STUDIA EKONOMICZNE

AKADEMIA EKONOMICZNA
im. Karola Adamieckiego
w Katowicach

53
ZESZYTY NAUKOWE

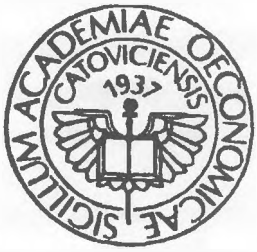
METODY WNIOSKOWANIA STATYSTYCZNEGO W BADANIACH EKONOMICZNYCH (METHODS OF STATISTICAL INFERENCE IN ECONOMIC SURVEYS)

ZESZYTY NAUKOWE

AKADEMII EKONOMICZNEJ IM. KAROLA ADAMIECKIEGO

„Studia Ekonomiczne"

METODY WNIOSKOWANIA STATYSTYCZNEGO W BADANIACH EKONOMICZNYCH (METHODS OF STATISTICAL INFERENCE IN ECONOMIC SURVEYS)



Katowice 2009

Editions

Józef Kolonko
Janusz L. Wywiał

Reviewers

Czesław Bracha
Witold Miszczak
Krzysztof Piasecki
Aleksandras Plikusas
Andrzej Sokołowski
Józef Stawicki
Jacek Wesołowski

Edition

Patrycja Keller

ISBN 978-83-7246-580-1

Publisher of The Karol Adamiecki University of Economics in Katowice

ul. 1 Maja 50, 40-287 Katowice, tel. +48 032 25 77 635, fax +48 032 25 77 643
www.ae.katowice.pl, e-mail: wydawnictwo@ae.katowice.pl

CONTENTS

INTRODUCTION

The papers are prepared by the employees of the Department of Statistics at the Faculty of Management of the University of Economics in Katowice. In general the following topics are considered: survey sampling, time series analysis, economic statistics, classification and clustering, demography, time series analysis, robust statistical inference.

In this volume there are ten papers presented. Four of them are connected with survey sampling. The first paper is about non response problem where some composite estimators using regression and classification methods are considered by Gamrot. In the second paper Wywiał considers a problem of application of the well-known Poisson sampling scheme to the modeling Internet survey. The response probabilities are explained by logit model. The approximate mean square error of a total estimator is derived. In the third paper by Wywiał a sampling proportional to a positive function of sample quantiles of an auxiliary variable is defined. Its sampling scheme and inclusion probabilities are evaluated. In the paper by Żądło some problem of small area estimation is studied. In this paper the problem of prediction of totals for domains specified by random attributes is presented.

Next papers are close to analysis of robust statistical methods. The paper by Trzpiot is a review of estimation methods of quantile regression parameters. In the next paper Trzpiot and Majewska present sensitivity analysis of selected robust estimators of volatility and the classification of generated investment portfolios with respect to chosen robust estimators. The authors try to convince that applying robust estimation in portfolio analysis ensures better method for effective investment decision-making than classical portfolio analysis. The purpose of the paper by Trzpiot and Krężołek is to present some quantiles ratio risk measures of financial assets. These measures are based on the VaR approach. The assumption of stable distributed log-returns is used. It allows to estimate investment risk more accurate.

Next two papers are connected with time series analysis. Paper by Ganczarek-Gamrot is about application of vector autoregressive model to analysis of Polish Electric Energy Market. Then a nonparametric test for linearity of trend is proposed by Kończak.

Some special taxonomic methods are considered by Rozmus. The main aim of her article is to compare the right class structure recognizing ability of classical and ensemble clustering methods. The performances of the new and existing taxonomy algorithms were compared on the basis of simulated and real data sets.

Janusz L. Wywiał

Wojciech Gamrot

ON COMPOSITE ESTIMATION UTILIZING REGRESSION AND CLASSIFICATION METHODS

Introduction

Let U denote finite population of size N. Consider some characteristic Y taking fixed values $y_1, \ldots, y_N$. The objective of the survey is to estimate its population mean:

$$\overline{Y} = \frac{1}{N} \sum_{i \in U} y_i \tag{1}$$

The sampling procedure involves two phases. In the first phase a simple random sample s of size n is drawn without replacement from U. Assume stochastic nonresponse: each i-th unit responds with some unknown probability ρ_i. Due to nonresponse the sample s splits into two subsets s_1 and s_2 of sizes n_1 and n_2 such that units from s_1 respond in the survey whereas units from s_2 do not. The second phase of the survey is then carried out to acquire some knowledge about non--responding population units. In this second phase a simple subsample s' of size $n' = cn_2$ (where $0 < c < 1$) is drawn without replacement from the nonrespondent subset s_2. It is assumed that all subsampled units respond in the second phase so the response probabilities ρ_i correspond only to the first phase of the survey.

1. Straightforward estimators

Consider the statistic:

$$\overline{y}(\alpha) = \alpha \overline{y}_{s_1} + (1-\alpha)\overline{y}_{s'} \tag{2}$$

where

$$\overline{y}_{s_1} = \frac{1}{n_1} \sum_{i \in s_1} y_i \tag{3}$$

$$\overline{y}_{s'} = \frac{1}{n'} \sum_{i \in s'} y_i \tag{4}$$

When $\alpha = n_1/n$ it takes the well-known form:

$$\overline{y}_{STD} = \frac{n_1}{n}\overline{y}_{s_1} + \frac{n_2}{n}\overline{y}_{s'} \tag{5}$$

and as indicated by Särndal et al. (1992) it is unbiased for the population mean under any possible set of individual response probabilities $\rho_1, \ldots, \rho_N$. In the following discussion it will be called the *standard estimator* and denoted by the symbol STD.

Wywiał (2001) suggests another way to construct the weight α in (2) using the classification algorithm. It relies on the assumption that the population consists of two strata U_1 and U_2 of sizes N_1 and N_2, such that $\rho_i = 1$ for $i \in U_1$ and $\rho_i = 0$ for $i \in U_2$. It is then assumed that the vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{ik}]'$ containing the values of k auxiliary variables $X_1, \ldots, X_k$ is observed for each i-th population unit and that it follows multivariate Gaussian distribution with different means in both strata. This is a special case of the well-known deterministic nonresponse model. However, estimators based on this model may also be used with good results for stochastic nonresponse. To identify the strata we employ the well-known quadratic Bayesian discrimination function minimizing the unit misclassification probability (Duda and Hart 2001). This leads to the division of population into two subsets U_1' and U_2' of sizes N_1' and N_2'. These subsets may differ from the original classes U_1 and U_2, but the ratio N_1'/N may be treated as an estimator of the true respondent fraction N_1/N. It is thus reasonable to set $\alpha = N_1'/N$ in the formula (2) and consider the statistic:

$$\overline{y}_{BAY} = \frac{N_1'}{N}\overline{y}_{s_1} + \frac{N_2'}{N}\overline{y}_{s'} \tag{6}$$

In the following discussion it will be denoted by the symbol BAY.

The classification estimator discussed above relies on dependencies between auxiliary characteristics and response probabilities. An alternative approach to mean value estimation under nonresponse postulates the use of direct relationship between auxiliary variables and the variable under study. This leads to the well known regression estimator:

$$\overline{y}_{REG} = \overline{y}_{s_1} + \widehat{\mathbf{b}}\,(\overline{\mathbf{x}}_s - \overline{\mathbf{x}}_{s_1}) \tag{7}$$

where

$$\widehat{\mathbf{b}} = \left(\sum_{i\in s_1}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i\in s_1}\mathbf{x}_i'y_i\right) \tag{8}$$

$$\overline{\mathbf{x}}_s = \frac{1}{n}\sum_{i\in s}\mathbf{x}_i \tag{9}$$

$$\overline{\mathbf{x}}_{s_1} = \frac{1}{n_1}\sum_{i\in s_1}\mathbf{x}_i \tag{10}$$

In further study the statistic (7) will be denoted by the symbol REG.

2. Composite estimator

Estimators REG and BAY both rely on the assumption that underlying models properly describe reality. When there is no possibility to decide in advance which model fits the data better, the sampler may try to assess the goodness of fit on the basis of sample data. In the case of the linear model it may be done using the determination coefficient:

$$R_{REG} = \frac{\sum_{i\in s_1}\left(\widehat{\mathbf{b}}\mathbf{x}_i - \frac{1}{n_1}\sum_{j\in s_1}\widehat{\mathbf{b}}\mathbf{x}_j\right)^2}{\sum_{i\in s_1}\left(y_i - \frac{1}{n_1}\sum_{j\in s_1}y_j\right)^2} \tag{11}$$

and in the case of discrimination function, its ability to correctly identify respondents and nonrespondents may be assessed by the expression:

$$R_{BAY} = 1 - R_M \tag{12}$$

where R_M is the initial sample misclassification ratio. These expressions enable the construction of a composite estimator in the form:

$$\overline{y}_{COM} = W_{REG} \cdot \overline{y}_{REG} + W_{BAY} \cdot \overline{y}_{BAY} \tag{13}$$

where $W_{REG} = R_{REG}/(R_{REG} + R_{BAY})$ and $W_{BAY} = 1 - W_{REG}$. It is worth emphasizing that weights W_{REG} and W_{BAY} depend on the sample. The estimator based on better fitting model will dominate the combination (13). Consequently, the composite estimator should behave in a similar way to the regression estimator when the linear model fits relatively well and more like the classification estimator when the deterministic model fits relatively well.

3. Simulations

A simulation study involving four experiments has been conducted to compare the accuracy of estimators: STD, REG, BAY, and COM. Each experiment was carried out by repeatedly drawing sample-subsample pairs from the predefined pseudo-random population and simulating response/nonresponse decisions of individual population units. The mean square error (MSE) of each estimator was evaluated by examining its empirical distribution. In each experiment four pseudo-random variables: Y, X_1, X_2, and X_3 of joint multivariate Gaussian distribution were generated. The mean value and standard deviation vectors were respectively set to $\mu = [0, 0, 0, 0]$ and $\sigma = [1, 1, 1, 1]$. The correlation matrices were different in each experiment and respectively equal to:

$$R_1 = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0.5 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix} \qquad R_2 = \begin{bmatrix} 1 & 0.7 & 0 & 0 \\ 0.7 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14}$$

$$R_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad R_4 = \begin{bmatrix} 1 & 0.7 & 0.5 & 0.5 \\ 0.7 & 1 & 0 & 0 \\ 0.5 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

The variable Y acted as a variable under study. The variable X_1 was used as an auxiliary variable for the regression estimator. The variables X_2 and X_3 determined the individual response probabilities according to the logistic model similar to the one considered by Ekholm and Laaksonen(1991):

$$p_i = \frac{1}{1 + exp(\beta_0 X_2 + \beta_1 X_3)} \tag{16}$$

with constants $\beta_0 = \beta_1 = 1$ chosen arbitrarily so that response probability decreases when X_2 or X_3 grows. It was also assumed that units respond independently. The functional form of the model (16) was assumed unknown. Instead, the variables X_2 and X_3 were used to compute the classification estimator. For each experiment a total of 50000 samples were drawn from among the population of 5000 units for $c = 0.3$ and any value of $n = 40, 80, \ldots, 200$. The assumptions of all four experiments are in some sense extreme. The matrix R_1 corresponds to the situation where only the deterministic nonresponse model fits well. The matrix R_2 corresponds to the situation where only the linear model fits well. The matrix R_3 corresponds to the situation where none of these models fits well and the matrix R_4 corresponds to the situation where both models fit well. The mean square errors observed in all four experiments are shown in tables 1 and 2. To facilitate the comparison, we also define the relative efficiency of any estimator T (where $T = \overline{y}_{REG}, \overline{y}_{BAY}, \overline{y}_{COM}$) with respect to the standard estimator as $eff(T) = MSE(T)/MSE(\overline{y}_{STD})$. They are shown on figures 1-4.

The results obtained in first two experiments are promising. If only one model fits well then the MSE of the composite estimator is the lowest one (for R_2) or very close to the lowest one (for R_1). This result may be explained by low or even negative correlation between REG and BAY estimates and by the fact that this estimator incorporates the information from two different sources. In both cases the

relative efficiency of the composite estimator is lower than one (it has lower MSE than the standard estimator) and remains relatively stable when the initial sample size n changes. Overall, these results suggest that the composite estimator is quite robust with respect to the model misspecification. For the third experiment and the matrix R_3 the composite estimator is the most accurate one for small samples ($n < 60$), but then its MSE steadily grows in relation to all other estimators, and for larger samples the REG estimator proves to be much more accurate, without even using subsample data. The composite estimator is however still more accurate than the standard estimator.

Table 1

Mean square error of estimators as a function of initial sample size n for correlation matrices R_1 and R_2

R_1					R_2				
n	STD	REG	BAY	COM	n	STD	REG	BAY	COM
40	0,0505	0,1709	0,0454	0,0434	40	0,0560	0,0393	0,0569	0,0376
60	0,0335	0,1554	0,0293	0,0285	60	0,0370	0,0257	0,0375	0,0244
80	0,0251	0,1477	0,0219	0,0215	80	0,0282	0,0192	0,0286	0,0184
100	0,0201	0,1430	0,0174	0,0171	100	0,0220	0,0152	0,0222	0,0143
120	0,0167	0,1409	0,0143	0,0142	120	0,0182	0,0126	0,0184	0,0118
140	0,0142	0,1391	0,0122	0,0121	140	0,0155	0,0107	0,0156	0,0100
160	0,0123	0,1370	0,0106	0,0105	160	0,0138	0,0094	0,0140	0,0089
180	0,0110	0,1356	0,0094	0,0094	180	0,0122	0,0083	0,0123	0,0078
200	0,0099	0,1348	0,0086	0,0085	200	0,0108	0,0075	0,0109	0,0069

Table 2

Mean square error of estimators as a function of initial sample size n for correlation matrices R_3 and R_4

R_3					R_4				
n	STD	REG	BAY	COM	n	STD	REG	BAY	COM
40	0,0542	0,0527	0,0553	0,0513	40	0,0520	0,1656	0,0454	0,0622
60	0,0362	0,0344	0,0369	0,0348	60	0,0344	0,1552	0,0292	0,0549
80	0,0274	0,0257	0,0278	0,0266	80	0,0258	0,1494	0,0217	0,0513
100	0,0216	0,0203	0,0220	0,0212	100	0,0208	0,1460	0,0173	0,0493
120	0,0179	0,0167	0,0182	0,0176	120	0,0173	0,1443	0,0144	0,0484
140	0,0155	0,0144	0,0156	0,0152	140	0,0147	0,1422	0,0122	0,0470
160	0,0134	0,0125	0,0136	0,0132	160	0,0129	0,1416	0,0107	0,0468
180	0,0120	0,0111	0,0121	0,0118	180	0,0114	0,1404	0,0095	0,0461
200	0,0108	0,0100	0,0109	0,0107	200	0,0102	0,1402	0,0084	0,0462

The results of first three experiments suggest that in the fourth experiment where the correlation matrix R_4 represents both models fitting well should prove the composite estimator to be very attractive in terms of MSE. However, simulation results paint a different picture. Only the BAY estimator competes successfully with the standard one in this case. The MSE of the REG estimator grows quickly with increasing n and this tendency is reflected by the MSE of the composite estimator. This outcome may be explained by the fact that the REG estimator does not utilize the subsample data and that the determination coefficient is only loosely related to the nonresponse behavior of this estimator. Also, arbitrarily constructed weights in the formula (13) fail to reflect accurately relative efficiencies of REG and BAY estimators.

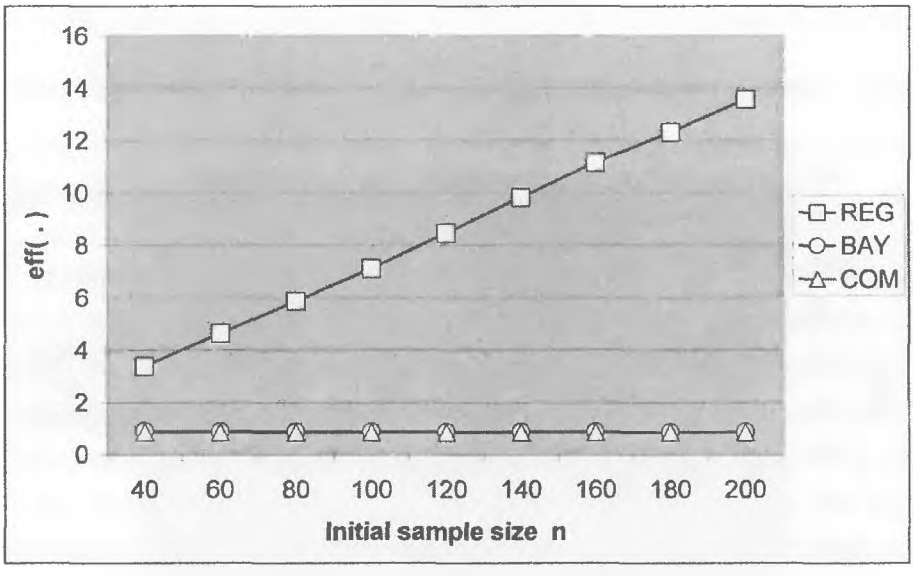


Figure 1. Relative efficiencies for R_1

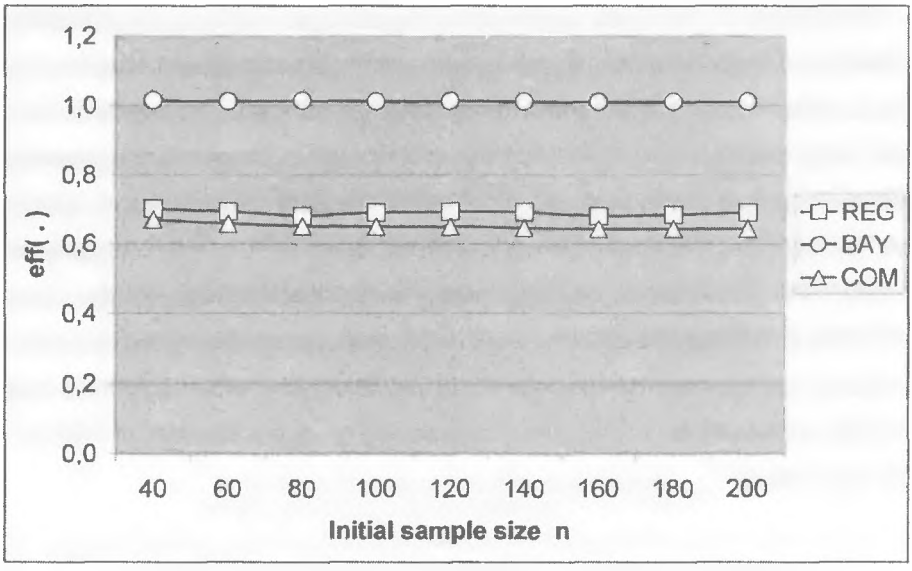


Figure 2. Relative efficiencies for R_2

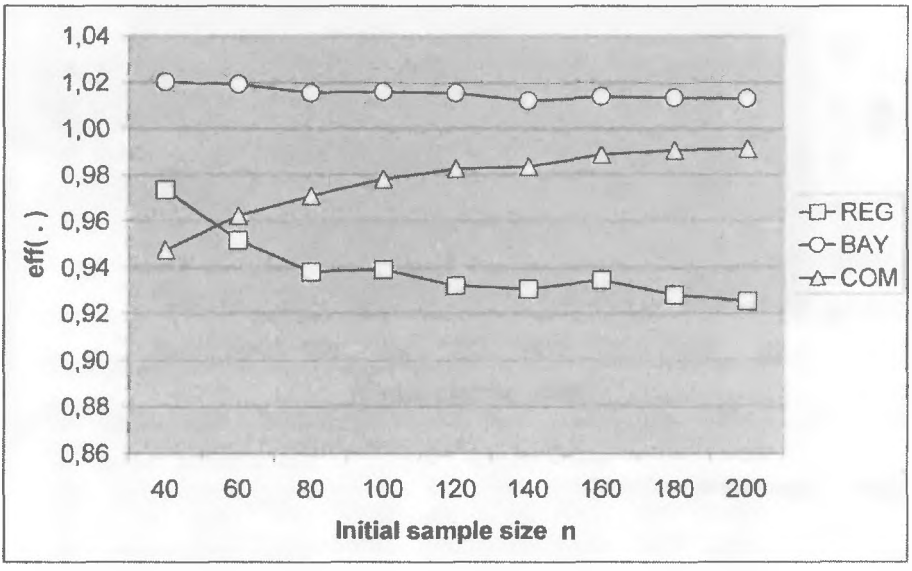


Figure 3. Relative efficiencies for R_3

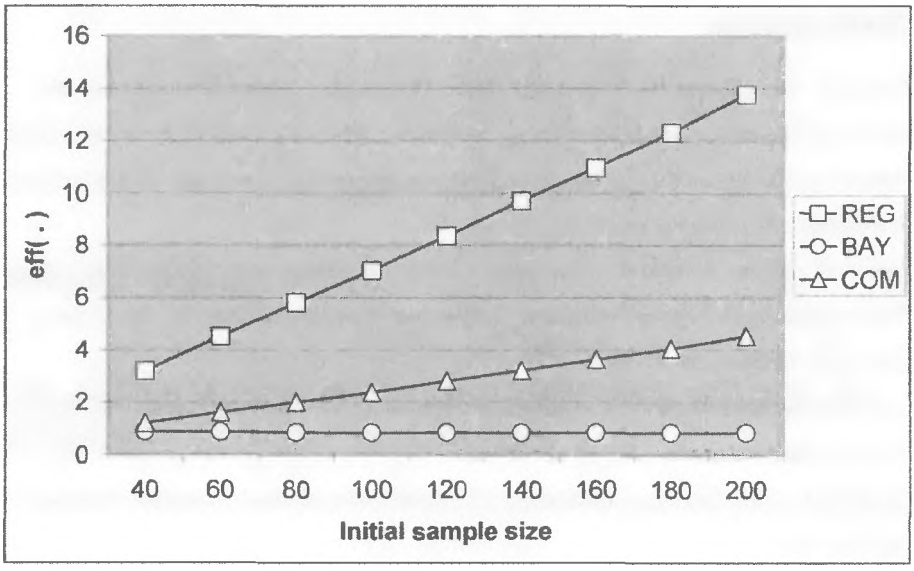


Figure 4. Relative efficiencies for R_4

Conclusions

Presented results of simulation experiments indicate, that the general idea of constructing composite estimators by combining two or more simple estimates is promising. The composite estimator incorporating the estimates based on different data and/or different approaches may be more robust to model misspecification than any of straightforward single-model-based estimators. In some situations it may provide a significant gain in accuracy. However, simulation results also show that in some cases the construction of composite estimators in a heuristic way based only on pure intuition may lead to dramatically inaccurate estimates. This stresses the need to develop more systematic approach to the construction of these estimators based on analytical evaluation of their stochastic properties.

References

Duda R.O., Hart P.E. and Stork D.G. (2001): Pattern Classification. Wiley & Sons, Inc., New York.

Gamrot W. (2003a): On Application of Some Discrimination Methods to Mean Value Estimation in the Presence of Nonresponse. In: J. Wywial (ed.): Metoda reprezentacyjna w Badaniach Ekonomiczno-Społecznych. AE, Katowice, pp. 37-50.

Gamrot W. (2003b): A Monte Carlo Comparison of Some Two-Phase Sampling. Strategies Utilizing Discrimination Methods in the Presence of Nonresponse. "Zeszyty Naukowe", No. 29, University of Economics, Katowice, pp. 41-54.

Hansen M.H., Hurwitz W.N. (1949): The Problem of Nonresponse in Sample Surveys. "Journal of the American Statistical Society", No. 41, pp. 517-529.

Särndal C.E., Swensson B. and Wretman J. (1997): Model Assisted Survey Sampling. Springer Verlag, New York.

Wywiał J. (2001): On Estimation of Population Mean in the Case When Nonrespondents Are Present. "Prace Naukowe", 8, 906, AE, Wrocław, pp. 13-21.

Abstract

Several estimation procedures have been developed to compensate for the deterioration in properties of parameter estimates resulting from sample data incompleteness. Most of them make use of available auxiliary data following one of two general approaches. The first approach relies on dependencies between auxiliary variables and the variable under study. This usually leads to the construction of various ratio and regression estimators. The second approach explores dependencies between auxiliary variables and response behavior of population units. This provides motivation to a broad range of methods such as weighting adjustments and classification estimators. In this paper a composite estimator of the population mean incorporating both approaches is considered. It is constructed as a combination of the well-known regression estimator and a classification estimator utilizing Bayesian quadratic discrimination function. The weights of the combination reflect the regression model's goodness of fit and the classification quality. Hence, greater weight is assigned to the estimator for which available observations of auxiliary variables are more useful. Simulation results exposing its properties are presented in the paper.

Janusz L. Wywiał

ON APPLICATION OF NON RESPONSE MODEL IN INTERNET SURVEY SAMPLING

Introduction and basic definition

We assume $U = \{1, 2, \ldots N\}$ is a fixed and finite population of size N. Let us suppose that each population element (people, firms) is identified and it can be observed by means of the Internet mail. We assume that all Internet addresses of respondents (population elements) are known. A questionnaire form is sent to all respondents. It is possible that some of them do not return the quastionnaire. So, the sample s consists only of the respondents who have returned the form. So, the size of the sample can be smaller than the population size. We can assume that a respondent returns the questionnaire with some probability π_k, $k = 1, \ldots, N$, which we call response probability.

Let us consider a sample as a vector $s = [s_1\ s_2 \ldots s_N]$. If a k-th population element is (is not) in the sample, $s_k = 1$ ($s_k = 0$). The sample size $n(s)$ is not fixed, so, $0 \leq n(s) \leq N$. The support (sampling space) is denoted by $\boldsymbol{S}$. The sampling design is a probability distribution of a sample s defined on a support $\boldsymbol{S}$: $P(s) \geq 0$ for $s \in \boldsymbol{S}$. The inclusion probability of the first order is: $\pi_k = \sum_{\{s:\ k \in s\}} P(s)$. The inclusion probability of the second order is: $\pi_{k,l} = \sum_{\{s:\ k \in s,\ l \in s\}} P(s)$. The

random sample is defined by Tillé (2006) as the vector $S = [S_1\ S_2 \dots S_N]$. The set $\boldsymbol{S}$ is the sample space of the random variable S and its probability distribution is: $P(S = s) = P(s)$ for $s \in \boldsymbol{S}$.

Let us assume that elements of the random sample S are independent and $P(S_k = 1) = \pi$ and $P(S_k = 0) = 1 - \pi$ for $k = 1, \dots, N$. So, $E(S_k) = \pi$, $D^2(S_k) = \pi(1 - \pi)$, $Cov(S_k S_l) = 0$, $k = 1, \dots, N$; $l = 1, \dots, N$ and $k \neq l$. Hence, the sample s can be treated as an outcome of Bernoulli trial. The Bernoulli sampling design without replacement is as follows:

$$P(s) = \prod_{k=1}^{N} \pi^{s_k}(1 - \pi)^{1 - s_k} \qquad \text{for} \qquad s \in \boldsymbol{S} \tag{1}$$

It is obvious that, $\pi_k = \pi$ and $\pi_{k,l} = \pi^2$ for all $k = 1, \dots, N; l = 1, \dots, N$ and $k \neq l$. Moreover, the distribution of the sample size is:

$$P(n(S) = n) = \binom{N}{n} \pi^n (1 - \pi)^{1-n} \qquad \text{for} \qquad s \in \boldsymbol{S} \tag{2}$$

Of course $E(n(S)) = N\pi$ and $D^2(n(S)) = N\pi(1 - \pi)$.

Let us stress that in our case the sample s consists only of those respondents who return a questionnaire by email. The probability that a respondent returns it is equal to π and it is the same in all the population. That is why we will name the parameter π as the response probability. Of course we assume that respondents decide to return the questionnaire independently. So, it is a simple model of an Internet survey.

The more realistic model of an Internet survey can based on the following well--known Poisson sampling design without replacement. Let the elements of the random sample S be independent and $P(S_k = 1) = \pi_k$, $P(S_k = 0) = 1 - \pi_k$ for $k = 1, \dots, N$. So:

$$P(s) = \prod_{k=1}^{N} \pi_k^{s_k}(1 - \pi_k)^{1 - s_k} \qquad \text{for} \qquad s \in \boldsymbol{S} \tag{3}$$

In this case the response probabilities π_k can be different and $E(S_k) = \pi_k$, $D^2(S_k) = \pi_k(1 - \pi_k)$, $Cov(S_k S_l) = 0$, $k = 1, \dots, N$, $l = 1, \dots, N$ and $k \neq l$.

Let a non random variable y be observed in the population U. Its possible values are elements of the vector $\mathbf{y} = [y_1\ y_2 \dots y_N]$. So the value y_k is attached to the k-th element of the population. Our purpose is estimation of the total $y_U = \sum_{k=1}^{N} y_k$ or mean value $\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k$.

1. Estimation under the Bernoulli model

1.1. Basic results

Under the Bernoulli model of generation of a sample the Horvitz-Thompson (1952) estimator of the mean $\bar{y}$ is as follows, respectively.

$$\bar{y}_{HTS} = \frac{1}{N\pi}\sum_{k=1}^{N} y_k S_k = \frac{1}{N\pi}\sum_{k\in S} y_k \tag{4}$$

It is unbiased estimator of the mean. Its variance is as follows:

$$D^2\left(\bar{y}_{HTS}\right) = \frac{1-\pi}{N^2\pi}\sum_{k=1}^{N} y_k^2 \tag{5}$$

The unbiased estimator of the variance is:

$$D_S^2\left(\bar{y}_{HTS}\right) = \frac{1-\pi}{N^2\pi^2}\sum_{k=1}^{N} y_k^2 S_k \tag{6}$$

The probability π in the above expressions is not known. The unbiased Horvitz-Thompson type estimator of this probability is:

$$\pi_S = \frac{n(S)}{N\pi}, \qquad n(S) = \sum_{k=1}^{N} S_k = \sum_{k\in S} 1 \tag{7}$$

When we substitute the estimator π_S for π in the expression (4) we obtain the following one (see the more general case considered by Bethlehem 1988):

$$\bar{y}_S = \frac{1}{n(S)}\sum_{k=1}^{N} y_k S_k = \frac{1}{n(S)}\sum_{k\in S} y_k \tag{8}$$

So, the sample mean $\bar{y}_S$ is the ratio of the estimators $\tilde{y}_S = \sum_{k=1}^{N} y_k S_k$ and $n(S)$. They are unbiased estimators of the parameters $y\pi$ and $N\pi$, respectively. The

variance of the statistic is:

$$D^2(\tilde{y}_S) = \pi(1-\pi)\sum_{k=1}^{N} y_k^2, \quad D^2(n(S)) = \pi(1-\pi)N, \quad Cov(\tilde{y}_S, n(S)) = \pi(1-\pi)y \tag{9}$$

Hence, the evaluation of the variance of the statistic $\bar{y}_S$ is as follows:

$$D^2(\bar{y}_S) \approx \frac{D^2(\tilde{y}_S)}{N^2\pi^2} + \left(\frac{y}{N^2\pi}\right)^2 D^2(n(S)) - 2\frac{y}{N^3\pi^2} Cov(\tilde{y}_S, n(S)),$$

$$D^2(\bar{y}_S) \approx \frac{1-\pi}{N^2\pi}\left(\sum_{k=1}^{N} y_k^2 - \frac{y^2}{N}\right) = \frac{1-\pi}{N\pi} v_{yy} \tag{10}$$

where $v_{yy} = \frac{1}{N}\sum_{k=1}^{N}(y_k - \bar{y})^2$.

The estimator of the variance is:

$$D_S^2(\bar{y}_S) = \frac{N - n(S)}{Nn(S)} v_{yyS} \tag{11}$$

where $v_{yyS} = \frac{1}{n(S)-1}\sum_{k=1}^{N}(y_k - \bar{y}_S)^2 S_k$.

1.2. Stratified sample

We assume that a fixed and finite population is divided into H non empty and mutually disjoint strata, so $U = \bigcup_{h=1}^{H} U_h$. The fraction of h-th stratum size is denoted by $w_h = N_h/N$ where N_h is the size of the h-th stratum. Let π_h, $h = 1, \ldots, H$, be the response probability for an h-th stratum, $h = 1, \ldots, H$. We assume that the response probability π_h is the same for all population elements in the stratum U_h, $h = 1, \ldots, H$. It is the particular case of the Response Homogeneity Groups model considered by e.g. Särndal, Swenson, and Wretman. The Bernoulli sample selected from an h-th stratum will be denoted by Z_h and its size by $n(Z_h.)$ The sample will be denoted by $n(Z_h)$. So, the sample is: $S = \bigcup_{h=1}^{H} Z_h$ and its size: $n(S) = \sum_{h=1}^{H} n(Z_h)$. The stratified sample mean is as follows:

$$\tilde{y}_S = \sum_{h=1}^{H} w_h \bar{y}_{Z_h} \tag{12}$$

where

$$\bar{y}_{Z_h} = \frac{1}{n(Z_h)}\sum_{k\in U_h} y_k S_k = \frac{1}{n(Z_h)}\sum_{k\in Z_h} y_k \tag{13}$$

The variance and its unbiased estimators are:

$$D^2(\bar{y}_S) = \sum_{h=1}^{H} w_h^2 D^2(\bar{y}_{Z_h}), \qquad D_S^2(\bar{y}_S) = \sum_{h=1}^{H} w_h^2 D_{Z_h}^2(\bar{y}_{Z_h}) \tag{14}$$

where

$$D^2(\bar{y}_{Z_h}) \approx \frac{1-\pi_h}{N_h \pi_h} v_{hyy} \tag{15}$$

where $v_{hyy} = \frac{1}{N_h}\sum_{k=1}^{N_h}(y_k - \bar{y}_h)^2$,

$$D_{Z_h}^2(\bar{y}_{Z_h}) = \frac{N - n(Z_h)}{N n(Z_h)} v_{yyZ_h} \tag{16}$$

where $v_{yyZ_h} = \frac{1}{n(Z_h)-1}\sum_{k=1}^{N}(y_k - \bar{y}_{Z_h})^2 S_k$.

It may be shown that in the case of the stratification of the sample after its selection, the above results are valid, too.

2. Estimation under the Poisson model

2.1. The case of known response probabilities

Under the earlier introduced Poisson model of the generation of a sample, the estimator of the mean $\bar{y}$ can be as follows:

$$\bar{y}_{HTPS} = \frac{1}{N}\sum_{k=1}^{N}\frac{y_k S_k}{\pi_k} = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k} \tag{17}$$

It is an unbiased estimator of the mean. Its variance is as follows:

$$D^2(\bar{y}_{HTPS}) = \frac{1}{N^2}\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)}{\pi_k} \tag{18}$$

The unbiased estimator of the variance is:

$$D_S^2(\bar{y}_{HTPS}) = \frac{1}{N^2}\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)}{\pi_k^2} S_k \tag{19}$$

But there is a problem because the probabilities π_k, $k = 1, \ldots, N$, in the above expressions are not known. Let us suppose that it is reasonable to assume that the population can be divided into strata and in each of them there are the same

response probabilities. Moreover, we have at least one observation in the sample selected from each stratum and the sizes of the strata are known. So, these assumptions let us adopt the estimation procedure analyzed at the end of the section 1.2. Other approaches using the well known logit model, see e.g. Chow (1983), is presented below.

2.2. Logit approximation of response probabilities

We assume that $\mathbf{x}_k$ is the row vector of m-auxiliary variables values attached to a k-the population element, $k = 1, \ldots, N$. Particularly, all values of the first variable will be equal to one. Let us consider the following logit model of the probabilities (see Ekholm and Laaksonen 1991):

$$\pi_k \approx \frac{1}{1 + exp\{-q_k\}}, \qquad k = 1, \ldots, N \tag{20}$$

where

$$q_k = \mathbf{x}_k \beta \tag{21}$$

and β is the column vector of parameters. The likelihood function is as follows:

$$l_{log}(S, \beta) = \prod_{k=1}^{N} \pi_k^{S_k} (1 - \pi_k)^{1 - S_k} \tag{22}$$

The vector of the first derivatives of the log likelihood function is:

$$\mathbf{h}_{log,S} = \frac{\partial ln(l_{log}(S, \beta))}{\partial \beta} = \sum_{k=1}^{N} (S_k - \pi_k) \mathbf{x}_k^T \tag{23}$$

The maximum likelihood β_S estimators of the parameters β is a solution to the equation $\mathbf{h}_{log,s} = 0$. In order to evaluate it, an appropriate computer program is needed because solving of non-linear equations is necessary. One of them is the well known method of Newton-Raphson, see e.g. Kelley (2003).

The estimator of the inclusion probabilities can be denoted in the following way:

$$\pi_{Sk} = \frac{1}{1 + exp\{-q_{Sk}\}}, \qquad q_{Sk} = \mathbf{x}_k \beta_S, \qquad k = 1, \ldots, N \tag{24}$$

This leads us to the construction of the following estimator of the mean $\bar{y}$:

$$\bar{y}_{logS} = \frac{1}{N}\sum_{k=1}^{N}\frac{y_k S_k}{\pi_{sk}} = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_{sk}}. \tag{25}$$

Let us define the following moment.

$$\mathbf{m}_{r,v,z}\left(y, \mathbf{c}, f(\pi)\right) = \frac{1}{N}\sum_{k=1}^{N} y_k^r, \mathbf{c}_k^v, f^z(\pi_k) \tag{26}$$

where y and π are treated as variables which take values y_k and π_k, respectively, for $k = 1, \ldots, N$. Similarly, let $\mathbf{c}_k$ be a k-th observation of a matrix $\mathbf{c}$, $k = 1, \ldots, N$. If the matrix $\mathbf{c}$ reduces to a scalar c, then $m_{r,v,z}\left(y, c, f^v(\pi)\right)$ is substituted for $\mathbf{m}_{r,v,z}\left(y, \mathbf{c}, f^v(\pi)\right)$.

