STUDIA EKONOMICZNE

UNIWERSYTET EKONOMICZNY
w Katowicach

STATISTICAL INFERENCE METHODS IN ECONOMIC RESEARCH

Redaktor naukowy
Janusz L. Wywiał

65
ZESZYTY NAUKOWE

STATISTICAL INFERENCE METHODS IN ECONOMIC RESEARCH

STUDIA EKONOMICZNE

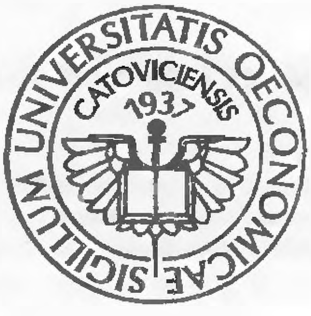
ZESZYTY NAUKOWE

UNIWERSYTETU EKONOMICZNEGO W KATOWICACH

STATISTICAL INFERENCE METHODS IN ECONOMIC RESEARCH

Redaktor naukowy

Janusz L. Wywiał



Katowice 2011

Redaktor
Beata Kwiecień

ISBN 978-83-7246-652-5

WYDAWNICTWO UNIWERSYTETU EKONOMICZNEGO W KATOWICACH
ul. 1 Maja 50, 40-287 Katowice, tel.: +48 32 257-76-35, faks: +48 32 257-76-43
www.ue.katowice.pl e-mail: wydawnictwo@ue.katowice.pl

CONTENTS

INTRODUCTION

The volume consists of papers dealing with statistics. In general, applications of statistics in economics are presented. The authors are employed in the Department of Statistics of the University of Economics in Katowice. The papers can be divided into two groups. The first, larger one is connected with time series analysis. The other group consists of the paper on miscellaneous topics.

Alicja Ganczarek-Gamrot has prepared the paper entitled *Assessing long memory characteristics of energy prices*. It deals with inference on characteristics of financial time series with very strong autocorrelation. On the basis of the presented statistical methods, the properties of the Polish electricity market are analyzed. In the paper by Grzegorz Kończak entitled *On the moving block bootstrap method for monitoring autocorrelated processes*, an application of the autocorrelated process for the quality control involving the well known control chart concept has been described. The next paper entitled *Bayesian quantile regression* by Grażyna Trzpiot is devoted to the relationship analysis based on the quantile regression. The Bayesian version of that regression is presented. Grażyna Trzpiot and Przemysław Jeziorski have written the paper *Application of asymmetric least squares method in estimation of CaViaR models.* In the paper, the regression function estimated by means of the special vague least square method is presented. That recently proposed method is applied to the analysis of Polish capital market. The paper *On predictionof linear combination of domains' totals in longitudinal analysis* by Tomasz Żądło is connected with inference on domain parameters. The considerations are based on the superpopulation model assumed for longitudinal data.

The first paper in the group of miscellaneous papers, written by Agnieszka Orwat-Acedańska is entitled *The classification of Polish mutual balanced funds according to the management style, using Andrews estimators.* It is on the risk managing on the investment market. In the paper, the problem of the estimation the style coefficients based on confidence intervals is considered. The next paper, written by E. Witek is *The use of finite mixture models in the classification of the EU member states.* For the first time, the basic properties of finite mixture models technique are presented. the idea is applied to the detection of inho-

mogeneities of the euro and non-euro countries characterized by the selected economic indicators and convergence criteria. The last paper is entitled *Test-estimator and double test* by Janusz Wywiał is connected with an appropriate choice estimator or a test statistic. It is based on preliminary assumption of considered inference procedures.

Janusz L. Wywiał

Alicja Ganczarek-Gamrot

ASSESSING LONG MEMORY CHARACTERISTICS OF ENERGY PRICES

Introduction

Time series from financial and commodity markets are characterized by very strong autocorrelation. Significant value of autocorrelation coefficient for large lag indicates the long memory in time series. Often time series' long memory is the cause of nonstationarity. Then often time series have unit roots. The unit root can be eliminated from time series through integer differencing. But a lot of empirical results show that integer differencing of empirical time series is unnecessary. Very often degree of difference of time series is real and smaller than one. Then time series are stationary.

In this paper the results of estimating long memory effects in time series on Polish electric energy market are presented. Methods, which are described by Zivot and Wang[1] are used to estimate long memory effects.

1. Long memory

The stationary time series y_t has long memory or long range dependence if:

$$\rho(k) \underset{k\to\infty}{\to} C_\rho k^{-\alpha}, \tag{1}$$

[1] E. Zivot, J. Wang: *Modeling Financial Time Series with S-PLUS*. Springer, New York 2006.

where:

$\rho(k)=\frac{\mathrm{cov}(y_t, y_{t-k})}{s^2(y_t)}$ – (ACF(k)) is a autocorrelation function, $C_\rho > 0$ (constant), $\alpha \in (0,1)$.

The autocorrelation function of long memory process decays slowly at a hyperbolic rate. So the autocorrelations are not summable: $\sum_{k=-\infty}^{\infty} \rho(k) = \infty$.

For a stationary process autocorrelation function contains equivalent information to its spectral density:

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k) e^{ik\omega}, \tag{2}$$

where ω is the Fourier frequency[2] (c.f. Hamilton, 1994).

The spectral density of long memory process tends to infinity at zero frequency:

$$f(\omega) \underset{\omega \to 0}{\to} C_f \omega^{\alpha-1}, \tag{3}$$

where $C_f > 0$ (constant), $\alpha \in (0,1)$.

Two convergences (1) and (3) are equivalent. In practice very often the Hurst coefficient (H) is used[3] instead of α. For time series with long memory H is a real number from the (0.5;1) interval. The larger H the longer memory of the stationary process. Coefficients H and α satisfy the relation:

$$H = 1 - \frac{\alpha}{2}. \tag{4}$$

[2] J.D. Hamilton: *Time Series Analysis*. Princeton University Press, New Jersey 1994.

[3] H.E. Hurst: *Long Term Storage Capacity of Reservoirs*. "Transactions of the American Society of Civil Engineers" 1951, No. 116, p. 770-799.

Using properties (1) and (3) Granger and Joyeux[4] and Hosking[5] have shown that long memory process can be described parametrically by *fractionally integrated process*:

$$(1-L)^d(y_t-\mu)=\mu_t, \tag{5}$$

where $L^k y_t = y_{t-k}$ is a lag operator, d is a fractional difference parameter ($d > -1$), μ is the expected value of y_t, u_t is a stationary process with short memory and $E(u_t)=0$, $(1-L)^d = \sum_{k=0}^{\infty}\binom{d}{k}(-1)^k L^k$ is a fractional difference filter,

$$\binom{d}{k}=\frac{d!}{k!(d-k)!}=\frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}.$$

If $|d| > 1/2$ then y_t is non-stationary. If $0 < d < 1/2$ then y_t is stationary and has long memory. If $-1/2 < d < 0$ then y_t is stationary and has short memory. Often the Hurst coefficient is used instead of d in equation (5) following the relation:

$$d = H - 1/2. \tag{6}$$

Hence if $0 < H$ or $H > 1$ then y_t is nonstationary. If $1/2 < H < 1$ then y_t is stationary and has long memory so the time series is characterized by positive autocorrelation and small risk, because there is a high probability, that trend direction will not be changed. If $0 < H < 1/2$ then y_t is stationary and has short memory, time series is characterized by nonpositive autocorrelation. and small risk, because y_t will react on new information from the market. If $H = 0.5$ then y_t is a random process (for example Wiener's process or random walk)[6].

---

[4] C.W.J. Granger, R. Joyeux: *An Introduction to Long-Memory Time Series Models and Fractional Differencing*. "Journal of Time Series Analysis" 1980, No. 1, p. 15-29.

[5] J.R.M. Hosking: *Fractional Differencing*. "Biometrika" 1981, No. 68, p. 165-176.

[6] J. Stawicki, I. Frączek-Miller: *Różnicowanie fraktalne szeregów czasowych*. W: *Dynamiczne modele ekonometryczne*. Materialy na V Ogólnopolskie Seminarium Naukowe. TNOiK „Dom Organizatora", Toruń 1997.

1.1. Assessing long memory estimation

In this part of the paper three long memory estimation methods are presented: R/S statistics, GPH statistics and periodogram method.

Range over standard deviation (R/S) is the most known test of long memory. It was proposed by Hurst[7] and modified by Mandelbrot[8] and finally by Lo[9] to take the form:

$$\mathrm{R/S} = \frac{1}{\hat{\sigma}_T(q)}\left[\max_{1\le k\le T}\sum_{j=1}^{k}(y_j - \bar{y}) - \min_{1\le k\le T}\sum_{j=1}^{k}(y_j - \bar{y})\right], \tag{7}$$

where $\bar{y}$ mean of y_t, s_T standard deviation of y_t,

$\hat{\sigma}_T(q)$ is a root of long term variance with bandwidth q:

$\hat{\sigma}_T^2(q) = s_T^2 + \frac{2}{T}\sum_{j=1}^{q}\omega_j(q)\left[\sum_{i=j+1}^{T}(y_i - \bar{y})(y_{i-j} - \bar{y})\right]$, $\omega_j(q) = 1 - \frac{j}{q+1}$ – Bartlett's weights

$q = \left[4\left(\frac{T}{100}\right)^{\frac{1}{4}}\right]$, [x] – integral part of x.

Lo[10] showed that if stationary time series y_t has short memory, then R/S statistic converges to a Brownian bridge at rate $T^{1/2}$. Mandelbrot[11] showed that, if stationary time series y_t has long memory, then R/S statistic converges to Brownian bridge at rate T^H. The R/S statistic is used to estimate Hurst coefficient in the following way:

1) k_1 observations from empirical time series (large k_1) are chosen,
2) values of R/S statistics for $k_i = f^* k_{i-1}$ where $i = 2,\ldots, s$ are calculated,
3) a line fit of all these R/S statistics versus k_i, $i = 1,\ldots, s$, on the log-log scale yields an estimate of the Hurst coefficient H.

---

[7] H.E. Hurst: Op. cit.

[8] B.B. Mandelbrot: *Limit Theorems on the Self-Normalized Range for Weakly and Strongly Dependent Processes*. "Zeitschrift f'ur Wahrscheinlichkeitstheorie und verwandte Gebiete" 1975, No. 31, p. 271-285.

[9] A.W. Lo: *Long Term Memory in Stock Market Prices*. "Econometrica" 1951, No. 59, p. 1279-1313.

[10] Ibidem.

[11] B.B. Mandelbrot: Op. cit.

The weakness of the above procedure is the choice of k_l. On one hand large k_l is desired, on the other hand for large k_l only few values of the R/S statistic can be calculated. To mitigate this problem the Least Absolute Deviation (LAD) is used to estimate parameters of the line.

Geweke and Porter-Hudak[12] proposed semi-parametric test (GPH) based on spectral density function. The spectral density function for fractal differenced process is given by:

$$f(\omega) = [4\sin^2(\tfrac{\omega}{2})]^{-d} f_{ut}(\omega), \tag{8}$$

where ω is a Fourier frequency, $f_{ut}(\omega)$ is a spectral density function of the process (5).

