metropolitan areas of southern Poland and population migration movement], Ekonomia i Prawo, Tom XII, nr 3, pp. 498–506.

POCIECHA, J., (Ed.), (2003). Ekonomiczne konsekwencje osiągania wieku emerytalnego przez generancje powojennego wyżu demograficznego [Economic consequences of reaching retirement age by the generations of post-war population explosion], Cracow: University of Economics Publishing House.

PRSKAWETZ, A., LINDH, T., (2011). The relationship between demographic change and economic growth in the EU, Wien: Austrian Academy of Sciences Publishing House.

ROSSET, E., (1959). Proces starzenia się ludności [Population aging process], Warsaw: PWE Publishers.

STRZELECKI, Z., (Ed.), (2009). Sytuacja demograficzna Polski. Raport 2008-2009 [Demographic situation in Poland. The report for the period 2008-2009], Warsaw: Governmental Population Council.

UHLENBERG, P., (Ed.), (2009). International Handbook of Population Aging, International Handbooks of Population 1, Springer.

WILK, J., BARTŁOMOWICZ, T., (2012). Wielowymiarowa analiza zmian demograficznych w Polsce w świetle koncepcji zrównoważonego rozwoju [Multivariate analysis of demographic changes in Poland in the light of sustainable development concept], Demographic Studies, No. 2(162), pp. 55–86.

WILK, J., PIETRZAK, M. B., MATUSIK, S., (2013). Sytuacja społeczno-gospodarcza jako determinanta migracji wewnętrznych w Polsce [Socio-economic situation as the determinant of internal migration in Poland], [in: K. Jajuga, M. Walesiak (ed.), Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania], Wrocław: PN UE we Wrocławiu No. 278, pp. 330–342.

WILK, J., PIETRZAK, M. B., (2013). Analiza migracji wewnętrznych w kontekście aspektów społeczno-gospodarczych – podejście dwuetapowe [The analysis of internal migrations in the context of socio-economic aspects – two-step approach], [in: J. Dziechciarz (ed.), Ekonometria], 2(40), Wrocław: Wyd. UE we Wrocławiu, pp. 62–73.

STATISTICS IN TRANSITION new series, Winter2014
Vol. 15, No. 1, pp. 171–173

BOOK REVIEW

Jerzy Witold Wiśniewski: *Correlation and regression of economic qualitative features,* Lambert Academic Publishing, 2013, ISBN 9783659512780, 72 pp., EUR 16, 26.

by Jan Kordos[1]

The purpose of this book is to present research methods suitable for qualitatively characterized phenomena, along with illustrations of their applications. The book consists of an introduction, four chapters, a summary and a bibliography.

The first chapter, *The specificity of the economic measurement*, starts with the concepts of measurement, metrology, economic metrology and the measurement scales.

The second chapter, *The features and quality processes in economics*, is devoted to the specificity of the qualitative characteristics, measurement by weak scales, arithmetic operations on numbers in a variety of scales, notes on the measurement of economic values, and transformation of the measurement results.

The third chapter, *Correlation of features and quality processes*, deals with the nature of correlation coefficients, correlation of the dummy variables, and the coefficient of association attributes.

The last, fourth chapter, *The regression model in the analysis of the attributes and quality processes*, analyses the nature of the regression model, a linear function of the probability, and models with transformed limited variables.

In brief, the following topics are covered:

1) simple methods of measurement of the quality features based on the weak Stevens scales;
2) common errors in the application of the statistical methods for the analysis of the results of the measurement in ordinal scale which have not been presented in any book before;
3) the risk of applying the well-known Spearman's correlation coefficient;
4) a new coefficient of association of quality features developed by the book's author which is equivalent to Pearson's correlation coefficient (this coefficient of association presented in the book can be tested, for instance, by a simple t-Student test);

[1] Warsaw School of Management, and Central Statistical Office of Poland.

5) the possibilities of applying the association coefficient which can be used to management decisions in an enterprise;
6) the possibilities of applying the linear probability function in an enterprise with advantages of its application;
7) the econometric models of limited dependent variables with general formula for logit transformation and possible applications in managerial decision-making at an enterprise.

The author considers shortly metrology, economic metrology, and the features and quality processes in economics. Metrology is the science of measurement and includes all theoretical and practical aspects of the measurement. I would like to add for considerations the philosophy of statistical thinking.[2]

Statistical thinking is the philosophy of learning and action based on the following fundamental principles:

a) all work occurs in a system of interconnected processes - a process being a chain of activities that turns inputs into outputs;

b) variation, which gives rise to uncertainty, exists in all processes; and

c) understanding and reducing variation are keys to success.

All three principles work together to create the power of statistical thinking.

Since the 1980s statistical thinking has been discussed in the literature, applied in the workplace, and formally taught at some university.[3] While there has been some resistance from those who prefer a more traditional, mathematically oriented approach, the profession has gradually accepted the need for readers to think deeply before calculating.

The definition highlights several key components: process thinking, understanding and managing uncertainty, and using data whenever possible to guide actions and improve decision-making. Statistical thinking is a philosophy of an overall approach to improvement and, therefore, more broadly applicable than statistical methods. It is a way of thinking, behaving, working, taking action and interacting with others. In addition, the focus of the statistical thinking process provides the context and the relevancy for broader and more effective use of statistical methods[4].

---

[2] See: http://www.statoo.com/en/statistical.thinking; eKPIsolutions; email: eu@ekpisolutions.com; http://www.asq.org.

[3] Hoerl, R. and Snee, R. D., (2012). *Statistical Thinking: Improving Business Performance*, New York: Wiley & Sons; Wild C. J. and Pfannkuch, M. (1999), *Statistical Thinking in Empirical Enquiry*, International Statistical Review, 67, 223–265.

[4] Abert, J. ,Ruud, H., (2007), *Statistical thinking in Sports*, Publisher Chapman & Hall/CRC; Britz G. C., Emerling, D. W., Hare, L. B., Hoerl, R. W., Janis, S. J., (2000). *Improving Performance Through Statistical Thinking,* ASQ Quality Press; Snee, R. D. (1990). *Statistical Thinking and its contribution to Total Quality,* Am. Statist. 44, 116-121; Yu-Kang Tu, (2012). *Statistical Thinking in Epidemiology*, Chapman and Hall/CRC.

Statistical thinking uses the scientific method to develop subject matter knowledge and to gather data to evaluate and revise hypotheses. First, statistical thinking recognizes that results are produced by a process and that the process must be understood and developed to improve the results. The second difference is the emphasis on variation on statistical thinking. The scientific method can be applied without any awareness of the concept of variation, which may lead to misinterpretation of the results. The key similarities are that they both are sequential approaches that integrate data and subject matter knowledge.