In the appendix, the following approximate formula for the variance is evaluated:

$$D^2\left(\bar{y}_{logS}\right) = D^2\left(\bar{y}_{HTPS}\right) + \sum_{i=1}^{5} a_i \tag{27}$$

where $D^2\left(\bar{y}_{HTPS}\right)$ is given by the expression (18) or by

$$D^2\left(\bar{y}_{HTPS}\right) = \frac{1}{N} m_{2,1}\left(y, \frac{1-\pi}{\pi}\right) \tag{28}$$

where

$$m_{2,1}\left(y, \frac{1-\pi}{\pi}\right) = \frac{1}{N}\sum_{k=1}^{N} y_k^2 \frac{1-\pi_k}{\pi_k}$$

and

$$a_1 = -\frac{2}{N^2} m_{2,1,1}\left(y, \mathbf{x}\mathbf{m}_{1,1,1}^{-1}(\mathbf{x}^T, \mathbf{x}, \pi(1-\pi))\mathbf{x}^T, (\pi^{-1}-1)(1-3\pi-2\pi^2)\right), \tag{29}$$

$$a_2 = -\frac{2}{N}\mathbf{m}_{1,1,1}(y, \mathbf{x}\mathbf{m}_{1,1,1}^{-1}(\mathbf{x}^T, \mathbf{x}, \pi(1-\pi)), 1-\pi)\mathbf{m}_{1,1,1}(y, \mathbf{x}^T, 1-\pi), \tag{30}$$

$$\begin{aligned} a_3 = \frac{1}{N^2}\Bigg(&\frac{1}{N} m_{2,2,1}\left(y, \mathbf{x}\mathbf{m}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right)\mathbf{x}^T, \pi^{-1}(1-\pi)^3(3\pi^2-3\pi+1)\right) + \\ &+ m_{2,1,1}\left(y, \mathbf{x}\mathbf{m}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right)\mathbf{x}^T, \pi^{-1}(1-\pi)^3\right) + \\ &- \frac{3}{N} m_{2,2,4}\left(y, \mathbf{x}\mathbf{m}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right)\mathbf{x}^T, 1-\pi\right) + \\ &+ m_{1,1,2}^2\left(y, \mathbf{x}\mathbf{m}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right)\mathbf{x}^T, 1-\pi\right)\Bigg), \end{aligned} \tag{31}$$

$$a_4 = \frac{2}{N^2}\mathbf{m}_{1,1,1}\Big(y, \mathbf{xm}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right)\mathbf{x}^T\mathbf{xm}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right), (1-\pi)^2(2\pi-1)\Big) \cdot \mathbf{m}_{1,1,1}\left(y, \mathbf{x}^T, 1-\pi\right), \tag{32}$$

$$a_5 = \frac{1}{N^2}\Big(-\frac{1}{N} m_{2,2,1}\left(y, \mathbf{xm}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right)\mathbf{x}^T, \pi(1-\pi)^3\right) + 2\mathbf{m}_{1,1,1}\left(y, \mathbf{xm}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right)\mathbf{x}^T\mathbf{xm}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}, (1-\pi)\pi\right), (1-\pi)^2\pi\right) \cdot \mathbf{m}_{1,1,1}\left(y, \mathbf{x}^T, 1-\pi\right) \tag{33}$$

If π_k is close to $\frac{n}{N}$, for $k = 1, \ldots, N$

$$D^2\left(\bar{y}_{logS}\right) = D^2\left(\bar{y}_{HTPS}\right) - 2\frac{N-n}{Nn}\mathbf{m}_{1,1}(y, \mathbf{xm}_{1,1}^{-1}(\mathbf{x}^T, \mathbf{x}))\mathbf{m}_{1,1}(y, \mathbf{x}^T) + \frac{1}{n^2}\Big(m_{2,1}\left(y, \mathbf{xm}_{1,1}^{-1}(\mathbf{x}^T, \mathbf{x})\mathbf{x}^T\right) + m_{1,1}^2\left(y, \mathbf{xm}_{1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}\right)\right)\mathbf{x}^T\right) + \mathbf{m}_{1,1}\left(y, \mathbf{xm}_{1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}\right)\mathbf{x}^T\mathbf{xm}_{1,1,1}^{-1}\left(\mathbf{x}^T, \mathbf{x}\right)\right)\mathbf{m}_{1,1}\left(y, \mathbf{x}^T\right)\Big) + O(n^{-1}N^{-1})$$

$$D^2\left(\bar{y}_{logS}\right) = \frac{N-n}{Nn}\left(\mathbf{m}_2(y) - 2\mathbf{m}_{1,1}(y, \mathbf{xm}_{1,1}^{-1}(\mathbf{x}^T, \mathbf{x}))\mathbf{m}_{1,1}(y, \mathbf{x}^T)\right) + O(n^{-2}). \tag{34}$$

The estimator of the variance $D_S^2\left(\bar{y}_{logS}\right)$ can be obtained through substitution the appropriate sample moments $\mathbf{m}_{r,v,z,S}\left(y, \mathbf{c}, f(\pi)\right)$ for the moments $\mathbf{m}_{r,v,z}\left(y, \mathbf{c}, f(\pi)\right)$ in the expressions (27)-(34) where the sample moment is defined as follows.

$$\mathbf{m}_{r,v,z,S}\left(y, \mathbf{c}, f^v(\pi)\right) = \frac{1}{\sum_{k=1}^N S_k}\sum_{k=1}^N y_k^r, \mathbf{c}_k^v, f^z(\pi_k)S_k \tag{35}$$

Conclusions

Several estimators of mean values are considered in the case when the sample is equal to the whole population. Each respondent is supplied with a questionnaire by means of the Internet, e.g. by email. But there exist respondents who do not take part in the survey. So, it is the problem of non response. The two models of sampling were fitted to the considered problem. It was the Poisson sampling design and its particular case – the Bernoulli sampling design. Properties of those

sampling models let us adopt well-known estimators to estimate the mean value. Those estimators are unbiased or almost unbiased and their approximate variances are evaluated. Respective variance estimators are also given.

The inference on the basis of the stratified sample is proper when response probabilities are homogenous inside each stratum. In other cases the logit estimator involving the Poisson model should be used although it is quite complicated. In a separate paper a comparative analysis of the accuracy of the considered estimators should be developed in order to explain their properties under several assumed artificial distributions of a variable under study and auxiliary variables.

The analyzed problem can be generalized in several ways. The estimation of a mean value can be straightforwardly used to estimate the population total.

Appendix

Evaluation of the expression (27).

The expressions (24) and (23) lead to the following matrix of the second derivatives of the log-likelihood function:

$$\frac{\partial^2 ln(l_{log}(s,\beta))}{\partial\beta\beta^T} = -\sum_{k=1}^{N}\pi_k(1-\pi_k)\mathbf{x}_k^T\mathbf{x}_k = -\mathbf{H}_{log} \tag{36}$$

Let

$$\mathbf{H}_{log} = N\mathbf{m}_{1,1,1}(\mathbf{x}^T, \mathbf{x}, \pi(1-\pi)) \tag{37}$$

where

$$\mathbf{m}_{1,1,1}(\mathbf{x}^T, \mathbf{x}, \pi(1-\pi)) = \frac{1}{N}\sum_{k=1}^{N}\pi_k(1-\pi_k)\mathbf{x}_k^T\mathbf{x}_k,$$

The first derivative $h_{log,s} = h_{log,s}(\beta)$ shown by the expression (17) is the function of the parameters β because $\pi = \pi(\beta)$. So, $h_{log,s} = h_{log,s}(\beta_s) = \mathbf{0}$ because β_s is maximum likelihood estimator of β. his leads to the following Taylor's expansion of the vector of the first derivative $\mathbf{h}_{log,s}$ as a function of β in the neighborhood of β_s.

$$\mathbf{h}_{log,s}(\beta) \approx h_{log,s}(\beta_s) + \frac{\partial\mathbf{h}_{log,s}(\beta)}{\partial\beta\beta^T}(\beta-\beta_s) = \frac{\partial^2 ln(l_{log}(s,\beta))}{\partial\beta\beta^T}(\beta-\beta_s)$$

or

$$\mathbf{h}_{log,s} \approx \mathbf{H}_{log}(\beta_s - \beta)$$

So,

$$\beta_s - \beta \approx \mathbf{H}_{log}^{-1}\mathbf{h}_{log,s}.$$

This and the expressions (23) and (36) lead to the following:

$$\beta_S - \beta \approx \mathbf{H}_{log}^{-1}\sum_{k=1}^{N}(S_k - \pi_k)\mathbf{x}_k^T \tag{38}$$

$$\beta_S - \beta \approx \left(\sum_{k=1}^{N}\pi_k(1-\pi_k)\mathbf{x}_k^T\mathbf{x}_k\right)^{-1}\sum_{k=1}^{N}(S_k - \pi_k)\mathbf{x}_k^T. \tag{39}$$

It is easy to show that $E(\beta_S - \beta)(\beta_S - \beta)^T = \mathbf{H}_{log}^{-1}$.

The equation (24) let us rewrite the expression (25) in the following way:

$$\bar{y}_{log,S} = \frac{1}{N}\sum_{k=1}^{N} y_k S_k \left(1 + e^{-x_k\beta_s}\right)$$

Under the assumption that $\pi_k^{-1} = 1 + e^{-x_k\beta}$, $k = 1, \ldots, N$, see the expression (20), the Taylor's expansion of the estimator $\bar{y}_{logS}$ as the function of β_S in the neighborhood of the parameters β is as follows:

$$\bar{y}_{log,S}(\beta_S) = \bar{y}_{HT,S} - \frac{1}{N}\sum_{k=1}^{N} y_k S_k e^{-q_k}\mathbf{x}_k(\beta_S - \beta)$$

where $\bar{y}_{HT,S} = N^{-1}\sum_{k=1}^{N} y_k S_k \pi_k^{-1}$ is the Horvitz-Thompson estimator. The expansion is derived with accuracy to the linear elements. This lets us infer that the statistic $\bar{y}_{logS}$ is an approximately unbiased estimator of the mean value. The above expression and $e^{-\mathbf{x}_k\beta_S} = \pi_k^{-1} - 1$ lead to the following one:

$$\bar{y}_{log,S} = \bar{y}_{HT,S} - \frac{1}{N}\sum_{k=1}^{N}\frac{y_k(1-\pi_k)S_k}{\pi_k}(\beta_S - \beta)$$

This let us make the following derivation of the mean square error.

$$MSE\left(\bar{y}_{log,S}\right) =$$
$$= E\left(\frac{2}{N}\sum_{k=1}^{N}\frac{y_k(S_k-\pi_k)}{\pi_k} - \frac{1}{N}\sum_{k=1}^{N}\frac{y_k(1-\pi_k)S_k}{\pi_k}\mathbf{x}_k(\beta_s-\beta)\right)^2 =$$
$$= E\Bigg(\frac{1}{N}\sum_{k=1}^{N}\frac{y_k(S_k-\pi_k)}{\pi_k} - \frac{1}{N}\sum_{k=1}^{N}\frac{y_k(1-\pi_k)(S_k-\pi_k)}{\pi_k}\mathbf{x}_k(\beta_s-\beta)+$$
$$- \frac{1}{N}\sum_{k=1}^{N}y_k(1-\pi_k)\mathbf{x}_k(\beta_s-\beta)\Bigg)^2 = E(c_1+c_2+c_3)^2 + D^2\left(\bar{y}_{HT,S}\right) + \sum_{i=1}^{5}a_i \tag{40}$$

where

$$a_1 = 2E(c_1c_2) = -\frac{2}{N^2}E\sum_{k=1}^{N}\frac{y_k(S_k-\pi_k)}{\pi_k}\sum_{k=1}^{N}\frac{y_k(1-\pi_k)(S_k-\pi_k)}{\pi_k}\mathbf{x}_k(\beta_s-\beta) =$$
$$= \frac{2}{N^2}E\Bigg(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)(S_k-\pi_k)^2}{\pi_k^2}\mathbf{x}_k(\beta_s-\beta)+$$
$$+\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq k}}^{N}\frac{y_ky_h(1-\pi_k)(S_k-\pi_k)(S_h-\pi_h)}{\pi_k\pi_h}\mathbf{x}_k(\beta_s-\beta)\Bigg)$$

On the basis of the expression (38) we have:

$$a_1 = -\frac{2}{N^2}E\Bigg(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)(S_k-\pi_k)^2}{\pi_k^2}\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{k=1}^{N}(S_k-\pi_k)\mathbf{x}_k^T+$$
$$+\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq k}}^{N}\frac{y_ky_h(1-\pi_k)(S_k-\pi_k)(S_h-\pi_h)}{\pi_k\pi_h}\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{k=1}^{N}(S_k-\pi_k)\mathbf{x}_k^T\Bigg) =$$
$$= -\frac{2}{N^2}\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)E(S_k-\pi_k)^3}{\pi_k^2}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T+$$
$$-\frac{2}{N^2}\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq k}}^{N}\frac{y_k^2(1-\pi_k)E(S_k-\pi_k)^2E(S_h-\pi_h)}{\pi_k^2}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T+$$
$$-\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq k}}^{N}\frac{y_ky_h(1-\pi_k)E(S_k-\pi_k)^2E(S_h-\pi_h)}{\pi_k\pi_h}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T+$$

$$-\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_k)E(S_k-\pi_k)E(S_h-\pi_h)^2}{\pi_k\pi_h}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T+$$

$$-\frac{2}{N^2}\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq h}}^{N}\sum_{\substack{t=1\\t\neq h\\t\neq k}}^{N}\frac{y_k y_h(1-\pi_k)E(S_k-\pi_k)E(S_h-\pi_h)E(S_t-\pi_t)}{\pi_k\pi_h}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_t^T=$$

$$=-\frac{2}{N^2}\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)(1-3\pi_k-2\pi^2)}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T.$$

This and the expression (36) lead to the following one:

$$\begin{aligned}a_1&=-\frac{2}{N^2}\sum_{k=1}^{N}y_k^2\frac{(1-\pi_k)(1-3\pi_k-2\pi^2)}{\pi_k}\mathbf{x}_k\left(\sum_{h=1}^{N}\pi_h(1-\pi_h)\mathbf{x}_k^T\mathbf{x}_k\right)^{-1}\mathbf{x}_k^T=\\&=-\frac{2}{N^3}\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)(1-3\pi_k-2\pi_k^2)}{\pi_k}\mathbf{x}_k\mathbf{m}_{1,1,1}^{-1}(\mathbf{x}^T,\mathbf{x},\pi(1-\pi))\mathbf{x}_k^T\end{aligned}\tag{41}$$

This leads to the expression (29):

$$a_2=2E(c_1c_3)=\frac{2}{N^2}E\sum_{k=1}^{N}\frac{y_k(S_k-\pi_k)}{\pi_k}\sum_{k=1}^{N}y_k(1-\pi_k)\mathbf{x}_k(\beta_s-\beta)=$$

$$=\frac{2}{N^2}E\Bigg(\sum_{k=1}^{N}\frac{y_k^2(S_k-\pi_k)(1-\pi_k)\mathbf{x}_k(\beta_s-\beta)}{\pi_k}+$$

$$+\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_h)(S_k-\pi_k)}{\pi_k}\mathbf{x}_h(\beta_s-\beta)\Bigg),$$

$$\begin{aligned}a_2=&-\frac{2}{N^2}E\Bigg(\sum_{k=1}^{N}\frac{y_k^2(S_k-\pi_k)(1-\pi_k)}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{h=1}^{N}(S_h-\pi_h)\mathbf{x}_h^T+\\&+\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_h)(S_k-\pi_k)}{\pi_k}\mathbf{x}_h\mathbf{H}_{log}^{-1}\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t^T\Bigg)=\\=&-\frac{2}{N^2}E\Bigg(\sum_{k=1}^{N}\frac{y_k^2(S_k-\pi_k)^2(1-\pi_k)}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T+\\&+\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k^2(1-\pi_k)(S_k-\pi_k)(S_h-\pi_h)}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T+\\&+\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_h)(S_k-\pi_k)^2}{\pi_k}\mathbf{x}_h\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\Bigg)=\end{aligned}\tag{42}$$

$$= -\frac{2}{N^2}\left(\sum_{k=1}^{N} y_k^2(1-\pi_k)^2\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T + \sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N} y_k y_h(1-\pi_h)(1-\pi_k)\mathbf{x}_h\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right) =$$

$$= -\frac{2}{N^2}\sum_{h=1}^{N} y_h(1-\pi_h)\mathbf{x}_h\mathbf{H}_{log}^{-1}\sum_{k=1}^{N} y_k(1-\pi_k)\mathbf{x}_k^T =$$

$$= -\frac{2}{N^3}\sum_{h=1}^{N} y_h(1-\pi_h)\mathbf{x}_h\mathbf{m}_{1,1,1}^{-1}(\mathbf{x}^T,\mathbf{x},\pi(1-\pi))\sum_{k=1}^{N} y_k(1-\pi_k)\mathbf{x}_k^T$$

This leads to the expression (30):

$$a_3 = E(c_2^2) = \frac{1}{N^2}E\left(\sum_{k=1}^{N}\frac{y_k(1-\pi_k)(S_k-\pi_k)}{\pi_k}\mathbf{x}_k(\beta_s-\beta)\right)^2 =$$

$$= \frac{1}{N^2}E\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)^2}{\pi_k^2}\mathbf{x}_k(\beta_s-\beta)(\beta_s-\beta)^T\mathbf{x}_k^T + \right.$$

$$\left. + \sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_k)(1-\pi_h)(S_k-\pi_k)(S_k-\pi_h)}{\pi_k\pi_h}\mathbf{x}_k(\beta_s-\beta)(\beta_s-\beta)^T\mathbf{x}_h^T\right),$$

$$a_3 = \frac{1}{N^2}E\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)^2}{\pi_k^2}\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{h=1}^{N}(S_h-\pi_h)\mathbf{x}_h^T\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_k^T + \right.$$

$$+ \sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_k)(1-\pi_h)(S_k-\pi_k)(S_h-\pi_h)}{\pi_k\pi_h}\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{p=1}^{N}(S_p-\pi_p)\mathbf{x}_p^T\cdot$$

$$\left.\cdot\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\right) =$$

$$= \frac{1}{N^2}E\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)^3}{\pi_k^2}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_k^T + \right.$$

$$+ \sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)^2(S_h-\pi_h)}{\pi_k^2}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_k^T +$$

$$+ 2\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_k)(1-\pi_h)(S_k-\pi_k)^2(S_h-\pi_h)}{\pi_k\pi_h}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\cdot$$

$$\left.\cdot\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\right) = \frac{1}{N^2}E\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)^4}{\pi_k^2}\left(\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right)^2 + \right.$$

$$+ \sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)^2(S_h-\pi_h)^2}{\pi_k^2}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\mathbf{x}_h\mathbf{H}_{log}^{-1}\mathbf{x}_k^T +$$

$$\left. + 2\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k y_h(1-\pi_k)(1-\pi_h)(S_k-\pi_k)^2(S_h-\pi_h)^2}{\pi_k\pi_h}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\mathbf{x}_h\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\right),$$

$$a_3 = \frac{1}{N^2}\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^3(3\pi_k^2-3\pi_k+1)}{\pi_k}\left(\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right)^2+\right.$$
$$+\sum_{k=1}^{N}\sum_{\substack{h=1\\h\neq k}}^{N}\frac{y_k^2(1-\pi_k)^3(1-\pi_h)\pi_h}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\mathbf{x}_h\mathbf{H}_{log}^{-1}\mathbf{x}_k^T+$$
$$\left.+2\sum_{k=1}^{N}\sum_{h=1,h\neq k}^{N}y_ky_h(1-\pi_k)^2(1-\pi_h)^2\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\mathbf{x}_h\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\right),$$

$$a_3 = \frac{1}{N^2}\left(Nm_{2,2,1}\left(y,\mathbf{x}\mathbf{H}_{log}^{-1}\mathbf{x}^T,\pi^{-1}(1-\pi_k)^3(3\pi_k^2-3\pi_k+1)\right)+\right.$$
$$+\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^3}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\left(\sum_{h=1}^{N}(1-\pi_h)\pi_h\mathbf{x}_h^T\mathbf{x}_h\right)\mathbf{H}_{log}^{-1}\mathbf{x}_k^T-$$
$$\left.+3\sum_{k=1}^{N}y_k^2(1-\pi_k)^4\left(\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right)^2+2\left(\sum_{k=1}^{N}y_k(1-\pi_k)^2\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right)^2\right)$$

On the basis of this result and the equations (36) and (37) we derive the expression (31):

$$a_4 = 2E(c_2c_3) = \frac{2}{N^2}E\sum_{k=1}^{N}\frac{y_k(1-\pi_k)(S_k-\pi_k)}{\pi_k}\mathbf{x}_k(\boldsymbol{\beta}_s-\boldsymbol{\beta})\sum_{k=1}^{N}y_k(1-\pi_k)\mathbf{x}_k(\boldsymbol{\beta}_s-\boldsymbol{\beta}) =$$
$$= \frac{2}{N^2}E\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)}{\pi_k}\mathbf{x}_k(\boldsymbol{\beta}_s-\boldsymbol{\beta})(\boldsymbol{\beta}_s-\boldsymbol{\beta})^T\mathbf{x}_k^T+\right.$$
$$\left.+\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq h}}^{N}\frac{y_ky_h(1-\pi_k)(1-\pi_h)(S_k-\pi_k)}{\pi_k}\mathbf{x}_k(\boldsymbol{\beta}_s-\boldsymbol{\beta})(\boldsymbol{\beta}_s-\boldsymbol{\beta})^T\mathbf{x}_h^T\right) =$$
$$= \frac{2}{N^2}E\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^2(S_k-\pi_k)}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{h=1}^{N}(S_h-\pi_h)\mathbf{x}_h^T\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_k^T+\right.$$
$$+\sum_{k=1}^{N}\sum_{h=1,k\neq h}^{N}\frac{y_ky_h(1-\pi_k)(1-\pi_h)(S_k-\pi_k)}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\cdot$$
$$\left.\cdot\sum_{h=1}^{N}(S_h-\pi_h)\mathbf{x}_h^T\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\right) =$$
$$= \frac{2}{N^2}\left(\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)^2E(S_k-\pi_k)^3}{\pi_k}\left(\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right)^2+\right. \tag{43}$$

$$+\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq h}}^{N}\frac{y_k y_h(1-\pi_k)(1-\pi_h)E(S_k-\pi_k)^3}{\pi_k}\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\Bigg)=$$

$$=-\frac{2}{N^2}\sum_{k=1}^{N}y_k(1-\pi_k)^2(2\pi_k-1)\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{h=1}^{N}y_h(1-\pi_h)\mathbf{x}_h^T$$

This leads to the expression (43):

$$a_5=E(c_3^2)=\frac{1}{N^2}E\left(\sum_{k=1}^{N}y_k(1-\pi_k)\mathbf{x}_k(\beta_s-\beta)\right)^2=$$

$$=\frac{1}{N^2}E\Bigg(\sum_{k=1}^{N}y_k^2(1-\pi_k)^2\mathbf{x}_k(\beta_s-\beta)(\beta_s-\beta)^T\mathbf{x}_k^T+$$

$$+\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq h}}^{N}y_k y_h(1-\pi_k)(1-\pi_h)\mathbf{x}_k(\beta_s-\beta)(\beta_s-\beta)^T\mathbf{x}_h^T\Bigg)=$$

$$=\frac{1}{N^2}E\Bigg(\sum_{k=1}^{N}y_k^2(1-\pi_k)^2\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{h=1}^{N}(S_h-\pi_h)\mathbf{x}_h^T\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_k^T+$$

$$+\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq h}}^{N}y_k y_h(1-\pi_k)(1-\pi_h)\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{h=1}^{N}(S_h-\pi_h)\mathbf{x}_h^T\sum_{t=1}^{N}(S_t-\pi_t)\mathbf{x}_t\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\Bigg)=$$

$$=\frac{1}{N^2}\Bigg(\sum_{k=1}^{N}y_k^2(1-\pi_k)^2E(S_k-\pi_k)^2\left(\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right)^2+$$

$$+2\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq h}}^{N}y_k y_h(1-\pi_k)(1-\pi_h)E(S_k-\pi_k)^2\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\Bigg)=$$

$$=\frac{1}{N^2}\Bigg(\sum_{k=1}^{N}y_k^2\pi_k(1-\pi_k)^3\left(\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\right)^2+$$

$$+2\sum_{k=1}^{N}\sum_{\substack{h=1\\k\neq h}}^{N}y_k y_h\pi_k(1-\pi_k)^2(1-\pi_h)\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_h^T\Bigg)=$$

$$=\frac{1}{N^2}\Bigg(-\frac{1}{N}m_{2,2,1}\left(y,\mathbf{x}\mathbf{m}_{1,1,1}^{-1}\left(\mathbf{x}^T,\mathbf{x},(1-\pi)\pi\right)\mathbf{x}^T,\pi_k(1-\pi_k)^3\right)+$$

$$+2\sum_{k=1}^{N}y_k\pi_k(1-\pi_k)^2\mathbf{x}_k\mathbf{H}_{log}^{-1}\mathbf{x}_k^T\mathbf{x}_k\mathbf{H}_{log}^{-1}\sum_{h=1}^{N}y_h(1-\pi_h)\mathbf{x}_h^T\Bigg) \tag{44}$$

References

Bethlehem J.G. (1988): Reduction of Nonresponse Bias Through Regression Estimation. "Journal of Official Statistics", Vol. 4, No. 3, pp. 251-260.

Chow G. C. (1983): Econometrics. McGraw Hill Book Company, New York.

Ekholm A., Laaksonen S. (1991): Weighting via Response Modelling in the Finish Household Budget Survey. "Journal of Official Statistics", Vol. 7, No. 3, pp. 325-338.

Horvitz D.G., Thompson D.J. (1952): A Generalization of the Sampling without Replacement from Finite Universe. "Journal of the American Statistical Association", Vol. 47, pp. 663-685.

Kelley C.T. (2003): Solving Nonlinear Equations with Newton's No. 1 in Fundamentals of Algorithms. SIAM, Philadelphia.

Särndal C.E., Swensson B., Wretman J. (1992): Model Assisted Survey Sampling. Springer Verlag, New York.

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

Abstract

The paper deals with a problem of estimating the total on the basis of data observed in Internet sample. The Poisson sampling design without replacement is used as a basic model of generation of Internet sample. Its particular case is so called the Bernoulli sampling design without replacement when all the response probabilities are the same. Some estimators (including logit type one) of the population mean as well as of the total are considered. Their variances are evaluated and their estimators, too.

Janusz L. Wywiał*

SAMPLING DESIGN PROPORTIONAL TO POSITIVE FUNCTION OF ORDER STATISTICS OF AUXILIARY VARIABLE

Introduction

The sampling designs dependent on sample moments of auxiliary variables are well known. Except mentioned Lahiri's sampling design Sing and Srivastava's (1980) sampling design is proportionate to a sample variance while Wywiał's (1999) one is proportionate to a sample generalized variance of auxiliary variables. Some other sampling designs dependent on moments of an auxiliary variable were considered e.g. by Wywiał (2000, 2003, 2003a) where accuracy of some sampling strategies were compared, too.

As it was mentioned Wywiał (2004, 2007) proposed the sampling design proportional to the value of an order statistic of a positive auxiliary variable observed in the simple sample selected without replacement. This sampling designs can be useful in the case when there are some censored observations of the auxiliary variable. Moreover, it cannot be too much sensitive to outliers observations. Its particular cases as well as its conditional version were considered, too. The

*The research was supported by the grant No. 1 H02B 018 27 from the Ministry of Science and Higher Education.

sampling scheme implementing this sampling design was proposed. The inclusion probabilities of the first and second orders were evaluated. The well known Horvitz-Thompson estimator was taken into account. A ratio estimator dependent on an order statistic was constructed. It is an unbiased estimator of the population mean when the sample is drawn according to the proposed sampling design dependent on the appropriate order statistic.

1. Basic definitions and notation

Let $U = (1, \ldots, i, \ldots, N)$ be a fixed population of size N. The observation of a variable under study and an auxiliary variable are identifiable and denoted by y_i and $x_i, i = 1, \ldots, N$, respectively. Firstly, we assume that $x_i < x_{i+1}, i = 1, \ldots, N-1$. Our problem is how to construct a sample strategy in order to estimate the population average $\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$.

Let us consider the sample space **S** of the samples s of the fixed effective size $1 < n < N$. The sampling design is denoted by $P(S = s)$ or more simply by $P(s)$ or $P(S)$. We assume that $P(s) > 0$ for all $s \in \mathbf{S}$ and $\sum_{s\in\mathbf{S}} P(s) = 1$.

Let $s = \{i_j : i_j < i_{j+1}, j = 1, \ldots, n-1\}$. Moreover, let

$$s = \{s_1, i_{r_1}, s_2, i_{r_2}, \ldots, s_j, i_{r_j}, \ldots, s_{h-1}, i_{r_{h-1}}, s_h, i_{r_h}, s_{h+1}\}$$

where $s_1 = \{i_1, \ldots, i_{r_1-1}\}$, $s_2 = \{i_{r_1+1}, \ldots, i_{r_2-1}\}$, $\ldots$, $s_j = \{i_{r_{j-1}+1}, \ldots, i_{r_j-1}\}$, $\ldots$, $s_h = \{i_{r_{h-1}+1}, \ldots, i_{r_h-1}\}$, $s_{h+1} = \{i_{r_h}, \ldots, i_{n-r_h}\}$. If $r_1 = 1$ then $s_1 = \varnothing$. When $r_h = n$, $s_{h+1} = \varnothing$. So, x_i is one of the possible observations of the order statistic $X_{(r_j)}$ of the rank r_j $(j = 1, \ldots, h \leq n)$ of the auxiliary variable from the sample s. In order to simplify the notation we state that $i_j = i_{r_j}$ for $j = 1, \ldots, h \leq n$. Let $G(r_1, \ldots, r_h; i_1, \ldots, i_h) = \{s : X_{(r_1)} = x_{i_1}, \ldots, X_{(r_j)} = x_{i_j}, \ldots, X_{(r_h)} = x_{i_h}\}$ be the set of all samples whose order statistics of ranks $r_1, \ldots, r_h$ of the auxiliary variable are equal to $x_{i_1}, \ldots, x_{i_h}$, respectively, where $i_j < i_{j+1}$ and $r_j \leq i_j \leq N - n + r_j$ for $j = 1, \ldots, h$. Particularly, $G(r_1; i_1) = \{s : X_{(r_1)} = x_{i_1}\}$. Hence,

$\bigcup_{\{i_1,\ldots,i_h\}} G(r_1,\ldots,r_h;i_1,\ldots,i_h) = \mathbf{S}$ or more precisely:

$$\bigcup_{i_1=1}^{N-n+r_1} \bigcup_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \bigcup_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} \cdots \bigcup_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \bigcup_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} G(r_1,\ldots,r_h;i_1,\ldots,i_h) = \mathbf{S} \tag{1}$$

In the Appendix the following expression has been derived:

$$g_h = g(r_1,\ldots,r_h;i_1,\ldots,i_h) = \prod_{j=1}^{h+1} \binom{i_j - i_{j-1} - 1}{r_j - r_{j-1} - 1} \tag{2}$$

where $r_0 = i_0 = 0$, $r_{h+1} = n+1$, $i_{h+1} = N+1$.

As it is well known the simple sampling design is defined as follows: $P_0(s) = \binom{N}{n}^{-1}$ for all $s \in \mathbf{S}$. *Wilks (1962, pp. 243-244) shows that the probability*, that the order statistics of ranks $r_1,\ldots,r_h$ of the auxiliary variable from simple sample (drawn without replacement) of an auxiliary variable takes values $x_{i_1},\ldots,x_{i_h}$, is as follows (see Guenther, 1975, Hogg and Craig, 1970, too).

$$P_0\left(X_{(r_1)} = x_{i_1},\ldots,X_{(r_h)} = x_{i_h}\right) = P_0\left(s \in G(r_1,\ldots,r_h;i_1,\ldots,i_h)\right) = \frac{g(r_1,\ldots,r_h;i_1,\ldots,i_h)}{\binom{N}{n}} \tag{3}$$

where $r_j \le i_j \le N-n+r_j$, $j = 1,\ldots,h \le n$.

Particularly,

$$P_0\left(X_{(r_1)} = x_{i_1}, X_{(r_2)} = x_{i_2}\right) = \frac{\binom{i_1-1}{r_1-1}\binom{i_2-i_1-1}{r_2-r_1-1}\binom{N-i_2}{n-r_2}}{\binom{N}{n}} \tag{4}$$

$$P_0\left(X_r = x_i\right) = P_0\left(s \in G(r;i)\right) = \frac{\binom{i-1}{r-1}\binom{N-i}{n-r}}{\binom{N}{n}} \tag{5}$$

$$E_0\left(X_{(r)}\right) = \sum_{i=r}^{N-n+r} x_i P_0\left(X_{(r)} = x_i\right) = \frac{1}{\binom{N}{n}} \sum_{i=r}^{N-n+r} x_i g(r,i). \tag{6}$$

If $x_i = i$ for $i = 1,\ldots,N$, the distribution function expressed by the equation (5) is called the negative hypergeometrical distribution (see Johnson and Kotz 1969, p. 157) or the inverse hypergeometrical one (see Patil and Joshi 1968, pp. 27-28) and $E_0\left(X_{(r)}\right) = \frac{1}{\binom{N}{n}} \sum_{i=r}^{N-n+r} i g(r,i) = \frac{r(N+1)}{n+1}$.

The sample quantile of order $\alpha \in (0;1)$ can be defined by the equation:

$$Q_\alpha = X_{(r)} \tag{7}$$

where $r = [n\alpha]+1$ is the integer part of the value $n\alpha$, (see e.g. Fisz 1963). Hence, $r = 1, 2, \ldots, n$ and $X_{(r)} = Q_\alpha$ for $\frac{r-1}{n} \leq \alpha < \frac{r}{n}$.

2. Sampling design proportional to positive function of order statistics

Let $W(X_{r_1}, \ldots, X_{r_h})$ be a positive function of the order statistics $X_{(r_1)}, \ldots, X_{(r_h)}$. The value of the statistic $W = W(X_{r_1}, \ldots, X_{r_h})$ will be denoted by $w(i_1, \ldots, i_h)$ or by w where $r_j \leq i_j \leq N - n + r_j$, $j = 1, \ldots, h$.

The expressions (1) and (38) lead to the following ones:

$$z_{r_1,\ldots,r_h} = \sum_{\{s \in \mathbf{S}\}} w(i_1, \ldots, i_h) = \sum_{\{s \in A\}} w$$

where

$$A = \bigcup_{i_1=1}^{N-n+r_1} \mathbf{S}(U(1, \ldots, i_1 - 1), s_1) \times \{i_1\} \times \ldots$$

$$\ldots \times \bigcup_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \mathbf{S}(U(i_{h-1} + 1, \ldots, i_h), s_h) \times \{i_h\} \times \mathbf{S}(U(i_h + 1, \ldots, N), s_{h+1})\}$$

$$z_{r_1,\ldots,r_h} = \sum_{\{s \in \bigcup_{i_1=1}^{N-n+r_1} \bigcup_{i_2=i_1+r_2-r_1}^{N-n+r_2} \ldots \bigcup_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} G(r_1,\ldots,r_h;i_1,\ldots,i_h)\}} w.$$

Finally, we have:

$$z_{r_1,\ldots,r_h} = \sum_{i_1=1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} g(r_1, \ldots, r_h; i_1, \ldots, i_h) w \tag{8}$$

The above expression leads to the following one:

$$E_0\left(W(X_{r_1}, \ldots, X_{r_h})\right) = \frac{z_{r_1,\ldots,r_h}}{\binom{N}{n}} \tag{9}$$

Definition 1. *The sampling design proportional to the values* $w(i_1,\ldots,i_h)>0$ *of the statistic* $W(r_1,\ldots,r_h)$ *is as follows:*

$$P_{r_1,\ldots,r_h}(s)=\frac{w(i_1,\ldots,i_h)}{z_{r_1,\ldots,r_h}} \tag{10}$$

for $s\in G(r_1,\ldots,r_h;i_1,\ldots,i_h)$ *or equivalently:*

$$P_{r_1,\ldots,r_h}(S)=\frac{W(r_1,\ldots,r_h)}{z_{r_1,\ldots,r_h}}$$

The important case of the function $W(r_1,\ldots,r_h)$ is the following positive linear function:

$$T=T(r_1,\ldots,r_h)=\sum_{j=1}^{h}a_jX_{(r_j)} \tag{11}$$

where a_j, $j=1,\ldots,h$ be such real and non-random values that at least one value $a_j\neq 0$. Let $t=t(i_1,\ldots,i_h)=\sum_{j=1}^{h}a_jx_{i_j}$ be a value of the random variable $T(r_1,\ldots,r_h)$ which expected value is as follows:

$$E_0\left(T(r_1,\ldots,r_h)\right)=\sum_{j=1}^{h}a_jE_0(X_{(r_j)}) \tag{12}$$

where

$$E_0(X_{(r_j)})=\frac{1}{\binom{N}{n}}\sum_{i=r_j}^{N-n+r_j}x_i\binom{i-1}{r_j-1}\binom{N-i}{n-r_j}$$

Moreover:

$$\begin{aligned}z_{r_1,\ldots,r_h}&=\sum_{s\in\mathbf{S}}t(r_1,\ldots,r_h)=\binom{N}{n}E_0\left(T(r_1,\ldots,r_h)\right)=\\&=\binom{N}{n}\sum_{j=1}^{h}a_jE_0\left(X_{(r_2)}\right)=\sum_{j=1}^{h}a_j\sum_{i=r_j}^{N-n+r_j}\binom{i_j-1}{r_j-1}\binom{N-i_j}{n-r_j}x_{i_j}\end{aligned} \tag{13}$$

Definition 2. *The sampling design proportional to the values* $t(i_1,\ldots,i_h)>0$ *of the statistic* $T(r_1,\ldots,r_h)$ *is as follows:*

$$P_{r_1,\ldots,r_h}(s)=\frac{t(i_1,\ldots,i_h)}{z_{r_1,\ldots,r_h}} \tag{14}$$

for $s\in G(r_1,\ldots,r_h;i_1,\ldots,i_h)$ *or equivalently*

$$P_{r_1,\ldots,r_h}(S)=\frac{T(r_1,\ldots,r_h)}{\binom{N}{n}E_0\left(T(r_1,\ldots,r_h)\right)}$$

When $a_j = 1/n$ for all $j = 1, \ldots, n$, the sampling design $P_{r_1,\ldots,r_1}(s)$ reduces to the one proportional to sample mean of the auxiliary variable which was considered e.g by Lahiri (1951).