The long memory parameter d (8) is estimated based on regression function:

$$\ln f(\omega_j) = \beta - d[4\sin^2(\frac{\omega_j}{2})] + \varepsilon_j \tag{9}$$

for $j = 1,2,\ldots n_f(T)$; $n_f(T) = T^{\alpha}$.

Geweke and Porter-Hudak[13] showed, that if $0 < \alpha < 1$, then d has normal distribution for long time series:

$$\hat{d} \sim N\left(d, \frac{\pi^2}{6\sum_{j=1}^{n_f}(U_j - \bar{U})^2}\right); \qquad U_j = \ln[4\sin^2(\tfrac{\omega_j}{2})].$$

By assuming that $d = 0$ (process has no long memory), GPH test statistic is given by:

$$t_{d=0} = \hat{d}\left(\frac{\pi^2}{6\sum_{j=1}^{n_f}(U_j - \bar{U})^2}\right)^{-\frac{1}{2}} \tag{10}$$

[12] J. Geweke, S. Porter-Hudak: *The Estimation and Application of Long Memory Time Series Models*. "Journal of Time Series Analysis" 1983, No. 4, p. 221-237.

[13] Ibidem.

The GPH statistic $t_{d=0}$ has standard normal distribution.

The periodogram can be used to estimate long memory too. Based on convergence (3) the log-log plot of periodogram versus the frequency should scatter around a straight line with slope $1-2H$ for frequencies close to zero.

1.2. Models describing long memory in time series

A lot of empirical research showed that long memory exist in financial time series. Lobato and Savin[14], Ray and Tsay[15]. Andersen, Bollerslev, Diebold and Labys[16] suggested to use FARIMA models to forecast daily volatility. In this part of paper models describing long memory in time series are presented.

The traditional approach to modeling time series with short memory is to use the ARIMA model:

$$\phi(L)(1-L)^d(y_t-\mu)=\theta(L)\varepsilon_t, \tag{11}$$

where polynomials $\phi(L)$, $\theta(L)$ are given by $\phi(L)=1-\sum_{i=1}^{p}\phi_i L^i$, $\theta(L)=1-\sum_{j=1}^{q}\theta_j L^j$, and ε_t is a white noise.

If d is a real number then ARIMA model is the fractionally integrated moving average ARFIMA (p, d, q) model, or simply, fractional ARIMA (p, d, q) FARIMA (p, d, q) model. Beran[17] extended the estimation of FARIMA models for any $d>-1/2$ by considering the following variation of the FARIMA model:

$$\phi(L)(1-L)^{\delta}((1-L)^m y_t-\mu)=\theta(L)\varepsilon_t, \tag{12}$$

where $-0.5 < \delta < 0.5$, $\varphi(L)$ and $\theta(L)$ are defined as above, m is the number of times that y_t must be differenced to achieve stationarity, $d=m+\delta$.

---

[14] I.N. Lobato, N.E. Savin: *Real and Spurious Long-Memory Properties of Stock-Market Data*. "Journal of Business and Economic Statistics" 1998, No. 16(3), p. 261-268.

[15] B.K. Ray, R.S. Tsay: *Long-Range Dependence in Daily Stock Volatilities*. "Journal of Business and Economic Statistics" 2001, No. 18, p. 254-262.

[16] T. Andersen, T. Bollerslev, F.X. Diebold, P. Labys: *(Understanding, Optimizing, Using and Forecasting) Realized Volatility and Correlation*. Manuscript. Northwestern University, Duke University and University of Pennsylvania, Pennsylvania 1999.

[17] J. Beran: *Maximum Likelihood Estimation of the Differencing Parameter for Invertible Short and Long Memory ARIMA Models*. "Journal of Royal Statistical Society Series B" 1995, No 57(4), p. 659-672.

In empirical research parameter m very often equals to 0 or 1:
- $m = 0$ $\mu = E(y_t)$,
- $m = 1$ $\mu = f(t)$.

Beran, Feng and Ocker[18] proposed – *Semiparametric Fractional Autoregressive* (SEMIFAR) FARIMA (p, d, 0):

$$\phi(L)(1-L)^{\delta}((1-L)^{m} y_t - g(i_t)) = \varepsilon_t, \tag{13}$$

where $i_t = t / T$ for $t = 1,\ldots, T$, $g(i_t)$ is a smooth trend function on [0, 1].

Bayesian Information Criteria BIC can be used to choose the short memory autoregressive order p.

Any GARCH (p, q) model can be written by ARMA (m, q) model:[19]

$$\phi(L)\varepsilon_t^2 = \sigma^2 + b(L)u_t, \tag{14}$$

where $u_t = \varepsilon_t^2 - \sigma_t^2$, $\phi(L) = 1 - \sum_{i=1}^{m} \phi_i L^i$, $b(L) = 1 - \sum_{j=1}^{q} b_j L^j$, $\phi_i = a_i + b_i$, $m = \max(p,q)$.

The high persistence in GARCH models suggests that the polynomial $\varphi(z) = 0$ may have a unit root. In this case the GARCH model becomes the integrated GARCH (IGARCH) model. The ARMA (m, q) process extends to a FARIMA (m, d, q) process to allow for high persistence and long memory in the conditional variance as follows[20]:

$$\phi(L)(1-L)^{d}\varepsilon_t^2 = \sigma^2 + b(L)u_t, \tag{15}$$

where all the roots of $\phi(z) = 0$ $b(z) = 0$ are outside the unit circle.

---

[18] J. Beran, Y. Feng, D. Ocker: *SEMIFAR Models*. "Technical Report" 3/1999, SFB 475, University of Dortmund, Dortmund 1999; J. Beran, D. Ocker: *SEMIFAR Forecasts, with Applications to Foreign Exchange Rates*. "Journal of Statistical Planning and Inference" 1999, No. 80, p. 137-153; J. Beran, D. Ocker: *Volatility of Stock Market Indices – An Analysis Based on SEMIFAR Models*. "Journal of Business and Economic Statistics" 2001, No. 19(1), p. 103-116.

[19] GARCH(p,q) $\sigma_t^2 = \sigma^2 + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} b_i \sigma_{t-j}^2$

[20] E. Zivit, J. Wang: Op. cit.

When

- $d = 0$ – model (15) reduces to the usual GARCH model,
- $d = 1$ – model (15) becomes the IGARCH model,
- $0 < d < 1$ – model (15) becomes the FIGARCH model.

However, to guarantee that a general FIGARCH model is stationary and the conditional variance is always positive, usually complicated and intractable restrictions have to be imposed on the model coefficients[21]. Noting that an EGARCH model can be represented as an ARMA process in terms of the logarithm of conditional variance and thus always guarantees that the conditional variance is positive, Bollerslev and Mikkelsen[22] proposed the following fractionally integrated EGARCH (FIEGARCH) model:

$$\ln \sigma_t^2 = \sigma^2 + (1 - a(L) - b(L))^{-1}(1 - L)^{-d}(1 + a(L))g(\xi_{t-1}), \qquad (16)$$

where $g(\xi_t) \equiv \gamma_1 \xi_t + \gamma_2(|\xi_t| - E|\xi_t|)$ depend on white noise's sign (first component) in time t and white noise's value of standard deviation in time t (second component). Bollerslev and Mikkelsen[23] showed that FIEGARCH is stationary if $0 < d < 1$.

2. Long memory on Polish electric energy market

In this part of the paper time series of electric energy prices from 29th March to 24th October 2009 are taken into consideration. The analysis focuses on prices from two day ahead markets: Day Ahead Market (RDN) of Polish Power Exchange and Conventional Energy Spot Market (RDK) of Internet Electricity Trading Platform. On both markets fixing prices are established two times a day. On RDN the first fixing (RDN1) is established at 8:00 o'clock and second one (RDN2) at 10:30 o'clock. Each fixing price from RDN is established one day before a delivery of the electric energy. On RDK the first fixing (RDK1) is established eleven minutes later then the first fixing on RDN on one day before a delivery of the energy. Second fixing (RDK2) on RDK is establish at 11:51 o'clock in holiday and at 14:01 o'clock in weekday on two days before a deliv-

[21] R.T. Baillie, T. Bollerslev, H.O. Mikkelsen: *Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity.* "Journal of Econometrics" 1996, No. 74, p. 3-30; T. Bollerslev, H.O. Mikkelsen: *Modeling and Pricing Long Memory in Stock Market Volatility.* "Journal of Econometrics" 1996, No. 73, p. 151-184; E. Zivot, J. Wang: Op. cit.

[22] T. Bollerslev, H.O. Mikkelsen: Op. cit.

[23] Ibidem.

ery of the electric energy. Each of four fixing prices is established independently for each one of 24 hours a day. The price for each hour balances the aggregate supply and demand for this hour.

The aim of this paper is to estimate memory of time series from Polish day ahead markets. In analysis four time series of fixing prices were used:

1) fixing prices: RDK1, RDK2, RDN1, RDN2,
2) hourly rates of return: z_t RDK1, z_t RDK2, z_t RDN1, z_t RDN2,
3) daily rates of return: z_t 24RDK1, z_t 24RDK2, z_t 24RDN1, z_t 24RDN2,
4) weekly rates of return: z_t 168RDK1, z_t 168RDK2, z_t 168RDN1, z_t 168RDN2,
5) z_t 192RDK1, z_t 192RDK2, z_t 192RDN1, z_t 192RDN2 rates of return, which include influence of daily and weekly seasonality.

2.1. Estimation

In this part of the paper Hurst coefficient was estimated for considered time series. Obtained coefficients are shown in Table 1.

Table 1

The Hurst coefficient for several time series

Szereg	R/S	GPH	Zbieżność (3)
	H (średni błąd szacunku)	H = d + 0,5 (średni błąd szacunku)	H = d + 0,5 (średni błąd szacunku)
RDK1	0.7934 (0.0407)	1.1614 (0.0991)	1.0102 (0.0790)
RDK2	0.7841 (0.0423)	1.0938 (0.1107)	0.9828 (0.0827)
RDN1	0.7864 (0.0408)	1.1252 (0.1161)	0.9934 (0.0819)
RDN2	0.7898 (0.0409)	1.0918 (0.1116)	0.9680 (0.0752)
z_t RDK1	0.3662 (0.0409)	0.5979 (0.0936)	0.5209 (0.0707)
z_t RDK2	0.3511 (0.0380)	0.5017 (0.0933)	0.4459 (0.0762)
z_t RDN1	0.3484 (0.0390)	0.5611 (0.1015)	0.4756 (0.0764)
z_t RDN2	0.3668 (0.0423)	0.6078 (0.1101)	0.4081 (0.0754)
z_t 24RDK1	0.5305 (0.0542)	0.0751 (0.0927)	0.1105 (0.0792)
z_t 24RDK2	0.5172 (0.0567)	0.0914 (0.1018)	0.1034 (0.0833)
z_t 24RDN1	0.5245 (0.0555)	0.0754 (0.0937)	0.1004 (0.0813)
z_t 24RDN2	0.5366 (0.0529)	0.0661 (0.0946)	0.0828 (0.0737)
z_t 168RDK1	0.8216 (0.0219)	1.0135 (0.1017)	0.9955 (0.0696)
z_t 168RDK2	0.8053 (0.0202)	0.9450 (0.1048)	0.9499 (0.0651)
z_t 168RDN1	0.8114 (0.0229)	0.9490 (0.1048)	0.9478 (0.0650)
z_t 168RDN2	0.8142 (0.0206)	0.8949 (0.0934)	0.9061 (0.0639)
z_t 192RDK1	0.7332 (0.0243)	0.8163 (0.1249)	0.7637 (0.0832)
z_t 192RDK2	0.7155 (0.0253)	0.7786 (0.1191)	0.7210 (0.0872)
z_t 192RDN1	0.7218 (0.0258)	0.7711 (0.1153)	0.7279 (0.0865)
z_t 192RDN2	0.7346 (0.0236)	0.7117 (0.1041)	0.6876 (0.0757)

Three methods of estimation lead to different results. The GPH (10) statistics and convergence (3) resulted in most similar coefficient values. But these results have large error estimate. The coefficients, which were estimated by R/S statistics, have the smallest error estimate, but they differ widely from others values. The result of R/S statistics depends on the choice of k_l. In this analysis $k_l = 9$ and $f = 2$. This way ten groups of value (7) was obtained. Additionally the convergences (1) and (3) are true for $0.5 < H < 1$.

Based on the value of H estimated by R/S statistic and convergence (3) one can say, that time series of electric energy prices: RDK1, RDK2, RDN1, RDN2 are stationary and have long memory. So they are characterized by positive autocorrelation. In risk analysis, one can say, that this time series are characterized by small risk. Large values of H suggest nonstationarity. Based on GPH result, we can say, that time series of electric energy prices are nonstationary.

Time series of linear rates of return of electric energy prices: z_t RDK1, z_t RDK2, z_t RDN1, z_t RDN2 are stationary and have short memory. They are characterized by negative autocorrelation. In this case probability, that time series rates of return will respond to new market information, is greater, than the probability, that these time series will change like in the past. For many cases values of H are close to 0.5, which means, that these time series change randomly. But these time series are not random. This result is an effect of seasonal autocorrelation on electric energy market[24] (Figure 2).

Based on R/S statistics the time series: z_t 24RDK1, z_t 24RDK2, z_t 24RDN1, z_t 24RDN2 without daily seasonality are random. But based on GPH statistics and convergence (3) we can say, that these time series are characterized by short memory. The reason for such an outcome is weekly seasonality[25].

The time series: z_t 168RDK1, z_t 168RDK2, z_t 168RDN1, z_t 168RDN2 without weekly seasonality are stationary and have long memory. The time series: z_t 192RDK1, z_t 192RDK2, z_t 192RDN1, z_t 192RDN2 without daily and weekly seasonality are stationary and have long memory too.

[24] A. Ganczarek: *Klasyfikacja Polskiego Rynku Energii.* W: Materiały Ogólnopolskiej Konferencji Naukowej pt. *Inżynieria ekonomiczna w badaniach społeczno-gospodarczych.* Oficyna Wydawnicza Politechniki Rzeszowskiej, Rzeszów 2003. 51-66; Idem: *Weryfikacja modeli z grupy GARCH na dobowo-godzinnych rynkach energii elekrycznej w Polsce.* W: *Rynek Kapitałowy. Skuteczne inwestowanie.* Studia i Prace WNEiZ, nr 9, s. 524-536; Idem: *Choosing between the Skewed Distribution and the Skewed Model in General Autoregressive Conditional Heteroscedasticity.* Proceedings of 26th International Conference Mathematical Methods in Economics 2008. Technical University of Liberec, Liberec 2008, p. 132-139.

[25] Ibidem.

Based on varying results of H estimation in Table 1 we couldn't clearly say which of fixing is more or less risky (have short or long memory).

On Figure 1-5 ACF(k) (k = 250) and periodogram for morning fixing on RDK: RDK1, z_t RDK1, z_t 24RDK1, z_t 168RDK1, z_t 192RDK1 are presented.

Considering the result of ACF and periodograms one can say, that results of H estimation depend on seasonality of electric energy. However, for time series: z_t 168RDK1, z_t 192RDK1 seasonal effect is weak. Significant ACF coefficient for large lag and high value of periodogram at zero frequency informs about long memory in these time series. These results confirm findings from Table 1.

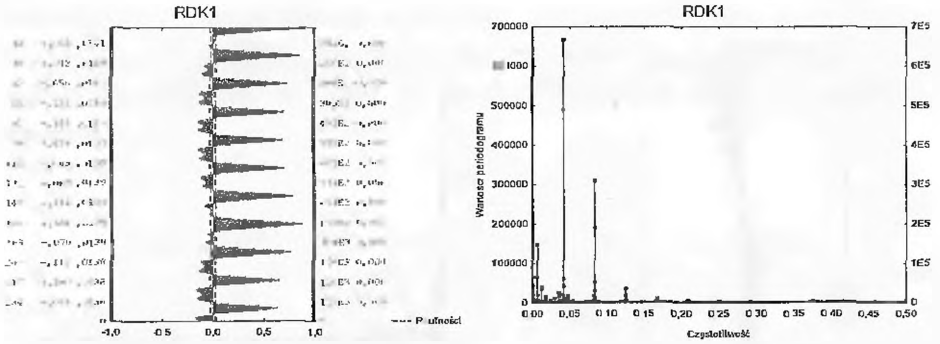


Figure 1. ACF and periodogram for RDK1

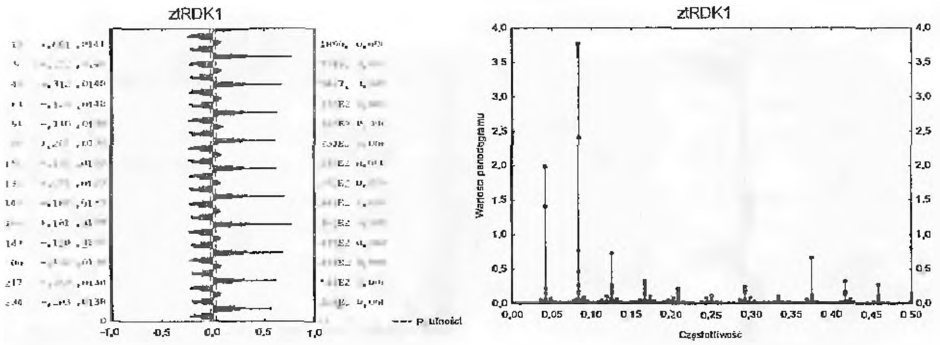


Figure 2. ACF and periodogram for z_t RDK1

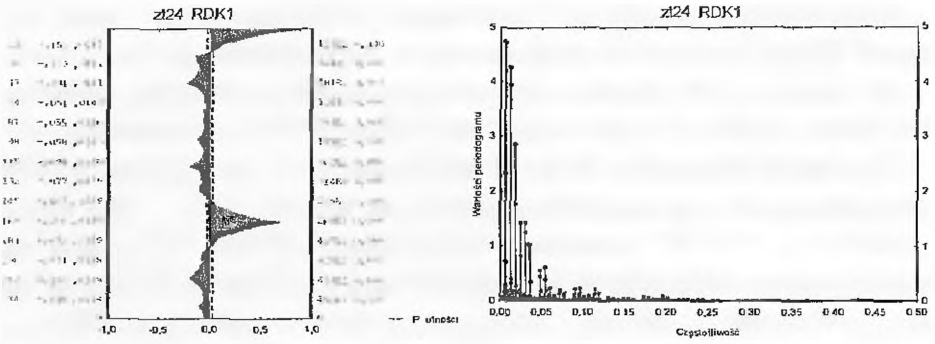


Figure 3. ACF and periodogram for *z*, 24RDK1

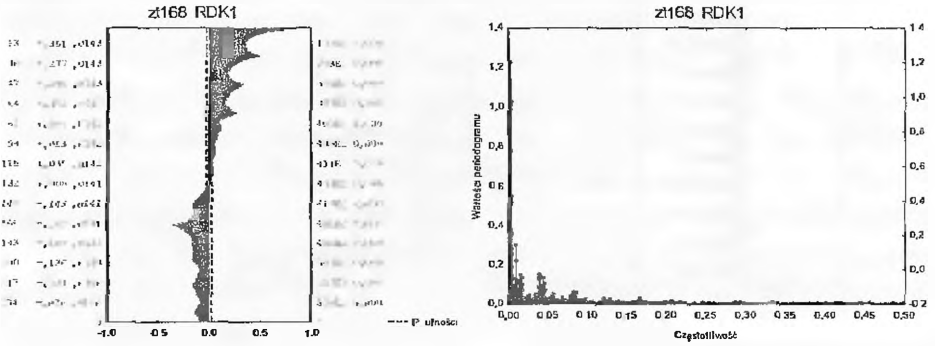


Figure 4. ACF and periodogram for *z*, 168RDK1

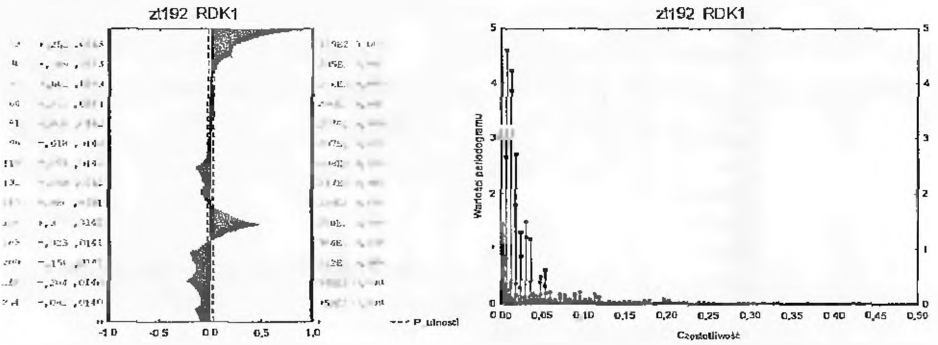


Figure 5. ACF and periodogram for *z*, 192RDK1

2.2. Day ahead markets

Empirical research[26] as well as the value of ACF and periodograms (Figure 1-5) indicate, that time series from day ahead markets are characterized by daily and weekly seasonality. Period of this research is restricted to summer time, so in this analysis yearly seasonality was omitted. This part of the paper focuses on long memory models for considered time series. To describe time series with seasonality: RDK1, RDK2, RDN1, z_t RDK1, z_t RDK2, z_t RDN1, z_t RDN2, z_t 24RDK1, z_t 24RDK2, z_t 24RDN1, z_t 24RDN2, z_t 168RDK1, z_t 168RDK2, z_t 168RDN1, z_t 168RDN2, SARIMA models should be used. These models do not describe long memory, but only seasonality[27], therefore this part of the paper focused on time series without seasonality effect: z_t 192RDK1, z_t 192RDK2, z_t 192RDN1, z_t 192RDN2. For time series mentioned above SEMIFAR-GARCH models are estimated. In Tables 2-5 result of SEMIFAR-GARCH estimation is presented.