In the case when $h = 1$ the expression (14) reduces to the following one proposed by Wywiał (2004, 2007):

$$P_r(s) = \frac{x_i}{\binom{N}{n} E_0\left(X_{(r)}\right)} \tag{15}$$

for $s \in G(r; i)$ or equivalently

$$P_r(S) = \frac{X_{(r)}}{\binom{N}{n} E_0\left(X_{(r)}\right)}$$

Wywiał (2004) considered the following particular case of the definition 2.

Definition 3. *The sampling design proportional to the value $x_{i_2} - x_{i_1}$ of the difference $X_{(r_2)} - X_{(r_1)}$ between two order statistics is as follows:*

$$P_{r_1,r_2}(s) = \frac{x_{i_2} - x_{i_1}}{z_{r_1,r_2}} \tag{16}$$

for $s \in G(r_1, r_2; i_1, i_2)$ where

$$\begin{aligned} z_{r_1,r_2} &= \binom{N}{n}\left(E_0\left(X_{(r_2)}\right) - E_0\left(X_{(r_1)}\right)\right) = \\ &= \sum_{i_2=r_2}^{N-n+r_2} \binom{i_2-1}{r_2-1}\binom{N-i_2}{n-r_2} x_{i_2} - \sum_{i_1=r_1}^{N-n+r_1} \binom{i_1-1}{r_1-1}\binom{N-i_1}{n-r_1} x_{i_1} \end{aligned} \tag{17}$$

or equivalently

$$P_{r_1,r_2}(S) = \frac{X_{(r_2)} - X_{(r_1)}}{\binom{N}{n}\left(E_0\left(X_{(r_2)}\right) - E_0\left(X_{(r_1)}\right)\right)}$$

Let us note that the particular case of the sampling design is the sample range of an auxiliary variable: $P_{r_1,r_n}(s) = \left(X_{(r_n)} - X_{(r_1)}\right)/z_{r_1,r_n}$.

The spread of values of an univariate variable can be analyzed by means of so called second L-moment which is alternative coefficient to the standard deviation, see e.g. Sillito (1969) or Elanir and Seheult (2003). In the case of the auxiliary variable the sample second L-moment is as follows:

$$D_S(x) = \frac{1}{n(n-1)} \sum_{j=1}^{n} (n - 2j + 1) X_{(r_j)} \tag{18}$$

or

$$D_S(x) = \frac{1}{n(n-1)} \sum_{j=1}^{[n/2]} (n-2j+1) \left(X_{(r_{n-j+1})} - X_{(r_j)} \right) \tag{19}$$

This leads to the following possible particular case of the definition 2.

Definition 4. *The sampling design proportional to the value $d_s(x)$ of the statistic $D_S(x)$ is as follows:*

$$P_{r_1,\ldots,r_h}(s) = \frac{d_s(x)}{z_{r_1,\ldots,r_h}} \tag{20}$$

for $s \in G(r_1, \ldots, r_h; i_1, \ldots, i_h)$ where

$$z_{r_1,\ldots,r_h} = \binom{N}{n} E_0(D_S(x)) =$$

$$= \binom{N}{n} \left(\frac{1}{n(n-1)} \sum_{j=1}^{n} (n-2j+1) \sum_{i_j=r_j}^{N-n+r_j} x_{i_j} \binom{i_j-1}{r_j-1} \binom{N-i_j}{n-r_j} \right) \tag{21}$$

or equivalently

$$P_{r_1,\ldots,r_h}(S) = \frac{D_S(x)}{z_{r_1,\ldots,r_h}}$$

Let us consider the following statistic:

$$L\left(Q_{\alpha_1}, Q_{\alpha_2}, Q_{\alpha_3}\right) = \left(Q_{\alpha_2} - Q_{\alpha_1}\right)\left(Q_{\alpha_3} - Q_{\alpha_2}\right) \tag{22}$$

where the quantile Q_α is defined by means of the order statistic in the expression (7) and $0 < \alpha_1 < \alpha_2 < \alpha_3 < 1$.

Particularly,

$$L\left(Q_{0,25}, Q_{0,50}, Q_{0,75}\right) = \left(Q_{0,50} - Q_{0,25}\right)\left(Q_{0,75} - Q_{0,50}\right)$$

or

$$L\left(X_{(r_1)}, X_{([n/2]+1)}, X_{(r_n)}\right) = \left(X_{([n/2]+1)} - X_{(r_1)}\right)\left(X_{(r_n)} - X_{([n/2]+1)}\right)$$

The statistic $L = L\left(Q_{\alpha_1}, Q_{\alpha_2}, Q_{\alpha_3}\right)$ can be treated as spread coefficient of the auxiliary variable.

The equation (7) let us write that $X_{(r_e)} = Q_{\alpha_1}$, $X_{(r_j)} = Q_{\alpha_2}$ and $X_{(r_k)} = Q_{\alpha_3}$. So, the expression (22) can be rewritten to the following one:

$$L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right) = \left(X_{(r_j)} - X_{(r_e)}\right)\left(X_{(r_k)} - X_{(r_j)}\right) \tag{23}$$

On the basis of the statistic L we can construct the following sampling design which is the possible particular case of the Definition 1.

Definition 5. *Under the assumption that $r_e < r_j < r_k$ the sampling design proportional to the value $l\left(x_{(i_e)}, x_{(i_j)}, x_{(i_k)}\right)$ of the statistic $L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right)$ is as follows:*

$$P_{r_1,\ldots,r_h}(s) = \frac{l\left(x_{(i_e)}, x_{(i_j)}, x_{(i_k)}\right)}{z_{r_1,\ldots,r_h}} \tag{24}$$

for $s \in G(r_e, r_j, r_k; i_e, i_j, i_k)$ or equivalently

$$P_{r_1,\ldots,r_h}(S) = \frac{L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right)}{z_{r_1,\ldots,r_h}}$$

where

$$z_{r_1,\ldots,r_h} = \binom{N}{n} E_0\left(L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right)\right) \tag{25}$$

$$\begin{aligned} &E_0\left(L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right)\right) = \\ &= E_0\left(X_{(r_j)}X_{(r_k)}\right) - E_0\left(X_{(r_e)}X_{(r_k)}\right) - E_0\left(X^2_{(r_j)}\right) + E_0\left(X_{(r_e)}X_{(r_j)}\right) \end{aligned} \tag{26}$$

On the basis of the expression (4) and (5) we have:

$$E_0\left(X^2_{(r_j)}\right) = \frac{1}{\binom{N}{n}} \sum_{i_j=r_j}^{N-n+r_j} \binom{i_j - 1}{r_j - 1}\binom{N - i_j}{n - r_j} x^2_{i_j}, \tag{27}$$

$$\begin{aligned} &E_0\left(X_{(r_a)}X_{(r_b)}\right) = \\ &= \frac{1}{\binom{N}{n}} \sum_{i_a=r_a}^{N-n+r_a} \sum_{i_b=i_a+r_b-r_a}^{N-n+r_b} \binom{i_a - 1}{r_a - 1}\binom{i_b - i_a - 1}{r_b - r_a - 1}\binom{N - i_b}{n - r_b} x_{i_a} x_{i_b} \end{aligned} \tag{28}$$

for $(a, b) = (j, k), (e, k), (e, j)$.

3. Inclusion probabilities

As it is well known the inclusion probability of the first order is determined by the following equations: $\pi_k^{(r_1,\ldots,r_h)} = \sum_{\{s:k\in s\}} P(s)$ for $k = 1,\ldots,N$.

Let $\delta(x)$ be such the function that if $x \leq 0$ then $\delta(x) = 0$ otherwise $\delta(x) = 1$. Let us note that $\delta(x)\delta(x-1) = \delta(x-1)$.

Theorem 1. *Under the sampling design $P_{r_1,\ldots,r_h}(s)$ introduced by the Definition 1 the first order inclusion probabilities are as follows:*

$$\pi_k^{(r_1,\ldots,r_h)} =$$

$$= \frac{1}{z_{r_1,\ldots,r_h}}\Bigg(\delta(r_1-1)\sum_{i_1=r_1}^{N-n+r_1}\sum_{\delta(r_1-k)+\delta(k-r_1+1)\delta(N-n+r_1-k)(k+1)}^{N-n+r_2}\sum_{i_2=i_1+r_2-r_1}\cdots$$

$$\cdots\sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j}\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}}\cdots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}\frac{\binom{i_1-2}{r_1-2}}{\binom{i_1-1}{r_1-1}}g(r_1,\ldots,r_h;i_1,\ldots,i_h)w(i_1,\ldots,i_h)+$$

$$+\sum_{j=2}^{h}\delta(k-r_j)\delta(k-r_{j-1})\delta(r_j-r_{j-1}-1)\sum_{i_1=r_1}^{N-n+r_1}\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2}\cdots\sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{k-1}$$

$$\sum_{i_j=k+1}^{N-n+r_j}\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}}\cdots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}$$

$$\frac{\binom{i_j-i_{j-1}-2}{r_j-r_{j-1}-2}}{\binom{i_j-i_{j-1}-1}{r_j-r_{j-1}-1}}\delta(i_j-i_{j-1}-r_j+r_{j-1})$$

$$\delta(i_j-i_{j-1}-r_j+r_{j-1})(r_1,\ldots,r_h;i_1,\ldots,i_h)w(i_1,\ldots,i_h)+$$

$$+\delta(n-r_h)\sum_{i_1=r_1}^{N-n+r_1}\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2}\cdots\sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j}\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}}\cdots$$

$$\cdots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{\delta(N-n+r_h-k+1)\delta(k-r_h)(k-1)+\delta(k-N+n-r_h)(N-n+r_h)}$$

$$\frac{\binom{N-i_h-1}{n-r_h-1}}{\binom{N-i_h}{n-r_h}}g(r_1,\ldots,r_h;i_1,\ldots,i_h)w(i_1,\ldots,i_h)+$$

$$+\delta(N-n+r_1-k+1)\delta(k-r_1+1)\sum_{i_2=k+r_2-r_1}^{N-n+r_2}\sum_{i_3=i_2+r_3-r_2}^{N-n+r_3}\cdots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}\binom{k-1}{r_1-1}\binom{i_2-k-1}{r_2-r_1-1}\prod_{e=3}^{h+1}\binom{i_e-i_{e-1}-1}{r_e-r_{e-1}-1}w(i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_h)+$$

$$+\sum_{j=2}^{h-1}\delta(N-n+r_j-k+1)\delta(k-r_j+1)\sum_{i_1=r_1}^{N-n+r_1}\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2}\cdots$$
$$\cdots\sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}\sum_{i_{j+1}=k+r_{j+1}-r_j}^{N-n+r_{j+1}}\cdots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}$$
$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}\binom{k-i_{j-1}-1}{r_j-r_{j-1}-1}\binom{i_{j+1}-k-1}{r_{j+1}-r_j-1}\prod_{e=1}^{j-1}\binom{i_e-i_{e-1}-1}{r_e-r_{e-1}-1}$$
$$\prod_{e=j+2}^{h+1}\binom{i_e-i_{e-1}-1}{r_e-r_{e-1}-1}w(i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_h)+ \tag{29}$$
$$+\delta(N-n+r_h-k+1)\delta(k-r_h+1)\sum_{i_1=r_1}^{N-n+r_1}\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2}\cdots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}$$
$$\binom{k-i_{h-1}-1}{r_h-r_{h-1}-1}\binom{N-k}{n-r_h}\prod_{e=1}^{h-1}\binom{i_e-i_{e-1}-1}{r_e-r_{e-1}-1}w(i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_h)\Big)$$

where $i_0=r_0=0$, $r_{h+1}=n+1$, $i_{h+1}=N+1$.

The proof of the theorem is in the Appendix.

4. Sampling scheme

The construction of the sampling scheme implementing the sampling design proposed by the definition 1 is as follows. Firstly, we evaluate the values of the all possible values $w(i_1,\ldots,i_h)$. Next, we calculate the following values:

$$p_w(i_1,\ldots,i_h)=P\left(W(r_1,\ldots,r_h)=w(i_1,\ldots,i_h)\right)=\frac{w(i_1,\ldots,i_h)g(r_1,\ldots,r_h)}{z_{r_1,\ldots,r_h}} \tag{30}$$

where $r_j\le i_j\le N-n+r_j$, $j=1,\ldots,h$ and $i_j<i_{j+1}$, $j=1,\ldots,h-1$. Now the set of the auxiliary variables $(x_{i_1},\ldots,x_{i_h})$ is drawn with the probability $p_w(i_1,\ldots,i_h)$. Let us note that $(x_{i_1},\ldots,x_{i_h})$ is the value of the order statistics $(X_{(r_1)},\ldots,X_{(r_h)})$. In the next step the sequence of the samples $(s_1,\ldots,s_j,\ldots,s_{h+1})$ are drown in the following way. The sample s_j is the simple sample of the size r_j-r_{j-1} drawn from the subpopulation $U(i_{j-1},\ldots,i_j)$ where $j=1,\ldots,h+1$ and $r_0=0$, $i_0=0$, $r_{h+1}=n$, $i_{h+1}=N$. Hence, we drawn the sample $s=s_1\cup\{i_1\}s_2\cup\{i_2\}\cup\ldots\cup s_h\cup\{i_h\}s_{h+1}$. It is drawn with probability determined in the definition 1 because

$$p_w(i_1,\ldots,i_h)P(s_1)\ldots P(s_{h+1})=P_{r_1,\ldots,r_h}(s)$$

Let us note that the number of the values $w = w(i_1, \ldots, i_h)$ which we have to calculate is very large and it is equal to $\binom{N}{h}$. So, the proposed sampling design will be quite difficult to implementing in the practise because we should have very speed computers which can work with very high accuracy.

5. Some sampling strategies

The well known Horvitz-Thompson (1952) estimator $t_{HT,s} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}$ is the unbiased estimator of the population mean $\bar{y}$ if $\pi_k > 0$ for $k = 1, \ldots, N$. So, under this condition the strategy $(t_{HT,s}, P_{r_1,\ldots,r_h}(s))$ is unbiased for $\bar{y}$.

Let us remember that the order ratio estimator is: $\bar{y}_{R,s} = \frac{\bar{y}_s}{\bar{x}_s}\bar{x}$ where $\bar{y}_s = \frac{1}{n}\sum_{i\in s} y_i$, $\bar{x}_s = \frac{1}{n}\sum_{k\in s} x_k$ and $\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$.

Wywiał (2004) proposed the similar estimator defined in the following way:

$$t_s^{(r)} = \bar{y}_s \frac{E_0(X_{(r)})}{X_{(r)}} \tag{31}$$

where $E_0(X_{(r)})$ is given by the expression (6). Moreover, Wywiał (2004, 2007) proved the following.

Theorem 2. *Under the sampling design stated in the definition 1 the strategy* $\left(t_s^{(r)}, P_r(s)\right)$ *is the unbiased strategy of the population mean.*

Properties of the strategy $\left(t_s^{(r)}, P_r(s)\right)$ is analysed by Wywiał (2007a). His considerations connected with the both strategies $\left(t_s^{(r)}, P_r(s)\right)$ and $(t_{HT,s}, P_r(s))$ lead to the following conclusions. Generally, these strategies can be more precise than the simple sample means when the degree of the order statistic is large and the sample size is small. The $\left(t_s^{(r)}, P_r(s)\right)$ strategy can be preferred especially in the case when outliers (too large values) of a variable under study and an auxiliary variable exist. In this case, its precision can be even better than $(\bar{y}_{R,s}, P_0(s))$ strategy, but only for a small size of the sample and the large degree r of the order statistic. Moreover, let us note that the strategies: $\left(t_s^{(r)}, P_r(s)\right)$ and $(t_{HT,s}, P_r(s))$ do not depend on the shortest or largest values of the auxiliary variable. Hence,

they are useful when there are right or left censored observations of the auxiliary variable.

Similarly, to the construction of the estimator we can define the following one:

$$\bar{y}_{R,S}^{(r_1,\ldots,r_h)} = \bar{y}_s \frac{E_0(W(r_1,\ldots,r_h))}{W(r_1,\ldots,r_h)} \tag{32}$$

where $E_0(W(r_1,\ldots,r_h))$ is given by the expression (9). The Theorem 2 can be generalized into the following one.

Theorem 3. *Under the sampling design stated in the definition 1 the strategy* $\left(\bar{y}_{R,S}^{(r_1,\ldots,r_h)}, P_{r_1,\ldots,r_2}(s)\right)$ *is the unbiased strategy of the population mean.*

The proof of the above is in the Appendix.

It seems that not all particular cases of the statistics $\bar{y}_{R,S}^{(r_1,\ldots,r_h)}$ is sensible. The following propositions seems to be interesting.

$$\bar{y}_{R,S}^{(r_j,r_k)} = \bar{y}_s \frac{E_0(X_{(r_k)}) - E_0(X_{(r_j)})}{X_{(r_k)} - X_{(r_j)}} \tag{33}$$

where from the expression (17) we have:

$$\begin{aligned} &E_0\left(X_{(r_2)}\right) - E_0\left(X_{(r_1)}\right) = \\ &= \frac{1}{\binom{N}{n}} \sum_{i_2=r_2}^{N-n+r_2} \binom{i_2-1}{r_2-1}\binom{N-i_2}{n-r_2} x_{i_2} - \sum_{i_1=r_1}^{N-n+r_1} \binom{i_1-1}{r_1-1}\binom{N-i_1}{n-r_1} x_{i_1} \end{aligned} \tag{34}$$

Particularly, we can assume that $X_{(r_k)} = Q_{0,75}$ and $X_{(r_j)} = Q_{0,25}$ or $X_{(r_k)} = X_{(r_n)}$ and $X_{(r_j)} = X_{(r_1)}$. So, in this case the estimator depend on the sample quantile range else sample range of the auxiliary variable.

The next proposition is as follows:

$$\bar{y}_{R,S}(D(x)) = \bar{y}_s \frac{E_0(D_S(x))}{D_S(x)} \tag{35}$$

where the estimator of the standard deviation of the auxiliary variable $D_S(x)$ is given by the expression (18) or (19) and

$$E_0(D_S(x)) = \frac{1}{\binom{N}{n}} \left(\frac{1}{n(n-1)} \sum_{j=1}^{n} (n-2j+1) \sum_{i_j=r_j}^{N-n+r_j} x_{i_j} \binom{i_j-1}{r_j-1}\binom{N-i_j}{n-r_j} \right) \tag{36}$$

The defined by the expressions (22) or (23) the coefficient of the auxiliary variable spread can by used in the following way:

$$\bar{y}_{R,S}^{(r_j,r_k)} = \bar{y}_s \frac{E_0\left(L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right)\right)}{L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right)} \tag{37}$$

where $L\left(X_{(r_e)}, X_{(r_j)}, X_{(r_k)}\right)$ is given by the expressions (26), (27), (28).

The defined estimators depends on some sample coefficients of spread of the auxiliary variable. So, they can be useful when values of the variable under study are dependent on spread of the auxiliary variable. This case is similar to that dealing with ratio estimator dependent on sample variance of the auxiliary variable considered e.g. by Das and Tripathi (1980) or Srivastava and Jhajj (1981).

Finally, let us note that Wywiał (2004) proposed sampling strategy dependent on some regression estimator and the sampling design defined by the expression (16).

The considered strategies are rather complicated and that is why analysis of their accuracy could be rather difficult. So, we will try to it on the basis of computer simulation, but in a separate paper.

Conclusions

The sampling design belonging to the class of the sampling designs dependent on the sample parameters of an auxiliary variable has been proposed. It is proportional to the non-negative function of order statistic of an auxiliary variable. It has several particular variants. The inclusion probabilities of the first and second degrees were derived. The sampling scheme implementing the sampling design has been constructed, too.

The new version of the ratio estimator has been proposed, too. It is an unbiased estimator of the population mean. Several, particular version of the ratio estimator has been considered.

Finally, let us note that the strategies dependent on quantiles do not depend on the values which are between two neighboring quantiles of the auxiliary variable.

Hence, they can be useful when there are censored observations of the auxiliary variable.

It seems that without an additional large analysis it is not possible to determine precisely how the sampling strategies depend on the parameters of the conditional sampling design as well as on the joint distribution of a variable under study and an auxiliary variable. It seems that in the next paper a computer simulation will be useful in this case.

Appendix

1. Derivation of the expression 2

A subpopulation of the population $U = (1, \ldots, N)$ will be denoted by $U(i_{j-1}+1, \ldots, i_j - 1)$. It is of size $i_j - i_{j-1} - 1$. Let $\mathbf{S}(U(i_{j-1}+1, \ldots, i_j - 1), s_j)$ be the space of a simple sample s_j of size $r_j - r_{j-1} - 1$ drawn without replacement from the subpopulation $U(i_{j-1}+1, \ldots, i_j - 1)$. This lead to the following:

$$\begin{aligned} G(r_1, \ldots, r_h; i_1, \ldots, i_h) = {} & \mathbf{S}(U(1, \ldots, i_1 - 1), s_1) \times \{i_1\} \times \\ & \times \mathbf{S}(U(i_1 + 1, \ldots, i_2 - 1), s_2) \times \{i_2\} \times \ldots \\ & \times \{i_{j-1}\} \times \mathbf{S}(U(i_{j-1} + 1, \ldots, i_j - 1), s_j) \times \{i_j\} \times \\ & \times \mathbf{S}(U(i_j + 1, \ldots, i_{j+1} - 1), s_{j+1}) \times \{i_{j+1}\} \times \ldots \\ & \times \{i_{h-1}\} \times \mathbf{S}(U(i_{h-1} + 1, \ldots, i_h - 1), s_h) \times \{i_h\} \times \\ & \times \mathbf{S}(U(i_h + 1, \ldots, N), s_{h+1}) \end{aligned}$$

or

$$G(r_1, \ldots, r_h; i_1, \ldots, i_h) = \bigcap_{i=1}^{h+1} \times \mathbf{S}(U(i_{j-1} + 1, \ldots, i_j - 1), s_j) \times \{i_j\}$$

where $\times$ is the symbol of the Kartezian product. Now the equation (1) can be rewritten in the following way:

$$\bigcup_{i_1=r_1}^{N-n+r_1} \mathbf{S}(U(1,\ldots,i_1-1),s_1)\times\{i_1\}\times$$
$$\bigcup_{i_2=i_1+r_2-r_1}^{N-n+r_2} \mathbf{S}(U(i_1+1,\ldots,i_2-1),s_2)\times\{i_2\}\times\ldots$$
$$\ldots \bigcup_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} \mathbf{S}(U(i_{j-2}+1,\ldots,i_{j-1}-1),s_{j-1})\times\{i_{j-1}\}\times$$
$$\bigcup_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} \mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1),s_j)\times\{i_j\}\times \tag{38}$$
$$\bigcup_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} \mathbf{S}(U(i_j+1,\ldots,i_{j+1}-1),s_{j+1})\times\{i_{j+1}\}\times\ldots$$
$$\ldots \bigcup_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \times\mathbf{S}(U(i_{h-2}+1,\ldots,i_{h-1}-1),s_{h-1})\times\{i_{h-1}\}\times$$
$$\bigcup_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \times\mathbf{S}(U(i_{h-1}+1,\ldots,i_h-1),s_h)\times\{i_h\}\times$$
$$\times\,\mathbf{S}(U(i_h+1,\ldots,N),s_{h+1})=\mathbf{S}$$

Let $g(r_1,\ldots,r_h;i_1,\ldots,i_h)$ be the size of the set $G(r_1,\ldots,r_h;i_1,\ldots,i_h)$. So,

$$g(r_1,\ldots,r_h;i_1,\ldots,i_h)=Card(G(r_1,\ldots,r_h;i_1,\ldots,i_h))=$$
$$=Card(\mathbf{S}(U(1,\ldots,i_1-1),s_1))Card(\{i_1\})\ldots Card(\{i_{j-1}\})\cdot$$
$$\cdot Card(\mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1),s_j))Card(\{i_j\})\ldots Card(\{i_h\})\cdot$$
$$\cdot Card(\mathbf{S}(U(i_h+1,\ldots,N),s_h))=Card(\mathbf{S}(U(1,\ldots,i_1-1),s_1))\ldots$$
$$\ldots Card(\mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1),s_j))\ldots Card(\mathbf{S}(U(i_h+1,\ldots,N),s_h))=$$
$$=\binom{i_1-1}{r_1-1}\ldots\binom{i_j-i_{j-1}-1}{r_j-r_{j-1}-1}\ldots\binom{N-i_h}{n-r_h}.$$

Hence:

$$g(r_1,\ldots,r_h;i_1,\ldots,i_h)=\prod_{j=1}^{h+1}Card(\mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1),s_j))Card(\{i_j\})=$$
$$=\prod_{j=1}^{h+1}Card(\mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1),s_j))$$

or

$$g_h = g(r_1, \ldots, r_h; i_1, \ldots, i_h) = \prod_{j=1}^{h+1} \binom{i_j - i_{j-1} - 1}{r_j - r_{j-1} - 1}$$

where $r_0 = i_0 = 0$, $r_{h+1} = n+1$, $i_{h+1} = N+1$.

Moreover, $\sum_{\{i_1,\ldots,i_h\}} g(r_1, \ldots, r_h; i_1, \ldots, i_h) = \binom{N}{n}$ or more precisely:

$$\sum_{i_1=r_1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} \cdots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} g(r_1, \ldots, r_h; i_1, \ldots, i_h) = \binom{N}{n}$$

2. Proof of the theorem 1

In order to simplify the derivations we assume that $u_i = r_i, v_i = N - n + r_i$ for $i = 1, \ldots, h$. On the basis of the equation (38) we have:

$$\begin{aligned}
\pi_k^{(r_1,\ldots,r_h)} &= P_{r_1,\ldots,r_h}(s : k \in s) = \\
&= P_{r_1,\ldots,r_h}\left(s : k \in \left(s_{h+1} \cup \bigcup_{j=1}^{h}(s_j \cup (X_{(r_j)} = x_{i_j}))\right)\right) = \\
&= P_{r_1,\ldots,r_h}\left(s : (k \in s_{h+1}) \cup \bigcup_{j=1}^{h}((k \in s_j) \cup (X_{(r_j)} = x_k))\right) = \\
&= \sum_{j=1}^{h+1} P_{r_1,\ldots,r_h}(s : k \in s_j) + \sum_{j=1}^{h} P_{r_1,\ldots,r_h}(s : X_{(r_j)} = x_k) = \\
&= \sum_{j=1}^{h+1} \sum_{\{s:k\in s_j\}} P_{r_1,\ldots,r_h}(s) + \sum_{j=1}^{h} \sum_{\{s:X_{(r_j)}=x_k\}} P_{r_1,\ldots,r_h}(s) = \\
&= \sum_{\{s:k\in s_1, k<r_1\}} P_{r_1,\ldots,r_h}(s) + \sum_{\{s:k\in s_1, k\geq r_1\}} P_{r_1,\ldots,r_h}(s) + \sum_{j=2}^{h} \sum_{\{s:k\in s_j\}} P_{r_1,\ldots,r_h}(s) + \\
&+ \sum_{\{s:k\in s_{h+1}, k\leq N-n+r_h\}} P_{r_1,\ldots,r_h}(s) + \sum_{\{s:k\in s_{h+1}, k>N-n+r_h\}} P_{r_1,\ldots,r_h}(s) + \\
&+ \sum_{j=1}^{h} \sum_{\{s:X_{(r_j)}=x_k\}} P_{r_1,\ldots,r_h}(s)
\end{aligned} \tag{39}$$

$$P_{r_1,\ldots,r_h}(s : k \in s_1, k < r_1) = \sum_{\{s:k\in s_1, k<r_1\}} P_{r_1,\ldots,r_h}(s) =$$

$$= \frac{\delta(r_1-k)\delta(r_1-1)}{z_{r_1,\ldots,r_h}} Card\Bigg(\bigcup_{i_1=r_1}^{N-n+r_1} \mathbf{S}(U(1,\ldots,i_1-1)-\{k\}, s_1-\{k\}) \times \{i_1\} \times$$

$$\bigcup_{i_2=i_1+r_2-r_1}^{N-n+r_2} \mathbf{S}(U(i_1+1,\ldots,i_2-1), s_2) \times \{i_2\} \times \ldots$$

$$\ldots \bigcup_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} \mathbf{S}(U(i_{j-2}+1,\ldots,i_{j-1}-1), s_{j-1}) \times \{i_{j-1}\} \times$$

$$\bigcup_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} \mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1), s_j) \times \{i_j\} \times$$

$$\bigcup_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} \mathbf{S}(U(i_j+1,\ldots,i_{j+1}-1), s_{j+1}) \times \{i_{j+1}\} \times \ldots$$

$$\ldots \bigcup_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \times \mathbf{S}(U(i_{h-2}+1,\ldots,i_{h-1}-1), s_{h-1}) \times \{i_{h-1}\} \times$$

$$\bigcup_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \times \mathbf{S}(U(i_{h-1}+1,\ldots,i_h-1), s_h) \times \{i_h\} \times$$

$$\times \mathbf{S}(U(i_h+1,\ldots,N), s_{h+1}) \Bigg) w(i_1,\ldots,i_h) =$$

$$= \frac{\delta(r_1-k)\delta(r_1-1)}{z_{r_1,\ldots,r_h}} \sum_{i_1=r_1}^{N-n+r_1} Card(\mathbf{S}(U(1,\ldots,i_1-1)-\{k\}, s_1-\{k\}) \times \{i_1\})$$

$$\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} Card(\mathbf{S}(U(i_1+1,\ldots,i_2-1), s_2) \times \{i_2\}) \ldots$$

$$\ldots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} Card(\mathbf{S}(U(i_{j-2}+1,\ldots,i_{j-1}-1), s_{j-1}) \times \{i_{j-1}\})$$

$$\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} Card(\mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1), s_j) \times \{i_j\})$$

$$\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} Card(\mathbf{S}(U(i_j+1,\ldots,i_{j+1}-1), s_{j+1}) \times \{i_{j+1}\}) \ldots$$

$$\dots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} Card(\mathbf{S}(U(i_{h-2}+1,\dots,i_{h-1}-1),s_{h-1})\times\{i_{h-1}\})$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} Card(\mathbf{S}(U(i_{h-1}+1,\dots,i_h-1),s_h)\times\{i_h\}\times$$

$$\mathbf{S}(U(i_h+1,\dots,N),s_{h+1})w(i_1,\dots,i_h)=$$

$$=\frac{\delta(r_1-k)\delta(r_1-1)}{z_{r_1,\dots,r_h}}\sum_{i_1=r_1}^{N-n+r_1} Card(\mathbf{S}(U(1,\dots,i_1-1)-\{k\},s_1-\{k\}))$$

$$\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} Card(\mathbf{S}(U(i_1+1,\dots,i_2-1),s_2))\dots$$

$$\dots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} Card(\mathbf{S}(U(i_{j-2}+1,\dots,i_{j-1}-1),s_{j-1}))$$

$$\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} Card(\mathbf{S}(U(i_{j-1}+1,\dots,i_j-1),s_j))$$

$$\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} Card(\mathbf{S}(U(i_j+1,\dots,i_{j+1}-1),s_{j+1}))\dots$$

$$\dots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} Card(\mathbf{S}(U(i_{h-2}+1,\dots,i_{h-1}-1),s_{h-1}))$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} Card(\mathbf{S}(U(i_{h-1}+1,\dots,i_h-1),s_h))$$

$$Card(\mathbf{S}(U(i_h+1,\dots,N),s_{h+1}))w(i_1,\dots,i_h)=$$

$$=\frac{\delta(r_1-k)\delta(r_1-1)}{z_{r_1,\dots,r_h}}\sum_{i_1=r_1}^{N-n+r_1}\binom{i_1-2}{r_1-2}\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2}\binom{i_2-i_1-1}{r_2-r_1-1}\dots$$

$$\dots\sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}\binom{i_{j-1}-i_{j-2}-1}{r_{j-1}-r_{j-2}-1}\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j}\binom{i_j-i_{j-1}-1}{r_j-r_{j-1}-1}$$

$$\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}}\binom{i_{j+1}-i_j-1}{r_{j+1}-r_j-1}\dots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}\binom{i_{h-1}-i_{h-2}-1}{r_{h-1}-r_{h-2}-1}$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}\binom{i_h-i_{h-1}-1}{r_h-r_{h-1}-1}\binom{N-i_h}{n-r_h}w(i_1,\dots,i_h)=$$

$$= \frac{\delta(r_1-k)\delta(r_1-1)}{z_{r_1,\ldots,r_h}} \sum_{i_1=r_1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}$$

$$\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} \sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} \cdots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}$$

$$\binom{i_1-2}{r_1-2}\binom{i_2-i_1-1}{r_2-r_1-1}\binom{i_{j-1}-i_{j-2}-1}{r_{j-1}-r_{j-2}-1}\binom{i_j-i_{j-1}-1}{r_j-r_{j-1}-1}$$

$$\binom{i_{j+1}-i_j-1}{r_{j+1}-r_j-1}\cdots\binom{i_{h-1}-i_{h-2}-1}{r_{h-1}-r_{h-2}-1}\binom{i_h-i_{h-1}-1}{r_h-r_{h-1}-1}\binom{N-i_h}{n-r_h} w(i_1,\ldots,i_h) =$$

$$= \frac{\delta(r_1-k)\delta(r_1-1)}{z_{r_1,\ldots,r_h}} \sum_{i_1=r_1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}$$

$$\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} \sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} \cdots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}$$

$$\frac{\binom{i_1-2}{r_1-2}}{\binom{i_1-1}{r_1-1}}\binom{i_1-1}{r_1-1}\binom{i_2-i_1-1}{r_2-r_1-1}\cdots\binom{i_{j-1}-i_{j-2}-1}{r_{j-1}-r_{j-2}-1}\binom{i_j-i_{j-1}-1}{r_j-r_{j-1}-1}$$

$$\binom{i_{j+1}-i_j-1}{r_{j+1}-r_j-1}\cdots\binom{i_{h-1}-i_{h-2}-1}{r_{h-1}-r_{h-2}-1}\binom{i_h-i_{h-1}-1}{r_h-r_{h-1}-1}\binom{N-i_h}{n-r_h} w(i_1,\ldots,i_h)$$

This result and the expression (2) lead to the following one:

$$P_{r_1,\ldots,r_h}(s: k\in s_1, k<r_1) = \sum_{\{s:k\in s_1,k<r_1\}} P_{r_1,\ldots,r_h}(s) =$$

$$= \frac{\delta(r_1-k)\delta(r_1-1)}{z_{r_1,\ldots,r_h}} \sum_{i_1=r_1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} \sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j}$$

$$\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} \cdots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \frac{\binom{i_1-2}{r_1-2}}{\binom{i_1-1}{r_1-1}}$$

$$g(r_1,\ldots,r_h;i_1,\ldots,i_h)w(i_1,\ldots,i_h) \tag{40}$$

Next derivation is as the follows:

$$P_{r_1,\ldots,r_h}(s:k\in s_1,k\geq r_1)=\sum_{\{s:k\in s_1,k\geq r_1\}}P_{r_1,\ldots,r_h}(s)=\frac{\delta(k-r_1+1)\delta(r_1-1)}{z_{r_1,\ldots,r_h}}$$

$$Card\Bigg(\bigcup_{i_1=k+1}^{N-n+r_1}\mathbf{S}(U(1,\ldots,i_1-1)-\{k\},s_1-\{k\})\times\{i_1\}\times$$

$$\bigcup_{i_2=i_1+r_2-r_1}^{N-n+r_2}\mathbf{S}(U(i_1+1,\ldots,i_2-1),s_2)\times\{i_2\}\times\ldots$$

$$\ldots\bigcup_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}\mathbf{S}(U(i_{j-2}+1,\ldots,i_{j-1}-1),s_{j-1})\times\{i_{j-1}\}\times$$

$$\bigcup_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j}\mathbf{S}(U(i_{j-1}+1,\ldots,i_j-1),s_j)\times\{i_j\}\times$$

$$\bigcup_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}}\mathbf{S}(U(i_j+1,\ldots,i_{j+1}-1),s_{j+1})\times\{i_{j+1}\}\times\ldots$$

$$\ldots\bigcup_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}\times\mathbf{S}(U(i_{h-2}+1,\ldots,i_{h-1}-1),s_{h-1})\times\{i_{h-1}\}\times$$

$$\bigcup_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}\times\mathbf{S}(U(i_{h-1}+1,\ldots,i_h-1),s_h)\times\{i_h\}\times$$

$$\times\,\mathbf{S}(U(i_h+1,\ldots,N),s_{h+1})\Bigg)w(i_1,\ldots,i_h)$$

Now after similar operation as it was in the case of evaluation of the equation (40) we obtain the following one:

$$P_{r_1,\ldots,r_h}(s:k\in s_1,k\geq r_1)=\sum_{\{s:k\in s_1,k\geq r_1\}}P_{r_1,\ldots,r_h}(s)=$$

$$=\frac{\delta(k-r_1+1)\delta(r_1-1)\delta(N-n+r_1-k)}{z_{r_1,\ldots,r_h}}\sum_{i_1=k+1}^{N-n+r_1}\sum_{i_2=i_1+r_2-r_1}^{N-n+r_2}\ldots$$

$$\ldots\sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j}\sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}}\ldots\sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}\frac{\binom{i_1-2}{r_1-2}}{\binom{i_1-1}{r_1-1}}g(r_1,\ldots,r_h;i_1,\ldots,i_h)w(i_1,\ldots,i_h)\tag{41}$$

Hence, the results (40) and (41) can be simultaneously expressed in the following way.

$$P_{r_1,\ldots,r_h}(s : k \in s_1) = \sum_{\{s:k\in s_1\}} P_{r_1,\ldots,r_h}(s) = \frac{\delta(r_1-1)}{z_{r_1,\ldots,r_h}}$$

$$\sum_{i_1=r_1\delta(r_1-k)+\delta(k-r_1+1)\delta(N-n+r_1-k)(k+1)}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}}$$

$$\sum_{i_j=i_{j-1}+r_j-r_{j-1}}^{N-n+r_j} \sum_{i_{j+1}=i_j+r_{j+1}-r_j}^{N-n+r_{j+1}} \cdots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}$$

$$\frac{\binom{i_1-2}{r_1-2}}{\binom{i_1-1}{r_1-1}} g(r_1,\ldots,r_h;i_1,\ldots,i_h)w(i_1,\ldots,i_h) \tag{42}$$

The similar derivations lead to the expressions $P_{r_1,\ldots,r_h}(s : k \in s_j)$, $j = 2,\ldots,h+1$.

Finally, we make the following derivation:

$$P_{r_1,\ldots,r_h}(s : X_{(r_j)} = x_k) =$$

$$= \sum_{\{s:X_{(r_j)}=x_k\}} P_{r_1,\ldots,r_h}(s) = \frac{\delta(N-n+r_j-k+1)\delta(k-r_j+1)}{z_{r_1,\ldots,r_h}}$$

$$Card\Big(\bigcup_{i_1=r_1}^{N-n+r_1} \mathbf{S}(U(1,\ldots,i_1-1),s_1)\times$$

$$\times\{i_1\}\times \bigcup_{i_2=i_1+r_2-r_1}^{N-n+r_2} \mathbf{S}(U(i_1+1,\ldots,i_2-1),s_2)\times\{i_2\}\times\ldots$$

$$\ldots \bigcup_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} \mathbf{S}(U(i_{j-2}+1,\ldots,i_{j-1}-1),s_{j-1})\times\{i_{j-1}\}\times$$

$$\bigcup_{i_j=k}^{k} \mathbf{S}(U(i_{j-1}+1,\ldots,k-1),s_j)\times\{k\}\times$$

$$\bigcup_{i_{j+1}=k+r_{j+1}-r_j}^{N-n+r_{j+1}} \mathbf{S}(U(k+1,\ldots,i_{j+1}-1),s_{j+1})\times\{i_{j+1}\}\times\ldots$$

$$\cdots \bigcup_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}} \times \mathbf{S}(U(i_{h-2}+1,\ldots,i_{h-1}-1),s_{h-1}) \times \{i_{h-1}\} \times$$

$$\bigcup_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \times \mathbf{S}(U(i_{h-1}+1,\ldots,i_h-1),s_h) \times \{i_h\} \times$$

$$\times\, \mathbf{S}(U(i_h+1,\ldots,N),s_{h+1})\Big) t(i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_h)$$

The above expression and similar transformations to those developed during derivation of the equation (40) lead to the following one:

$$P_{r_1,\ldots,r_h}(s:X_{(r_j)}=x_k) = \sum_{\{s:X_{(r_j)}=x_k\}} P_{r_1,\ldots,r_h}(s) =$$

$$= \frac{\delta(N-n+r_j-k+1)\delta(k-r_j+1)}{z_{r_1,\ldots,r_h}} \sum_{i_1=r_1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots$$

$$\cdots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} \sum_{i_j=k}^{k} \sum_{i_{j+1}=k+r_{j+1}-r_j}^{N-n+r_{j+1}} \cdots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} g(r_1,\ldots,r_h;i_1,\ldots,i_h) t(i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_h)$$

or

$$P_{r_1,\ldots,r_h}(s:X_{(r_j)}=x_k) = \sum_{\{s:X_{(r_j)}=x_k\}} P_{r_1,\ldots,r_h}(s) =$$

$$= \frac{\delta(N-n+r_j-k+1)\delta(k-r_j+1)}{z_{r_1,\ldots,r_h}} \sum_{i_1=r_1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots$$

$$\cdots \sum_{i_{j-1}=i_{j-2}+r_{j-1}-r_{j-2}}^{N-n+r_{j-1}} \sum_{i_{j+1}=k+r_{j+1}-r_j}^{N-n+r_{j+1}} \cdots \sum_{i_{h-1}=i_{h-2}+r_{h-1}-r_{h-2}}^{N-n+r_{h-1}}$$

$$\sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \binom{k-i_{j-1}-1}{r_j-r_{j-1}-1} \binom{i_{j+1}-k-1}{r_{j+1}-r_j-1} \prod_{e=1}^{j-1} \binom{i_e-i_{e-1}-1}{r_e-r_{e-1}-1}$$

$$\prod_{e=j+2}^{h+1} \binom{i_e-i_{e-1}-1}{r_e-r_{e-1}-1} t(i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_h)$$

The above derivations compleat the proof of the theorem 1.

3. Proof of the theorem 3

The expressions (8), (9), and (10) lead to the following:

$$E\left(\bar{y}_{R,S}^{(r_1,\ldots,r_h)}\right) =$$

$$\sum_{i_1=1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \sum_{s\in G(r_1,\ldots,r_h;i_1,\ldots,i_h)} \bar{y}_s \frac{E_0\left(W(r_1,\ldots,r_h)\right)}{w(r_1,\ldots,r_h)} P_{r_1,\ldots,r_h}(s) = \sum_{i_1=1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h}$$

$$\sum_{s\in G(r_1,\ldots,r_h;i_1,\ldots,i_h)} \bar{y}_s \frac{E_0\left(W(r_1,\ldots,r_h)\right)}{w(r_1,\ldots,r_h)} \frac{w(r_1,\ldots,r_h)}{\binom{N}{n} E_0\left(W(r_1,\ldots,r_h)\right)} =$$

$$= \sum_{i_1=1}^{N-n+r_1} \sum_{i_2=i_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{i_h=i_{h-1}+r_h-r_{h-1}}^{N-n+r_h} \sum_{s\in G(r_1,\ldots,r_h;i_1,\ldots,i_h)} \bar{y}_s =$$

$$\sum_{s\in \mathbf{S}} \bar{y}_s = E_0(\bar{y}_s) = \bar{y}$$

References

Das A.K., Tripathi T.P. (1980): Sampling Strategies for Population Mean When the Coefficient of Variation of Auxiliary Characer is Known. "Sankhya", Vol. C42.

Elamir E.A.H., Seheult A.H. (2003): Trimmed L-moments. "Computational Statistics and Data Analysis", Vol. 43, pp. 299-314.

Fisz M. (1963): Probability Theory and Mathematical Statistics. Wiley and Sons Inc., New York.

Guenther W. (1975): The Inverse Hypergeometric – a Useful Model. "Statistica Neerlandica", Vol. 29, pp. 129-144.

Hogg R.V., Craig A.T. (1970): Introduction to Mathematical Statistics. 3rd edition. MacMillian, New York.

Horvitz D.G., Thompson D.J. (1952): A Generalization of the Sampling without Replacement from Finite Universe. "Journal of the American Statistical Association", Vol. 47, pp. 663-685.

Johnson N., Kotz S. (1969): Distributions in Statistics: Discrete Distributions. Houghton Mifflin Company, Boston.

Lahiri G.W. (1951): A Method for Sample Selection Providing Unbiased Ratio Estimator. "Bulletin of

the International Statistical Institute", Vol. 33, pp. 133-140.

Patil G.G., Joshi S.W. (1968): A Dictionary and Bibliography of Discrete Distributions. Hafner, New York.

Sillito G.P. (1969): Derivation of Approximations to the Inverse Distribution Function of a Continuous Univariate Population from the Order Statistics of a Sample. "Biometrika", Vol. 56, pp. 641-650.

Singh P., Srivastava A.K. (1980): Sampling Schemes Providing Unbiased Regression Estimators. "Biometrika", Vol. 67, 1, pp. 205-209.

Srivastava S.K., Jhajj H.S. (1981): A Class of Estimators of the Population Mean in Survey Sampling Using Auxiliary Information. "Biometrika", Vol. 68.

Tillé Y. (1998): Estimation in Surveys Using Conditional Inlusion Probabilities: Simple Random Sampling. "International Statistical Review", Vol. 66, 3, pp. 303-322.

Wilks S.S. (1962): Mathematical Statistics. John Wiley and Sons, Inc., New York-London.

Wywiał J. (1999): Sampling Designs Dependent on the Sample Generalized Variance of Auxiliary Variables. "Journal of the Indian Statistical Association", Vol. 37, pp. 73-87.

Wywiał J. (2000): On Precision of Horvitz-Thompson Strategies. "Statistics in Transition", Vol. 4, No. 5, pp. 779-798.

Wywiał J. (2003). Some Contributions to Multivariate Methods in Survey Sampling. Karol Adamiecki University of Economics, Katowice.

Wywiał J. (2003a): On Conditional Sampling Strategies. "Statistical Papers", Vol. 44, 3, pp. 397-419.

Wywiał J. (2004): Quantile Regression Sampling Strategy. In: Metoda Reprezentacyjna w Badaniach Ekonomiczno-Społecznych. Red. J. Wywiał. AE, Katowice, pp. 32-42.

Wywiał J. (2007): Sampling Design Proportional to Order Statistic of Auxiliary Variable. "Statistical Papers" April, Vol. 49, No. 2, pp. 277-289.

Wywiał J. (2007a): Simulation Analysis of Accuracy Estimation of Population Mean on the Basis of Strategy Dependent on Sampling Design Proportionate to the Order Statistic of an Auxiliary Variable. "Statistics in Transition-New series", Vol. 8, No. 1, pp. 125-137.

Abstract

The sampling design proportional to some positive function of an auxiliary variable is considered. Its characteristics are derived on the basis of well known combinatoric definitions and theorems. For instance, it is well known the sampling design proportional to the sample mean, see Lahiri (1951). The sampling design proportional to the value of an order statistic of an auxiliary variable was prepared by Wywiał (2004, 2007). In this paper that sampling design is generalized in the following way. The sampling design proportional to the positive function of order statistics of the auxiliary variable is defined and its basic properties are considered. Its inclusion probabilities are derived. This let to use the Horvitz-Thompson statistic to estimation population mean value of an variable under study. The sampling scheme implementing the sampling design is proposed, too. Particular cases of the proposed sampling design are as follows. The sampling design proportional to the sample second L-statistic of the auxiliary variable. Sampling design proportional to the sample range of the auxiliary variable. It is useful to construction sampling strategy using estimators of the regression coefficients based on order statistics of the auxiliary variable.

Tomasz Żądło

ON PREDICTION OF TOTALS FOR DOMAINS DEFINED BY RANDOM ATTRIBUTES

1. Motivating example

In many countries during referendum voters present their opinion on more than one problem. Hence, one voter may be proponent of several issues at the same time. Let suppose that the purpose of some sample survey is to estimate total value of income of proponents of certain issue. In this case one population element (one voter) may belong to many such defined domains (may be proponent of many issues at the same time). In the paper it is assumed that population elements belong to such defined domains at random. In the discussed example this can be explained by the fact that people are uncertain about their final voting decision. Moreover, in a opinion poll sampled voters may be asked to present their support in percents. In this case the probabilities that one element belong to different domains are known for all of sampled population elements. On the other case, voters may declare their support as 0-1 variable in the sample survey. In this case probabilities that one element belong to different domains may be estimated based on sampled data (some auxiliary variables may be used in this case too). It is important to note that some voters may vote "no" for all of questions.

2. Basic notations

Let us consider population Ω of size N and a sample s of size n drawn from the population. The set of nonsampled elements of the population is denoted by Ω_r and its size by $N_r = N - n$.

In the classic approach population is divided into D domains denoted by Ω_d, the dth of size N_d (where $d = 1, \ldots, D$) such that:

$$\bigcup_{d=1}^{D} \Omega_d = \Omega \tag{1}$$

$$\Omega_d \cap \Omega_{d'} = \emptyset \tag{2}$$

for $d \neq d'$.

In this paper we consider some random sets (random domains) $\Omega_d \subset \Omega$, where $d = 1, \ldots, D$, but we do not assume (1) and (2) (as in the motivating example).

3. General linear mixed model

In the model approach in survey sampling when we use the general linear mixed model it is assumed for a random vector $\mathbf{Y} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_N \end{bmatrix}^T$ that (e.g. Rao 2003):

$$\begin{cases} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \\ E_\xi(\mathbf{e}) = \mathbf{0} \\ E_\xi(\mathbf{v}) = \mathbf{0} \\ D^2_\xi \begin{bmatrix} \mathbf{v} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \end{cases} \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known $N \times p$ and $N \times h$ matrices, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and random vectors $\mathbf{v}$ and $\mathbf{e}$ are $h \times 1$ and $N \times 1$, respectively. If the population elements are rearranged so that the first n elements of $\mathbf{Y}$ are those in the sample, and the first n rows of $\mathbf{X}$ and $\mathbf{Z}$ are for units in the sample, then $\mathbf{Y}$,

$\mathbf{e}$, $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{R}$ and $D^2_\xi(\mathbf{Y}) = \mathbf{V}$, can be expressed as $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_s \\ \mathbf{Y}_r \end{bmatrix}$, $\mathbf{e} = \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_s & \\ & \mathbf{Z}_s \end{bmatrix}$, $\mathbf{R} = \begin{bmatrix} \mathbf{R}_{ss} & \mathbf{R}_{sr} \\ \mathbf{R}_{rs} & \mathbf{R}_{rr} \end{bmatrix}$ $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$, where $\mathbf{Y}_s$ and $\mathbf{e}_s$ are $n \times 1$, $\mathbf{Y}_r$ and $\mathbf{e}_r$ are $N_r \times 1$, $\mathbf{X}_s$ is $n \times p$, $\mathbf{X}_r$ is $N_r \times p$, $\mathbf{Z}_s$ is $n \times h$, $\mathbf{Z}_s$ is $N_r \times h$, $\mathbf{R}_{ss}$ is $n \times n$, $\mathbf{R}_{rr}$ is $N_r \times N_r$, $\mathbf{R}_{sr}$ is $n \times N_r$ and $\mathbf{R}_{rs} = \mathbf{R}^T_{sr}$, $\mathbf{V}_{ss}$ is $n \times n$, $\mathbf{V}_{rr}$ is $N_r \times N_r$, $\mathbf{V}_{sr}$ is $n \times N_r$, and $\mathbf{V}_{rs} = \mathbf{V}^T_{sr}$. We assume that $\mathbf{V}$ is positive definite.

Under (3) we can express variance-covariance matrix of $\mathbf{Y}$ as:

$$D^2_\xi(\mathbf{Y}) = \mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T \tag{4}$$

and variance-covariance matrix of $\mathbf{Y}_s$ as:

$$D^2_\xi(\mathbf{Y}_s) = \mathbf{V}_{ss} = \mathbf{R}_{ss} + \mathbf{Z}_s\mathbf{G}\mathbf{Z}^T_s \tag{5}$$

Matrices $\mathbf{R}$ and $\mathbf{G}$ (and hence matrix $\mathbf{V}$) may depend on some variance parameters.

The equations of the BLUP and its ξ-MSE are presented by Henderson (1950).

Theorem 1 (Henderson 1950). *Assume that the population data obey the GLMM (see equation (3)). Among linear, model-unbiased predictors $\hat{\theta}^s = \mathbf{a}^T\mathbf{Y}_s + b$ of linear combination of β and the realization of $\mathbf{v}$ given by $\theta^s = \mathbf{l}^T\beta + \mathbf{m}^T\mathbf{v}$ (for specified vectors, $\mathbf{l}$ and $\mathbf{m}$, of constants) the MSE is minimized by:*

$$\hat{\theta}^s_{BLU} = \mathbf{l}^T\hat{\beta} + \mathbf{m}^T\hat{\mathbf{v}} \tag{6}$$

where

$$\hat{\beta} = \left(\mathbf{X}^T_s\mathbf{V}^{-1}_{ss}\mathbf{X}_s\right)^{-1}\mathbf{X}^T_s\mathbf{V}^{-1}_{ss}\mathbf{Y}_s \tag{7}$$

$$\hat{\mathbf{v}} = \mathbf{G}\mathbf{Z}^T_s\mathbf{V}^{-1}_{ss}\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}\right) \tag{8}$$

The MSE of $\hat{\theta}^s_{BLU}$ is given by

$$MSE_\xi(\hat{\theta}^s_{BLU}) = Var_\xi(\hat{\theta}^s_{BLU} - \theta^s) = g^s_1(\delta) + g^s_2(\delta) \tag{9}$$

where

$$g_1^s(\delta) = \mathbf{m}^T \left(\mathbf{G} - \mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\mathbf{Z}_s\mathbf{G}\right)\mathbf{m} \tag{10}$$

$$g_2^s(\delta) = \left(\mathbf{l}^T - \mathbf{m}^T\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)\left(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\left(\mathbf{l}^T - \mathbf{m}^T\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^T \tag{11}$$

The proof of the theorem is presented in details for example by Rao (2003, pp. 112-113).

4. Superpopulation model

Let us define random variables which realizations inform if the i-th population element has the attribute d (belongs to the dth random domain). Hence, we consider D random vectors $\mathbf{Q}_{(d)}$ for $d = 1, \ldots, D$ each of size $N \times 1$ where the i-th element of $\mathbf{Q}_{(d)}$ equals:

$$Q_{(d)i} = \begin{cases} 0 & \text{if} \quad i \notin \Omega_d \\ 1 & \text{if} \quad i \in \Omega_d \end{cases} \tag{12}$$

but – what is important – we do not assume that random sets Ω_d meet assumptions (1) and (2). Define a $N \times 1$ vector $\boldsymbol{\kappa}_{(d)}$ which i-th element equals $\kappa_{(d)i} = Pr(Q_{(d)i} = 1)$. Suppose that the values $\kappa_{(d)i}$ are known for all the population elements. Hence, $\kappa_{(d)i}$ for $i = 1, \ldots, N$ is the known value of the probability that the i-th element has the attribute d. In addition the probability that the i-th element does not have the attribute d equals $Pr(Q_{(d)i} = 0) = 1 - \kappa_{(d)i}$. Suppose also that $Q_{(d)i}$ and $Q_{(d)j}$ for $i \neq j$ are independent and that $Q_{(d)i}$ and $Q_{(d')j}$ for $i \neq j$ and $d \neq d'$ are independent too. Let us assume that $Q_{(d)i}$ and $Q_{(d')i}$ for $d \neq d'$ do not have to be independent and let $\kappa_{(dd')i} = Pr(Q_{(d)i} = 1 \wedge Q_{(d')i} = 1)$. In the special case, when $Q_{(d)i}$ and $Q_{(d')i}$ for $d \neq d'$ are independent, we obtain that $\kappa_{(dd')i} = \kappa_{(d)i}\kappa_{(d')i}$. Generally, the probability that the i-th element of population belong to domains $\Omega_d, \Omega_{d'}, \Omega_{d''}, \ldots$ will be denoted by $\kappa_{(dd'd''\ldots)i} = Pr(Q_{(d)i} = 1 \wedge Q_{(d')i} = 1 \wedge Q_{(d'')i} = 1 \wedge \ldots)$.

Consider joint distribution of $Q_{(d)i}$ which will be denoted by q. We introduce the notation q similarly to response distribution q considered by Cassel, Särndal, and

Wretman (1983) what we discuss below. Finally, we can write the assumptions as follows:

$$E_q(Q_{(d)i}) = \kappa_{(d)i} \tag{13}$$

$$cov_q(Q_{(d)i}, Q_{(d')j}) = \begin{cases} \kappa_{(d)i}(1-\kappa_{(d)i}) & \text{if} \quad i=j \wedge d=d' \\ \kappa_{(dd')i} - \kappa_{(d)i}\kappa_{(d')i} & \text{if} \quad i=j \wedge d \neq d' \\ 0 & \text{if} \quad i \neq j \end{cases} \tag{14}$$

where all of the values the of $\kappa_{(d)i}$ are known for $i = 1, \ldots, N$ and $d = 1, \ldots, D$. In empirical analysis they are usually unknown (especially for unsampled elements of population) and should be estimated.

It should be stressed that a similar approach is used when nonresponse occurs (e.g. Cassel et al. 1983). In this case the two subpopulations of respondents and nonrespondents can be treated as domains (the number of domains $D = 2$), but in this case one element of population may belong only to one domain (to the domain of respondents or to the domain of nonrespondents) and assumptions (1) and (2) are met. Moreover, it is often assumed that the elements of the population respond independently. Cassel et al. (1983) assume the form of the response distribution to propose imputation methods and to modify forms of classical estimators. In this paper the ξ and q distributions are assumed to derive the formula of the BLUP.

Let us define $N \times N$ matrices $\mathbf{\Gamma}_{(dd)}$ and $\mathbf{\Gamma}_{(dd')}$ which (i,j)th elements are respectively given by:

$$\Gamma_{(dd)ij} = \begin{cases} \kappa_{(d)i}(1-\kappa_{(d)i}) & \text{if} \quad i=j \\ 0 & \text{if} \quad i \neq j \end{cases}, \tag{15}$$

$$\forall_{d \neq d'} \Gamma_{(dd')ij} = \begin{cases} \kappa_{(dd')i} - \kappa_{(d)i}\kappa_{(d')i} & \text{if} \quad i=j \\ 0 & \text{if} \quad i \neq j \end{cases} \tag{16}$$

Hence, (13) and (14) may be written as follows:

$$E_q(\boldsymbol{Q}_{(d)}) = \boldsymbol{\kappa}_{(d)} \tag{17}$$

$$cov_q(\boldsymbol{Q}_{(d)}, \boldsymbol{Q}_{(d')}) = \mathbf{\Gamma}_{(dd')} \tag{18}$$

It should also be noticed that, for $d = d'$, (18) may be expressed as follows:

$$V_q(\boldsymbol{Q}_{(d)}) = \boldsymbol{\Gamma}_{(dd)} \tag{19}$$

We consider the two distributions ξ and q. Assume (13), (14) and that (compare with (3)):

$$\begin{cases} \mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}_Q\mathbf{v} + \mathbf{e} \\ E_{\xi|q}(\mathbf{e}) = \mathbf{0} \\ E_{\xi|q}(\mathbf{v}) = \mathbf{0} \\ D^2_{\xi|q}\begin{bmatrix} \mathbf{v} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \end{cases} \tag{20}$$

where $\mathbf{v}$ and $\mathbf{e}$ are independent, $\mathbf{Z}_Q$ and $E_q(\mathbf{Z}_Q) = \mathbf{Z}_\kappa$ are $N \times D$ matrices with idth elements equal $z_{(d)i}Q_{(d)i}$ and $z_{(d)i}\kappa_{(d)i}$ respectively ($i = 1, ..., N; d = 1, ..., D$), where $z_{(d)i}$ is a value of known function of some auxiliary variables (e.g. $\forall_{i,d} z_{(d)i} = x_{(d)i}$ or $\forall_{i,d} z_{(d)i} = 1$).

Note that model with assumption (20) may be written as the general linear mixed linear model with assumption (3) using: $E_{q\xi}(\mathbf{Z}_Q\mathbf{v}) = 0$, $E_{q\xi}(\mathbf{e}) = 0$ and $E_{q\xi}(\mathbf{Z}_Q\mathbf{v}\mathbf{e}) = 0$ what gives uncorrelated $\mathbf{Z}_Q\mathbf{v}$ and $\mathbf{e}$.

Thus,

$$E_{q\xi}(\mathbf{Y}) = E_q E_{\xi|q}(\mathbf{X}\beta + \mathbf{Z}_Q\mathbf{v} + \mathbf{e}) = E_q(\mathbf{X}\beta) = \mathbf{X}\beta \tag{21}$$

$$\begin{aligned} \mathbf{V} = V_{q\xi}(\mathbf{Y}) &= E_q(V_{\xi|q}(\mathbf{Y})) + V_q(E_{\xi|q}(\mathbf{Y})) = \\ &= E_q(\mathbf{R} + \mathbf{Z}_Q\mathbf{G}\mathbf{Z}_Q^T) + V_q(\mathbf{X}\beta) = \mathbf{R} + \mathbf{Z}_\kappa\mathbf{G}\mathbf{Z}_\kappa^T + \boldsymbol{\Sigma} \end{aligned} \tag{22}$$

where

$$\begin{aligned} \boldsymbol{\Sigma} &= E_q(\mathbf{Z}_Q - \mathbf{Z}_\kappa)\mathbf{G}(\mathbf{Z}_Q - \mathbf{Z}_\kappa)^T = \\ &= diag_{1\le i\le N}\left(\sum_{d=1}^{D} z^2_{(d)i}G_{dd}\Gamma_{(dd)ii} + \sum_{d=1}^{D}\sum_{d\neq k=1}^{D} z_{(d)i}z_{(k)i}G_{dk}\Gamma_{(dk)ii}\right) \end{aligned}$$

If the population elements are rearranged so that the first n elements of $\mathbf{Q}_d$ and κ_d are those in the sample and the first n rows of $\boldsymbol{\Gamma}_{(dd')}$, $\boldsymbol{\Sigma}$, $\mathbf{Z}_Q$ and $\mathbf{Z}_\kappa$ are for units in the sample, then these matrices may decomposed as follows: $\mathbf{Q}_{(d)} =$

$\begin{bmatrix} \mathbf{Q}_{s(d)} \\ \mathbf{Q}_{r(d)} \end{bmatrix}, \kappa_d = \begin{bmatrix} \kappa_{s(d)} \\ \kappa_{r(d)} \end{bmatrix}, \mathbf{\Gamma}_{(dd')} = \begin{bmatrix} \mathbf{\Gamma}_{ss(dd')} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_{rr(dd')} \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{ss} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{rr} \end{bmatrix},$
$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}, \mathbf{Z}_Q = \begin{bmatrix} \mathbf{Z}_{Qs} \\ \mathbf{Z}_{Qr} \end{bmatrix}$ and $\mathbf{Z}_\kappa = \begin{bmatrix} \mathbf{Z}_{\kappa s} \\ \mathbf{Z}_{\kappa r} \end{bmatrix}$, where $\kappa_{s(d)}$ and $\mathbf{Q}_{s(d)}$ are $n \times 1$, $\kappa_{r(d)}$ and $\mathbf{Q}_{r(d)}$ are $N_r \times 1$, $\mathbf{\Gamma}_{ss(dd')}$, $\mathbf{\Sigma}_{ss}$ and $\mathbf{V}_{ss}$ are $n \times n$, $\mathbf{\Gamma}_{rr(dd')}$, $\mathbf{\Sigma}_{rr}$ and $\mathbf{V}_{rr}$ are $N_r \times N_r$, $\mathbf{V}_{sr}$ is $n \times N_r$, $\mathbf{V}_{rs} = \mathbf{V}_{sr}^T$, $\mathbf{Z}_{Qs}$ and $\mathbf{Z}_{\kappa s}$ are $n \times D$, $\mathbf{Z}_{Qr}$ and $\mathbf{Z}_{\kappa r}$ are $N_r \times D$.

It was mentioned that in practical applications probabilities $\kappa_{(d)i}$ may be treated as known for all of sampled elements of population. For unsampled elements of population these probabilities have to be estimated.

Let us also consider the following superpopulation model which is a special case of the above introduced model with assumptions (20). Let us assume that:

$$E_q(Q_{(d)i}) = \kappa_{(d)}, \tag{23}$$

$$cov_q(Q_{(d)i}, Q_{(d')j}) = \begin{cases} \kappa_{(d)}(1 - \kappa_{(d)}) & \text{if } i = j \wedge d = d' \\ 0 & \text{elsewhere} \end{cases}, \tag{24}$$

$$Y_i = \mu + \sum_{d=1}^{D} Q_{(d)i} v_d + e_i \tag{25}$$

where v_d are iid and $v_d \sim (0, \sigma_v)$ $(d = 1, ..., D)$, e_i are iid and $e_i \sim (0, \sigma_e)$ $(i = 1, ..., N)$ and v_d and e_i are independent $(i = 1, ..., N; d = 1, ..., D)$. Note that we have inter alia assumed that $\forall_{d,d'} \forall_i \kappa_{(dd')i} = \kappa_{(d)i} \kappa_{(d')i}$, $\forall_d \forall_i \kappa_{(d)i} = \kappa_{(d)}$, $\forall_{i,d} x_{(d)i} = 1$ and $\forall_{i,d} z_{(d)i} = 1$.

5. Predictor and its MSE

We consider the problem of prediction of the total value for population elements with attribute $d*$ using a linear predictor of the form $\hat{\theta} = \mathbf{g}_s^T \mathbf{Y}_s$. We consider two cases, where the total of random domain is defined as $\theta_{d*}^Q = \mathbf{Q}_{(d*)}^T (\mathbf{X}\beta + \mathbf{Z}_Q \mathbf{v})$ or $\theta_{d*}^\kappa = \kappa_{(d*)}^T (\mathbf{X}\beta + \mathbf{Z}_Q \mathbf{v})$. Note, that both cases are more general then Henderson's case $\theta^s = \mathbf{l}^T \beta + \mathbf{m}^T \mathbf{v}$. Even in simpler case, i.e. $\theta_{d*}^\kappa = (\kappa_{(d*)}^T \mathbf{X})\beta + (\kappa_{(d*)}^T \mathbf{Z}_Q)\mathbf{v}$,

the expression $(\kappa_{(d*)}^T \mathbf{Z}_Q)$ is random, while in Henderson's case both $\mathbf{l}^T$ and $\mathbf{m}^T$ are fixed. Note that this additional source of variability will have influence on the formula of MSE.

The forms of the BLU predictors are derived by conditional minimization of their error variances with respect to both the ξ and q distributions (denoted by $Var_{q\xi}(\hat{\theta}-\theta)$). The constraint is introduced to ensure the predictor's $q\xi$-unbiasedness and it is given by $E_{q\xi}(\hat{\theta}-\theta)=0$.

Theorem 2. *Assume that the population data obey the model with the assumptions* (20). *Among linear, $q\xi$-unbiased predictors $\hat{\theta}=\mathbf{g}_s^T\mathbf{Y}_s$ of $\theta_{d*}^Q=\mathbf{l}_Q^T\beta+\mathbf{m}_Q^T\mathbf{v}$ (where $\mathbf{l}_Q^T=\mathbf{Q}_{(d*)}^T\mathbf{X}$ and $\mathbf{m}_Q^T=\mathbf{Q}_{(d*)}^T\mathbf{Z}_Q$) the error variance (which equals $q\xi$-MSE under $q\xi$-ubiasedness) is minimized by:*

$$\hat{\theta}_{BLUP}=\mathbf{p}^T\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s-\mathbf{X}_s\hat{\beta})+\mathbf{l}_\kappa^T\hat{\beta} \tag{26}$$

where

$$\mathbf{l}_\kappa^T=\kappa_{(d*)}^T\mathbf{X} \tag{27}$$

$$\begin{aligned}\mathbf{p}&=E_q(\mathbf{Z}_{Qs}\mathbf{G}\mathbf{Z}_Q^T\mathbf{Q}_{(d*)})=\\&=col_{1\le j\le n}\left(\sum_{d=1}^{D}\sum_{k=1}^{D}\sum_{i=1}^{N}z_{(d)j}z_{(k)i}G_{(dk)}E_q(Q_{(d)j}Q_{(k)i}Q_{(d*)i})\right)\end{aligned} \tag{28}$$

where

$$E_q(Q_{(d)j}Q_{(k)i}Q_{(d*)i})=\begin{cases}\kappa_{(d*)i} & \text{if } i=j\wedge d=k=d*\\ \kappa_{(dd*)i} & \text{if } i=j\wedge d\neq k\wedge k=d*\\ \kappa_{(dkd*)i} & \text{if } i=j\wedge d\neq k\neq d*\\ \kappa_{(d*)i}\kappa_{(d)j} & \text{if } i\neq j\wedge k=d*\\ \kappa_{(kd*)i}\kappa_{(d)j} & \text{if } i\neq j\wedge k\neq d*\end{cases} \tag{29}$$

$$\hat{\beta}=(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{Y}_s \tag{30}$$

$$\mathbf{V}_{ss}=\mathbf{Z}_{\kappa s}\mathbf{G}\mathbf{Z}_{\kappa s}^T+\mathbf{\Sigma}_{ss}+R_{ss} \tag{31}$$

$$\boldsymbol{\Sigma}_{ss} = E_q(\mathbf{Z}_{Qs} - \mathbf{Z}_{\kappa s})\mathbf{G}(\mathbf{Z}_{Qs} - \mathbf{Z}_{\kappa s})^T =$$
$$= diag_{1\le i\le n}\left(\sum_{d=1}^{D} z_{(d)i}^2 G_{(dd)}\Gamma_{(dd)ii} + \sum_{d=1}^{D}\sum_{d\neq k=1}^{D} z_{(d)i}z_{(k)i}G_{(dk)}\Gamma_{(dk)ii}\right) \tag{32}$$

$G_{(dk)}$ *is* dk*-th element of* $\mathbf{G}$ *matrix.*

The $q\xi$*-MSE of* $\hat{\theta}_{BLUP}$ *is given by:*

$$MSE_{q\xi}(\hat{\theta}_{BLUP}) = Var_{q\xi}(\hat{\theta}_{BLUP} - \theta_{d*}^{Q}) = g_{1\natural}(\boldsymbol{\delta},\boldsymbol{\kappa}) + g_{2\natural}(\boldsymbol{\delta},\boldsymbol{\kappa}) + (\mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}_{(d*d*)}\mathbf{X}\boldsymbol{\beta} \tag{33}$$

where

$$g_{1\natural}(\boldsymbol{\delta},\boldsymbol{\kappa}) = r - \mathbf{p}^T\mathbf{V}_{ss}^{-1}\mathbf{p} \tag{34}$$

$$g_{2\natural}(\boldsymbol{\delta},\boldsymbol{\kappa}) = (\mathbf{l}_{\kappa}^T - \mathbf{p}^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s)\left(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}(\mathbf{l}_{\kappa}^T - \mathbf{p}^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^T \tag{35}$$

$$r = E_q(\mathbf{Q}_{(d*)}^T\mathbf{Z}_Q\mathbf{G}\mathbf{Z}_Q^T\mathbf{Q}_{(d*)}) =$$
$$= \sum_{d=1}^{D}\sum_{k=1}^{D}\sum_{i=1}^{N}\sum_{j=1}^{N} z_{(d)i}z_{(k)j}G_{(dk)}E_q(Q_{(d)i}Q_{(d*)i}Q_{(k)j}Q_{(d*)j}) \tag{36}$$

where

$$E_q(Q_{(d)i}Q_{(d*)i}Q_{(k)j}Q_{(d*)j}) =$$
$$= \begin{cases} \kappa_{(d*)i} & \text{if } i=j \wedge d=k=d* \\ \kappa_{(dd*)i} & \text{if } i=j \wedge d=k\neq d* \\ \kappa_{(dkd*)i} & \text{if } i=j \wedge d\neq k \wedge d\neq d* \wedge k\neq d* \\ \kappa_{(dd*)i} & \text{if } i=j \wedge d\neq k \wedge d\neq d* \wedge k=d* \\ \kappa_{(dd*)i}\kappa_{(kd*)j} & \text{if } i\neq j \wedge d\neq d* \wedge k\neq d* \\ \kappa_{(dd*)i}\kappa_{(d*)j} & \text{if } i\neq j \wedge d\neq d* \wedge k=d* \\ \kappa_{(d*)i}\kappa_{(d*)j} & \text{if } i\neq j \wedge d=d* \wedge k=d* \end{cases} \tag{37}$$

The proof of the theorem is presented in the part 1 of the appendix.

Theorem 3. *Assume that the population data obey the model with the assumptions (20). Among linear,* $q\xi$*-unbiased predictors* $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$ *of* $\theta_{d*}^{\kappa} = \boldsymbol{\kappa}_{(d*)}^T(\mathbf{X}\boldsymbol{\beta} +$

$\mathbf{Z}_Q\mathbf{v}$), *the error variance (which equals* $q\xi$*-MSE under* $q\xi$*-ubiasedness) is minimized by:*

$$\hat{\theta}_{BLUPa} = \mathbf{p}_a^T\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}) + \mathbf{l}_\kappa^T\hat{\beta} \tag{38}$$

where

$$\mathbf{p}_a = (\mathbf{Z}_{\kappa s}\mathbf{G}\mathbf{Z}_{\kappa s}^T + \mathbf{\Sigma}_{ss})\kappa_{s(d*)} + (\mathbf{Z}_{\kappa s}\mathbf{G}\mathbf{Z}_{\kappa r}^T)\kappa_{r(d*)} \tag{39}$$

The $q\xi$*-MSE of* $\hat{\theta}_{BLUPa}$ *is given by:*

$$MSE_{q\xi}(\hat{\theta}_{BLUPa}) = Var_{q\xi}(\hat{\theta}_{BLUPa} - \theta_{d*}^{\kappa}) = g_{1\sharp a}(\delta,\kappa) + g_{2\sharp a}(\delta,\kappa) \tag{40}$$

where

$$g_{1\sharp a}(\delta,\kappa) = r_a - \mathbf{p}_a^T\mathbf{V}_{ss}^{-1}\mathbf{p}_a \tag{41}$$

$$g_{2\sharp a}(\delta,\kappa) = (\mathbf{l}_\kappa^T - \mathbf{p}_a^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s)\left(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}(\mathbf{l}_\kappa^T - \mathbf{p}_a^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^T \tag{42}$$

$$\begin{aligned} r_a &= \kappa_{(d*)}^T(\mathbf{Z}_\kappa\mathbf{G}\mathbf{Z}_\kappa^T + \mathbf{\Sigma})\kappa_{(d*)} = \\ &= \kappa_{s(d*)}^T(\mathbf{Z}_{\kappa s}\mathbf{G}\mathbf{Z}_{\kappa s}^T + \mathbf{\Sigma}_{ss})\kappa_{s(d*)} + \kappa_{r(d*)}^T(\mathbf{Z}_{\kappa r}\mathbf{G}\mathbf{Z}_{\kappa r}^T + \mathbf{\Sigma}_{rr})\kappa_{r(d*)} + \\ &+ 2\kappa_{s(d*)}^T(\mathbf{Z}_{\kappa s}\mathbf{G}\mathbf{Z}_{\kappa r}^T)\kappa_{r(d*)} \end{aligned} \tag{43}$$

The proof of the theorem is presented in the part 2 of the appendix. The proof of the theorem 3 may also be obtained using theorem 1 and fact that model (20) may be written as (3).