Table 2

Result of SEMIFAR(1, 0.39)-GARCH(1, 1) estimation for z_t 192RDK1

	coefficient	standard error	p-value
d-ARFIMA (δ)	0.3942	0.0472	< 0.01
ϕ_1	0.5258	0.0535	< 0.01
σ^2	0.0006	0.0001	< 0.01
a_1	0.6738	0.1002	< 0.01
b_1	0.2508	0.0458	< 0.01

Table 3

Result of SEMIFAR(1, 0.33)-FIGARCH(1, 0.52, 1) estimation for z_t 192RDK2

	coefficient	standard error	p-value
d-ARFIMA (δ)	0.3320	0.0777	< 0.01
ϕ_1	0.5871	0.0757	< 0.01
d-FIGARCH	0.5199	0.0698	< 0.01
a_1	0.9071	0.0313	< 0.01
b_1	0.9831	0.0035	< 0.01

[26] Ibidem.

[27] A. Ganczarek: *Weryfikacja modeli*..., op. cit., Idem: *Choosing between*..., op. cit.

Table 4

Result of SEMIFAR(1, 0.42)-FIEGARCH(1, 0.64, 1) estimation for z_t 192RDN1

	coefficient	standard error	p-value
d-ARFIMA (δ)	0.4191	0.0570	<0.01
ϕ_1	0.5203	0.0501	<0.01
d-FIEGARCH	0.6444	0.0379	<0.01
a_1	-0.8413	0.0435	<0.01
b_1	0.7118	0.0500	<0.01
γ_1	0.2263	0.0379	<0.01
γ_2	0.7655	0.0671	<0.01

Table 5

Result of SEMIFAR(1, 0.44)-FIEGARCH(1, 0.66, 1) estimation for z_t 192RDN2

	coefficient	standard error	p-value
d-ARFIMA (δ)	0.4412	0.0601	< 0.01
ϕ_1	0.4757	0.0713	< 0.01
d-FIEGARCH	0.6571	0.0399	< 0.01
a_1	-0.8204	0.0519	< 0.01
b_1	0.7133	0.0577	< 0.01
γ_1	0.1802	0.0340	< 0.01
γ_2	0.6840	0.0601	< 0.01

Obtained results acknowledge the presence of long memory in time series. Besides, long memory effect is present in expectation and variance of time series (Table 2). The exception is only z_t 192RDK1, which has long memory only in expectation. In every model value of d belongs to the interval (0; 0.5), which means, that all models are stationary and have long memory. Moreover greater value of d for time series from RDN than for time series from RDK indicate, that time series from RDN have longer memory than time series from RDK. So the obtained models for time series show, that on RDN there is a little lower risk than on RDK, but differences between estimated parameters are modest. Also residuals of obtained models are characterized by significant autocorrelation and have not normal distribution, so parameter estimates can be biased.

Conclusion

The obtained result show, that time series from day ahead electric energy markets in Poland have long memory. The long memory should be modeled taking under consideration seasonality, autocorrelation and heteroscedasticity of time series from electric energy markets.

Presented models can be used to predict expected value and expected volatility of time series from discussed markets. So the results can be used to estimate expected profits and – in case of risk analysis – estimated future loss.

EFEKT DŁUGIEJ PAMIĘCI W SZEREGACH CZASOWYCH CEN ENERGII ELEKTRYCZNEJ

Streszczenie

W pracy dokonano przeglądu metod, służących do estymacji długiej pamięci szeregów czasowych. Za pomocą wykładnika Hursta, szacowanego niezależnie trzema metodami, oraz modeli uwzględniających ułamkową integrację szeregów czasowych wykonano ocenę długiej pamięci szeregów czasowych na polskim rynku energii elektrycznej. W estymacji długiej pamięci uwzględniono wpływ cykliczności, autokorelacji oraz heteroskedastyczności szeregów czasowych cen energii elektrycznej. Otrzymane wyniki pokazały, że w szeregach czasowych polskiego dobowo-godzinowego rynku energii elektrycznej obecny jest efekt długiej pamięci.

Grzegorz Kończak

ON THE MOVING BLOCK BOOTSTRAP METHOD FOR MONITORING AUTOCORRELATED PROCESSES

Introduction

The statistical control chart concept was developed in 1924 by Walter A. Shewhart. The control chart is a graphical display of a quality characteristic such as sample mean, standard deviation or range.

The main assumptions that are usually cited in the use of control charts are that the data generated by the under control process are normally and independently distributed with mean μ and standard deviation σ. The parameters μ and σ are usually unknown and should be estimated. An out-of-control condition is a change or shift in μ and/or σ to some different value. The assumption of uncorrelated or independent observation is not even approximately satisfied in some manufacturing processes.

The methods for monitoring autocorrelated processes were analyzed by some authors. Datta and McCormic[1] considered bootstrap method for a first-order autoregressive model. Liu and Tang[2] presented the construction of the control chart based on the bootstrap method. Kończak[3] proposed the use of Markov chains for monitoring autocorrelated processes.

---

[1] S. Datta, W.P. McCormic: *Bootstrap Inference for a First-Order Autoregression with Positive Innovations*. "Journal of the American Statistical Association" December 1995, Vol. 90, No. 432.

[2] R.Y. Liu, J. Tang: *Control Chart for Dependent and Independent Measurements Based on Bootstrap Methods*. "Journal of the American Statistical Association" 1996, No. 436.

[3] G. Kończak: *Łańcuchy Markowa w analizie własności procedur kontroli jakości*. W: *Klasyfikacja i analiza danych – teoria i zastosowania*. Red. K. Jajuga, M. Walesiak. „Taksonomia" vol. 11, Wydawnictwo Akademii Ekonomicznej, Wroclaw 2004.

The proposal of the control chart for monitoring autoregressive processes AR(1) is presented in the paper. The proposal is based on the moving blocks bootstrap method[4].

1. Autoregressive processes

The autoregressive model of order p[5] can be written as follows:

$$Z_t = \alpha + \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + ... + \varphi_p Z_{t-p} + \varepsilon_t \quad \text{for } t = 1,2,..., \tag{1}$$

where ε_t are independent and identically distributed with zero mean and finite variance σ^2. This model is usually denotes by AR(p).

In this paper we will consider the stationary autoregressive model of order 1:

$$Z_t = \alpha + \varphi Z_{t-1} + \varepsilon_t, \quad \text{for } t = 1,2,..., \tag{2}$$

where $-1 < \varphi < 1$.

The model (2) can be equally rewritten as follows:

$$Z_t = \mu + \varphi(Z_{t-1} - \mu) + \varepsilon_t, \quad \text{for } t = 1,2,..., \tag{3}$$

where $\mu = \dfrac{\alpha}{1-\varphi}$ is the process mean.

The parameter φ of the autoregressive model (2) can be estimated[6] using formula:

$$\hat{\varphi} = \frac{\sum_{t=2}^{k} (z_t - \bar{z})(z_{t-1} - \bar{z}_{(-1)})}{\sum_{i=2}^{k} (z_{t-1} - \bar{z}_{(-1)})^2}, \tag{4}$$

where:

$$\bar{z} = \frac{\sum_{t=2}^{n} z_t}{n-1} \text{ and } \bar{z}_{(-1)} = \frac{\sum_{t=2}^{n} z_{t-1}}{n-1}.$$

[4] B. Efron, R. Tibshirani: *An Introduction to the Bootstrap*. Chapman & Hall, New York 1993.

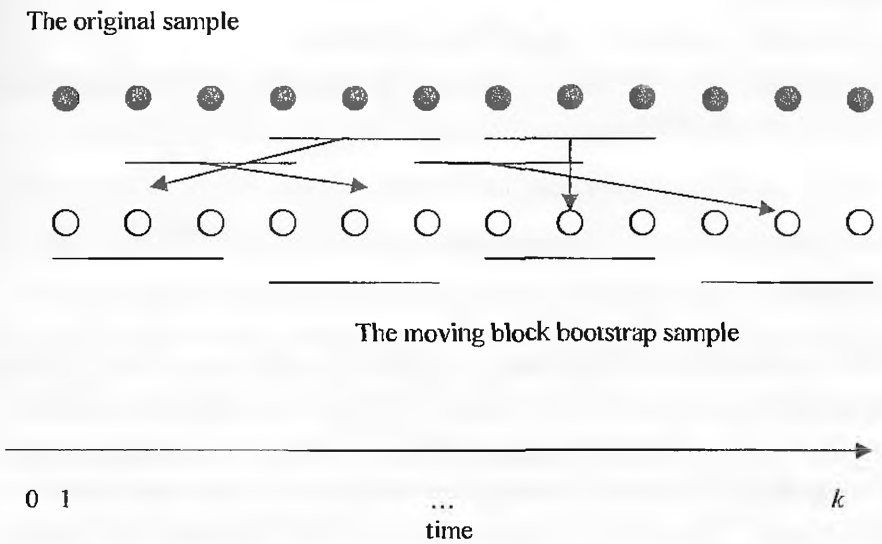
[5] G.E.P. Box, G.M. Jenkins: *Analiza szeregów czasowych. Prognozowanie i sterowanie*. PWN, Warszawa 1983.

[6] R.K. Rayner: *Bootstraping p Values and Power in the First Order Autoregression: A Monte Carlo Investigation*. "Journal of Business & Economic Statistics" April 1990, Vol. 8, No. 2.

2. Moving blocks bootstrap

Let us assume that $z_1, z_2, ..., z_k$ is the sample from the in-control AR(1) process. Let us denote by $z_1^*, z_2^*, ..., z_k^*$ the bootstrap sample which is sampled with replacement from the previous sample. Bootstrap sampling can be used for processes with independent observations. In the autoregressive cases the bootstrap sampling destroys the correlation in observed time series.

In the autocorrelated processes the moving block bootstrap[7] can be used instead of the classical bootstrap method. The main idea of the moving block bootstrap is presented in Figure 1. In this method we sample with replacement the entire blocks of lengths m and paste them together to form the bootstrap series. In Figure 1 m is equal to 3. The moving blocks bootstrap is less model dependent than the classical bootstrap method.



Source: Based on B. Efron, R. Tibshirani: *An Introduction to the Bootstrap*. Chapman&Hall, New York 1993

Figure 1. The diagram of the moving block bootstrap for time series

[7] B. Efron, R. Tibshirani: Op. cit.

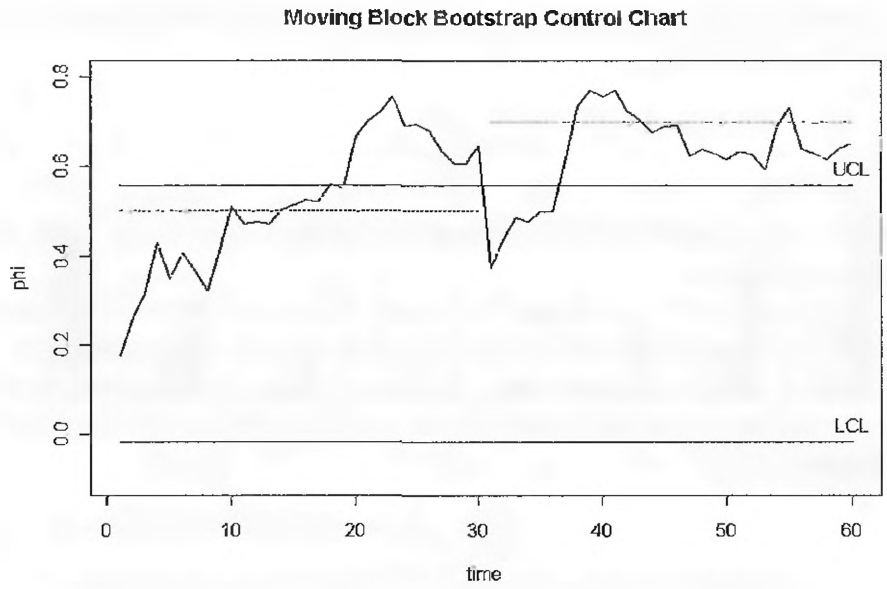
3. The construction of the moving block bootstrap control chart

Let us assume that the process AR(1) given by (3) with unknown parameter φ is monitored. The sample $(z_1, z_2, \ldots, z_k)$ from in-control process AR(1) was taken. The estimated values $\hat{\varphi}_t$ $(t = 1,2, \ldots)$ of the parameter φ for each sample of size k will be plotted in the control chart. To construct the control limits we need estimated quantilies of the distrubtion of $\hat{\varphi}$. The control limits should be calculated using the moving block bootstrap method. The procedure for determining the control limits under the assumed significance level α for the moving block bootstrap control chart is as follows:

1. The sample $(z_1, z_2,\ldots, z_k)$ is taken from the in-control AR(1) process.
2. The moving block bootstrap sample from $(z_1^*, z_2^*,\ldots, z_k^*)$ is taken from the original sample.
3. The parameter φ using the formula (4) is estimated.
4. The steps 2-3 are repeated B times. Let $\hat{\varphi}_1, \hat{\varphi}_2,\ldots,\hat{\varphi}_B$ denote the estimated values of the parameter φ.
5. Let $\varphi_{\alpha/2}$ and $\varphi_{1-\alpha/2}$ are quantilies of orders $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$. These quantiles denote the lower and upper control limits in the moving block bootstrap control chart.

Let us consider the sample $y_1, y_2,\ldots, y_k$ which is from in- or out-of-control process. In time t $(t = k + 1, k + 2,\ldots)$ the value of $\hat{\varphi}_t$ based on the sample $(y_{t-k+1}, y_{t-k+2},\ldots, y_t)$ is calculated and the point $(t, \hat{\varphi}_t)$ is plotted in the control chart.

The example of the control chart based on the moving block bootstrap is presented in Figure 2. Based on the first sample of size $n = 30$ the control lines were established. The second sample of size $n = 30$ was generated with shift in the parameter φ. The dotted lines show the true value of the parameter φ for the first in-control process sample and for the second sample.



Source: The result of the Monte Carlo study

Figure 2. The example of the moving block bootstrap control chart

4. The simulation study

The simulation study was prepared for determining the properties of the proposed moving block bootstrap control chart. There were analyzed blocks of length $m = 3$, 4 and 6. The steps of the simulation study was as follows:

1. The in-control sample $(z_1, z_2,\ldots, z_{30})$ from the AR(1) process with $\varphi = 0.5$ was generated.
2. Based on the sample from in-control process the control lines LCL and UCL were determined ($\alpha = 0.05$).
3. The sample $(y_1, y_2,\ldots, y_{130})$ from the autoregressive process AR(1) with parameter φ_s ($s = 1, 2,\ldots, 8$, see Table 1) was generated.
4. For $n = 100$ time series samples $(y_1, y_2,\ldots, y_{30})$, $(y_2, y_3,\ldots, y_{31})$,…, $(y_{101}, y_{102},\ldots, y_{130})$ the $\hat{\varphi}_1, \hat{\varphi}_2,\ldots, \hat{\varphi}_{100}$ were calculated.
5. The number of signals (v) is determined:

$$v = card\{t : \hat{\varphi}_t > UCL) + card\{t : \hat{\varphi}_t < LCL) \quad . \tag{5}$$

6. The steps 1-5 were repeated $B = 1000$ times.

The probability that the chart point falls outside the control limits is estimated using formula:

$$p = \frac{1}{B}\sum_{b=1}^{B} v_b \,, \tag{6}$$

where v_b is the number of points falling outside the control limits in the b-th bootstrap replication.

The properties of control charts are usually determined by the probability of falling points outside the control limits or by *ARL*. Average Run Length (*ARL*) is the expected number of samples from the start of process until a signal is given. *ARL* is connected to the probability that the point falls outside the control limits by formula[8]:

$$ARL = \frac{1}{p}. \tag{7}$$

The parameters φ_s and the estimated probabilities p_s are presented in Table 1. There were considered blocks of length of 3, 4, and 6 elements in each case.

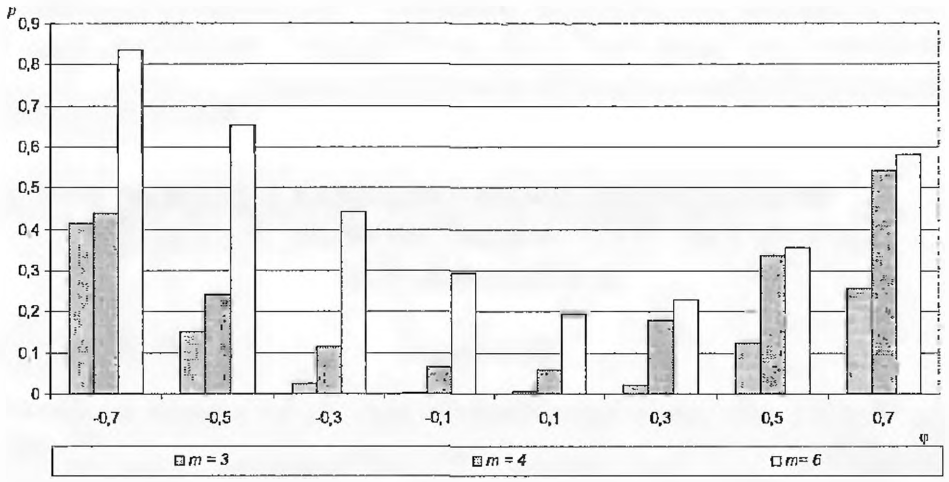
Table 1

The estimated probabilities that the chart point falls outside the control limits

s	φ_s	The block length (m)		
		$m=3$	$m=4$	$m=6$
1	- 0.7	0.4132	0.4404	0.8354
2	- 0.5	0.1517	0.2421	0.6541
3	- 0 3	0.0259	0.1153	0.4409
4	- 0.1	0.0043	0.0636	0.2937
5	0.1	0.0059	0.0581	0.1958
6	0.3	0.0228	0.1809	0.2304
7	0.5	0.1237	0.3356	0.3552
8	0.7	0.2569	0.5421	0 5820

Source: The result of the Monte Carlo study

[8] D.C. Montgomery: *Introduction to Statistical Quality Control*. John Wiley & Sons. New York 1996.



Source: The result of the Monte Carlo study (Table 1)

Figure 2. The estimated probabilities that the chart point falls outside the control limits

The Monte Carlo study has shown that the moving block bootstrap method can be used for monitoring autoregressive processes. The moving block bootstrap method destroys the correlation in the observed time series. The increase of the absolute value of the autoregressive φ parameter can be detected by the proposed control chart. This method is not as good in the case of decreasing the absolute value of the φ parameter as in the increasing case. The length of the moving blocks should not to be too small but the increase of the length leads to the increase of the probability of the conclude that the process is out of control when it is really in control (probability of type I error).

Conclusion

The classical control charts are constructed under normality and independence assumptions. In many situations we may have reason to doubt the validity of the independence assumption. The proposal of the control chart for monitoring autocorrelated processes is presented in the paper. The properties of this control chart is analyzed in the Monte Carlo study.

The Monte Carlo study has shown that the moving block bootstrap method can be used for monitoring autoregressive processes. If the length of the block is small then the method destroys the correlation in observed time series. The large

sizes of blocks are preferred in the monitoring of the autoregressive processes. This method gives better results in the case of detecting changes in the case of increasing the absolute value of the autoregressive parameter.

WYKORZYSTANIE METODY RUCHOMYCH BLOKÓW BOOTSTRAPOWYCH W MONITOROWANIU PROCESÓW Z AUTOKORELACJĄ

Streszczenie

Metody monitorowania procesów produkcyjnych z wykorzystaniem kart kontrolnych zostały wprowadzone w 1924 roku przez Waltera A. Shewharta. Klasyczne karty kontrolne wymagają spełnienia założenia normalności rozkładu oraz niezależności pomiarów. W praktyce bardzo często założenia te nie są spełnione. W artykule przedstawiono propozycję monitorowania procesów z występującą autokorelacją. Zaproponowana karta kontrolna wykorzystuje metodę ruchomych bloków bootstrapowych. Ponadto w artykule omówiono konstrukcję karty kontrolnej oraz zbadano jej właściowści z wykorzystaniem symulacji komputerowej.

Grażyna Trzpiot

BAYESIAN QUANTILE REGRESSION

Introduction

There is intensive research on the return forecasts for securities, and most of it has focused on the conditional mean estimation strategy. The classical least squares methods and maximum likelihood estimator provide attractive methods of estimation for Gaussian linear equation models with additive errors. However, these methods offer only a conditional mean view of the causal relationship, implicitly imposing quite restrictive location-shift assumptions on the way that covariates are allowed to influence the conditional distributions of the response variables. Quantile regression methods seek to broaden this view, offering a more complete characterization of the stochastic relationship among variables and providing more robust, and consequently more efficient, estimates in some non-Gaussian settings.

Since the seminal paper of Koenker and Bassett[1], quantile regression has gradually become a complimentary approach for the traditional conditional mean estimation methods. Most of applications of quantile regression in finance have been focused on conditional value at risk models. Engle and Manganelli[2] have elegantly laid out the natural interpretation of VaR models as the concept of quantile regression and considered the application of quantile regression to a nonlinear autoregressive VaR model. The estimation method and application on Polish capital market based on quantile regression was evaluated in some previous earliest publications[3].

---

[1] R. Koenker, G. Bassett: *Regression Quantiles*. "Econometrica" 1978, No. 46, p. 33-50.

[2] Engle, Manganelli: *CaViaR: Conditional Autoregressive Value at Risk by Regression Quantiles*. "Journal of Business and Economic Statistics" 2004, No. 22, p. 367-381.

[3] G. Trzpiot: *Regresja kwantylowa a estymacja VaR*. W: *Inwestycje finansowe i ubezpieczenia – tendencje światowe a polski rynek*. Red. W. Ronka-Chmielowiec, K. Jajuga. Wydawnictwo Akademii Ekonomicznej, Wrocław 2007, s. 465-471; Idem: *Implementacja metodologii regresji kwantylowej w estymacji VaR*. W: *Rynek kapitałowy. Skuteczne inwestowanie*. Studia i Prace, nr 9, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin 2008, s. 316-323; Idem: *Estimation Methods for Quantile Regression*. W: *Restrukturyzacja spółek handlowych. Zagadnienia ekonomiczne i prawne*. Red. J. Kolonko, J.L. Wywial. Studia Ekonomiczne, nr 53, Wydawnictwo Akademii Ekonomicznej, Katowice 2009, s. 81-90.