6. Special cases

Let us introduce special cases of theorems 2 and 3 under superpopulation model with assumptions (23), (24), and (25). Let us introduce following notations: $\mathbf{Q}$ and κ are $N \times D$ matrices with id-th elements given by $Q_{(d)i}$ and $\kappa_{(d)i}$, respectively. Under these assumptions we obtain that:

$$\mathbf{Z}_Q = \mathbf{Q} \tag{44}$$

$$\mathbf{Z}_\kappa = \kappa \tag{45}$$

$$\begin{aligned}\Sigma_{ss} &= E_q(\mathbf{Q}_s - \kappa_s)\mathbf{G}(\mathbf{Q}_s - \kappa_s)^T = \\ &= \sigma_v^2 \sum_{d=1}^{D} \mathbf{\Gamma}_{(dd)} = \sigma_v^2 diag_{1\le i\le n}\left(\sum_{d=1}^{D} \Gamma_{(dd)ii}\right) = \\ &= \sigma_v^2 \sum_{d=1}^{D} \kappa_{(d)}(1-\kappa_{(d)})\mathbf{I}_n\end{aligned} \tag{46}$$

$$\mathbf{V}_{ss} = \sigma_v^2 \kappa_s \kappa_s^T + \sigma_v^2 \sum_{d=1}^{D} \kappa_{(d)}(1-\kappa_{(d)})\mathbf{I}_n + \sigma_e^2 \mathbf{I}_n \tag{47}$$

Hence the ij-th element of $\mathbf{V}_{ss}$ is given by:

$$Cov_{q\xi}(Y_i, Y_j) = \begin{cases} \sigma_e^2 + \sigma_v^2 c_1 & \text{for} \quad i = 1 \\ \sigma_v^2 c_2 & \text{for} \quad i \neq 1 \end{cases} \tag{48}$$

where $c_1 = \sum_{d=1}^{D} \kappa_d$, $c_2 = \sum_{d=1}^{D} \kappa_d^2$ To obtain $\mathbf{V}_{ss}^{-1}$ the following equation (e.g. Rao 1982, p. 86) is used. If $(a_{ij}) = \begin{bmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \dots & \dots & \dots & \dots \\ b & b & \dots & a \end{bmatrix}$ then inverse matrix has the same form where $a_{ii} = \frac{a+(n-2)b}{(a+(n-1)b)(a-b)}$, $a_{ii} = \frac{-b}{(a+(n-1)b)(a-b)}$ for $i \neq j$. In our case the ii-th and ij-th elements of $\mathbf{V}_{ss}^{-1}$ are given by:

$$a_{ii} = k^{-1} - k^{-1}\frac{\sigma_v^2 c_2}{k + \sigma_v^2 c_2 n} \tag{49}$$

$$a_{ii} = -k^{-1}\frac{\sigma_v^2 c_2}{k + \sigma_v^2 c_2 n} \tag{50}$$

where

$$k = \sigma_e^2 + \sigma_v^2(c_1 - c_2) \tag{51}$$

Theorem 4. *Assume that the population data obey the model with the assumptions* (23), (24), *and* (25). *Among linear, $q\xi$-unbiased predictors $\hat{\theta} = \mathbf{g}_s^T \mathbf{Y}_s$ of $\theta_{d*}^Q = \mathbf{l}_Q^T \beta + \mathbf{m}_Q^T \mathbf{v}$ (where $\mathbf{l}_Q^T = \mathbf{Q}_{(d*)}^T \mathbf{X}$ and $\mathbf{m}_Q^T = \mathbf{Q}_{(d*)}^T \mathbf{Z}_Q$), the error variance (which equals $q\xi$-MSE under $q\xi$-ubiasedness) is minimized by:*

$$\hat{\theta}_{BLUP} = \kappa_{(d*)} N \bar{Y}_s \tag{52}$$

The $q\xi$-MSE of $\hat{\theta}_{BLUP}$ is given by:

$$\begin{aligned} MSE_{q\xi}(\hat{\theta}_{BLUP}) &= Var_{q\xi}(\hat{\theta}_{BLUP} - \theta^{Q}_{d*}) = \\ &= g_{1\sharp}(\delta,\kappa) + g_{2\sharp}(\delta,\kappa) + (\mathbf{X}\beta)^{T}\mathbf{\Gamma}_{(d*d*)}\mathbf{X}\beta \end{aligned} \tag{53}$$

where

$$\begin{aligned} g_{1\sharp}(\delta,\kappa) = \sigma_v^2 \left[\kappa_{(d*)}(1+c_1-\kappa_{(d*)}) + \kappa^2_{(d*)}N(N-1)(1+c_2-\kappa^2_{(d*)})\right] + \\ - n\sigma_v^4\kappa^2_{(d*)}\left[1+c_1+(N-2)\kappa_{(d*)}+(N-1)(c_2-\kappa^2_{(d*)})\right]^2 (k+\sigma_v^2c_2n)^{-1} \end{aligned} \tag{54}$$

$$\begin{aligned} g_{2\sharp}(\delta,\kappa) = n^{-1}(k+\sigma_v^2c_2n)^{-1}\left[N\kappa_{(d*)}(k+\sigma_v^2c_2n) - n\sigma_v^2\kappa_{(d*)}\times\right. \\ \left.\times\left(1+c_1+(N-2)\kappa_{(d*)}+(N-1)(c_2-\kappa^2_{(d*)})\right)\right]^2 \end{aligned} \tag{55}$$

$$(\mathbf{X}\beta)^{T}\mathbf{\Gamma}_{(d*d*)}\mathbf{X}\beta = \mu^2N\kappa_{(d*)}(1-\kappa_{(d*)}) \tag{56}$$

The proof of the theorem is the special case of the proof of the theorem 2.

Theorem 5. *Assume that the population data obey the model with the assumptions* (23), (24), *and* (25). *Among linear, $q\xi$-unbiased predictors $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$ of $\theta^{\kappa}_{d*} = \kappa^T_{(d*)}(\mathbf{X}\beta + \mathbf{Z}_Q\mathbf{v})$, the error variance (which equals $q\xi$-MSE under $q\xi$-ubiasedness) is minimized by:*

$$\hat{\theta}_{BLUPa} = \kappa_{(d*)}N\bar{Y}_s \tag{57}$$

The $q\xi$-MSE of $\hat{\theta}_{BLUP}$ is given by:

$$MSE_{q\xi}(\hat{\theta}_{BLUPa}) = Var_{q\xi}(\hat{\theta}_{BLUP} - \theta^{\kappa}_{d*}) = g_{1\sharp a}(\delta,\kappa) + g_{2\sharp a}(\delta,\kappa) \tag{58}$$

where

$$g_{1\sharp a}(\delta,\kappa) = \sigma_v^2\kappa^2_{(d*)}(c_1+(N-1)c_2)(k+\sigma_v^2c_2n)^{-1}(Nk+\sigma_v^2n(c_2-c_1)) \tag{59}$$

$$g_{2\sharp a}(\delta,\kappa) = n\kappa^2_{(d*)}(k+\sigma_v^2c_2n)^{-3}[Nk+n\sigma_v^2(c_2-c_1)]^2 \tag{60}$$

The proof of the theorem is the special case of the proof of the theorem 3.

7. Estimation of superpopulation model's parameters for special case

Proposed BLUPs (even in the presented special cases) are functions of unknown superpopulation model parameters. If they are replaced by their estimates we obtain the empirical best linear unbiased predictor (EBLUP). In this section the problem of estimation of superpopulation model parameters will be considered.

In the classic case, where only distribution is considered (Henderson 1950), different approximately unbiased MSE estimators of EBLUP were proposed (Prasad and Rao 1990; Datta and Lahiri 2000; Das, Jiang and Rao 2005). These estimators take into account the additional variability of EBLUP due to the estimation of superpopulation parameters. Because in the classic case in many practical issues MSEs of EBLUPs are slightly higher than MSEs of BLUPs, the naive MSE estimator of EBLUP (which has the form of MSE of BLUP, where superpopulation model parameters are replaced by their estimators) gives acceptable results (its bias is not high).

To obtain naive estimators of MSE of EBLUPs considered in this paper unknown superpopulation model parameters in (53) and (58) will be replaced by maximum likelihood estimators under normality assumption. Density function of $\mathbf{Y}$ may be written as follows:

$$f(\mathbf{y}) = \sum_{q} f(\mathbf{y}|\mathbf{Q}=\mathbf{q})p_{\mathbf{Q}}(\mathbf{q}) \tag{61}$$

where the sum on right side of (61) is over all realization $\mathbf{q}$ of $\mathbf{Q}$ qiven by (44) (in practice – what is important – only one realization $\mathbf{q}$ of $\mathbf{Q}$ is known),

$$f(\mathbf{y}|\mathbf{Q}=\mathbf{q}) = (2\pi)^{-\frac{n}{2}} det(\mathbf{V}_{qss})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{V}_{qss}^{-1}(\mathbf{y}-\mu)\right], \tag{62}$$

$$\mathbf{V}_{qss} = D^2_{\xi|q}(\mathbf{Y}_s) = \sigma_e^2\mathbf{I}_n + \sigma_v^2\mathbf{q}\mathbf{q}^T, \tag{63}$$

$$Cov_{\xi|q}(Y_i, Y_j) = \begin{cases} \sigma_e^2 + \sigma_v^2 \sum_{d=1}^{D} q_{(d)i} & \text{for} \quad i = 1 \\ \sigma_v^2 \sum_{d=1}^{D} q_{(d)i}q_{(d)j} & \text{for} \quad i \neq 1 \end{cases} \tag{64}$$

$$p_{\mathbf{Q}}(\mathbf{q}) = \prod_{d=1}^{D} \kappa_{(d)}^{\sum_{i=1}^{n} q_{(d)i}} (1-\kappa_{(d)})^{n-\sum_{i=1}^{n} q_{(d)i}} \tag{65}$$

Because only one realization $\mathbf{q}$ of $\mathbf{Q}$ is known, to obtain estimators of μ, σ_e^2, σ_v^2, and $\kappa_{(d)}$ $(d = 1, ..., D)$, we solve the following equalities:

$$\frac{\partial \ln f(\mathbf{y}|\mathbf{Q}=\mathbf{q})}{\partial \mu} = 0 \tag{66}$$

$$\frac{\partial \ln f(\mathbf{y}|\mathbf{Q}=\mathbf{q})}{\partial \sigma_e^2} = 0 \tag{67}$$

$$\frac{\partial \ln f(\mathbf{y}|\mathbf{Q}=\mathbf{q})}{\partial \sigma_v^2} = 0 \tag{68}$$

$$\frac{\partial \ln p_{\mathbf{Q}}(\mathbf{p})}{\partial \kappa_{(d)}} = \frac{1}{\kappa_{(d)}} \sum_{i=1}^{n} q_{(d)i} + \frac{1}{1-\kappa_{(d)}} (n - \sum_{i=1}^{n} q_{(d)i}) = 0 \tag{69}$$

Solutions of (66), (67), and (68) are obtained iteratively and they are denoted by $\hat{\mu}$, $\hat{\sigma}_e^2$, $\hat{\sigma}_v^2$, respectively. Solution of (69) is given by $\hat{\kappa}_{(d)} = \frac{1}{n} \sum_{i=1}^{n} q_{(d)i}$. Hence the naive estimators of (53) and (58) are given by (53) and (58) where μ, σ_e^2, σ_v^2 and $\kappa_{(d)}$ are replaced by $\hat{\mu}$, $\hat{\sigma}_e^2$, $\hat{\sigma}_v^2$, and $\hat{\kappa}_{(d)}$ $(d = 1, ..., D)$, respectively.

Additionally, in the next section the performance of EBLUPs and their naive MSEs estimators will be studied in the simulation study.

8. Simulation study

In the simulation study using R language (R Development Core Team 2007) $M = 10000$ of realization of $\mathbf{Y}$ and $\mathbf{Q}$ are generated based on (23), (24), and (25), where (parameters are chosen arbitrary) $\mu = 100$, v_d and e_i $(i = 1, ..., N;$ $d = 1, ..., D)$ are generated independently and $v_d \sim N(0, 6)$ and $e_i \sim N(0, 3)$, $N = 2000$, $n = 200$, $D = 3$, domain attributes are generated independently (see (24)) with $\kappa_{(1)} = 0,8$, $\kappa_{(2)} = 0,6$ and $\kappa_{(3)} = 0,4$.

Four predictors are studied:

- BLUP of θ_{d*}^{Q} given by (52) (denoted in further analysis by $BLUP_Q$),
- BLUP of θ_{d*}^{κ} given by (57) ($BLUP_\kappa$),
- EBLUP of θ_{d*}^{Q} given by (52) where $\kappa_{(d*)}$ is replaced by $\hat{\kappa}_{(d*)}$ ($EBLUP_Q$),

– EBLUP of θ_{d*}^{κ} given by (57) where $\kappa_{(d*)}$ is replaced by $\hat{\kappa}_{(d*)}$ $(EBLUP_{\kappa})$.

Let $\theta_{d*}^{Q}(i)$, $\theta_{d*}^{\kappa}(i)$, $BLUP_Q(i)$, $BLUP_{\kappa}(i)$, $EBLUP_Q(i)$, $EBLUP_{\kappa}(i)$ denote values of θ_{d*}^{Q}, θ_{d*}^{κ}, $BLUP_Q$, $BLUP_{\kappa}$, $EBLUP_Q$, $EBLUP_{\kappa}$ obtained in the ith step of the simulation study ($i = 1, ..., M$), respectively. In the simulation following statistics are considered:

– Relative bias of $BLUP_Q$ (denoted by $B(BLUP_Q)$ in %) given by

$$100\left(\frac{1}{M}\sum_{i=1}^{M}(\theta_{d*}^{Q}(i))\right)^{-1}\frac{1}{M}\sum_{i=1}^{M}\left(BLUP_Q(i)-\theta_{d*}^{Q}(i)\right) \tag{70}$$

Relative bias of $EBLUP_Q$ (denoted by $B(EBLUP_Q)$ in %) is given by (70), where $BLUP_Q$ is replaced by $EBLUP_Q$.

– Relative bias of $BLUP_{\kappa}$ (denoted by $B(BLUP_{\kappa})$ in %) given by

$$100\left(\frac{1}{M}\sum_{i=1}^{M}(\theta_{d*}^{\kappa}(i))\right)^{-1}\frac{1}{M}\sum_{i=1}^{M}(BLUP_{\kappa}(i)-\theta_{d*}^{\kappa}(i))) \tag{71}$$

Relative bias of $EBLUP_{\kappa}$ (denoted by $B(EBLUP_{\kappa})$ in %) is given by (71), where $BLUP_{\kappa}$ is replaced by $EBLUP_{\kappa}$.

– Relative root of prediction variance error of $BLUP_Q$ (denoted by $D(BLUP_Q)$ in %) given by

$$100\left(\frac{1}{M}\sum_{i=1}^{M}(\theta_{d*}^{Q}(i))\right)^{-1}\times \\ \times\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left[(BLUP_Q(i)-\theta_{d*}^{Q}(i))-\frac{1}{M}\sum_{i=1}^{M}(BLUP_Q(i)-\theta_{d*}^{Q}(i))\right]^2} \tag{72}$$

Relative root of prediction variance error of $EBLUP_Q$ (denoted by $D(EBLUP_Q)$ in %) is given by (72), where $BLUP_Q$ is replaced by $EBLUP_Q$.

– Relative root of prediction variance error of $BLUP_{\kappa}$ (denoted by $D(BLUP_{\kappa})$ in %) given by

$$100\left(\frac{1}{M}\sum_{i=1}^{M}(\theta_{d*}^{\kappa}(i))\right)^{-1}\times \\ \times\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left[(BLUP_{\kappa}(i)-\theta_{d*}^{\kappa}(i))-\frac{1}{M}\sum_{i=1}^{M}(BLUP_{\kappa}(i)-\theta_{d*}^{\kappa}(i))\right]^2} \tag{73}$$

Relative root of prediction variance error of $EBLUP_\kappa$ (denoted by $D(EBLUP_\kappa)$ in %) is given by (73), where $BLUP_\kappa$ is replaced by $EBLUP_\kappa$.

- Relative root of prediction MSE of $BLUP_Q$ (denoted by $RMSE(BLUP_Q)$ in %) given by

$$100\left(\frac{1}{M}\sum_{i=1}^{M}(\theta_{d*}^{Q}(i))\right)^{-1}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(BLUP_Q(i)-\theta_{d*}^{Q}(i)\right)^2} \tag{74}$$

Relative root of prediction MSE of $EBLUP_Q$ (denoted by $RMSE(EBLUP_Q)$ in %) is given by (74), where $BLUP_Q$ is replaced by $EBLUP_Q$.

- Relative root of prediction MSE of $BLUP_\kappa$ (denoted by $RMSE(BLUP_\kappa)$ in %) given by

$$100\left(\frac{1}{M}\sum_{i=1}^{M}(\theta_{d*}^{\kappa}(i))\right)^{-1}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(BLUP_\kappa(i)-\theta_{d*}^{\kappa}(i)\right)^2} \tag{75}$$

Relative root of prediction MSE of $EBLUP_\kappa$ (denoted by $RMSE(EBLUP_\kappa)$ in %) is given by (75), where $BLUP_\kappa$ is replaced by $EBLUP_\kappa$.

- Relative root of expectation of naive MSE estimator of $EBLUP_Q$ (denoted by $\widehat{\mathrm{E}RMSE}(EBLUP_Q)$ in %) given by

$$100\left(\frac{1}{M}\sum_{i=1}^{M}(\theta_{d*}^{Q}(i))\right)^{-1}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\widehat{MSE}(EBLUP_Q(i))\right)} \tag{76}$$

where $\widehat{MSE}(EBLUP_Q(i))$ is a value of naive MSE estimator of $EBLUP_Q$ obtained in the ith step of the simulation. Relative root of expectation of naive MSE estimator of $EBLUP_\kappa$ is given by (76) where $EBLUP_Q$ and θ_{d*}^{Q} are replaced by $EBLUP_\kappa$ and θ_{d*}^{κ}, respectively.

Table 1

Simulation results

	d=1	d=2	d=3
$B(BLUP_Q)$ in %	$-0,0081$	$0,0152$	$-0,0804$
$B(BLUP_\kappa)$ in %	$-0,0007$	$-0,0007$	$-0,0007$
$B(EBLUP_Q)$ in %	$0,0392$	$0,0601$	$-0,2000$
$B(EBLUP_\kappa)$ in %	$0,0466$	$0,0442$	$-0,1204$
$D(BLUP_Q)$ in %	$1,6977$	$3,0706$	$4,5469$
$D(BLUP_\kappa)$ in %	$0,3911$	$0,3911$	$0,3911$
$D(EBLUP_Q)$ in %	$3,6008$	$6,0346$	$9,1164$
$D(EBLUP_\kappa)$ in %	$3,6097$	$5,8212$	$8,8291$
$RMSE(BLUP_Q)$ in %	$1,6977$	$3,0707$	$4,5476$
$RMSE(BLUP_\kappa)$ in %	$0,3911$	$0,3911$	$0,3911$
$RMSE(EBLUP_Q)$ in %	$3,6010$	$6,0349$	$9,1186$
$RMSE(EBLUP_\kappa)$ in %	$3,6100$	$5,8214$	$8,8299$
$\mathrm{E}\widehat{RMSE}(EBLUP_Q)$ in %	$1,5774$	$2,8297$	$4,2288$
$\mathrm{E}\widehat{RMSE}(EBLUP_\kappa)$ in %	$1,0402$	$1,1025$	$1,0081$

Simulation results confirm that all of the considered predictors are model unbiased (simulation biases are very small). In the above table MSEs of $BLUP_Q$ are higher than MSEs of $BLUP_\kappa$ and their values are higher for domains with smaller $\kappa_{(d)}$. Due to the estimation of probabilities substantial increase of MSEs of EBLUPs is observed comparing with MSEs of BLUPs. What is more, naive MSE estimators do not explain additional variability of EBLUPs (due to estimation of unknown parameters) and further analysis have to be conducted to obtain better MSEs estimators.

Appendix 1

To prove theorem 2 we consider the problem of prediction of the total value in the $d*$th domain $\theta_{d*}^{Q} = \mathbf{l}_Q^T\beta + \mathbf{m}_Q^T\mathbf{v}$ (where $\mathbf{l}_Q^T = \mathbf{Q}_{(d*)}^T\mathbf{X}$ and $\mathbf{m}_Q^T = \mathbf{Q}_{(d*)}^T\mathbf{Z}_Q$) using a linear predictor of the form $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$. The error variance of $\hat{\theta}$ with respect to both distributions ξ and q is given by:

$$\begin{aligned} Var_{q\xi}(\hat{\theta} - \theta_{d*}^{Q}) &= Var_q(E_{\xi|q}(\hat{\theta} - \theta_{d*}^{Q})) + E_q(Var_{\xi|q}(\hat{\theta} - \theta_{d*}^{Q})) = \\ &= (\mathbf{X}\beta)^T\mathbf{\Gamma}_{(d*d*)}\mathbf{X}\beta + \mathbf{g}_s^T\mathbf{V}_{ss}\mathbf{g}_s + r - 2\mathbf{g}_s^T\mathbf{p} \end{aligned} \tag{77}$$

where $\mathbf{p}$ and r are given by (28) and (36), respectively. The condition of $q\xi$-unbiasedness is given by:

$$E_qE_{\xi|q}(\hat{\theta} - \theta_{d*}^{Q})) = (\mathbf{g}_s^T\mathbf{X}_s - \kappa_{(d*)}^T\mathbf{X})\beta = 0 \tag{78}$$

what gives:

$$\mathbf{g}_s^T\mathbf{X}_s = \kappa_{(d*)}^T\mathbf{X} \tag{79}$$

The optimal vector of weights of the predictor $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$ (and hence the formula of BLUP and its MSE presented in the theorem 2) is obtained by conditional minimization of the error variance (77), where the constraint is given by (79). The problem is solved using the method of Lagrange multipliers.

Appendix 2

To prove theorem 3 we consider the problem of prediction of the total value in the $d*$th domain $\theta_{d*}^{\kappa} = \kappa_{(d*)}^T(\mathbf{X}\beta + \mathbf{Z}_Q\mathbf{v})$ using a linear predictor of the form $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$. The error variance of $\hat{\theta}$ with respect to both distributions ξ and q is given by:

$$\begin{aligned} Var_{q\xi}(\hat{\theta} - \theta_{d*}^{\kappa}) &= Var_q(E_{\xi|q}(\hat{\theta} - \theta_{d*}^{\kappa})) + E_q(Var_{\xi|q}(\hat{\theta} - \theta_{d*}^{\kappa})) = \\ &= \mathbf{g}_s^T\mathbf{V}_{ss}\mathbf{g}_s + r_a - 2\mathbf{g}_s^T\mathbf{p}_a \end{aligned} \tag{80}$$

where $\mathbf{p}_a$ and r_a are given by (39) and (43), respectively. The condition of $q\xi$-unbiasedness is given by (79). The optimal vector of weights of the predictor $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$ (and hence the formula of BLUP and its MSE presented in the theorem 3) is obtained by conditional minimization of the error variance (80), where the

constraint is given by (79). The problem is solved using the method of Lagrange multipliers.

References

Cassel C.M., Särndal C.E., Wretman J.H. (1983): Some Uses of Statistical Models in Connection with the Nonresponse Problem, In: Incomplete Data in Sample Surveys. Vol. 3. W.G. Madow , I. Olkin (eds.). Proceedings of the Symposium, Academic Press, New York, pp. 143-170.

Das K., Jiang J., Rao J.N.K. (2004): Mean Squared Error of Empirical Predictor. "The Annals of Statistics", Vol. 32, No. 2, pp. 818-840.

Datta G. S., Lahiri P. (2000): A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems. "Statistica Sinica", 10, pp. 613-627.

Henderson C.R. (1950): Estimation of Genetic Parameters (Abstract). "Annals of Mathematical Statistics", 21, pp. 309-310.

Prasad N.G.N, Rao J.N.K. (1990): The Estimation of Mean the Mean Squared Error of Small Area Estimators. "Journal of the American Statistical Association", 85, pp. 163-171.

R Development Core Team (2007): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org.

Rao C.R. (1982): Modele liniowe statystyki matematycznej. Państwowe Wydawnictwo Naukowe, Warszawa.

Rao J.N.K. (2003): Small Area Estimation. Wiley & Sons, Inc., New York.

Żądło T. (2006): On Prediction of Total Value in Incompletely Specified Domains. "Australian and New Zealand Journal of Statistics", 48 (3), pp. 269-283.

Abstract

The problem of prediction of domain totals is widely discussed in the small area estimation literature (e.g. Rao 2003). In the classic approach it assumed that the population is divided into disjoint domains and sum of domains gives the whole set of population elements. In this paper we define random variables which realizations inform if the i-th population element has the attribute d (belongs to the d-th random domain). What is more, one population element may have no attribute or more than one attribute. The proposed model may be treated as the model assuming random overlapping domains. We present the problem of prediction of a domain total (or being more precise – total value for elements of population with some attribute) based on the general linear mixed model (GLMM). Different

model (assuming inter alia that one population element may belong at random only to one of domains) was considered by Żądło (2006). The main aim of this paper is to present the equation of the best linear unbiased predictor (BLUP) and its mean squared error (MSE) under the proposed model. Additionally the problem of estimation of model parameters will be studied and its influence on the predictor's accuracy will be considered in the simulation study.

Grażyna Trzpiot*

ESTIMATION METHOD FOR QUANTILE REGRESSION

Introduction

Quantile regression, as introduced by Koenker and Bassett (1978), is gradually evolving into a comprehensive approach to the statistical analysis of linear and non-linear response models for conditional quantile functions. Like a classical linear regression methods based on minimizing sums of squared residuals enable one to estimate models for conditional mean functions, quantile regression methods based on minimizing asymmetrically weighted absolute residuals over a mechanism for estimating models for the conditional median function, and the full range of other conditional quantile functions.

Like robust estimation, the quantile approach detects relationships missed by traditional data analysis. Robust estimates detect the influence of the bulk of the data, whereas quantile estimates detect the influence of co-variates on alternate parts of the conditional distribution. Unlike least squares however, the regression through the quantiles produces a different estimate than the regression quantiles.

*The research was supported by the grant number KBN: N111 003 32/0262.

1. Quantile regression

Quantile regression is a method for estimating functional relations between variables for all portions of probability distribution. Typically a response variable Y is some function of predictor variable X. Regression application focus in estimating rates of changes in the mean of the response variable distribution as some function of a set of predictor variables. In the other words the function is defined for the expected value of Y conditional X, $E(Y|X)$. Regression analysis gave incomplete picture of the relationships between variables especially for regression models with heterogeneous variances.

Quantile regression was developed as an extension of the linear model for estimating rate of change in all parts of the distribution of response variables. The estimates are semi parametric in the sense that no parametric distributional form (eg. normal, Poisson, negative binominal, etc.) is assumed for the random error part of the model ε, although a parametric form is assumed for the deterministic portion of the model (eg. $\beta_0 X_0 + \beta_1 X_1$). The conditional quantiles denoted by $Q_y(\tau|X)$ are the inverse of the conditional cumulative distribution function of the response variable $F_y^{-1}(\tau|X)$, where $\tau \in [0,1]$ denotes quantile.

The quantile model posits the τ^{th} quantile of Y conditional on x to be:

$$Q(\tau|x) = \alpha(\tau) + x\beta(\tau), \quad 0 < \tau < 1.$$

If $\beta(\tau)$ is a constant β, the model reduces to the standard conditional expectation model, $E(Y|x) = \alpha + x\beta$, with constant variance errors. When $\beta(\tau)$ depends on τ, the model allows the distribution of Y to depend on x in different ways at different parts of the distribution. The traditional linear model can be viewed as a summary of all the quantile effects; that is, $\int Q(\tau|x)d\tau = E(Y|x)$. Under this interpretation, traditional analysis loses information due to its aggregation of possibly disparate quantile effects. Many different quantile paths, for example, can lead to $\beta_k = 0$. On the one hand, $\beta_k = 0$ can mean x_k does not matter – does not affect the distribution of Y. But it can also mean there are important, but compensating quantile effects relating Y and x.

2. Estimation linear quantile regression function

Consider the following regression model:

$$y_i = g(x_i) + e_i \tag{1}$$

where the dependent variable $y = (y_1, y_2, \ldots, y_n)$ and independent $x = (x_1, x_2, \ldots, x_n)$ where $y \in \mathbb{R}$ and $x \in \mathbb{R}^p$, $g(\cdot)$ is real valued and unknown. We are interested in estimating the regression function $g(\cdot)$ given x_i.

In the parametric framework of the linear regression model when $g(x_i) = \beta(\tau)x_i$ the quantile regression was proposed as a solution of:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \varrho_\tau(y_i - x_i\beta) \tag{2}$$

where $\varrho_\tau(z) = |\tau - I(z < 0)| \cdot |z|$, I is the indicator function*.

The conditional quantile τ of y_i given x_i, by monotonicity of quantile function:

$$Q(\tau|x) = g(x) + D^{-1}(\tau|x) \equiv g_\tau(x) \tag{3}$$

where $D^{-1}(\tau|x)$ is conditional τ^{th} quantile of error term ε_i and $Q(\tau|x) \equiv \equiv \inf\{\lambda : P(y_i \leq \lambda|x) \geq \tau\}$. In equation (3) $g(x)$ and $D^{-1}(\tau|x)$ are not identified separately. However $g_\tau(x)$, the conditional τ^{th} quantile can be identified, then the equation (1) can be rewritten as:

$$y_i = g_\tau(x_i) + v_i \tag{4}$$

where $v_i = \varepsilon_i - D^{-1}(\tau|x)$ and v_i is a new error term which has a zero conditional quantile.

Given $(y_i; x_i)$, the quantile model can be estimated by regression quantiles which are defined by the minimization problem:

$$\beta^*(\tau) = \min_{b \in \mathbb{R}} \left\{ \sum_{y_i \geq x_i b} w_i(\tau|y_i - x_i b|) + \sum_{y_i < x_i b} w_i(1-\tau)|y_i - x_i b| \right\} \tag{5}$$

* $I[A] = 1$ if A is true, $I[A] = 0$ otherwise.

where the weights w_i are introduced to account for different variability of x_i and the different number of observations at each x_i.

To compare the different quantile estimates it is useful to express the data in terms of the empirical distribution:

$$\hat{F}_i(v) = \frac{1}{n_i} \sum I(y_{ij} < v) \tag{6}$$

and let $\hat{Q}_i(\tau)$ be the associated empirical quantile function. Notice that $\hat{Q}_i$ is an ordinary empirical quantile and hence is asymptotically normal with mean $\alpha(\tau) + x_i\beta(\tau)$ and variance $\frac{\sigma_i^2(\tau)}{n_i}$ where $\sigma_i^2(\tau) = \tau(1-\tau)/(f(Q_i(\tau))^2)$.

Hence the data can be written in familiar linear model form as:

$$\hat{Q}(\tau|x) = \alpha(\tau) + x\beta(\tau) + \varepsilon_i, \quad i = 1, \ldots, n \tag{7}$$

where the error terms are independent and asymptotically normal with mean zero and variance $\frac{\sigma_i^2(\tau)}{\lambda_i}$. The model is thus seen to be amenable to weighted least squares estimation. Instead of implementing regression quantile estimation on all the $(y_i; x_i)$ data, we can do weighted least squares on the smaller data set, $[\hat{Q}_i(\tau); x_i]$, $i = 1, \ldots, n$.

How does this "regression-through-the-empirical quantiles" estimate compare to regular regression quantiles? What does the usual regression quantile problem look like when expressed in terms of the empirical distributions? The answer turns out to be given by:

$$\beta^*(\tau) = \min_b \sum_{i=1}^{n} w_i \varrho_i(x_i b; \tau) \tag{8}$$

where

$$\varrho_i(v, \tau) = \int_{-\infty}^{v} \hat{F}_i(t)dt - v\tau \tag{9}$$

It can be verified that these result in the regression quantiles. This can be most easily verified by taking the (sub)derivative of (9) and noting that the summands reduce to

$$\frac{\partial \varrho_i(v, \tau)}{\partial v} = \hat{F}_i(v) - \tau$$

For a given τ, $\sqrt{N}(\beta^*(\tau) - \beta(\tau))$ is asymptotically normal under general dependence and heterogeneity.

3. Quantile regression model

In the general linear quantile regression model* specified as:

$$Q(\tau|x) = x^T\beta(\tau) \tag{10}$$

such models may be represented by the linear hypothesis:

$$\beta(\tau) = \alpha + \gamma F_0^{-1}(\tau) \tag{11}$$

for α and γ in $\mathbb{R}^p$ and F^{-1} a univariate quantile function. Thus, all p coordinates of the quantile regression coefficient vector are required to be affine functions of the same univariate quantile function, F_0^{-1}. Such models may be viewed as arising from linearly heteroscedastic model:

$$y_i = x_i^T\alpha + (x_i^T\gamma)u_i \tag{12}$$

with the $\{u_i\}$ iid from the distribution F_0.

The parameters are for a specified quantile $\tau \in [0, 1]$. The parameters vary with τ due to effects of the τ^{th} quantile of the unknown error distribution. Parameters estimated in linear quantile regression model have the same interpretation as those in any linear model. They are rate of change conditional on adjusting for the effects on the other variables in the model.

Quantiles are usually estimated by a process of ordering the sample data. Extension to the regression model was to recognition that quantiles could be estimated by an optimization function minimizing a sum of weighted absolute deviations, where the weights are asymmetric τ. They can be estimated by solving the linear programming problem:

$$\min_{b\in\mathbb{R}^p} \sum_{i=1}^{n} \varrho_\tau(y_i - x_i^T b) \tag{13}$$

When the vectors α and γ are fully specified under the null hypothesis tests may be formulated as suggested in Koenker and Machado (1999).

* Here we consider function of X that the linear in the parameters $Q_y(\tau|X) = \beta_0(\tau)X_0 + \beta_1(\tau)X_1 + \ldots + \beta_n(\tau)X_n$.

The simplest unconstrained form of regression quantiles estimates allows the predictor variables (X) to apply changes in central tendency, variance, and shape of the response variable (Y) distribution (Koenker, Machado 1999). When we estimate only a changes in central tendency (in mean) of the response variable (Y) distribution we have well known homogeneous variance regression model associated with least squares regression model. Ail the regression quantile slope $b_1(\tau)$ are for the common parameter, and any deviation among the regression quantiles estimates is simply due to sampling variation. An estimate of the rate of change in means from ordinary least squares regression is also an estimate of the same parameter as for regression quantiles. The intercept estimates $b_0(\tau)$ of the quantile regression model are for parametric quantile $\beta_0(\tau)$ of y when $X_1, X_2, \ldots, X_n = 0$, which differ across quantiles and for the mean μ.

When the predictor variables X exert changes in central tendency and in variance of the response variable (Y) distribution (Koenker, Machado 1999). We have a model with unequal variances (a location-scale model). As a consequence, changes in the quantiles of y across X cannot be the same for ail quantiles. Slope estimates $b_1(\tau)$ differ across quantiles because $\beta_1(\tau)$ differ, since the variance in y changes as a function of X. Note that the pattern of changes in estimates $b_0(\tau)$ mirror those for $b_1(\tau)$.

4. Nonparametric estimation of conditional quantiles

Denote, as earlier, the τ quantile of the distribution Y given $X = x$ as $Q_y(\tau|X)$ which solves:

$$F(Q_y(\tau|X)|x) = \tau \tag{14}$$

where $F(y|x)$ is the conditional cumulative distribution of Y given x evaluated at $Y = y$ an estimate $\hat{Q}(\tau|x)$ can be obtained from the observed pairs (X_i, Y_i) $(i = 1, \ldots, n)$ by solving (1) after replacing F with some estimate $\hat{F}$. One choice

of $\hat{F}$, which smoothes over X, is:

$$\hat{F}(y|x) = \frac{\sum_{i=1}^{n} K\{(X_i - x)/h\} I[Y_i \leq y]}{\sum_{i=1}^{n} K\{(X_i - x)/h\}} \tag{15}$$

where K is a kernel function and I is the indicator function, h is the bandwidth parameter. For chosen h we can use a cross validation approach to minimize the loss function:

$$L(h) = \sum_{i=1}^{n} \varrho_\tau(z)(Y_i - \hat{Q}_\tau^{(i)}(X_i)) \tag{16}$$

where $\varrho_\tau(z)$ can be interpreted as the loss function (Koenker, Bassett 1978) and $\hat{Q}_\tau^{(i)}$ denotes the estimate of $Q_\tau(X_i)$ using bandwidth h, where observation i has been dropped from a sample.

Equivalently, the nonparametric quantile regression estimator $\hat{Q}_\tau^{(i)}$ can be defined to minimize:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^p} K\left(\frac{x - x_i}{h}\right) \varrho_\tau\left(y_i - \hat{Q}_\tau^{(i)}\right) \tag{17}$$

over all Q_τ.

For nonparametric quantile regression Yu and Jones (1998) suggests the automatic bandwidth selection strategy for smoothing conditional quantiles, which minimizes mean squared error of the conditional quantile functions as follows:

$$h_\tau = h_{\text{mean}} \left\{\tau(1-\tau)/\phi(\Phi^{-1}(\tau))^2\right\}^{0,5} \tag{18}$$

where ϕ and Φ are the standard normal density and distribution functions.

Conclusion

While quantile regression and robust estimation are concerned with different aspects of data analysis, they have the shared objective of uncovering relationships missed by traditional data analysis. The robustness criterion translates into estimates that are unaffected by a small fraction of the data, and sampling distributions that stay good when hypothesized models are only approximately valid. Robust estimation is designed to deal with mistakes due to discrepant data. Quantile

regression is concerned with mistakes due to summarizing potentially disparate quantile effects into a single, potentially misleading, representation of the way y and x are related.

Quantile regression allows us to directly model conditional VaR, utilizing only the pertinent information that determines quantiles of interest. This is contrast with the traditional methods that use information on the central moments of conditional distribution – mean, variance, kurtosis, etc. – to construct the VaR estimates. From this point of view quantile regression is important for modeling intermediate and extremal conditional VaR.

References

Chernozhukoy V., Umantsev. L. (2001): Conditional Value-at-Risk: Aspects of Modeling and Estimation. "Empirical Economics", 26, pp. 271-292.

Koenker R., Bassett G. (1978): Regression Quantiles. "Econometrica", 46, pp. 33-50.

Koenker R., Hallok K.F. (2001): Quantile Regression. "Journal of Economic Perspective", 15, 4, pp. 143-156.

Koenker R., Machado G. (1999): Goodness of Fit and Related Inference Processes for Quantile Regression. "Journal of American Statistical Association", 94, 448, pp. 1296-1310.

Koenker R., Portnoy S. (1997): Quantile Regression. "Working Paper", 97-0100, University of Illinois, Urbana-Champaign.

Trzpiot G. (2003): Analiza portfelowa z wykorzystaniem metody momentów i dominacji stochastycznych – podejście kwantylowe. "Prace Naukowe", 990, AE, Wrocław, pp. 216-224.