In this article we present chosen method for estimation Bayesian quantile regression. As first we present a method quite flexible in a way of doing Bayesian inference about quantile regression models. This method can be implemented without using a potentially restrictive parametric likelihood[4]. The second method which we present is a quantile curve fitting with a Bayesian method. The Bayesian method applies the reversible jump Markov chain Monte Carlo approach. To fit conditional quantile functions, the model based on a well-defined likelihood function – the asymmetric Laplace likelihood function with both location and scale parameters.

1. Quantiles – Bayesian approach

Consider first Bayesian inference about quantiles. The τ 'th quantile θ_τ satisfies the moment restriction $E(g) = 0$, where:

$$g = \mathrm{I}(y \le \theta_\tau) - \tau, \tag{1}$$

where $\mathrm{I}(\cdot)$[5] is the indicator function.

First method which we described can using Bayesian exponentially tilted empirical likelihood, provides a likelihood for data y subject only to a set of m moment conditions of the form

$$\mathrm{E}g(y, \theta) = 0, \tag{2}$$

where θ is a k dimensional parameter of interest and k may be smaller, equal to or larger than m. The method may be thought of as construction of a likelihood supported on the n data points that is minimally informative, in the sense of maximum entropy, subject to the moment conditions. Specifically the probabilities $\{p_i\}$ attached to the n data points are chosen to solve:

$$\max_{p} \sum_{i=1}^{n} - p_i \log p_i ,$$

$$\text{subject to } \sum_{i=1}^{n} p_i = 1 \text{ and } \sum_{i=1}^{n} p_i \; g(y_i, \theta) = 0. \tag{3}$$

[4] S.M. Schennach: *Bayesian Exponentially Tilted Empirical Likelihood.* "Biometrika" 2005, No. 92 (1), p. 31-46.

[5] I[A] = 1 if A is true, I[A] = 0 otherwise

The solutions of this problem, is the maximum entropy and take the form:

$$p_i(\theta) = \frac{exp\{\lambda(\theta)' g(y_i, \theta)\}}{\sum_{i=1}^{n} exp\{\lambda(\theta)' g(y_i, \theta)\}}, \tag{4}$$

where m vector λ, dependent on θ, satisfies

$$\lambda(\theta) = arg\,min_{\eta} \frac{1}{n} \sum_{i=1}^{n} exp\{\eta' g(y_i, \theta)\}. \tag{5}$$

The $\{\lambda_i\}$ are the Lagrange multipliers corresponding to the m constraints in the problem (3) and for every θ the last equation is a convex minimization problem and computationally straightforward. The resulting likelihood for i.i.d. data is $\prod_{i=1}^{n} p_i(\theta)$ and this may be combined with a prior density on θ to yield the posterior density:

$$p(\theta|Y) = p(\theta) \prod_{i=1}^{n} p_i(\theta), \tag{6}$$

on a support such that θ is in the convex hull of the $g(y_i, \theta)$.

Given a random sample of size n and for any θ in [min y, max y] the Lagrange multiplier λ solving the problem (5) satisfies the equation:

$$\sum_{i=1}^{n} g_i e^{\lambda g_i} = 0,$$

with solution

$$e^{\lambda(\theta)} = \frac{\tau n_0}{(1-\tau) n_1},$$

where $n_1 = \#(y_i \le \theta_\tau)$ and $n_0 = n - n_1$.

Substituting this solution into the expression for the posterior density (6) and assuming a uniform prior gives:

$$p(\theta_\lambda|y) \propto \frac{\phi^{n_1}}{n_1^{n_1} n_0^{n_0}}, \text{ where } \phi = \frac{\tau}{1-\tau}. \tag{7}$$

This is a piecewise constant density supported on [min (y_i), max (y_i)]. The density (7) may be sampled as follows. There are $n - 1$ pieces, if the observations are all distinct, forming a partition of the interval from min (y_i) to max (y_i). Sample each piece according to its probability. If the sampled piece is bounded by $y_{(i)}$ and $y_{(i+1)}$ sample a random variable uniformly distributed on this interval.

1.1. Regression quantiles

The τ 'th quantile regression is such that:

$$Pr(Y \le \alpha(\tau) + \beta(\tau)X|X) = \tau$$

and so satisfies the moment conditions:

$$E(\mathbf{I}(Y \le \alpha(\tau) + \beta(\tau)X) - \tau\,|X) = 0$$
$$E(X\,\mathbf{I}(Y \le \alpha(\tau) + \beta(\tau)X) - \tau\,|X) = 0.$$

If we now define:

$$g_{1i} = \mathbf{I}(y_i \le \alpha + \beta x_i) - \tau$$
$$\text{and}\quad g_{2i} = x_i(\mathbf{I}\,(y_i \le \alpha + \beta x_i) - \tau),$$

we may compute the Lagrange multipliers λ_1 and λ_2 by:

$$\lambda = arg\,\min_{\eta} \frac{1}{n}\sum_{i=1}^{n} exp\{\eta_1 g_{1i} + \eta_2 g_{2i}\}$$

and then calculate the posterior density according to (4).

2. Bayesian quantile regression with asymmetric Laplace likelihood

To fit conditional quantile functions (curves), we use a well-defined likelihood function – the asymmetric Laplace likelihood function with both location and scale parameters:

$$g_\tau(y|\mu,\sigma) = \frac{\tau(1-\tau)}{\sigma} e^{-\frac{\tau(y-\mu)_+ + (1-\tau)(y-\mu)_-}{\sigma}}, \qquad (8)$$

where $(\cdot)+$ and $(\cdot)-$ represent the positive and negative part of the quantity, respectively. With this likelihood function, we are able to approximate the marginal likelihood ratio of the number of knots and their locations. This ratio, also called the Bayesian factor, plays an important rule in model selection through the accept/reject probability in RJMCMC as shown by DiMatteo[6] in the case of conditional mean.

First, instead of the check function used by Yu[7]:

$$g_\tau(y,\mu)=\tau(y-\mu)_+ +(1-\tau)(y-\mu)_-, \tag{9}$$

we use the proper likelihood function $g_\tau(y|\mu, \sigma)$ in (8), which is more flexible with the extra scale parameter σ.

Assume that (y_i, x_i), $i = 1,\ldots, n$, are independent bivariate observations from the pair of response-explanatory variables (Y, X). Our goal is to find the τ'th conditional quantile $Q_\tau(x)$ of Y given $X = x$. The method in this paper uses the model with the asymmetric Laplace likelihood:

$$f_\tau(y\Big|\mu,\sigma)=\frac{\tau(1-\tau)}{\sigma}e^{-\frac{\tau(y-Q_\tau(x))_+ +(1-\tau)(y-Q_\tau(x))_-}{\sigma}} \tag{10}$$

The standard asymmetric Laplace distribution has the likelihood (8). Its unique mode μ, which satisfies:

$$\int_{-\infty}^{\tau} g_\tau(y|\mu,\sigma)dy=\tau,$$

is the τ'th quantile. This property guarantees the consistency of the maximum likelihood estimator (MLE) as an estimator of the τ'th quantile. If σ is considered constant, the MLE of $Q\tau(x)$ is equivalent to the regression quantile. Thus, regardless of the original distribution of (Y, X), the asymmetric Laplace distribution of (10) is used to model the τ'th regression quantile. With the likelihood specified, Bayesian quantile regression provides posterior inference on parameters or functions of parameters in the model.

---

[6] I. DiMatteo, C.R. Genovese, R.E. Kass: *Bayesian Curve Fitting with Free-knot Splines.* "Biometrika" 2001, No. 33, p. 1055-1073.

[7] K. Yu: *Reversible Jump MCMC Approach Quantile Regression.* "Computational Statistics Data Analysis" 2002, No. 40(2), p. 303-315.

As in typical nonparametric regression, the τ'th conditional quantile of Y given $X = x$ is assumed to be a nonparametric function, which belongs to the closure of a linear function space – here the piecewise polynomials:

$$P_{k,l}(x) = \sum_{v=0}^{l} \beta_{v,0}(x - t_0)_+^v + \sum_{m=1}^{k} \sum_{v=l_0}^{l} \beta_{v,m}(x - t_m)_+^v , \qquad (11)$$

where t_i, $i = 0,\ldots, k+1$, indexed in ascending order, are the knot points with the boundary knots $t_0 = \min\{x_i, i = 1,\ldots, n\}$ and $t_{k+1} = \max\{x_i, i = 1,\ldots, n\}$. Without loss of generality, can be assume that x_i , $i = 1,\ldots, n$, are in ascending order. So, $t_0 = x_1$ and $t_{k+1} = x_n$. Then l ($l \geq 0$) is the order of the piecewise polynomials and l_0 ($l_0 \geq 0$) controls the degree of continuity at the knots. The special case with $l = l_0 = 3$ corresponds to the cubic splines. Piecewise polynomials $P_{k,l}$ were used for mean-based curve fitting.

Let assume that the nonparametric function $Q_\tau(x)$ is such a piecewise polynomial with l and l_0 predecided, while the coefficients $\beta_{v,m}$, the number of knots k, and their locations t_i are estimated with the data (y_i, x_i) by reversible jump Markov chain Monte Carlo approach (RJMCMC) of Green[8]. Let $\beta = \{\beta_{v,m}, 0 \leq v \leq l, 1 \leq m \leq k\}$, $t = \{t_i , 1 \leq i \leq k\}$. Then (k, t, β, σ) represents the full vector of parameters in the model. Then RJMCMC approach estimates $Q\tau(x)$ as a function of (k, t, β, σ).

2.1. Prior specification and likelihood ratio approximation

Now we describes the prior specification of (k, t, β, σ) the full vector of parameters in the model. We specify a prior for parameters (k, t, β, σ) hierarchically,

$$\pi_{k,t,\beta,\sigma}(k, t, \beta, \sigma) = \pi_{k,t}(k, t)\, \pi_{\beta}(\beta \mid k, t, \sigma)\, \pi_{\sigma}(\sigma). \qquad (12)$$

Firstly, can be specify a prior for the model space, which is characterized by the first two parameters k and t. Then, we specify a prior for the quantile function $Q\tau(x)$ in the specified model space. For the scale parameter σ, we use the non informative prior $\pi_\sigma(\sigma) = 1$. For the model space, a prior $\pi_{k,t}(k, t)$ can be further decomposed as:

[8] P.J. Green: *Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination*. "Biometrika" 1995, No. 82, p. 711-732.

$$\pi_{k,t}(k,t) = \pi_k(k)\,\pi_t(t \mid k). \tag{13}$$

So, we first need to specify a prior $\pi_k(k)$ for the number of knots k and next the sequence of knots t_i, for $i = 1,\ldots, k$, are considered order statistics from the uniform distribution with candidate knot sites $\{x_1,\ldots, x_n\}$ as the state space. We also consider the candidate knots from the continuous state space (x_1, x_n).