Trzpiot G. (2004): Kwantylowe miary ryzyka. "Prace Naukowe", 1022, AE, Wrocław, pp. 420-430.

Trzpiot G. (2006): O nieparametrycznych metodach estymacji VaR i ETL. W: Modelowanie preferencji a ryzyko /05, AE, Katowice, pp. 229-238.

Trzpiot G. (2007): Regresja kwantylowa a estymacja VaR. "Prace Naukowe", 1176, AE, Wrocław, pp. 465-471.

Trzpiot G. (2008): Implementacja metodologii regresji kwantylowej w estymacji VaR. "Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania", nr 9, Uniwersytet Szczeciński, Szczecin, pp. 316-323.

Yu K. and Jones M.C. (1998): Local linear quantile regression. "Journal of American Statistical Association", 93, 441, pp. 228-237.

Abstract

In this paper the estimation of linear quantile regression model is presented. That kind of model can be used for modeling conditional VaR using only the pertinent information that determines quantiles of interest. Moreover, both the classical quantile regression models and nonparametric estimation approaches are shown.

Grażyna Trzpiot, Justyna Majewska*

SENSITIVITY ANALYSIS OF SOME ROBUST ESTIMATORS OF VOLATILITY

Introduction

Leptocurtic tails of data distributions and contamination of data with outliers are two features which very often characterize the financial time series. Consequently, the standard estimators, which are optimal for uncontaminated multivariate normal distributions, have very little chance to correctly estimate statistical parameters. In order to achieve stable and accurate estimates of parameters the robust estimators are required.

In the process of assets selection and their allocation to the investment portfolio the most important is the accurate evaluation of the volatility of the return rate. Lots of robust estimators of volatility are presented and analyzed in literature, so we want to compare and use some of them in the process of asset selection and their allocation to the investment portfolio.

In this paper we consider estimators with explicit formulas, satisfactory efficiency, 50% breakdown point.

The main goal of this paper is sensitivity analysis of selected robust estimators

*Research supported by the grant number KBN: N111 003 32/0262.

of volatility and the classification of generated investment portfolios with respect to chosen robust estimators. Selected methods of cluster analysis were used for the classification. We have tried to isolate homogeneous groups of similar portfolios as well as reveal relations between these portfolios.

Also authors try to convince that applying robust estimation in portfolio analysis ensures better method for effective investment decision-making than classical portfolio analysis. We proceed as follows. In Section 1, we review some basic concepts of robust statistics. In Section 2 we characterize some robust estimators of volatility. In Section 3 we give brief overview of the minimum-risk portfolio selection problem. In section 4 applications to simulated time series are presented before we give some conclusions.

1. Basic concepts of robust statistics

The pioneering work of Tukey (1960), Huber (1964), and Hampel (1968) has laid the ground for the theory of robust statistics. As a generalization of classical theory, robust statistics takes into account the possibility of model misspecification (i.e. model deviation). This theory and its results are valid at the model as well as in a neighborhood of the model, which is not the case for classical statistics.

The aim of robust statistics is to provide tools not only to assess the robustness properties of classical procedures, but also to produce new estimators and tests that are robust to model deviations.

Let us define:

$$\{F_{\varepsilon,G} | F_{\varepsilon,G} = (1-\varepsilon)F + \varepsilon G\}$$

where G is an arbitrary probability function and $\varepsilon \in [0,1]$, the set of all distributions defining a neighborhood of the parametric model F.

The neighborhood $F_{\varepsilon,G}$ of F represents data contamination (not all data follow the pre-specified distribution, but ε-part of data can come from a different distribution G). An estimator is robust if it remains stable in a neighborhood $F_{\varepsilon,G}$ of F.

One particular case is when $G = \delta_x$, the distribution that gives a probability of one to a point x chosen arbitrarily. In this case, the neighborhood of the model featuring all local nonparametric departures from F is given by $F_{\varepsilon,\delta_x} = (1-\varepsilon)F+\varepsilon\delta_x$. Hence F_{ε,δ_x} generates observations from F with probability $(1-\varepsilon)$ and observations equal to an arbitrary point x with probability ε.

Two main concepts for robust measures analyze the sensitivity of an estimator T to infinitesimal deviations ($\varepsilon \to 0$) and to finite (large) deviations ($\varepsilon > 0$).

The influence of infinitesimal contamination on an estimator is characterized by the *influence function* that measures the relative change in estimates caused by an infinitesimally small amount ε of contamination at x. Clearly, the relative effect on T is desired to be small or at least bounded. A functional $T(\cdot)$ with bounded influence function is regarded as robust.

The influence function is defined as:

$$IF(x;T,F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)+\varepsilon\delta_x) - T(F)}{\varepsilon}$$

The influence function allows to define various desirable properties of an estimation method. First, the largest influence of contamination on estimates can be formalized by the *gross-error sensitivity*:

$$\gamma(T,F) = \sup_{x \in \mathbb{R}} IF(x,T,F)$$

which under robustness considerations be finite and small.

Second, the sensitivity to small changes in data, for example moving an observation from x to $y \in \mathbb{R}$, can be measured by the *local-shift sensitivity*:

$$\lambda(T,F) = \sup_{x \neq y} \frac{\|IF(x;T,F) - IF(y;T,F)\|}{\|x-y\|}$$

This quantity should be relatively small since we generally do not expect that small changes in data cause extreme changes in values or sensitivity of estimates.

Third, as an unlikely large or distant observations may represent data errors, their influence on estimates should become zero. Such a property is characterized by the *rejection point*:

$$\varrho(T,F) = \inf_{r>0} \{r : IF(x;T,F) = 0,\ \|x\| \geq r\}$$

which indicates the non-influence of large observations.

Alternatively, behavior of the estimator T can be studied for any finite amount ε of contamination. A very broad measures of global robustness of T at F is the so-called *maximum bias* defined as:

$$B(\varepsilon; T, F) = \sup_G \|T(1-\varepsilon)F + \varepsilon G) - T(F)\|$$

$B(\varepsilon; T, F)$ measures the worst case bias due to an ε amount contamination of the assumed distribution. T is regarded as robust if it has a moderate maximum bias for small ε.

The most prominent is the breakdown point (Hampel 1971), which is defined as the smallest amount ε:

$$\varepsilon^*(T) = \min\{\varepsilon : B(\varepsilon; T, F) = \infty\}$$

Clearly, the higher the breakdown point of an estimator, the more robust the estimator against outliers. Besides, the intuitive aim of this definition specifies the breakdown point $\varepsilon^*(T)$ as the smallest amount of contamination that makes the estimator T useless. Note that in most cases $\varepsilon^*(T) \leq 0,5$ (He and Simpson 1993).

The most popular measure of volatility – variance – is not robust estimator. It is not robust globally and locally – its influence function is unbounded (that is an infinitesimal point mass contamination can have an arbitrarily large influence) and his breakdown point is $1/n$, the lowest possible value. It has an unbounded maximum bias for any $\varepsilon > 0$ and hence is not robust in terms of this maximum bias measure.

We remark that robustness is one of the most important performance criteria of a statistical procedure. There are, however, other important performance criteria. For example, efficiency is always a very important performance measure for any statistical procedure. Also issue of equivariance concepts is very significant (if the estimator is affected by location or scale transformation).

2. Some robust scale estimators

In this paper we will pay attention to scale estimators that have the following properties:

- an explicit formula,
- 50% breakdown point,
- a bounded influence function,
- affine equivariant*,
- being easily computable, using at most $O(n\log n)$ time and $O(n)$ storage.

Most existing scale estimators proposed in the literature fail one or more of these requirements. Especially the condition of 50% breakdown cuts away most estimators such as the interquartile range (which has 25% breakdown) and the trimmed standard deviation.

We will concern on scale estimators, which can be written as combinations of medians. This includes L- and U-statistics, but excludes most M- and R-statistics. We also want the estimators to be consistent for the scale parameter of gaussian distributions**. We will concern also on location-free estimators. Locatio-free estimators have the advantage that they do not implicitly rely on a symmetric noise distribution.

The scale of $X = (x_1, \ldots, x_n)$ is typically estimated by standard deviation which is very efficient for the assumed normal distribution, but highly sensitive to deviation from normality in a sample or empirical distribution.

The most commonly used robust estimator with 50% breakdown point (but location-based) is median absolute deviation about median (Hampel 1974)

$$MAD_n = a_n 1,4286 \underset{i}{med}\left\{\left|x_i - \underset{j}{med}(x_j)\right|\right\} \tag{2.1}$$

Here, a_n is a small sample correction factor that can be chosen depending on the

* Scale estimator S is affine equivariant if and only if $S(ax_1 + b, \ldots, ax_n + b = |a|S(x_1, \ldots, x_n)$ for arbitrary constants a and b.

** Usually this is achieved by premultiplying S by an appropriate factor C such that $cS(x_1, \ldots, x_n) \to 1$ when the observations are drawn from the standard gaussian distribution.

sample size to achieve unbiasedness. The MAD has become quite popular because of its simplicity and extremely good robustness properties. Its gross-error sensitivity is the smallest possible for a Fisher – consistent estimator – $1,17$. The asymptotic efficiency of the MAD is 37%, which is unusually low. Collins (1999) noted that the discontinuity of the influence function of the MAD causes its asymptotic variance to increase in an ε-contamination neighborhood with arbitrarily small ε resulting in even smaller efficiency – $14,5$%.

We want to determine whether any of these possess additional properties which the MAD does not have such as a continuous influence function (IF) or a better efficiency.

2.1. Location free estimators

A classical U-statistic is defined as the average of the $\binom{n}{k}$ values $\{\xi(x_{i_1}, \ldots, x_{i_k}); i_1 < i_2 < \ldots < i_k\}$ where k is the order of the kernel ξ. For scale estimation, we can use the generalized L-estimators:

$$\sum_{k=1}^{\binom{n}{2}} a_k \{|x_i - x_j|; i < j\}_{(k)} \tag{2.2}$$

with the kernel $\xi(x_i, x_j) = |x_i - x_j|$ of the second order*. Replacing the absolute values by squares we obtain a similar class of estimators given by:

$$\left(\sum_{k=1}^{\binom{n}{2}} a_k \left\{(x_i - x_j)^2; i < j\right\}_{(k)} \right)^{\frac{1}{2}} \tag{2.3}$$

Choosing all the weights equal to $\binom{n}{2}^{-1}$ yields the classical standard deviation in (2.3) and Gini's estimator in the case of (2.2).

We can obtain the maximal breakdown point if we choose coefficients a_k for which $a_k = 0$ when $k > \binom{h}{2}$ and such that there exists $\binom{h-1}{2} < k \leq \binom{h}{2}$ with $a_k > 0$ **.

*The subscript (k) means the k-th order statistic of the $\binom{n}{2}$ values in $\{|x_i - x_j|; i < j\}$.

**Note that $h = \left[\frac{n}{2}\right] + 1$ (h standing for half).

An example of a 50% breakdown estimator of type (2.2) is:

$$\binom{h}{2}^{-1} \sum_{k=1}^{\binom{h}{2}} a_k \{|x_i - x_j| ; i < j\}_{(k)} \tag{2.4}$$

and relatively to (2.3):

$$\binom{h}{2}^{-1} \left(\sum_{k=1}^{\binom{h}{2}} a_k \left\{ (x_i - x_j)^2 ; i < j \right\}_{(k)} \right)^{\frac{1}{2}} \tag{2.5}$$

To obtain consistency in gaussian models (2.4) and (2.5) must be multiply by 17,904 and 7,7405, respectively. Besides, their efficiencies are 81,45% and 81,55%, and their gross-error sensitivities amount to 2,0340 and 2,0416. A important drawback of these estimators is that they need $O(n^2)$ computation time. Therefore the following estimator is preferred (Croux and Rousseeuw 1992):

$$Q_n = b_n 2,2219 \{|x_i - x_j| ; i < j\}_{\binom{h}{2}} \tag{2.6}$$

which is a special case of both (2.4) and (2.5) and still attains the optimal breakdown point. Its asymptotic efficiency 82,27% which is better than that of (2.4) and (2.5). Its gross-error sensitivity 2,069 is a bit worse. Croux and Rousseeuw (1992) construct an $O(n \log n)$ – time algorithm for computing (2.6). They also obtain finite-sample correction factor b_n to make Q_n unbiased at small samples. Here, b_n is a small sample correction factor that can be chosen depending on the sample size to achieve unbiasedness and constant 2,2219 succeeds in making Q_n approximately unbiased for finite samples (for details see Croux and Rous--seeuw 1992).

2.2. Nested L-estimators

Apart from the generalized L-estimators there is also another way to robustify U-statistics. Instead of processing the kernels as one homogeneous set of data, it is also possible to carry out "nested" operations which eliminate one argument at a time.

To define nested L-estimators for kernels of order 2, two steps are required: computation the L-statistic for each observation x_i $(i = 1, \ldots, n)$:

$$H(x_i) = \sum_{k=1}^{n} a_k(\xi(x_i, x_j); i \neq j)_{(k)} \tag{2.7}$$

and the second, computation an L-statistic based on the $H(x_i)$ values:

$$\sum_{k=1}^{n} b_k(H(x_i); i = 1, \ldots, n)_{(k)} \tag{2.8}$$

Nested L-estimator has the maximal breakdown point if $a_h > 0, a_k = 0$ for $k > h$ and $b_k = 0$ for $k > h$. Its computation time is quite high, but in the special case of (2.7) given by $H(x_i) = \underset{j \neq i}{med} |x_i - x_j|$ it is possible to compute each $H(x_i)$ in $O(\log n)$ (Shamos 1976; Croux and Rousseeuw 1992). This leads to the following type of nested L-estimator:

$$\sum_{k=1}^{h} b_k \left\{ \underset{j \neq i}{med} |x_i - x_j|; i = 1, \ldots, n \right\}_{(k)} \tag{2.9}$$

which can be computed in $O(n \log n)$ time and has maximal breakdown (if $b_h > 0$).

We take into consideration the class of estimators:

$$S_n^{\alpha} = s_{\alpha} \left\{ \underset{j \neq i}{med} |x_i - x_j|; i = 1, \ldots, n \right\}_{[\alpha n]}$$

where $0 \leq \alpha \leq 0.5$.

The highest efficiency – 58.23%, is attained at $\alpha = 0,5$ and corresponding to the estimator:

$$S_n = S_n^{\alpha} = c_n 1.1926 \underset{i}{med} \underset{j \neq i}{med} |x_i - x_j| \tag{2.10}$$

the influence function of S_n is discontinuous and its gross error sensitivity is $1,625$. An algorithm for computation of S_n in $O(n \log n)$ time is described by Croux and Rousseeuw (1992). Here, c_n is a small sample correction factor that can be chosen depending on the sample size to achieve unbiasedness.

Taking the average of all S_n^{α} with $0 \leq \alpha \leq 0,5$ yields the another location-free 50% breakdown point estimator-trimmed mean of median deviations defined as:

$$T_n = 1,3800 \frac{1}{h} \sum_{k=1}^{h} \left\{ \underset{j \neq i}{med} |x_i - x_j| \right\}_{(k)} \tag{2.11}$$

which is of type (2.9) with constant coefficients b_k. The asymptotic efficiency of this estimator is about 52% and its gross-error sensitivity becomes $1,4578$. Therefore, T_n is less efficient, but more robust than S_n, for which $\gamma = 1,625$. An important advantage of T_n over S_n is that its influence function is continuous and yields a finite local-shift sensitivity.

2.3. Scale estimators based on contiguous subsamples

A different type of location-free scale estimator is given by:

$$C_n^\alpha = d_n \left| x_{(i+[\alpha n]+1)} - x_{(i)} \right|_{([n/2]-[\alpha n])} \tag{2.12}$$

where $0 < \alpha < 0,5$ and $x_{(1)} \leq \ldots \leq x_{(n)}$ are the order statistics. The constant d_n is needed to make the estimator Fisher-consistent at gaussian distributions.

For α satisfying $[\alpha n] = [n/2] - 1$ the estimator C_n^α becomes:

$$LMS_n = 0,7413 \min_i \left| x_{(i+[n/2])} - x_{(i)} \right| \tag{2.13}$$

This estimator first occurs as a scale part of the least median of squares (LMS) regression estimator. We can interpret it as the length of the shortest half sample.

It had the same influence function and thus the same asymptotic efficiency as the MAD. For small samples its efficiency is larger than that of the MAD and its maximum bias in case of many outliers is smaller than that of the MAD.

Replacing the range of a subsample as in (2.12) by standard deviation, yielding the estimators:

$$D_n^\alpha = e_n std \left\{ x_{(i)}, \ldots, x_{(i+[n/2]+1)} \right\}_{([n/2]-[\alpha n]} \tag{2.14}$$

where $0 < \alpha < 0,5$. For α satisfying $[\alpha n] = [n/2] - 1$ we get least trimmed squares estimator:

$$LTS_n = 2,6477 \min_i std \left\{ x_{(i)}, \ldots, x_{(i+[n/2]+1} \right\} \tag{2.15}$$

For the scale estimator we obtain an asymptotic efficiency of $30,67\%$, which is less than the $36,74\%$ efficiency of the scale estimator.

3. The traditional approach to portfolio optimization

The fundamental goal of the portfolio theory is to optimally allocate investments to different assets. Mean-variance optimization is a quantitative tool, which allows to make this allocation by considering the trade-off between risk and return. However, since the covariance matrix can be estimated much more precisely than the expected returns the minimum variance portfolios are usually more stable because the composition of the minimum variance portfolio depends only on the covariance matrix of asset returns.

The classical Markowitz optimization problem, which constitutes the main theoretical background for the modern portfolio theory is widely described and analyzed in literature, so we will just briefly recall the minimum-variance problem.

For given n risky assets the minimum-variance portfolio is the portfolio of assets that minimizes risk measured by the variance of portfolio return for a given covariance matrix C. It is a solution to the following problem:

$$\min_{x=(x_1,\ldots,x_n)^T} x^T C x \quad \text{s.t. } x \in X \tag{3.1}$$

where $x \in \Re^n$ is the vector of portfolio weights.

The simplest non-empty and bounded set X of feasible portfolios are usually considered as:

$$X = \left\{ x \in \Re^n : \sum_{i=1}^{n} x_i = 1, x \geq 0 \right\}$$

4. Simulation results

The present section is devoted to a comparison of selected risk estimators* an analysis and classification of generated investment portfolios with respect to chosen robust estimation.

*Risk is measured by volatility of returns.

For our experiment we use market indexes: WIG20, WIRR, and MIDWIG* of the Polish Stock Exchange Members.

We have generated 1000 weekly return following t-student distribution (parameters were estimated weekly based on returns of each index WIG20, WIRR, MIDWIG). For this purpose we used the Monte Carlo simulation. In order to reduce estimation errors we have chosen a weekly periodicity for the rates of return (Simaan 1997).

We have considered the following types of database: without contamination, next we used 2%, 4%, 6%, 8%, and 10% percentage of contamination level. The point mass multiplicative contamination have been studied, that relies on random multiplication of 2%, 4%, 6%, 8%, and 10% of the asset returns by specific value – three times the estimated standard deviation. The contamination occurs for each of the three series at the same data points.

For each dataset we have established the risk estimators described in the Section 2 (see Table 1).

After analysing the results form Table 1 and Figure 1 we observed that MAD, Q_n, T_n, and S_n estimators offer a stable estimate of volatilities. Whereas, together with increasing the percentage of contamination the standard deviation changes substantially.

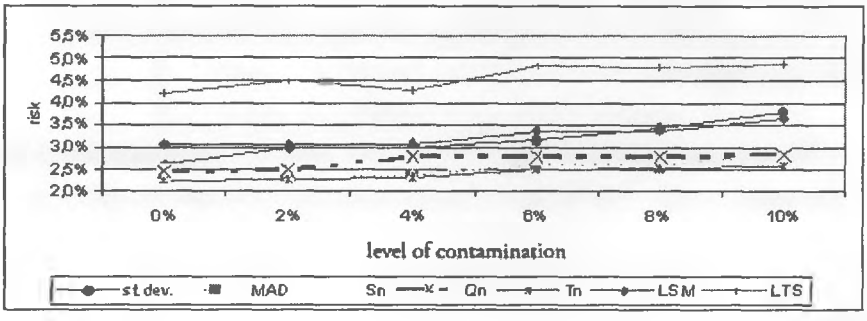


Figure 1. The impact of contamination on the risk level

*WIRR and MIDWIG were replaced by mWIG40 and sWIG80 on the Polish Stock Exchange Members from March 2007.

Table 1

Risk estimators calculated for six datasets

Amount of conta--mination	Indexes deviation	Standard	MAD_n	S_n	Q_n	T_n	LMS_n	LTS_n
0%	WIG20	0,0259	0,0232	0,0236	0,0243	0,0225	0,0305	0,0420
	MIDWIG	0,0216	0,0194	0,0204	0,0195	0,0194	0,0233	0,0397
	WIRR	0,0263	0,0254	0,0265	0,0261	0,0258	0,0331	0,0418
2%	WIG20	0,0299	0,0232	0,0237	0,0248	0,0227	0,0305	0,0348
	MIDWIG	0,0246	0,0194	0,0237	0,0218	0,0227	0,0239	0,0301
	WIRR	0,0298	0,0260	0,0274	0,0271	0,0260	0,0339	0,0453
4%	WIG20	0,0298	0,0237	0,0239	0,0278	0,0231	0,0309	0,0427
	MIDWIG	0,0241	0,0201	0,0206	0,0223	0,0198	0,0242	0,0375
	WIRR	0,0304	0,0271	0,0282	0,0301	0,0273	0,0345	0,0462
6%	WIG20	0,0316	0,0262	0,0273	0,0278	0,0250	0,0338	0,0482
	MIDWIG	0,0277	0,0210	0,0227	0,0223	0,0209	0,0257	0,0316
	WIRR	0,0340	0,0295	0,0308	0,0301	0,0293	0,0366	0,0399
8%	WIG20	0,0339	0,0258	0,0268	0,0276	0,0251	0,0338	0,0478
	MIDWIG	0,0271	0,0192	0,0227	70,0225	0,0217	0,0255	0,0389
	WIRR	0,0319	0,0277	0,0287	0,0292	0,0275	0,0354	0,0406
10%	WIG20	0,0380	0,0279	0,0277	0,0283	0,0256	0,0365	0,0487
	MIDWIG	0,0306	0,0211	0,0277	0,0249	0,0256	0,0277	0,0376
	WIRR	0,0367	0,0287	0,0310	0,0315	0,0283	0,0372	0,0421

Source: Own calculations.

In the next stage we solved the minimum risk model (3.1) changing every time the estimators of risk. The Tables 2-7 present optimal minimum-risk portfolios.

Table 2

Shares of optimal portfolios for dataset without contamination

Risk estimator	WIRR	WIG20	MIDWIG	Portfolio risk
Standard deviation	25,60%	40%	34,40%	1,86%
MAD	25,46%	40%	34,54%	1,72%
S_n	26,68%	40%	33,32%	1,79%
Q_n	24,92%	40%	35,08%	1,76%
T_n	27,80%	40%	32,20%	1,72%
LMS_n	25,83%	40%	34,17%	2,18%
LTS_n	26,72%	33,28%	40%	3,15%

Source: Own calculations.

Table 3

Shares of optimal portfolios for dataset with 2% of contamination

Risk estimator	WIRR	WIG20	MIDWIG	Portfolio risk
Standard deviation	20,58%	40%	39,42%	2,30%
MAD	25,51%	40%	34,49%	1,87%
S_n	26,47%	39,55%	33,98%	2,07%
Q_n	26,54%	40%	33,46%	2,02%
T_n	27,12%	39,13%	33,75%	1,97%
LMS_n	28,58%	40%	31,42%	2,40%
LTS_n	38,48%	40%	21,52%	2,96%

Source: Own calculations.

Table 4

Shares of optimal portfolios for dataset with 4% of contamination

Risk estimator	WIRR	WIG20	MIDWIG	Portfolio risk
Standard deviation	22,31%	40%	37,69%	2,29%
MAD	26,23%	40%	33,77%	2,18%
S_n	27,83%	40%	32,17%	1,98%
Q_n	26,63%	40%	33,37%	2,18%
T_n	27,68%	40%	32,32%	1,91%
LMS_n	28,96%	40%	31,04%	2,43%
LTS_n	25,87%	40%	34,13%	3,47%

Source: Own calculations.

Table 5

Shares of optimal portfolios for dataset with 6% of contamination

Risk estimator	WIRR	WIG20	MIDWIG	Portfolio risk
Standard deviation	27,69%	40%	32,31%	2,67%
MAD	27,07%	40%	32,93%	2,17%
S_n	27,89%	40%	32,11%	2,30%
Q_n	27,80%	40%	32,20%	2,28%
T_n	27,74%	40%	32,26%	2,13%
LMS_n	29,12%	40%	30,88%	2,72%
LTS_n	20,00%	40%	40,00%	3,30%

Source: Own calculations.

Table 6

Shares of optimal portfolios for dataset with 8% of contamination

Risk estimator	WIRR	WIG20	MIDWIG	Portfolio risk
Standard deviation	20%	40%	40%	2,61%
MAD	26,96%	40%	33,04%	2,04%
S_n	25,84%	40%	34,16%	2,22%
Q_n	26,09%	40%	33,91%	2,24%
T_n	27,70%	40%	32,30%	2,11%
LMS_n	24,84%	40%	35,16%	2,67%
LTS_n	20%	40%	40%	3,56%

Source: Own calculations.

Table 7

Shares of optimal portfolios for dataset with 10% of contamination

Risk estimator	WIRR	WIG20	MIDWIG	Portfolio risk
Standard deviation	20%	40%	40%	2,61%
MAD	26,96%	40%	33,04%	2,04%
S_n	25,84%	40%	34,16%	2,22%
Q_n	26,09%	40%	33,91%	2,24%
T_n	27,70%	40%	32,30%	2,11%
LMS_n	24,84%	40%	35,16%	2,67%
LTS_n	20%	40%	40%	3,56%

Source: Own calculations.

From the results analysis of Tables 2-7 some conclusions may be drawn:

- portfolios based on MAD-, Q_n-, T_n-, and S_n-estimators have the most stable weights in the presence of contamination of data,
- portfolios based on standard deviation is sensitive to the presence of contamination,

– changes in S_n-risk and T_n-risk portfolios weights are practically indistinguishable in every the cases,

– portfolios weights are the same for over 6% of contamination based on LTS_n-estimator.

Finally, we have tried to classify generated investment portfolios with respect to chosen robust estimators. The results are shown on the tree diagrams.

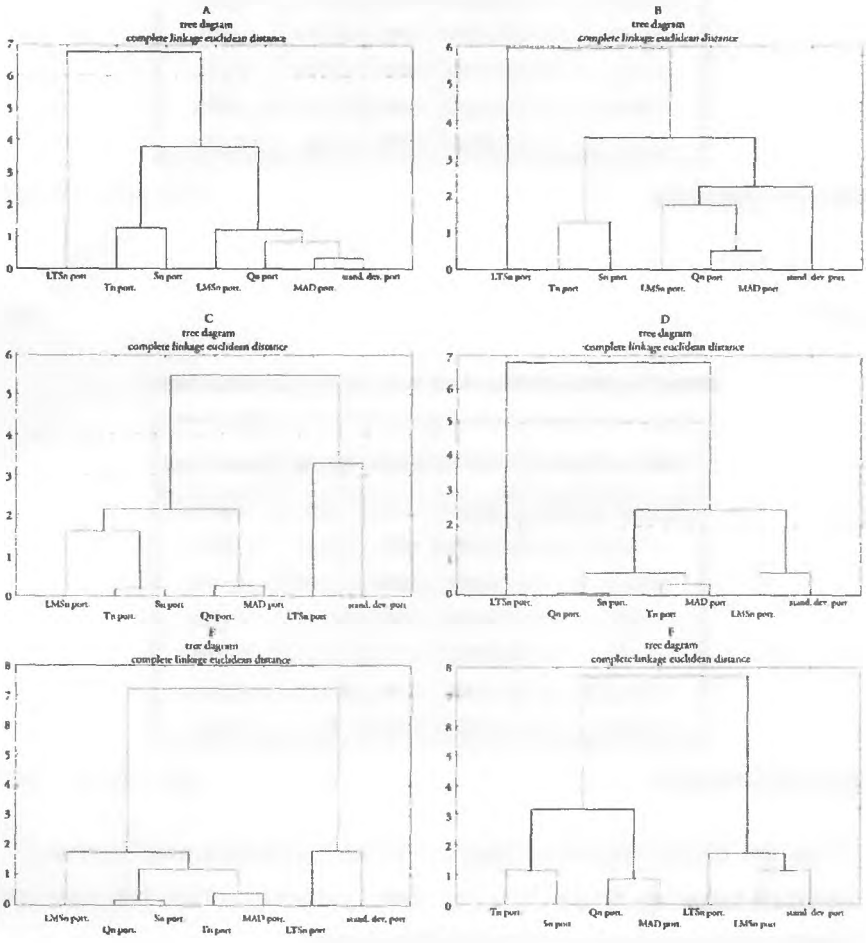


Figure 2. A–F) Tree diagram portfolios based on robust estimators in the case of 0%-10% contamination

One may notice homogeneous groups of similar portfolios. Note that in the case of 0% and 2% (Figure 2 A and B) there are the same 3 clusters. First cluster consists of portfolios based on standard deviation, MAD, Q_n, LMS_n. The second cluster consists of portfolios based on S_n- and T_n-estimators and the last cluster of portfolios based on the LTS_n estimator.

It is interesting that with increasing level of contamination portfolios based on MAD-, Q_n-, S_n-, and T_n-estimators belong to the same group. For the highest considered contamination there are only two clusters (see Figure 2F).

5. Concluding remarks

Atypical observations (outliers) and sudden changes in the financial time series can be detected if we additionally apply a reliable estimator of scale. Robust estimators are powerful tools for stable evaluation of statistical parameters. Especially in the process of assets selection and their allocation to the investment portfolio the most important is the accurate evaluation of the volatility of the return rate.

If volatility is a measure of risk, then the Q_n-, S_n-, and T_n-estimators, but also location-based MAD are seemed to be the most promising among analyzed volatility estimators. However, the final choice between Q_n, S_n, and T_n depends on personal taste because their advantages and disadvantages are hard to compare. For instance, Q_n, S_n have a higher efficiency than T_n. On the other hand, Q_n and T_n have a continuous influence function, unlike S_n. And finally, it turns out that our algorithms for S_n and T_n need only half as much computation time and storage space as Q_n.

In this paper homogenous groups of similar portfolios have been obtained as the result of classification. Portfolios based on Q_n-, S_n-, T_n-, and MAD--estimators of risk have the most stable weights in the presence of contaminated data. Achieved results can be used in the investment decisions-making process.

References

Collins J.R (1999): Robust M-estimators of Scale: Minimax Bias Versus Maximal Variance. "Canadian Journal of Statistics", 27.

Hampel F.R. (1971): A General Qualitative Definition of Robustness. "Annals of Mathematical Statistics", Vol. 42, No. 6.

Hampel F.R. (1974): The Influence Curve and Its Role in Robust Estimators. "JASA", 69.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986): Robust Statistics: The Approach Based on Influence Function. Wiley & Sons Inc., New York.

He X., Simpson D.G. (1993): Lower Bounds for Contamination Bias: Globally Minimax Versus Locally Linear Estimation. "The Annals of Statistics", 21(1).

Huber P. (1964): Robust Estimation of a Location Parameter. "Annals of Mathematical Statistics", 53.

Rousseeuw P.J., Croux C. (1991): Alternatives to the Median Absolute Deviation, Technical Report. Department of Mathematics and Computer Science, Universitaire Instelling, Antwerpen.

Rousseeuw P.J., Croux C. (1992): Explicit Scale Estimators with High Breakdown Point L_1 – Statistical Analysis and Related Methods. Y. Dodge (ed.). Amsterdam, North-Holland.

Shamos M.I. (1976): Geometry and Statistics: Problems at the Interface. In: New Directions and Recent Results in Algorithms and Complexity. J.F. Traub (ed.). Academic Press, New York.

Simaan Y. (1997): Estimation Risk in Portfolio Selection: The Mean Variance Model and the Mean-Absolute Deviation Model. "Management Science", 43.

Tukey J.W. (1960): A Survey of Sampling from Contaminated Distributions. In: Contributions to Probability and Statistics. I. Olkin (ed.). Stanford University Press, Palo Alto, pp. 448–485.

Abstract

Leptokurtotic tails of data distributions and contamination of data with outliers in financial time series are the reasons for adapting robust methods to constructing effective investment portfolios. In this paper we present the sensitivity analysis of selected robust estimators of volatility and the classification of generated investment portfolios with respect to chosen robust estimators.

Grażyna Trzpiot, Dominik Krężołek*

QUANTILES RATIO RISK MEASURE FOR STABLE DISTRIBUTIONS MODELS IN FINANCE

Introduction

The volatility analysis of financial assets becomes the new area in risk management. Unexpected and unpredictable events observed in financial markets exhibits higher risk level and therefore its forecasts can be biased. It becomes a big challenge not only for individual investors, but also for the companies and can produce very huge financial losses, including bankruptcy. The special role in risk management plays the Value-at-Risk (VaR) approach.

Classical financial models assume that the empirical log-returns are normally distributed, but the reality verifies this assumption and the classical statistical models are inappropriate. The most relevant features to be observed in time series of market returns are as follow: high volatility, clustering, fat tails, leptokurtosis, leverage effects, serial correlations and hence the normality assumption has to be rejected. In the beginning of 60s Mandelbrot proposed the class of distributions satisfying mentioned features – the stable distributions**. The term stable

*Research supported by the grant number KBN: N111 003 32/0262.

**In financial area; stable laws was introduced by Lévy in 1925 during the investigation of the behavior of sums of independent random variables.

refers to the property of these models*. However, due to the lack of closed form of stable densities (except Cauchy, Lévy, and Gauss distributions) its applicability in financial time series analysis is still limited. The solution arises from very advanced computer tools and technics, therefore the use of stable distributions to modeling financial data can be extended.

The purpose of this paper is to present some quantiles ratio risk measures of financial assets. These measures are based on the VaR approach. Additionally the assumption of stable distributed log-returns is used. It allows to estimate Investment risk more accurate.

1. Methodology

1.1. Stable distributions

Stable distributions (also called alfa-stable, stable Paretian) has been proposed as distributions to describe financial phenomena in the work of Mandelbrot (1963) and Fama (1965). While studying empirical distributions of stock returns they found the excess kurtosis and this find led them reject the assumption of normality. The strongest statistical argument for use stable models is the Central Limit Theorem which states that stable laws are only possible limit distributions for properly normalized and centered sums of independent and identically distributed random variables (although exist other heavy-tailed distributions as hyperbolic, Student or truncated stable) (Borak, Hrdle, Weron 2005-2008). This class of models accommodate to heavy-tailed financial time series capturing also skewness in a distribution.

As mentioned previously the history of stable distributions dates the beginning of 60s, but the main barrier in use of these models was the lack of its densities. Hence the estimation of all alfa-stable models is approximate in the sense that the

*The stability is considered in terms of probability scheme; the sum of two stable distributed random variables is also the stable distributed random variable (the summation scheme).

alfa-stable density function is approximated via Inverse Fourier Transform. The most common* characterization of stable distribution can be obtained through its characteristic function (Borak, Hrdle, Weron 2005-2008, p. 4):

$$\ln \varphi(t) = \begin{cases} i\delta t - \gamma^{\alpha}|t|^{\alpha}\left[1 - i\beta \operatorname{sign}(t)\operatorname{tg}\left(\frac{\alpha\pi}{2}\right)\right], & \text{dla} \quad \alpha \neq 1, \\ i\delta t - \gamma|t|\left[1 + i\beta \operatorname{sign}(t)\frac{2}{\pi}\ln|t|\right], & \text{dla} \quad \alpha = 1, \end{cases} \tag{1}$$

where

$$\operatorname{sign}(t) = \begin{cases} 1 & \Leftrightarrow t > 0 \\ 0 & \Leftrightarrow t = 0 \\ -1 & \Leftrightarrow t < 0 \end{cases}$$

To describe the density of stable distributions four parameters are used. The most relevant is the index of stability α, responsible for the thickness of tail and satisfying $(0, 2\rangle$. If $\alpha \to 0$ then the analyzed distribution and the normal one differs more (if normal, then $\alpha = 2$) (Racher, Hittnik 2000, p. 25). The last three parameters $\beta \in \langle -1, 1\rangle, \gamma > 0, \delta \in \mathbb{R}$ describe skewness, scale and location of stable density, respectively. The value of α satisfying $\alpha \in (1, 2)$ confirms heavy-tailed data. The important feature of stable distributions is the existence of moments of stable random variables. If X denotes stable random variable and $\alpha < 2$, then $E(X) = \delta$ and $D^2(X) = \infty$. In the case where $\alpha < 1$ even $E(X) = \infty$. Thus, the p-th moment of stable random variable exists if and only if $p < \alpha : E(X)^p < \infty \Leftrightarrow p < \alpha$.

The most common approach that gives rise to stability is based on probabilistic scheme, basically on the summation scheme, and can be expressed as follow: if $X_1, X_2, \ldots$ are independent and identically distributed real-valued random variables, then the summation scheme satisfies** $X_1 \stackrel{d}{=} a_n \sum_{i=1}^{n} X_i + b_n, a_n > 0, b_n \in \mathbb{R}$. Apart from summation, the geometric scheme is often in use. In this approach the number of terms $\nu(p)$ is geometrically distributed and the scheme satisfies $X_1 \stackrel{d}{=} a(p) \sum_{i=1}^{\nu(p)} X_i + b(p), a(p) > 0, b(p) \in \mathbb{R}$. Moreover $P(\nu(p) = k) = p(1-p)^{k-1}, k = 1, 2, \ldots$. The number of terms can be interpreted as the

*The approach proposed by J. Nolan in "Stable distribution" (2005).

**Notation $\stackrel{d}{=}$ denotes equality in distribution.

moment at which the probabilistic structure governing the asset returns breaks down*. Mittnik and Rachev (1991) show that geometric stable random variable can be considered in terms of alfa-stable random variable using its characteristic function:

$$\psi(t) = (1 - \log \varphi(t))^{-1} \tag{2}$$

where $\varphi(t)$ denotes alfa-stable random variable (Kozubowski, Rachev 1999, p. 178).