When the model space has been specified, the τ'th conditional quantile function:

$$Q_\tau(x) = \sum_{v=0}^{l} \beta_{v,0}(x - t_0)_+^v + \sum_{m=1}^{k}\sum_{v=l_0}^{l} \beta_{v,m}(x - t_m)_+^v \tag{14}$$

is specified through the coefficients β. Let z be the vector of the basis of piecewise polynomials evaluated at x, then:

$$Q_\tau(x) = z'\beta. \tag{15}$$

It can be (Yu and Moyeed[9]) considered a noninformative prior on R^d for β, where $d = l + 1 + k(l - l_0 + 1)$, and verified that the posterior of β is proper. A key step in RJMCMC is to decide the accept/reject probability for moves from one model (k, t) to another model (k', t'). As shown by Green[10], the acceptance probability for our problem is:

$$\alpha = min\{1, \frac{p(y|k',t')}{p(y|k,t)} \frac{\pi_{k,t}(k',t')}{\pi_{k,t}(k,t)} \frac{q(k,t|k',t')}{q(k',t'|k,t)}\}, \tag{16}$$

where:

$$p(y|k,t) = \iiint \prod_{i=1}^{n} f_\tau(y_i|x_i,\sigma)\pi_\beta(\beta|k,t,\sigma)\pi_\sigma(\sigma)d\beta d\sigma$$

is the marginal likelihood of (k, t) and $q(k, t \mid k', t')$ is the proposal probability of the equilibrium distribution. The prior ratio $\pi_{k,t}(k', t')/\pi_{k,t}(k, t)$ can be computed once the priors are specified. The proposal ratio $q(k, t \mid k', t')/q(k', t'|k, t)$ can be computed according to the moving strategy, which will be discussed in the next section. The hard part is to compute the marginal likelihood ratio $p(y| k', t')/p(y|k, t)$.

[9] K. Yu, R.A. Moyced: *Bayesian Quantile Regression.* "Statistics and Probability Letters" 2001, No. 54, p. 437-447.

[10] P.J. Green· Op. cit.

By using a noninformative prior of β, we are able to get the approximation:

$$\frac{p(y|k',t')}{p(y|k,t)} = \left(\left(\frac{n}{D_\tau(k',t')} \right)^{d-d'} \left(\frac{D_\tau(k,t)}{D_\tau(k',t')} \right)^{n-d} \right) \times (1 + o(1)), \quad (17)$$

where $D_\tau(k,t) = \sum_{i=1}^{n} \rho_\tau(y_i - z'\hat{\beta}_\tau(k,t))$,

$d' = l + 1 + k'(l - l_0 + 1)$, and $\hat{\beta}_\tau(k,t)$ is the τ'th regression quantile for the model (k, t).

Once this likelihood ratio is computed, the remaining work with the Metropolis–Hastings accept/reject probability α in RJMCMC is to compute the proposal ratio $q(k, t \mid k', t')/q(k', t'|k, t)$. This ratio acts as a symmetric correction for various moves.

2.2. Moving strategies in RJMCMC

Following the scheme of Green[11] and Denison et al.[12], we describe moves that involve knot addition, deletion, and relocation. For each move $(k, t) \rightarrow (k', t')$, the potential destination models (k', t') form a subspace, called *allowable space*. For the same type of moves, for example knot addition $(k' = k + 1, t' = (t, t_{k+1}))$, the subspace is defined by all possible choices of the $(k + 1)$st knot. Denote $M_k = (k, t)$ and $M_{k+1} = (k+1, (t, t_{k+1}))$.

Different ways to restrict the subspace provide different moving strategies in RJMCMC. We chose the candidate knot uniformly from data points and required that it is at least n_{sep} data points away from the current knots to avoid numerical instability (or *discrete proposal*).

We notice that we required $n_{sep} \geq l$ to avoid numerical instability. Using a different set of basis functions of piecewise polynomials rather than the explosive truncated power basis as in (11) significantly reduces the condition number of the design matrix, thus the numerical instability. However, in quantile regression, the lack of data between two knots may cause serious crossing. So, we usually require a larger n_{sep}, for example $n_{sep} \geq 2l$.

Experiences suggest that, in quantile regression curve fitting, the discrete proposal works as well as or better than the continuous proposal, especially with

[11] Ibidem.

[12] D. Denison, B. Mallick, A. Smith: *Automatic Bayesian Curve Fitting*. "Journal of the Royal Statistical Society" 1998, Series B, No. 60, p. 333-350.

middle or large data sets ($n \geq 200$). One explanation is that placing too many knots near a point, where data may form some suspicious patterns, would result in a chance of over fitting locally, inflating the design matrix and impairing the computational efficiency. The probabilities of addition, deletion, and relocation steps of the RJMCMC sampler are:

$$b_k = c \min\{1, \pi_k(k+1)/\pi_k(k)\}$$
$$d_k = c \min\{1, \pi_k(k-1)/\pi_k(k)\},$$
$$\eta_k = 1 - b_k - d_k$$

where c is a constant in (0, 1/2), which controls the rate of dimension change among these moves as illustrated by Denison[13]. These probabilities ensure that $b_k\pi_k(k) = d_k + 1\pi_k(k+1)$, which will be used to maintain the detailed balance requested by RJMCMC. With these probabilities, RJMCMC cycles among proposals of addition, deletion, and relocation.

Knot Addition

A candidate knot is uniformly selected from the allowable space. Assume that currently there are k knots from the n data points.
Then, the allowable space has $n - Z(k)$ data points where:

$$Z(k) = 2(n_{sep} + 1) + k(2n_{sep} + 1). \tag{18}$$

In this case the jump probability is:

$$q(M_{k+1}|M_k) = b_k \frac{n - Z(k)}{n}. \tag{19}$$

Knot Deletion

A knot is uniformly chosen from the existing set of knots and deleted. The jump probability from M_k to M_{k-1} is:

$$q(M_{k-1}|M_k) = d_k \frac{1}{k} \tag{20}$$

[13] Ibidem.

Knot Relocation

A knot t_{i*} is uniformly chosen from the existing set of knots and relocated within the allowable intervals between its two neighbors. Relocation does not change the order of the knots. Let M_C be the current model and M_R be the model after relocation. The jump probability from M_R to M_C is:

$$q(M_R|M_C) = \eta_k \frac{1}{k} \frac{n(t_{i*}) - 2n_{sep}}{n(t_{i*})}, \tag{21}$$

where $n(t_{i*})$ is the number of data points between the two neighboring knots of t_{i*}. Due to the symmetry.

2.3. The algorithm

At the end we describe details of the RJMCMC algorithm for quantile regression, especially the initialization of RJMCMC. The experience shows that the initialization has impact on the performance of RJMCMC.

To set up an initial model configuration, we choose λ locations between x_1 and x_n, where λ could be the prespecified mean of the distribution of the number of knots.

The λ locations take the values of $[hJ]$'th observations of x_i, $i = 1,\ldots, n$, where $h = [n/\lambda + 1]$ and $J = 1,\ldots, \lambda$. In this way, we evenly assign $h - 1$ observations between two neighboring knots. Compared with other initial knot assignment methods, this even observation assignment (EOA) is more natural for the implementation of our strategy to add a knot, which prefers certain symmetric distribution of observations between neighboring knots. From experiments we know that EOA performs better than other initial knot assignment methods, for example, evenly spacing on (x_1, x_n).

To implement a full Bayesian version of RJMCMC for quantile regression, we need to draw β from its posterior distribution, which does not have a closed form in our case.

Instead, we use the posterior mode $\hat{\beta}$, which is the regression quantile for the given model (k, t) and scale parameter σ. These regression quantiles create different Markov chains from the ordinary Markov chains by drawing samples from the posterior. Estimated quantile curves from these regression quantile chains have better convergence rates than those estimated from the ordinary

Markov chains sampled from the posterior. In addition, using regression quantiles improves computation efficiency, since the posterior mode $\hat{\beta}$ has already been computed when we compute the acceptance probability of (k, t) with the given σ.

The algorithm of RJMCMC for quantile regression is described as the following steps:

1. Sort the data by the independent variable and normalize the independent variable to interval [0, 1].
2. Assign initial knots according to the described method.
3. Run RJMCMC N_b iterations for the burn-in process from step a to e.

a. Take knot steps: addition, deletion, relocation. This recommends a new model (k, t).

b. *Compute regression quantile* $\hat{\beta}_\tau(k, t)$ *for model* (k, t).

c. Compute the acceptance probability α based on $\hat{\beta}_\tau(k, t)$.

d. Update the model according to the accept/reject scheme.

e. Draw σ with the Gibbs sampling method.

Run RJMCMC N_s iterations for the sampling process after the N_b iterations of burn-in. Within each iteration, in addition to the steps a to e in 3, sequentially run the following step f:

f. Using $\hat{\beta}_\tau(k, t)$, obtain the τ th regression quantile curve fit $\hat{Q}_\tau(x)$, objective function value $\hat{D}_\tau(k, t)$, number of modes of $\hat{Q}_\tau(x)$, and other interested summary statistics.

4. From the sampling process, obtain mean and median estimates of the quantile function values $Q_\tau(x)$ and means of the objective function value $D_\tau(k, t)$ and the number of modes of $Q_\tau(x)$, respectively.

It should be noted that the mean and median estimates obtained from the algorithm are the approximations to the posterior mode and median, respectively. The difference between these two estimates are not significant due to large number of samples used in the sampling process.

With the Laplace likelihood in (10), the posterior of σ given (k, t) and β follows an inverse gamma distribution. The Gibbs sampler draws σ from this inverse gamma distribution. The most computationally intensive part of the algorithm is computing the regression quantile $\hat{\beta}_\tau(k, t)$. The final evaluation of the fitted quantile regression curve can be taken on all the observed values or a grid of the independent variable. We measure the goodness-of-fit of the estimated Bayesian quantile regression curve based on the mean squared error on observed values:

$$mse = \frac{1}{n}\sum_{i=1}^{n}(\hat{Q}_\tau(x_i) - Q_\tau(x_i))^2 . \qquad (22)$$

It should be pointed out that to fit constrained quantile curves, we can add the constrains and fit the constrained regression quantile $\hat{\beta}\tau(k, t)$ in step b for each specified model (k, t). Bayesian average with the computed constrained curves usually will not break these constrains.

BAYESOWSKA REGRESJA KWANTYLOWA

Streszczenie

Poczynając od ważnej pracy Koenkera i Bassetta[14], regresja kwantylowa stała się uzupełniającym podejściem do kalsycznych metod estymacji regresji względem średniej. Wiele zastosowań regresji kwantylowej znajdujemy w finansach w estymacji modeli warunkowych w ocenie ryzyka. Engle i Manganelli[15] zaproponowali model, który w natrualny sposób wykorzystuje regresję kwantylową do opisu ryzyka z wykorzystaniem VaR, zapisując nieliniowy model autoregresyjny z wykorzystaniem VaR. Metody estymacji tego modelu oraz zastosowania na polskim rynku kapitałowym były tematem poprzednich publikacji autorki[16]. W tym artykule podjęto problem wybranej metody estymacji bayesowskiej regresji kwantylowej. Omówiono problem wnioskowania w tym podejściu, wykorzystując wcześniejsze pozycje, bez zastosowania założeń wynikających z metody największej wiarygodności. Ponadto zaprezentowano metodę przybliżenia krzywej rozkładu kwantyli z wykorzystaniem podejścia bayesowskiego.