The application of stable distributions in risk management theory is confirmed to be more well-founded than the normal ones. If the stability approach in Markovitz optimalization is considered, then the estimates of expected returns and risk are better fitted to the real level of these characteristics than in the normal case (Rachev, Han 2000, p. 341). In this paper risk is considered as a measure based on quantiles and the most popular is the Value-at-Risk. This risk measure can be defined as a level of financial loss where the probability of achieving or exceeding this level in a given horizon of time is equal to the given confidence level α. As its seen, VaR is a function of corresponding $1 - \alpha$ quantile in the distribution of analyzed series.

The methodology of estimate VaR is very wide and includes models, where the normality assumption is strongly supported. In this paper the calculation of VaR based on estimation of quantile in arbitrary distribution is used.

1.2. Quantiles ratio risk measures

Risk management in a financial distribution analysis is the basis in decision-making process. Its importance arises from the history of world's financial crashes, when the biggest companies suffered bankruptcy. Nonetheless the risk is not defined classically at the moment and has statistical interpretation. Statistically, the most popular risk measure is variance (standard deviation) and its variants (semi-variance, safety level, aspiration level or coefficient of variation). Al-

* As a result of new information, market crash, or catastrophe.

though these risk measures play very important role in financial analysis, their use is supported by the normal distribution. As was presented earlier, financial time series characterize kurtosis, asymmetry, or outliers and for that reason normal distribution is inappropriate. If focused on unexpected events (market volatility, political situation, catastrophes, terrorist threat), risk measurement based on normality is biased, as the probability that one of these events occur is higher than if the normality assumption is supported. Hence, considering heavy-tailed distribution, the proper risk measure should be analyzed using the tail of distribution. Thus, quantiles risk measures are of interest.

The analysis and risk assessment is carried out using quantiles ratio risk measures (including VaR and its derivates). Let X be a random variable representing the log-return of an asset or portfolio (equally-weighted) with cumulative distribution function (cdf) F. Let $VaR_\alpha(X)$ denotes one-period Value-at-Risk with α confidence level. Hence the *expected shortfall* $ES_\alpha(X)$ and *median shortfall* $MS_\alpha(X)$ have the form:

$$ES_\alpha(X) = CVaR_\alpha(X) = E\left[X - VaR_\alpha(X) | X > VaR_\alpha(X)\right] \tag{3}$$

$$MS_\alpha(X) = Median\left[X - VaR_\alpha(X) | X > VaR_\alpha(X)\right] \tag{4}$$

From (3) and (4) results that for any given α exists one-step-ahead prediction representing expected value* of loss beyond the VaR_α. Moreover using (3) and (4) the new measures can be defined:

$$m_\alpha(X) = \frac{VaR_\alpha(X) + ES_\alpha(X)}{VaR_\alpha(X)} \tag{5}$$

$$m^*_\alpha(X) = \frac{VaR_\alpha(X) + MS_\alpha(X)}{VaR_\alpha(X)} \tag{6}$$

which represent expected (in terms of expected value and median) total loss of a portfolio standardized by its VaR_α.

It can be shown (Vaz de Melo Bends, Martins de Souza 2004, p. 32) that if F is a standard normal cdf of random variable X, then $m_\alpha(X)$ tends to 1 as α tends

*In terms of expected value and median.

to 0. Moreover, if X has a t_n – Student distribution, then $m_\alpha(X) \to n/(n-1)$ as $\alpha \to 0$. In this case $m_\alpha(X)$ is greater than 1 even if $\alpha \to 0$. Additionally, it can be shown that if random variable Z represents the sum of iid random variables X and Y (which can represent the components of a portfolio) then:

$$D(X,Y) = VaR_\alpha(X) + VaR_\alpha(Y) - VaR_\alpha(X+Y) \tag{7}$$

This measure informs if there is some benefit in diversifying a portfolio.

2. Empirical analysis

The assessment of risk level using the concept of VaR based on estimation of quantile in arbitrary distribution is modeled in the WSE. The daily log-returns of WIG and WIG20 in the period of 3^{rd} of January 2000 – 29^{th} of December 2006 (1756 observations) are considered. The analysis is based on three scenarios: for both indexes separately and for the portfolio (considered as a linear combination of equally-weighted log-returns of WIG and WIG20).

Table 1

The Kolmogorov-Smirnov test of normality

Statistics	WIG	WIG20	PORTFOLIO
Expected value	0,00056	0,00033	0,00044
Standard deviation	0,01278	0,01567	0,01373
K-S statistics	1,84430	2,07255	1,66274
p-value	0,00222	0,00037	0,00794

Source: Own calculations.

The K-S test confirms discrepancy with the normal distribution (at 0,01 significance level). The parameters of stable distributions for empirical log-returns are estimated using MME. The results are presented in Table 2.

Table 2

The parameters of stable distributions

Parameter	WIG	WIG20	PORTFOLIO
α	1,78739	1,80872	1,84480
β	0,06005	0,03081	0,01224
γ	0,00794	0,00996	0,00889
δ	0,00073	0,00031	0,00047

Source: Own calculations.

These results show that both indexes and portfolio are heavy-tailed and moreover have right-side asymmetry. The location parameter representing the expected return has the highest level for WIG.

The goodness of fit for stable distributions to empirical data is presented graphically (using histograms and QQ-plots). As it's seen, the fitting to the stable models is good and compiles with leptokurtosis and heavy tails. Additionally it's confirmed by the QQ-plots.

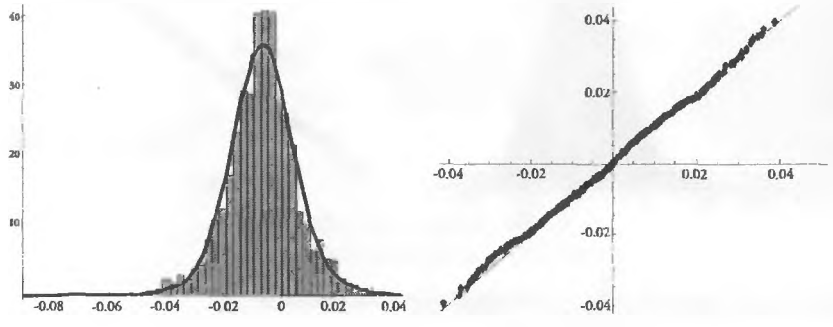


Figure 1. Stable distribution (left) and QQ-plot (right) – WIG

Source: Own calculations.

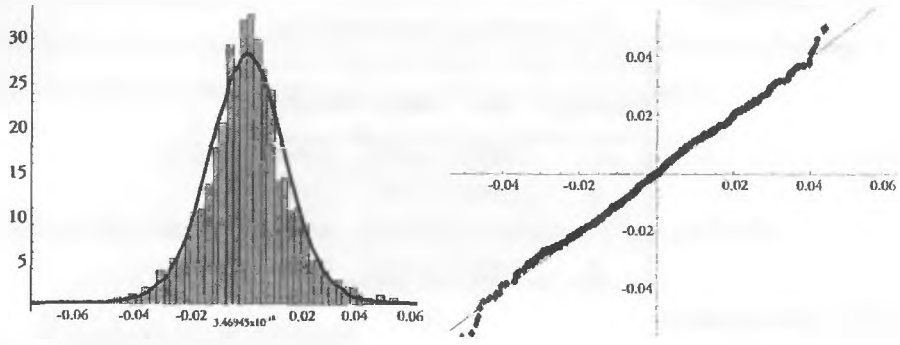


Figure 2. Stable distribution (left) and QQ-plot (right) – WIG20

Source: Own calculations.

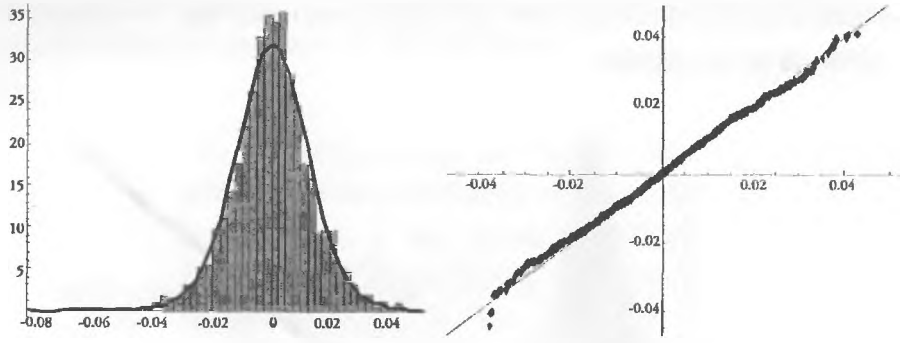


Figure 3. Stable distribution (left) and QQ-plot (right) – PORTFOLIO

Source: Own calculations.

The main purpose of this analysis is to use some quantiles ratio risk measures based on stable distributions for quantiles risk measures. The quantiles of stable distributions for both indexes and portfolio are noticed in Table 3.

Table 3

Stable quantiles

Quantile	WIG	WIG20	PORTFOLIO
0,001	-0,100564	-0,119033	-0,092484
0,01	-0,033301	-0,041289	-0,034715
0,05	-0,019169	-0,024492	-0,023403

Source: Own calculations.

The results presented above are then used to calculate the VaR. If the ap-proach of arbitrary distribution for VaR is considered, the risk measures (2)-(6) which was described as were calculated (Tables 4-6).

Table 4

Quantiles ratio measures – WIG

Risk measure	Quantiles		
	0,05	0,01	0,001
VaR_α	0,0192	0,0333	0,1006
ES_α	0,0265	0,0411	0,0699
MS_α	0,0237	0,0390	0,0699
$m_\alpha(X)$	2,3828	2,2328	1,6953
$m^*_\alpha(X)$	2,2360	2,1713	1,6953

Source: Own calculations.

Table 5

Quantiles ratio measures – WIG20

Risk measure	Quantiles		
	0,05	0,01	0,001
VaR_α	0,0245	0,0413	0,1190
ES_α	0,0334	0,0497	0,0760
MS_α	0,0297	0,0425	0,0760
$m_\alpha(X)$	2,3635	2,2042	1,6384
$m^*_\alpha(X)$	2,2142	2,0290	1,6384

Source: Own calculations.

Table 6

Quantiles ratio measures – PORTFOLIO

Risk measure	Quantiles		
	0,05	0,01	0,001
VaR_α	0,0234	0,0347	0,0925
ES_α	0,0329	0,0626	0,0211
MS_α	0,0264	0,0366	0,0211
$m_\alpha(X)$	2,4050	2,8044	1,2284
$m^*_\alpha(X)$	2,1284	2,0529	1,2284
$D(X,Y)$	0,0203	0,0399	0,1271

Source: Own calculations.

For calculation all risk measures presented in Tables 4-6 one-period forecast horizon is used. Considering risk in terms of the classical VaR, the smallest loss in one-period VaR gains WIG (with tolerance level 0,01 and 0,05, respectively). The similar conclusions give conditional VaR, considered in terms of expected value and median. In this case also investments in WIG represent the lowest level of risk. But if the 0,001 tolerance level is of interest, the lowest losses are generated by diversification (portfolio investment). It confirms the classical approach given by Markovitz in the portfolio theory. The ratios representing total loss of a portfolio standardized by its value of VaR have the lowest level for WIG20 (in the case of expected shortfall) and for a portfolio (in the case of median shortfall).

Conclusions

Analyzing empirical stock returns in terms of their statistical distributions it's confirmed that outliers play very significant role in risk management. Therefore, it's necessary to use the distributions with heavier tails than the normal one. Empirical investigations support the alfa-stable distribution as a describing financial time series (see, for example, Mittnik, Paolella, Rachev 2000). As a part of the

family of stable distributions, the geometric stable once are used also. This subclass of distributions allows for unexpected events considered as a market break down. Moreover, geo-stable distributions exhibit very good fitting to empirical data, even if consider emerging markets (see Kozubowski 1999 – the case of exchange rates, Krężołek 2007 – the case of portfolio analysis), characterized by high volatility of price changes.

In this paper the unclassical approach for risk measure is presented. To calculate the VaR value the estimation of quantiles in arbitrary distribution is used (it allows to use the alfa-stable models). Stable models, as is confirmed in this paper, shows better fit to empirical data than the normal ones, complying with leptokurtosis, heavy tails and asymmetry. These features are very important in terms of risk assessment. Moreover, risk measures are considered as a quantiles ratios based on VaR approach (expected shortfall and median shortfall). Every of these measures have defects and advantages and should be interpreted very carefully. However, the results confirm the advantage of diversifying a portfolio in terms of risk level reduction.

References

Borak Sz., Hrdle W., Weron R. (2005-2008): Stable Distributions. SFB 649 Economic Risk, Discussion Paper, Berlin.

Fama E. (1965): The Behavior of Stock Market Prices. "Journal of Business", No. 38, pp. 34-105.

Kozubowski T. (1999): Geometric Stable Laws: Estimation and Application. "Mathematical and Computer Modelling", No. 29, pp. 241-253.

Kozubowski T., Rachev S. (1999): Univariate Geometric Stable Laws. "Journal of Computational Analysis and Applications", Vol. 1, No. 2, pp. 177-217.

Krężołek D. (2007): Zastosowanie asymetrycznego rozkładu Laplace'a do budowy portfeli inwestycyjnych na polskim rynku kapitałowym. "Dynamiczne Modele Ekonometryczne", pp. 245-252.

Mandelbrot B. (1963): The Variation of Certain Speculative Prices. "Journal of Business", No. 36, pp. 394-419.

Mittnik S., Paolella M., Rachev S. (2000): Diagnosis and Treating the Fat Tails in Financial Returns Data. "Journal of Empirical Finance", pp. 389-416.

Mittnik S., Rachev S. (1991): Alternative Multivariate Stable Distributions and Their Applications to Financial Modeling. Stable Processes and Related Topics, Birkhaser, Boston, pp. 107-119.

Rachev S., Han S. (2000): Portfolio Management with Stable Distribution. "Mathematical Methods of Operation Research", No. 50, pp. 341-352.

Rachev S., Mittnik S. (2000): Stable Paretian Models in Finance. Series in Financial Economics and Quantitative Analysis, John Wiley & Sons, Ltd., England.

Vaz de Melo Bends B., Martins de Souza R. (2004): Measuring Financial Risk with Copulas. "International Review of Financial Analysis", No. 13, pp. 27-45.

Abstract

This article presents some quantile risk ratio measures based on unclassical VaR approach (expected and median shortfall). The stable distributed log-returns of Polish indexes WIG and WIG20 are used. The results shows clear lead of stable distribution over the normal one (especially in rerms of VaR calculation).

Alicja Ganczarek-Gamrot

VECTOR AUTOREGRESSIVE MODELS ON THE POLISH ELECTRIC ENERGY MARKET

Introduction

For the last few years the Polish energy market has developed significantly. The Polish Power Exchange was established in July 2000. It was the most important event in change of Polish energy market. The Day Ahead Market (DAM) was the first market which was established on the Polish Power Exchange. At the beginning of 2001 the Internet Electricity Trading Platform (IETP) functions on Polish electric energy market.

The Platform is an Internet-based trading tool which may be used to buy or sell electricity, green certificates, and CO2 emission allowances. Analogically to DAM, the prices are quoted 24 times per day.

The Balance Market (BM) exists from September 2001. This is technical market, which looks after balance on Polish energy market. From July 2002 BM introduced additional prices, Price Accounting Deviations of sale PADs and Price Accounting Deviations of purchase PADp. These prices should discipline market participants to precisely anticipate their future demand for energy. The aim of this article was compare these three whole-day markets. Vector Autoregressive (VAR)

models were applied for this purpose.

1. Vector Autoregressive Models

VAR is one of the most successful, flexible, and easy to use models for the analysis of multivariate time series. VAR model in economics was popularized by Sims (1980) who proposed theory which is opposite to structural models.

For a set of k time series $y_t = (y_{1t}; \ldots; y_{kt})^t$ a basic VAR model of the order p (VAR(p)) has a form (Sims 1980; Lütkepohl, Krätzig 2004):

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + \varepsilon_t \quad (1)$$

where the A_is are coefficient matrices, $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{kt})'$ is a zero-mean independent white noise process with time-invariant, positive definite covariance matrix; $\varepsilon_t \sim N(0; E(\varepsilon_t \varepsilon_t'))$. The process is stable if the eigenvalues of the matrix $A_{kp \times kp}$ satisfy:

$$|I\lambda^p - A_1\lambda^{p-1} - A_2\lambda^{p-2} - \ldots - A_p| = 0 \quad (2)$$

Hence, VAR(p) is covariance stationary as long as $|\lambda| < 1$ for all values of λ satisfying the above condition (Hamilton 1994). Basic VAR(p) models are usually necessary to represent trend and seasonality of time series:

$$y_t = A_0 D_t + BQ_t + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + \varepsilon_t \quad (3)$$

where D_t is a deterministic trend and Q_t is a matrix of seasonal dummy variables.

The VAR(p) models may contain independent time series X_t. It is referred to as VARX(p):

$$y_t = A_0 D_t + BQ_t + C_1 X_{1t} + \ldots + C_r X_{rt} + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + \varepsilon_t \quad (4)$$

Parameters of these k equations of VAR model may be estimated separately by ordinary least squares (OLS).

Each equation must have the same order of the lag p which may be chosen using model selection criteria. The three most common information criteria are (Zivot, Wang 2006):

– Akaike (AIC):

$$AIC(p) = \ln \left| \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t' \right| + \frac{2}{T} pk^2 \tag{5}$$

– Schwarz-Bayesian (BIC):

$$SC(p) = \ln \left| \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t' \right| + \frac{\ln T}{T} pk^2 \tag{6}$$

– Hannan-Quinn (HQC):

$$HQ(p) = \ln \left| \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t' \right| + \frac{2 \ln \ln T}{T} pk^2 \tag{7}$$

The AIC usually leads to evaluations of p which are too high and as a result to inclusion of insignificant parameters in the model. Consequently, the other two criteria are preferred. The significance of the VAR(p) model is verified by the F-test.

The general VAR(p) model has many parameters, and they may be difficult to interpret. As a result, properties of a VAR(p) are often explained with various types of structural analysis: Granger causality tests, impulse response function, forecast error variance decompositions (Osińska 2006).

The VAR(p) (1) model like the univariate AR(p) model have MA(∞) (Moving Average) representation (Hamilton 1994):

$$y_t = A_0 + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots \tag{8}$$

where matrix $\psi_\varepsilon (k \times k)$, has interpretation:

$$\frac{\partial y_{t+s}}{\partial \varepsilon_t'} = \psi_\varepsilon \tag{9}$$

The element in the row i and column j of matrix ψ_ε identifies the consequences of a one unit increase in the j^{th} variables innovation at date $t(\varepsilon_{j,t})$ for the value of the i^{th} variable at time $t + s(y_{i,t+s})$, holding all other innovations at all dates constant.

If the first element of ε_t changed by δ_1, at the same time second element changed by δ_2, the k^{th} element by δ_k then the combined effect of these changes on the

value of the vector would by given by (Hamilton 1994):

$$\Delta y_{t+s} = \frac{\partial y_{t+s}}{\partial \varepsilon_{1t}}\delta_1 + \frac{\partial y_{t+s}}{\partial \varepsilon_{2t}}\delta_2 + \ldots + \frac{\partial y_{t+s}}{\partial \varepsilon_{kt}}\delta_k = \psi_s \delta \tag{10}$$

By doing the separate simulation for impulses to each of the k-innovations, all of the columns of matrix ψ_s can be calculated. A plot of the row i and column j : $\frac{\partial y_{i,t+s}}{\partial \varepsilon_{j,t}}$ as a function of s is called the impulse-response function. It describes the response of $y_{i,t+s}$ to a one-time impulse in $y_{j,t}$ with ail other variables dated t or earlier held constants (Hamilton 1994).

2. Vector Autoregressive Models on the Polish electric energy market

To describe dependencies between prices on DAM, IETP, and BM markets from July to September 2007 VAR models were used. The prices on each of these markets are quoted 24 times a day. The advantage of the DAM and the IETP is that ail participants can buy and sell electric energy, irrespective of whether they are producers or receivers. The quotations on each market are calculated simultaneously, so the hypothesis of their interdependence shall be tested. Each financial decision involves risk. The knowledge of dependencies may be useful for risk management. The electric energy cannot be stored; it is delivered once there is demand for it. The BM has ensured the balance on the Polish energy market. Hence, the second hypothesis verified in this paper states, that prices on DAM and IETP depend on real demand for electric energy.

Hourly time series ($T = 2208$) of prices of electric energy from every market were analyzed. For ail time series three VAR models were constructed containing seasonal variables and exogenous variables of volumes (4) (Tables 1-3). Based on Schwartz-Bayesian criterion for each VAR model the lag order $p = 3$ was chosen. All three models are stationary $|\lambda| < 1$ (Figure 1).

Figure 1. The eigenvalues of the matrix $A_{kp\times kp}$

The first VAR(3) model describes dependence between three prices on different markets (Table 1). The model incorporates hourly seasonality and exogenous variables. All parameters are significant. Criterion function values indicate high goodness of fit of the VAR model (Figure 2-4).

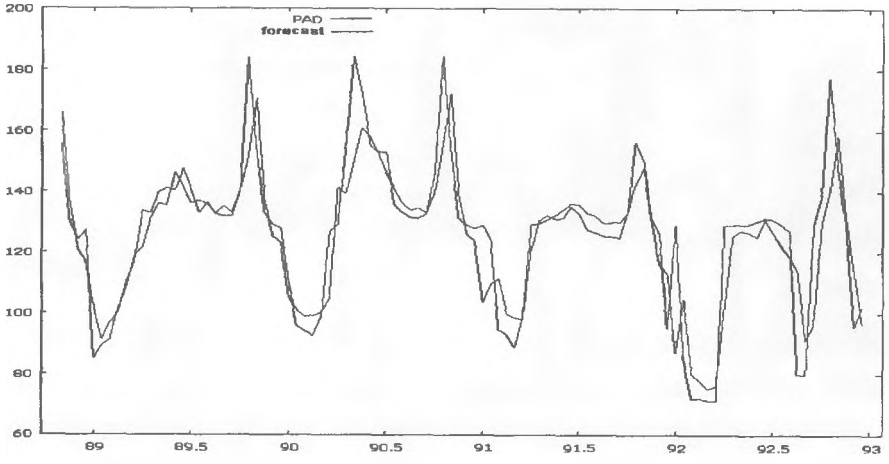


Figure 2. Theoretical and empirical PAD noted on BM from 24 to 30 September 2007

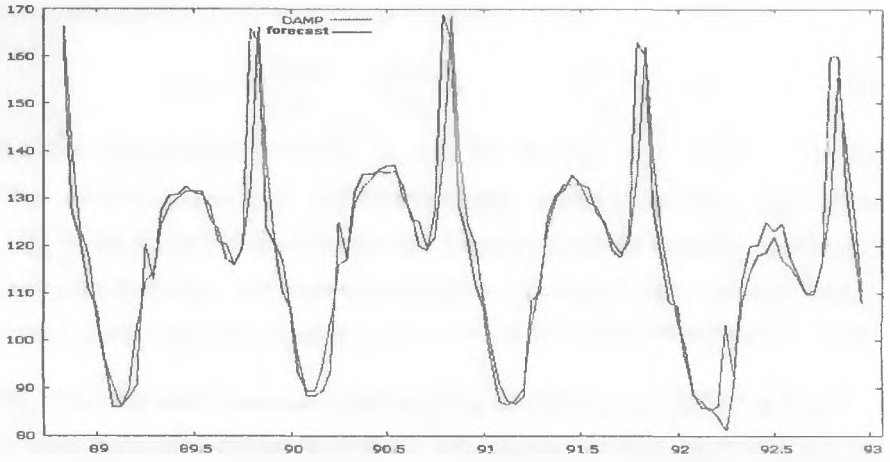


Figure 3. Theoretical and empirical prices noted on DAM from 24 to 30 September 2007

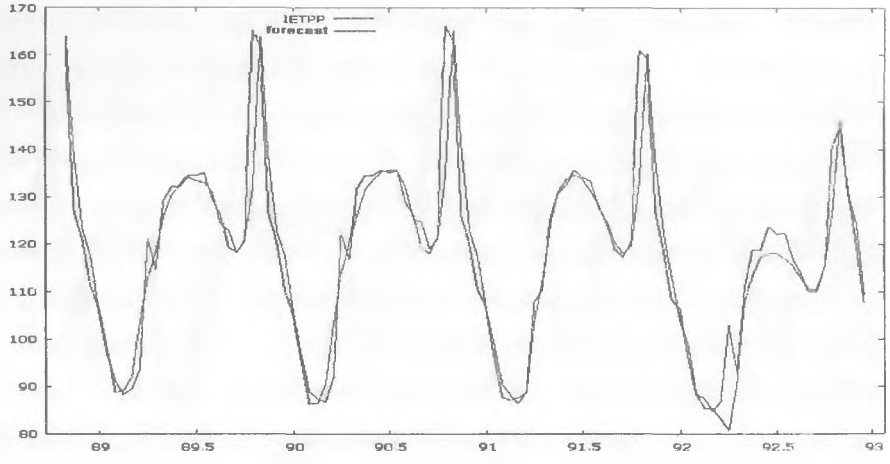


Figure 4. Theoretical and empirical prices noted on IETP from 24 to 30 September 2007

Table 1

The VAR(3) of prices of electric energy on BM, DAM, and IETP

Endogenous variables/prices	PAD y_{1t}	DAMP y_{2t}	IETPP y_{3t}
Parameters			
Mean	117,21	112,40	112,40
Standard deviation	23,30	18,17	17,51
Standard deviation of residuals	11,41	5,69	4,98
R^2	0,76	0,90	0,92
D-W	1,95	2,01	2,01
p-value			
PAD y_{1t-3}	0,0000	0,0173	0,0018
DAMP y_{2t-3}	0,0276	0,0000	0,0403
IETPP y_{3t-3}	0,0005	0,0000	0,0000
F	0,0000	0,0000	0,0000
Linear trend	<0,01	<0,01	<0,01
Seasonal variable	<0,01	<0,01	<0,01
Exogenous variable/volumes x_{1t}, x_{2t}, x_{3t}	<0,01	<0,01	<0,01
AIC	18,5469		
BIC	18,8337		
HQC	18,6517		

Source: Own calculations.

The determination coefficients are close to one. Residuals of separate models are not correlated and have very small variance. Hence, every price on whole--day market can be a cause of change for another one. Empirical time series observed in last week of the available data and the theoretical values of VAR(3) are presented below on Figures 2-4. Comparison of Figure 3 and Figure 4 leads to conclusion, that time series of prices on DAM and IETP are very similar.

Figure 5 shows a length and a size of impulse describing the influence of PAD on DAM and IETP. After fifteen hours influence of impulse vanishes. The impulse coming from DAM causes increase of prices on BM and IETP. The impulse from

PAD yields the same result but its influence vanishes later. The impulse from IETP has the longest duration, but its influence is the most weak.

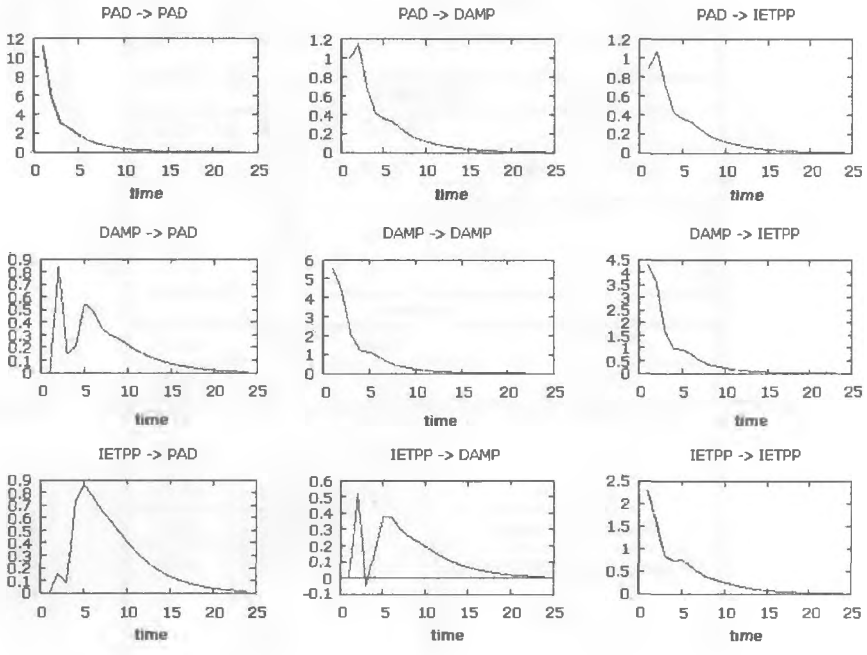


Figure 5. Impulse response function

The second VAR(3) model describes dependence between three prices noted on BM and prices noted on DAM (Table 2). Ali four series are seasonal, but the trend is significant only for PAD and DAM. PADs and PADp depend only on the volume of BM and on DAM prices. Criterion function values indicate high goodness of fit of the VAR model. The price on DAM can be a cause of every price on BM, and the opposite. But the PADs can not be the cause of PAD or PAPp and the opposite. Empirical time series noted in last week of the data and the theoretical values of VAR(3) are presented on Figures 6-7. The time series corresponding to PADs and PADp shown on Figures 2-4 are significantly different from PAD, DAM and IETP.

Table 2

The VAR(3) of prices of electric energy on BM and DAM

Endogenous variables/prices	PAD y_{1t}	PADs y_{2t}	PADp y_{3t}	DAMP y_{4t}
Parameters				
Mean	117,21	150,98	76,56	112,40
Standard deviation	23,30	3,07	6,96	18,17
Standard deviation of residuals	11,46	2,31	3,42	5,70
R^2	0,76	0,44	0,76	0,90
D-W	1,95	1,99	1,99	2,00
p-value				
PAD y_{1t-3}	0,0000	0,4606	0,7340	0,0004
PADs y_{2t-3}	0,9684	0,0000	0,8366	0,0000
PADp y_{3t-3}	0,0162	0,6192	0,0000	0,0003
DAMP y_{4t-3}	0,0042	0,0009	0,0000	0,0000
F	0,0001	0,0005	0,0005	0,0000
Linear trend	<0,01	0,39	0,89	<0,01
Seasonal variable	<0,01	<0,01	<0,01	<0,01
Exogenous variable/volumes x_{1t}, x_{2t}	<0,01	<0,01 BM	<0.01 BM	<0,01
AIC	23,6030			
BIC	24,0061			
HQC	23,7503			

Source: Own calculations.

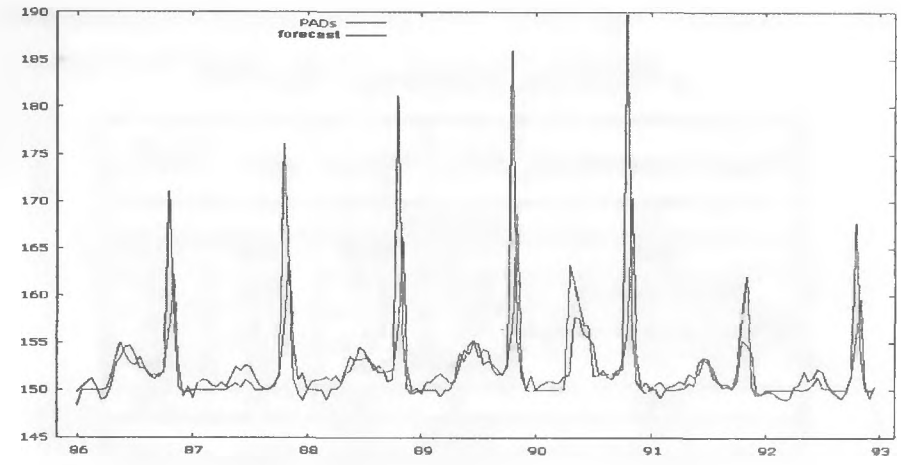


Figure 6. Theoretical and empirical PADs noted on BM from 24 to 30 September 2007

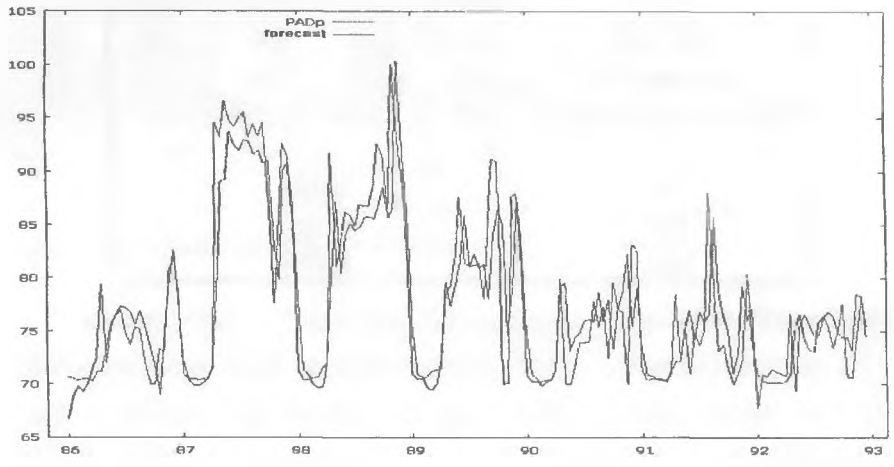


Figure 7. Theoretical and empirical PADp noted on BM from 24 to 30 September 2007

Impulse response function shown on Figure 8 indicates that the impulses from PAD and PADs cause PADp to decrease on short time lag and increase on long time lag. PADp influences DAM in the same way.

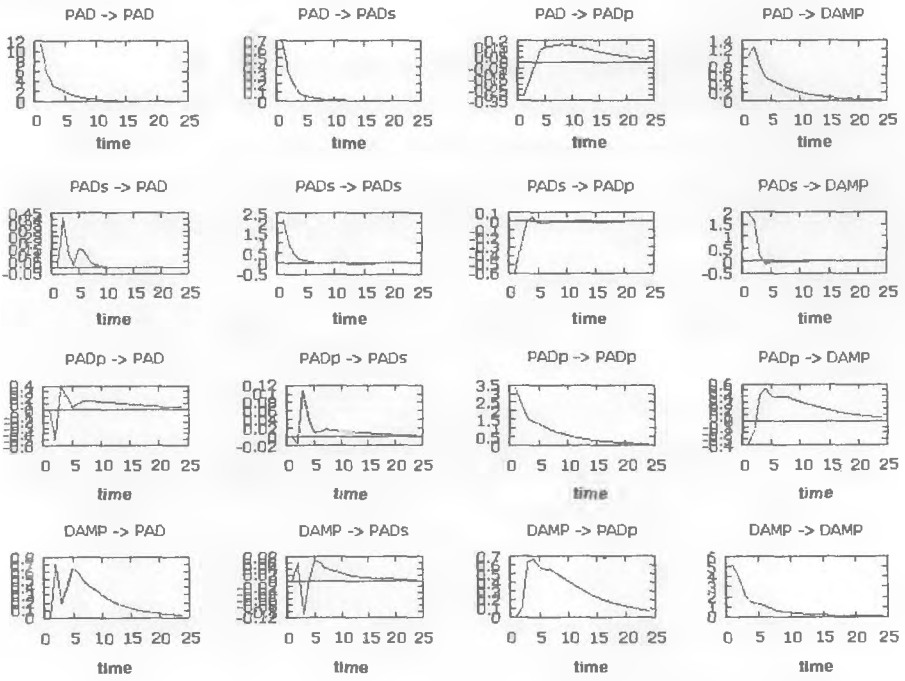


Figure 8. Impulse response function

The last VAR(3) model describes the dependence between three prices noted on BM and prices noted on IETP (Table 3). Results of estimation resemble the results obtained for the second model. The price structure on PADs and PADp is different from the structure on IEPT, but there are significant causal relationships between these prices.

Impulse response function shown on Figure 9 indicates that the impulse from PAD and PADs causes a decrease of the PADp for short lags and the increase of PADp for longer lags. The impulse from IETP causes an increase of PAD and PADp, and the decrease of PADs.

Table 3

The VAR(3) of prices of electric energy on BM and IETP

Endogenous variables/prices	PAD y_{1t}	PADs y_{2t}	PADp y_{3t}	IETPP y_{4t}
Parameters				
Mean	117,21	150,98	76,56	112,40
Standard deviation	23,30	3,07	6,96	17,51
Standard deviation of residuals	11,43	2,31	3,42	4,93
R^2	0,76	0,45	0,76	0,92
D-W	1,95	1,99	1,99	
p–value				
PAD y_{1t-3}	0,0000	0,5664	0,8195	0,0000
PADs y_{2t-3}	0,7219	0,0000	0,8511	0,0000
PADp y_{3t-3}	0,0310	0,7254	0,0000	0,0013
IETP y_{4t-3}	0,0009	0,0010	0,0000	0,0000
F	0,0001	0,0004	0,0005	<0,01
Linear trend	<0,01	0,26	0,66	<0,01
Seasonal variable	<0,01	<0,01	<0,01	<0,01
Exogenous variable/volumes	<0,01	<0,01	<0,01	<0,01
AIC	23,2578			
BIC	23,6609			
HQC	23,4051			

Source: Own calculations.

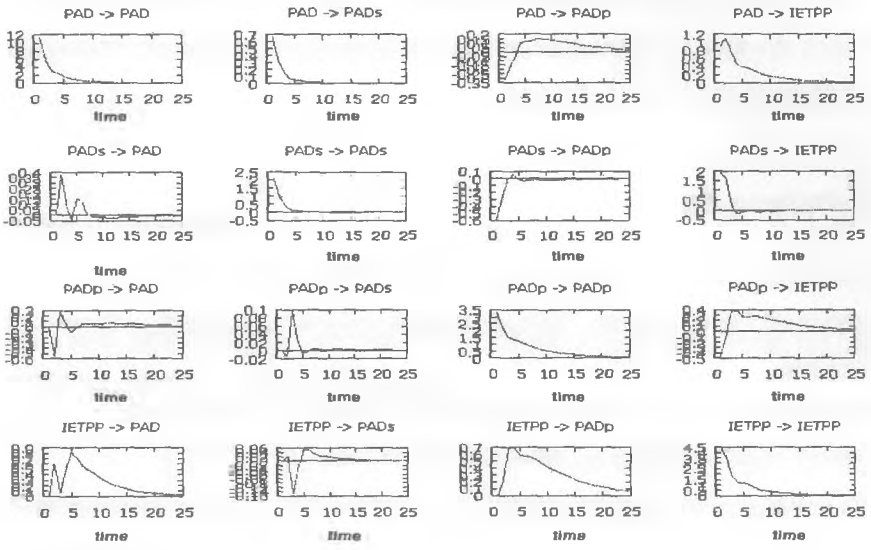


Figure 9. Impulse response function

Conclusions

Finally, we conclude that the time series on whole-day electric energy market are correlated. Analysis of the impulse response function suggests that if the electric energy prices on DAM increase then prices on IETP increase for next few hours and the opposite. It also indicates, that changing prices on BM cause prices on DAM and IETP to change. If there is too little volume of energy on the market then participants of this market can buy it at very high PADs prices on BM. So the prices on DAM and IETP and the PAD price rise simultaneously. Presented results show that Vector Autoregressive Models are well suited for describing multivariate time series on the electric energy market. They may be applied to predict the futures prices on this market in short investment horizon and to manage risk on the whole-day electric energy market. However, one should remember about assumptions of VAR. Unfortunately, the hypothesis that residuals of analyzed three

models have multivariate normal distribution is rejected. The residuals are not correlated, but they are heteroscedastic. More research is needed on the properties of VAR models in such a situation.

References

Hamilton J.D. (1994): Time Series Analysis. Princeton University Press, Princeton.

Lütkepohl H., Krätzig M. (2004): Applied Time Series Econometrics. Cambridge University Press, Cambridge.

Osińska M. (2006): Ekonometria Finasowa. Wydawnictwo Naukowe PWN, Warszawa.

Sims C.A. (1980): Macroeconomic and Reality. "Econometrica", 48, pp. 1-48.

Zivot E., Wang J. (2006): Modeling Financial Time Series with S-plus. Springer Science and Business Media, New York.

Abstract

In this article the relation between three whole-day markets from Polish electric energy market was presented. Vector Autoregressive models of prices and volumes of electric energy from the Day Ahead Market (DAM), the Internet Electricity Trading Platform (IETP), and Balance Market (BM) were applied to describe similarities and dependence between them.

Grzegorz Kończak

ON THE METHOD OF DETECTION LINEAR TREND IN STOCHASTIC PROCESSES

Introduction

Various physical, technical, biological and economic processes can be modeled using stochastic processes. A physical example of a stochastic process is the brownian motion and an economic examples are production processes. The method of modeling stochastic processes are widely used in analysis of properties of statistical quality control procedures.

M. Chao (2000) showed that Markov chains may be used to calculate performance of control charts. Author considered various quality control procedures and in each case indicates why such a Markov chain can be constructed. He evaluated exact average run of length for discrete observations and approximation of ARL for continuous observations. M. Chao concentrated on control charts but showed that Markov chain method can be applied to other quality systems.

M.F. Ramalhoto (2000) used stochastic models to describe real-life queuing systems. Author considered Poisson process and Markov chains.

G. Kończak (2004) proposed to use the Markov chains for monitoring technological processes. The main assumptions when control chart are used are that

the process data are normally and independently distributed. Author presented the method which can be used when the process with autocorrelated data is monitoring.

The methods of analysis of stochastic processes are adapted to evaluation properties of acceptance sampling plans as finite Markov chains were presented in E.B. Wilson and A.R. Burgess (1971). They calculated characteristics of acceptance plans using the methods of analysis Markov chains. G. Kończak (2007) presented the method of analysis properties of the continuous sampling plan CSP-1 and the modification.

One of the most common problems in monitoring real processes in quality control is to test the stability of the process. The proposition of the test to verify the hypothesis about stability of discrete stochastic process is presented in the paper.

1. Process monitoring in quality control procedures

W.A. Shewhart introduced control chart in 1924. Control charts are the graphical procedures for monitoring production processes. The control chart is a graphical display of a quality characteristic versus the sample number or time. A point that plots outside of the control limits is interpreted as evidence that the process is out of control. There are many sets of rules which help to detect instability of processes. In quality control analysis programs (e.g. QI Analyst) users can define sets of such rules. Some of the sensitizing rules that are used in practice are shown in Table 1.

Table 1

Some sensitizing rules for Shewhart control charts

A set	A set of rules	Source
1	One or more points outside of the control limits	W.A. Shewhart
2	One or more points outside of the control limits Two of three consecutive points outside the 2-sigma warning limits, but still inside the control limits Four of five consecutive points on one side of the center line A run of eight consecutive points on one side of the center line	Western Electric Rules
3	Western Electric Rules (as above) Six point in a row steadily increasing or decreasing Fifteen points in a row between $m - s$ and $m + s$ lines Fourteen points in a row alternating up and down An unusual or nonrandom pattern in the data One or more points near a warning or control limit	D.C. Montgomery (1996)
4	One or more points outside of the control limits Two successive points at or beyond +/– 1.5 SD	A.M. Hurwitz, M. Mathur (1996)

Source: Montgomery (1996), Hurwitz, Mathur (1996).

A basic assumption in quality control procedures is that the process should be stationary. D.C. Montgomery (1996) presents a set of rules for Shewhart control charts (see Table 1). One of these rules is "six points in a row steadily increasing or decreasing". This rule can be used to detect trend in the production process. We will concentrate on the procedure for detecting a trend in stochastic processes.

2. Stationary processes

Let $X_t(t \in T)$ be a stochastic process. If T is real axis then X_t is called a continuous-time process. If T is the set of integers, then X_t is a discrete-time process. X_t is a discrete-state process if its values are countable and otherwise it is continuous-state process. We will concentrate on the discrete-time and continuous-state processes. In this case we can write the stochastic process as $X(t)$ where $t = 1, 2, \ldots, k$ or equivalently $X(1), X(2), \ldots, X(k)$.

A stochastic process $X(t)$ is called strict-sense-stationary if its statistical properties are invariant to a shift of the origin. It can be say that $X(t)$ and $X(t+c)$ have the same parameters for any c. A stochastic process $X(t)$ is called wide sense stationary if (Kowalenko et al. 1980):

$$E(X^2(t)) < \infty$$

$$E(X(t)) = m$$

$$R(t,s) = E(X^0(t)X^0(s)) = R(t-s)$$

where

$$X^0(t) = X(t) - E(X(t))$$

In many production processes we expect that the statistic properties of the process will be invariant to a shift of the origin. We can test the hypothesis that process is stationary, that's mean the probability characteristic do not change in time. We will test the hypothesis that random variables $X(1), X(2), \ldots, X(k)$ have the same cumulative distributions. This hypothesis can be written as follows:

$$H_0 : F_1(x) = F_2(x) = ... = F_k(x) \tag{1}$$

against the alternative:

$$H_1 : F_i(x) \neq F_j(x)$$

for any $i \neq j$.

Let us assume that $x_1, x_2, \ldots, x_k$ are the observed independent realizations of the stochastic process $X(t)$. The example of the realization of the stochastic process is presented in Figure 1.

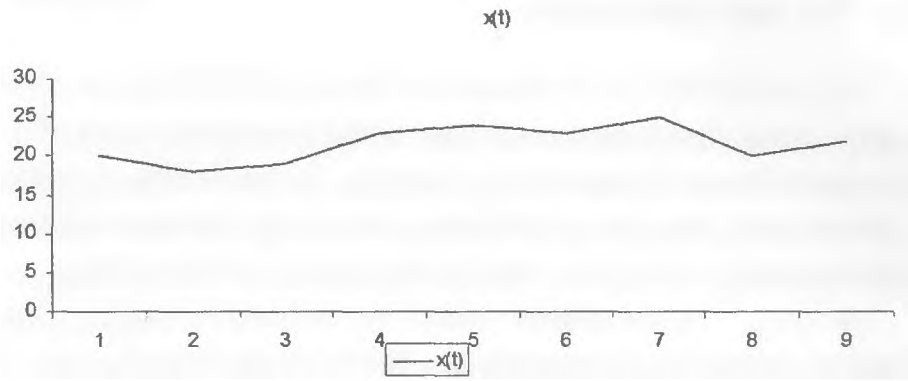


Figure 1. The stochastic process – observed realization

Let us assume that the realizations of the stochastic process are independent then under H_0 we can obtain several possible realizations of the process permuting observed realizations. The observed realization of the stochastic process and examples of the possible realizations are shown in the Figure 2.

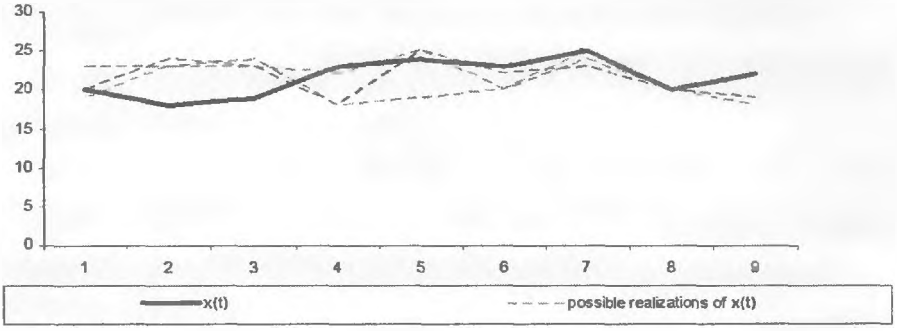


Figure 2. The stochastic process – observed realizations $x(t)$ and examples of possible realizations of this process under H_0

3. The test procedure

Let us assume that $X(t)$ is a discrete stochastic process with continuous set of values. We will test the hypothesis (1) about stability the stochastic process $X(t)$. A proposed test will be based on indexes analysis. As a test statistic will be considered average index and two modifications of this index's. The first modification was proposed by I. Timofiejuk (1994) and the second by G. Kończak (1995).

Let $x_1, x_2, \ldots, x_k$ are observed values of the stochastic process $X(t)$. The classical average index is calculated as a geometric mean of chain indexes. It can be written as follows:

$$\bar{i} = \sqrt[k-1]{\prod_{i=2}^{k} \frac{x_i}{x_{i-1}}} = \sqrt[k-1]{\frac{x_k}{x_1}} \tag{2}$$

The value $\bar{i}$ depends on the first and the last values of the series $x_1, x_2, \ldots, x_k$. To obtain information about changes in the middle of this time series can be used modifications of classical index (2).

I. Timofiejuk (1994) proposed an average index which depends on all values of time series. This index can be written as follows:

$$\bar{i}^* = \sqrt[s]{\prod_{i=2}^{k} \frac{x_i}{x_1}} \tag{3}$$

where $s = \frac{k(k-1)}{2}$.

The modification of Timofiejuk's index proposed by G. Kończak (1995) can be written:

$$\bar{\bar{i}} = \sqrt[t-1]{\prod_{i=1}^{k-2} g_i} \tag{4}$$

where

$$t = \frac{(k-2)(k-1)(k+3)}{6}$$

$$g_1 = \sqrt[k-1]{\frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \ldots \cdot \frac{x_k}{x_{k-1}}}$$

$$g_2 = \sqrt[2(k-2)]{\frac{x_3}{x_1} \cdot \frac{x_4}{x_2} \cdot \ldots \cdot \frac{x_k}{x_{k-2}}}$$

and generally:

$$g_i = \sqrt[i(k-i)]{\frac{x_{i+1}}{x_1} \cdot \frac{x_{i+2}}{x_2} \cdot \ldots \cdot \frac{x_k}{x_{k-i}}} \text{ for } i = 1, 2, \ldots k-2$$

The index (4) can be used in the case when we want to describe all changes in the time series. This index measures average changes in the whole time series (Kończak 1995). Statistics (2), (3), and (4) can be used as a test statistics to verify the hypothesis (1). These statistics especially can be used to determine a trend in time series.

Let us assume that $x_1, x_2, \ldots, x_k$ are observed realizations of the stochastic process. We will use permutation test (Efron, Tibshirani 1993) introduced by R.A. Fisher. Having observed value of the statistic $\widehat{i}$, the achieved significance level ASL is defined to be the probability of observing at least that large a value when the null hypothesis is true:

$$ASL = P(i \geq \widehat{i} | H_0)$$

where i is the test statistic. The smaller the value of ASL, the stronger the evidence against H_0.

To obtain critical values for each statistic let us consider all the possible permutations of the set $T = \{1, 2, \ldots, n\}$.

Let $T_i = (t_{i1}, t_{i2}, \ldots, t_{in})$ be the i-th permutation of the set T. Let Γ be the set of all permutations T_i of T. From each permutation we get a time series $x_{t_{i1}}, x_{t_{i2}}, \ldots, x_{t_{ik}}$. For these time series we calculate:

a) average index $\bar{i}$ (2),

b) Timofiejuk's modification $\bar{i}^*$ (3),

c) index $\bar{\bar{i}}$ (4).

The quantiles of statistics under H_0 (2), (3), and (4) we accept as critical values.

4. Monte Carlo study

If the length of the time series is greater than 10, then the number of permutation is very large. In these cases we can estimate critical values. From the set Γ we take an N (B. Efron and R. Tibshirani suggest that N should be at least 1000) element random sample of permutations of set T. For each set T_i we calculate indexes $\bar{i}$, $\bar{i}^*$, and $\bar{\bar{i}}$. Then we accept empirical quantiles as critical values.

The approximate values of the probabilities of rejection the hypothesis (2) in the case of use the statistics (2), (3), and (4) are obtained in Monte Carlo analysis. There were following steps in Monte Carlo study:

1. Observations $x_1, x_2, \ldots, x_k$ $(k = 20)$ were generated from normal distribution with the following scheme:
   (a) $X \sim N(100, 5)$, for $t = 1, 2, \ldots, 10$
   $X \sim N(100 + b, 5)$, for $t = 11, 12, \ldots, 20$
   where $b = 1, 2, \ldots, 5$.
   (b) $X \sim N(100 + at, 5)$, for $t = 1, 2, \ldots, 20$
   where $a = 0.2, 0.4, \ldots, 1$.
   The expected values of $X(t)$ are shown in Figure 3.
2. The values of indexes $\bar{i}$, $\bar{i}^*$ and $\bar{\bar{i}}$ were calculated.
3. The sample of 10 000 elements was taken from the set of permutations of $x_1, x_2, \ldots, x_k$. For each case indexes $\bar{i}$, $\bar{i}^*$, and $\bar{\bar{i}}$ were calculated and the empirical quantiles of order 0.025 and 0.975 of these statistics were accepted as critical values for the significance level $\alpha = 0.05$.

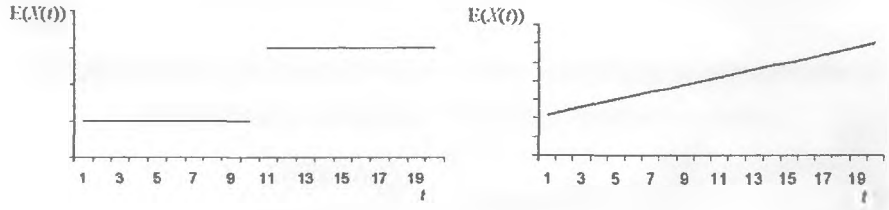


Figure 3. Expected values of $X(t)$ in Monte Carlo study

The results of Monte Carlo study and the probabilities for Shewhart's limits are presented in tables 2, 3 and in the figures 4 and 5.

Table 2

The estimated values of probabilities of rejection the nuli hypothesis (one step shift) for significance level $\alpha = 0.05$ and the correspondent probabilities for Shewhart's limits

Parameter b	Shewhart's limits	The test statistic		
		$\bar{i}$ (1)	$\bar{i}^*$ (2)	$\bar{\bar{i}}$ (3)
1	0.059	0.0443	0.0421	0.0702
2	0.090	0.0657	0.0496	0.1130
3	0.147	0.0562	0.0563	0.1765
4	0.235	0.0802	0.0678	0.2913
5	0.355	0.0760	0.0778	0.4184

Source: Monte Carlo study.

Table 3

The estimated values of probabilities of rejection the null hypothesis (linear trend) for significance level $\alpha = 0.05$ and the correspondent probabilities for Shewhart's limits

Parameter a	Shewhart's limits	The test statistic		
		$\bar{i}$ (1)	$\bar{i}^*$ (2)	$\bar{\bar{i}}$ (3)
0.2	0.114	0.0862	0.0693	0.1712
0.4	0.386	0.1528	0.0893	0.5103
0.6	0.805	0.2413	0.1070	0.8581
0.8	0.986	0.3152	0.1519	0.9713
1.0	1.000	0.4160	0.1555	1.000

Source: Monte Carlo study.

The value of the average index depends on the first and the last observations from time series. The value of the index given by (4) depends on all changes in time series. The results of Monte Carlo study showed that the average index (4) may be used as a statistic to test the hypothesis (1) especially in the case of linear trend detecting.

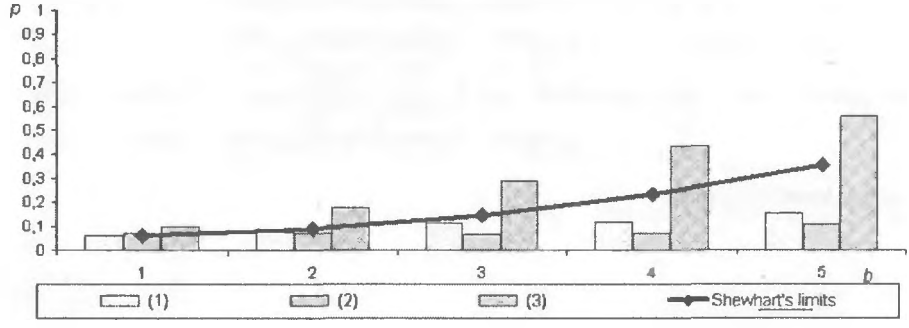


Figure 4. The estimated values of probabilities of rejection the null hypothesis (one step shift) for significance level $\alpha = 0.05$

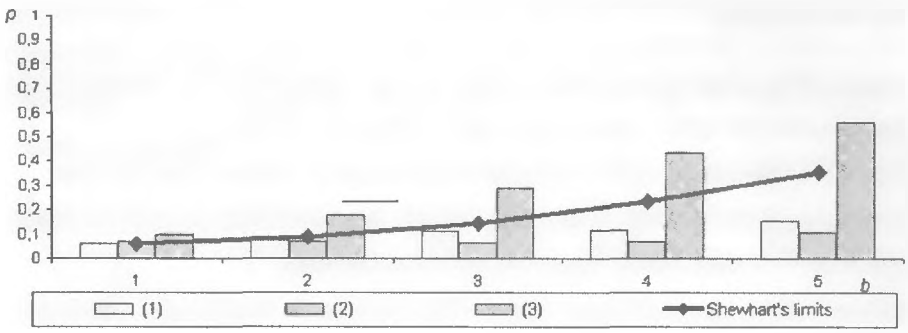


Figure 5. The estimated values of probabilities of rejection the null hypothesis (linear trend) for significance level $\alpha = 0.05$

Conclusions

In quality control procedures such as control charts the main assumption is that the process is stationary. The methods for detecting the trend in stochastic processes are proposed in the paper. There are three methods analyzed which are based on the average indexes. There are used classical average index, Timofiejuk's modification and the modification proposed by G. Kończak (1995). Two cases of no stationary processes were analyzed: one step shift and linear trend.

The index proposed by G. Kończak (1995) measures the average changes in the whole time series. This index can be used as a test statistic for detecting a trend in time series. The proposed test is based on the permutation tests idea introduced by R.A. Fisher. The Monte Carlo study have been made. The results of the simulation study have shown that the proposed test can be used to verify the hypothesis about the stability of stochastic process.

References

Chao, M. (2000): Applications of Markov Chains in Quality-Related Matters. In: Statistical Process Monitoring and Optimization. Marcel Dekker, New York-Basel, pp. 175-188.

Efron B. and Tibshirani R. (1993): An Introduction to the Bootstrap. Chapman & Hall, New York.

Hurwitz A.M., Mathur M. (1996): A Very Simple Set of Process Control Rules. In: Statistical Applications in Process Control. Marcell Dekker, New York-Basel-Hong Kong.

Kończak G. (2004): Łancuchy Markowa w analizie własności procedur kontroli jakości. "Taksonomia", 11, AE, Wrocław.

Kończak G. (2007): Metody statystyczne w sterowaniu jakością produkcji. AE, Katowice.

Kończak G. (1995): On the Modification of Timofiejuk's Index. In: Proceedings of 14th International Conference on Multivariate Statistical Analysis MSA'95. Uniwersytet Łódźki, Łódź, pp. 175-185.

Kowalenko I.N., Kuzniecow N.J., Szurienkow W.M. (1989): Procesy stochastyczne. PWN, Warszawa.

Montgomery D.C. (1996): Introduction to Statistical Quality Control. John Wiley & Sons, Inc., New York.

Ramalhoto M.F. (2000): Stochastic Modeling for Quality Improvement in Processes. Marcel Dekker, New York-Basel.

Thompson J.R., Koronacki J. (2001): Statistical Process Control: The Deming Paradigm and Beyond. Chapman and Hall-CRC, New York-London.

Timofiejuk I. (1994): O liczeniu średniego tempa wzrostu. "Wiadomości Statystyczne", nr 12.

Wilson E.B., Burgess A.R. (1971): Multiple Sampling Plans Viewed as Finite Markov Chains. "Technometrics", Vol. 13, pp. 371-383.

Abstract

Variaus physical, technical, biological, and economic processes can be modeled using stochastic processes. A physical example of a stochastic process is the brownian motion and an economic examples are production processes. The method of modeling stochastic processes are widely used in analysis of properties of statistical quality control procedures. One of the most common problems in monitoring real processes in quality control is to test the stability of the process. The methods for detecting the trend in stochastic processes are presented in the paper. There are three methods analyzed which are based on the

average indexes. Two cases of no stationary processes were analyzed: one step shift and linear trend. The Monte Carlo study have been made. The results of the simulation study have shown that the proposed test can be used to verify the hypothesis about the stability of stochastic process.

Dorota Rozmus

USING BAGGING AGGREGATION METHOD IN TAXONOMY

Introduction

Resampling methods such as bagging (Breiman 1996) and boosting (Freund 1990; Freund and Schapire 1995) have been applied successfully in the area of supervised learning to improve prediction accuracy in classification and regression. In order to get an aggregated model (ensemble) in first step we build many different single models and then we combine them by means of some aggregation operator. For example in bagging method we construct single models on bootstrap samples* chosen from the original learning data set and then we aggregate theoretical values of dependent variable gained on the basis of these models. In regression the most popular aggregation operator is mean of all theoretical values of dependent variable and in classification we apply majority voting – we classify the observation to the most often predicted class. It appears that ensemble approach can be successfully also applied in taxonomy (unsupervised learning). The interest in cluster ensembles has been growing in the past few years (Ayad and Kamel 2003; Fern and Brodley 2003; Fischer and Buhmann 2003; Fred and

*Bootstrap sample is constructed by choosing N elements with replacement from a set that counts N elements.

Jain 2002; Monti et al. 2003; Strehl and Ghosh 2002). The aim of combining several partitions into a single one is to improve the quality and robustness of the result.

A new resampling method, inspired from bagging in classification and regression, was proposed to improve the accuracy of a given clustering procedure (Dudoid and Fridlyand 2003). The main aim of the article is to compare the right class structure recognizing ability of classical and ensemble clustering methods. The performances of the new and existing taxonomy algorithms were compared using simulated and real data sets. It appears that the bagged clustering procedures were in general at least as accurate and often more accurate than a single application of the partitioning clustering procedure.

The rest of the paper is organized as follows. Section 1 explains the bagging algorithm in taxonomy. The chosen data sets, the experimental set-up and measure of clustering algorithms quality are detailed in Section 2. This section contains also the empirical results. Section 3 concludes the study.

1. The bagging algorithm in taxonomy

In this ensemble method, a partitioning clustering procedure is applied to bootstrap learning sets and the resulting multiple partitions are combined by voting. As in prediction, the motivation behind bagging is to reduce variability in the partitioning results via averaging. Partitioning methods are typically based on iterative optimization techniques, thus additional sources of variability in the results include the sensitivity to starting conditions and the possibility of convergence to local minima (or maxima, depending on the objective function). In a recent manuscript, Leisch (1996) proposed a bagged clustering method which is a combination of partitioning and hierarchical procedures. A partitioning method is applied to bootstrap learning sets and the resulting partitions are combined by performing hierarchical clustering of the cluster centers. This procedure compared favorably to existing partitioning methods for a variety of simulated and real data sets considered by

the author. A new bagging procedure proposed by Dudoid and Fridlyand (2003) is similar in spirit to that of Leisch (1996), however, different approach based on voting is proposed to combine multiple partitioning results.

The algorithm works as follow. For a fixed number of clusters K:

1. Apply the partitioning clustering procedure C to the original learning set S to obtain cluster labels $C(x_i; S) = \hat{y}_i$ for each observation x_i, $i = 1, \ldots, n$.

2. Form the b-th bootstrap sample $S^b = (x_1^b, \ldots, x_n^b)$.

3. Apply the clustering procedure C to the bootstrap learning set S^b and obtain cluster labels $C(x_i^b; L^b)$ for each observation in S^b.

4. Permute the cluster labels assigned to the bootstrap learning set S^b so that there is maximum overlap with the original clustering of these observations. Specifically, let P_K denote the set of all permutations of the integers $1, \ldots, K$. Find the permutation $\tau^b \in P_K$ that maximizes:

$$\sum_{i=1}^{n} I(\tau(C(x_i^b; S^b)) = C(x_i^b; S)) \tag{1}$$

where $I(\cdot)$ is the indicator function, equal 1 if the condition in parentheses is true and 0 otherwise.

5. Repeat Steps 2-4 B times and assign a bagged cluster label for each observation i by majority vote, that is, the cluster label corresponding to x_i is:

$$\text{argmax}_{1 \le k \le K} \sum_{b: x_i \in S^b} I(\tau^b(C(x_i; S^b)) = k) \tag{2}$$

Also, it is possible to record a cluster vote, which is the proportion of votes in favor of the winning cluster assignment, that is:

$$CV(x_i) = \frac{\max_{1 \le k \le K} \sum_{b: x_i \in S^b} I(\tau^b(C(x_i; S^b)) = k)}{|b : x_i \in S^b|} \tag{3}$$

2. Empirical results

In empirical part of the research I wanted to compare the right class structure recognizing ability of classical and ensemble clustering methods. Among classical taxonomy algorithms I used popular k-means method and c-means method which is fuzzy version of the k-means method introduced by Bezdek (1981). In an aggregated approach as base algorithms I used the same methods and the number of bootstrap samples (B) was equal 50. All computations were made in R*, using k-means algorithm from stats library, c-means algorithm from e1071 library; the aggregated approach is implemented in clue library as cl-bagg function.

As a measure of the correctness of the algorithm I used popular Silhouette Index (Kaufman, Rousseeuw 1990). This measure is included in the $[-1, 1]$ interval and the higher value it takes the stronger class structure was found by the algorithm. It is computed as:

$$Sil(k) = \sum_{i \in k} \frac{Sil(i)}{n_k} \tag{4}$$

where:

$$Sil(i) = \frac{b(i) - a(i)}{\max[a(i); b(i)]} \tag{5}$$

and: $a(i)$ is mean distance of the ith object from other objects in kth cluster,

$a(i) = \sum_{j \in [k \setminus i]} \frac{d_{ij}}{(n_k - 1)}$,

$b(i) = \min_{k \neq k'} [d_{ik'}]$,

$d_{ik'}$ – mean distance of the ith object from objects in k'th cluster:

$d_{ik'} = \sum_{j \in k'} \frac{d_{ij}}{n'_k}$,

$k, k' = 1, 2, \ldots, K$,

K – number of clusters,

$i = 1, 2, \ldots, n_k$,

n_k – number of observations in kth cluster.

In my research I used artificial data sets that are commonly used in taxonomy**. Their structure is illustrated on Figure 1, Figure 2, and Figure 3.

*This program is free available in Internet on website: www.r-project.org.

**These data sets are in *mlbench* library in R.

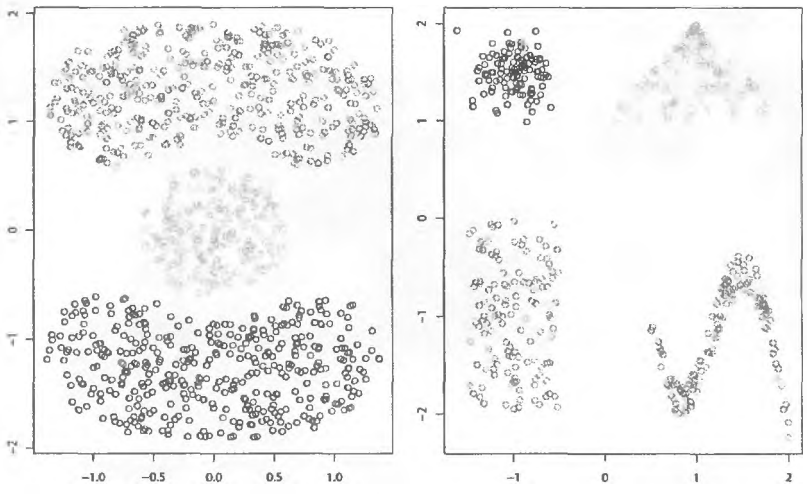


Figure 1. The *Cassini* and *Shapes* data sets

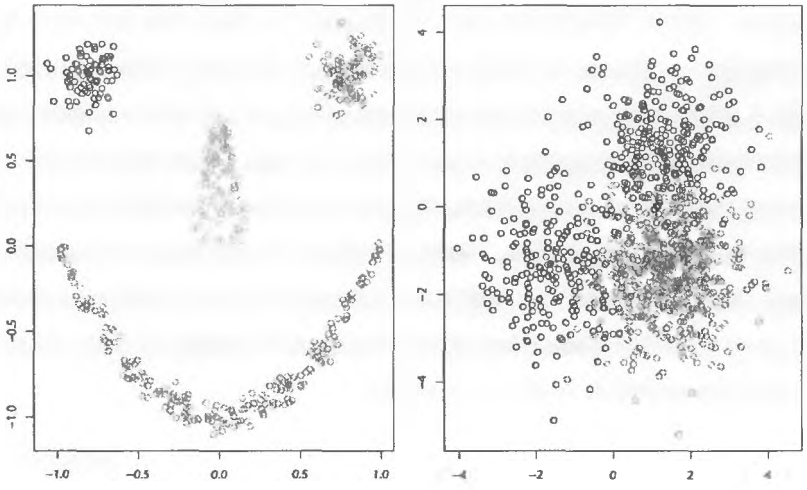


Figure 2. The *Smiley* and *Threenorm* data sets

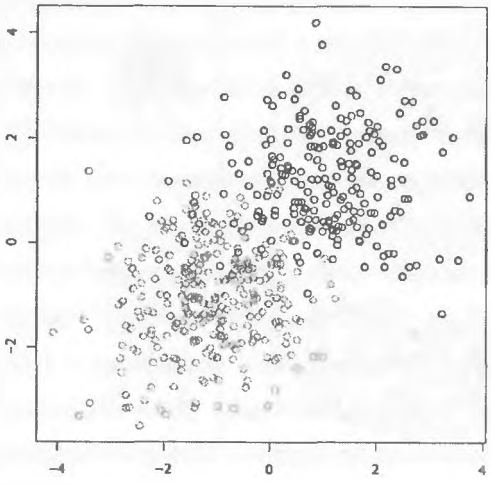


Figure 3. The *2dnormals* data set

Moreover I used real data sets that are used in classification (supervised learning) for model construction and its evaluation; so these are data sets where the adherence of objects to right class is known in advance. This information is used as *a priori* information about number of groups. Such solution is also often used by researchers in taxonomy. Among real data sets I used *Boston** that is widely used in comparative researches; this data set is made available by University of California (Blake, Keogh, Merz 1988). The next two data sets were prepared on the basis of household income research and are used for classification of observations from the point of view of their income rating and material situation valuation.

Results are shown in Table 1 and Table 2.

*This data set is used for real estate valuation; for the need of this research one nominal variable was deleted.

Table 1

Silhouette index for classical and bagged taxonomy algorithms (aritficial data sets)

Data set	kmeans	kmeans_bagg	cmeans	cmeans_bagg
Cassini	0,41	0,44	0,39	0,40
Shapes	0,59	0,70	0,70	0,71
Smiley	0,55	0,60	0,56	0,59
Threenorm	0,36	0,37	0,36	0,39
2dnormals	0,44	0,47	0,44	0,47

Table 2

Silhouette index for classical and bagged taxonomy algorithms (real data sets)

Data set	kmeans	kmeans_bagg	cmeans	cmeans_bagg
Boston	0,42	0,49	0,43	0,56
Material situation	0,30	0,31	0,21	0,24
Income rating	0,30	0,32	0,21	0,24

In the case of most data sets k-means method gives better results than c-means method in both – classical and aggregated approach. Exeption are only *Shapes* and *Boston* data sets (in classical and aggregated approach), *Smiley* in classical approach and *Threenorm* in aggregated approach. The highest differences between k-means and c-means method we observe in the case of data sets based on household income research. Generally, we can notice advantage from using aggregated approach because the Silhouette Index is higher for ensembles.

Summary

Resampling methods such as bagging and boosting have been applied successfully in a supervised learning context to improve prediction accuracy. An idea of bagging method is used to generate and aggregate multiple clusterings. The bagged clustering procedure was proposed by Dudoid and Fridlyand (2003) where the clustering procedure is repeatedly applied to each bootstrap sample and the

final partition is obtained by plurality voting. For the real and simulated data sets considered in this study, the clusterings produced by bagging procedure was in general at least as accurate, and often more accurate, than the clusterings resulting from a single application of classical taxonomy algorithms. Although the bagging was illustrated using k-means and c-means it is applicable to any clustering procedure and it would be worthwhile to evaluate the improvement in accuracy for methods such as e.g. k-medoids or self-organizing maps. It is suspected that, as in prediction, the increase in accuracy observed with used classical algorithms is due to a decrease in variability achieved by aggregating multiple clusterings. It would be interesting to carry out a more thorough study of the bias and variance properties of different clustering methods, as was done for classifiers in Breiman (1998). Other ongoing research directions include the investigation of different resampling schemes, similar in spirit to the adaptive resampling schemes used in boosting.

It is also worth to add that selecting a good clustering algorithm is more difficult than selecting a good classifier. The difficulty comes from the fact that in clustering there is no supervision, i.e. data have no labels against which to match the partition obtained through the clustering algorithm. Therefore, instead of running the risk of picking an unsuitable clustering algorithm, a cluster ensemble can be used (Strehl and Ghosh 2002). The presumption is that even a basic off-the-shelf cluster ensemble will outperform a randomly chosen clustering algorithm.

References

Ayad H., Kamel M. (2003): Finding Natural Clusters Using Multi-Clusterer Combiner Based on Shared Nearest Neighbors. "Proceedings of the Fourth International Workshop on Multiple Classifier Systems", MCS'03, Vol. 2709 of Lecture Notes in Computer Science, Springer Verlag, Guildford, UK, pp. 166-175.

Bezdek J.C. (1981): Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York.

Blake C., Keogh E., Merz C.J. (1988): UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California, Irvine.

Breiman L. (1996): Bagging Predictors, "Machine Learning", 26(2), pp. 123-140.

Breiman L. (1998): Arcing classifiers. "Annals of Statistics", 26, pp. 801-824.

Dudoit S., Fridlyand J. (2003): Bagging to Improve the Accuracy of a Clustering Procedure. "Bioinformatics", Vol. 19. No. 9, pp. 1090-1099.

Fern X.Z., Brodley C.E. (2003): Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach. "Proceedings of the Twentieth International Conference on Machine Learning", ICML, pp. 186-193.

Fischer B., Buhmann J.M. (2003): Bagging for Path-Based Clustering. "IEEE Transactions on Pattern Analysis and Machine Intelligence", 25(11), pp. 1411-1415.

Fred A., Jain A. K. (2002): Data Clustering Using Evidence Accumulation. "Proceedings of the Sixteenth International Conference on Pattern Recognition", ICPR, Canada, pp. 276-280.

Freund Y. (1990): Boosting a Weak Learning Algorithm by Majority. "Proceedings of the Third Annual Workshop on Computational Learning Theory", pp. 202-216.

Freund Y., Schapire R. E. (1995): A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. "Proceedings of the Second European Conference on Computational Learning Theory", Springer Verlag, pp. 23-27.

Kaufman, L., Rousseeuw P.J. (1990): Finding Groups in Data: An Introduction to Cluster Analysis, Wiley & Sons, Inc., New York.

Leisch F. (1996): Bagged Clustering. Technical Report. SFB Adeptive Information Systems and Modelling in Economics and Management Science, University of Economics and Business, Vienna, http://www.ci.tuwen.ac.at/~teisch/papers/fl-techrep.htm.

Monti S., Tamayo P., Mesirov J., Golub T. (2003): Consensus Clustering: A Resampling Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. "Machine Learning", 52, pp. 91-118.

Strehl A., Ghosh J. (2002): Cluster Ensembles – a Knowledge Reuse Framework for Combining Multiple Partitions. "Journal of Machine Learning Research", 3, pp. 583-618.

Abstract

Ensemble approach based on aggregated models has been successfully applied in the context of supervised learning in order to increase the accuracy and stability of classification. Recently, analogous techniques for cluster analysis have been suggested. Research has proved that, by combining a set of different clusterings, an improved solution can be obtained.

In the literature a resamplimg method, inspired from bagging in classification, was proposed to improve the accuracy and stability of clustering procedures. In the ensemble method, a partitioning clustering method is applied to bootstrap learning sets and the resulting different partitions are combined by majority voting. Similarly as in prediction, the motivation behind bagging is to reduce variability in the partitioning results via averaging. The performances of the new and existing methods were compared using real and artificial data sets. Generally the bagged clustering procedure was at least as accurate and ofter even much more accurate than a single application of the partitioning clustering method.

Informacja o Kated

Pracownicy K
dydaktycznej pro
Ważne rezultaty nauk
niach dotyczących
nych. W szczególno
macji wartości globalnej cechy w populacji
na podstawie prób
uzyskano na polu t
wnioskowania w
danych w próbie o
polu wyniki
tylko polskich P
jakości metod pre
szeroki anali
Własności m
z wykorzystaniem
symulacji komput
empiryczne d
modelowania
analizy regresji
wielowymiarowej

ISBN 978-83-7246-590-1