[14] R. Koenker, G. Bassett: Op. cit.

[15] R.F. Engle, S. Magnatelli: Op. cit.

[16] G. Trzpiot: *Regresja kwantylowa*..., op. cit.; Idem: *Implementacja metodologii*..., op. cit.; Idem: *Estimation Mmethods for Quantile*..., op. cit.

Grażyna Trzpiot, Przemysław Jeziorski

APPLICATION OF ASYMMETRIC LEAST SQUARES METHOD IN ESTIMATION CaViaR MODELS

Introduction

Value at risk (*VaR*) measures the maximum potential loss of a given portfolio over a prescribed holding period at a given confidence level, which is typically chosen to be 1% or 5%. Therefore, assessing *VaR* amounts to estimating tail quantiles of the conditional distribution of a series of financial returns. A recent development in the *VaR* literature is the Conditional Autoregressive Value at Risk (*CaViaR*) class of models[1]. This approach to *VaR* estimation has strong appeal in that it provides a modeling framework and does not rely on distributional assumptions.

Manganelli and Engle[2] divide *VaR* methods into three categories: parametric, semiparametric, and nonparametric. Parametric approaches involve a parameterization of the behavior of prices. Conditional quantiles are estimated using a conditional volatility forecast with an assumption for the shape of the distribution. The distribution is typically chosen to be Gaussian or the Student's *t*-distribution. Included in the semiparametric *VaR* category are methods based on extreme value theory (*EVT*) or quantile regression. The most widely used nonparametric *VaR* method is historical simulation, which requires no distributional assumptions and estimates the *VaR* as the quantile of the empirical distribution of historical returns from a moving window of the most recent periods.

---

[1] See: R.F. Engle, S. Manganelli: *CaViaR: Conditional Autoregressive Value at Risk by Regression Quantiles*. "Journal of Business and Economic Statistics" 2004, No. 22, p. 367-381.

[2] Ibidem.

Application on polish capital market based on semiparametric *VaR* can be find in some previous articles published by authors[3]. A recent proposal using quantile regression is the class of *CaViaR* models introducedby Engle and Manganelli[4]. The approach involves the use of asymmetric least squares (ALS) regression, which is the least squares analogue of quantile regression.

1. Asymmetric least squares

Asymmetric least squares method (ALS) is the combination of two well known methods: the classical method of least squares (LS) and quantile regression (QR). Classical least squares method solves the following optimization problem:

$$\min_{b\in R^k}\left\{\sum_t (y_t - x_t b)^2\right\}. \tag{1}$$

The approach of quantile regression search vector parameters of b to achieve a minimum for the following expression[5]:

$$\min_{b\in R^k}\left\{\sum_{t\in\{t:y_t\geq x_t b\}} \theta|y_t - x_t b| + \sum_{t\in\{t:y_t<x_t b\}} (1-\theta)|y_t - x_t b|\right\}. \tag{2}$$

Asymmetric least squares method is a combination of these two methods. ALS approach gives different weights for positive residuals and negative residuals of model – an analogy to quantile regression. Positive residuals get weight equal to θ, while negative residuals get weight equal to $1-\theta$. The distances between the fitted values and empirical values – in contrast to the quantile regression – were raised to the square – an analogy to the method of least squares.

[3] See: G. Trzpiot: *Regresja kwantylowa a estymacja VaR.* W: *Inwestycje finansowe i ubezpieczenia – tendencje światowe a polski rynek.* Red. W. Ronka-Chmielowiec, K. Jajuga. Wydawnictwo Akademii Ekonomicznej, Wrocław 2007, s. 465-471, Idem: *Implementacja metodologii regresji kwantylowej w estymacji VaR.* W: *Rynek kapitałowy. Skuteczne inwestowanie.* Studia i Prace WNEiZ, nr 9, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin, 2008. s. 316-323; G. Trzpiot, P. Jeziorski: *Implementacja modeli CaViaR z wykorzystaniem rolowanej regresji kwantylowej.* W: *Inwestycje finansowe i ubezpieczenia – tendencje światowe a polski rynek.* Red. K. Jajuga, W. Ronka-Chmielowiec. Wydawnictwo Akademii Ekonomicznej, Wrocław 2010, s. 431-440; G. Trzpiot, J. Majewska: *Estimation of Value at Risk: Extreme Value and Robust Approaches.* "Badania Operacyjne i Decyzje" 2010, nr 1, Oficyna Wydawnicza Politechniki Wrocławskiej. Wrocław 2010. s. 131-143.

[4] R.F. Engel, S. Manganelli: Op. cit.

[5] R. Koenker, G. Bassett: *Regression Quantiles.* "Econometrica" 1978, Vol. 46, No. 1, p. 33-50.

$$\min_{b\in R^k}\left\{\sum_{t\in\{t:\,y_t\geq x_t b\}}\theta(y_t-x_t b)^2+\sum_{t\in\{t:\,y_t<x_t b\}}(1-\theta)(y_t-x_t b)^2\right\} \tag{3}$$

The solution of an ALS regression is known as an expectile. This name was given by Newey and Powell[6] who note that the ALS solution is determined by the properties of the expectation of exceedances beyond the solution.

On the Figure 1 we present the example of the fitted values for linear model obtained using three types of regression:

$$r_t = b_0 + b_1 r_{t-1}. \tag{4}$$

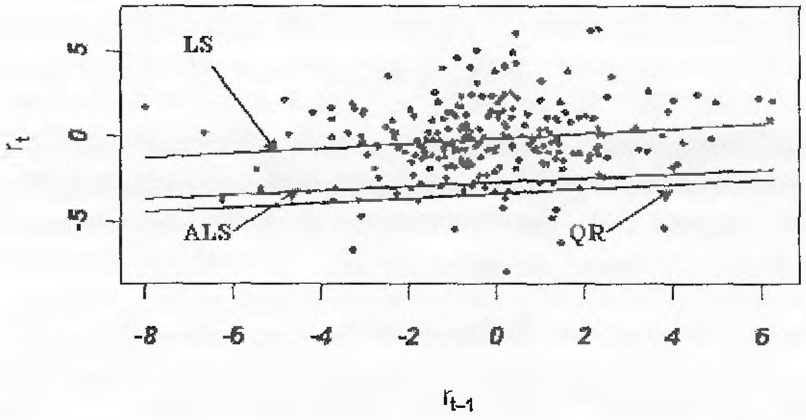


Figure 1. Regression models: LS, ALS ($\theta = 0.05$), QR ($\theta = 0.05$)

In least squares method, fitted values are estimated based on the conditional average. By contrast, asymmetric least squares regression and qunatile regression give weight for estimated residuals and this causes that estimated values move up or down depending on the level of quantile θ.

[6] W.K. Newey, L. Powell: *Asymmetric Least Squares Estimation and Testing*. "Econometrica" 1987, Vol. 55, p. 819-847.

2. *CaViaR* models

CaViaR models we can use for estimation *Value at Risk* based on behavioral property of the financial market – autocorrelation. *VaR* is a measure based on the quantile of rates distribution (the level θ is established as we mentioned in introduction). The value of the quantile describes loss, which investor can accept assuming confidence level. *CaViaR* models don't make restrictive assumptions about the distribution of return, we can estimate any models based on historical values of rates of returns. Estimation *CaViaR* models is preceded by the determination of estimates of *VaR*. Value of *VaR* will be obtained using the variance–covariance approach (VC), in which the *VaR* is quantile of Gaussian distribution. The model, which estimates the *VaR* with α significance level has the following form:

$$VaR_{\alpha,t} = u_{\alpha}\sigma_{t-1}. \tag{5}$$

Obtained estimators of standard deviation (σ), based on an variance–covariance model (5), allows as for doing forecasts of *VaR*, which are necessary for estimating *CaViaR* models. *VaR* forecasts with the significance level α are dependent on the current standard deviation:

$$VaR_{\alpha,t+1} = u_{\alpha}\sigma_{\alpha,t}. \tag{6}$$

Engle and Manganelli[7] were appiled autoregression models for estimating *VaR*, where the explanatory variables are delayed returns and delayed estimate *VaR*.

Adaptive:

$$VaR_t = VaR_{t-1} + \beta_1\left[I\left(r_{t-1} \le VaR_{t-1}\right) - \alpha\right] \tag{7}$$

Proportional Symmetric Adaptive:

$$VaR_t = VaR_{t-1} + \beta_1\left(\left|r_{t-1}\right| - VaR_{t-1}\right)^{+} - \beta_2\left(\left|r_{t-1}\right| - VaR_{t-1}\right)^{-} \tag{8}$$

Symmetric Absolute Value:

$$VaR_t = \beta_0 + \beta_1 VaR_{t-1} + \beta_2\left|r_{t-1}\right| \tag{9}$$

[7] R.F. Engel, S. Manganelli: Op. cit.

Asymmetric Absolute Value:

$$VaR_t = \beta_0 + \beta_1 VaR_{t-1} + \beta_2 |r_{t-1} - \beta_3| \tag{10}$$

Asymmetric slope:

$$VaR_t = \beta_0 + \beta_1 VaR_{t-1} + \beta_2 (r_{t-1})^+ + \beta_3 (r_{t-1})^- \tag{11}$$

Adaptive model forecast *VaR* by using the *VaR* of the previous period and the indicator function, which check that rate in the previous period exceeded *Value at Risk*. Proportional Symmetric Adaptive model uses information about *VaR* of the previous period and the two components including positive and negative deviations the absolute rates of return in the previous period form *Value at Risk*. Symmetric Absolute Value model uses information about the *VaR* in the previous period and the absolute value of returns on previous period. Asymmetric Absolute Value model also takes into account the absolute value of the *VaR* of the previous period, but the second term measures the absolute deviation returns form some level, which is estimated. Asymmetric slope model when estimating *VaR* uses this value (*VaR*) from the previous period and formula includes positive and negative returns.

Properties of ALS model will be testing by Kupiec test[8]. This test checks the number of exceedances is equal to significance level of *VaR*. *LR* formula has asymptotic χ^2 distribution with 1 degree of freedom and has the form:

$$LR = -2\ln\left[(1-\alpha)^{T-N}\alpha^N\right] + 2\ln\left[\left(1-\frac{N}{T}\right)^{N-T}\left(\frac{N}{T}\right)^N\right], \tag{12}$$

where:

α – significance level of *VaR*,

T – sample size,

N – number of exceedances of *VaR*.

[8] W.K. Newey, L. Powell: *Asymmetric Least Squares Estimation and Testing*. "Econometrica" 1987, Vol. 55, p. 819-847.

3. Empirical result

In this part, we evaluate efficiency of *VaR* estimation by using *CaViaR* model. Effectiveness of estimation will be reviewed by the Kupiec test. Parameters of *CaViaR* models were estimated using the asymmetric least squares method, where θ corresponds to the significance level of the estimated *VaR*. In the first step of analysis we use the historical *VaRs*, which are necessary to estimate *CaViaR* models, were estimated using model (6). Next we use this *VaR* estimators to estimation *CaViaR* models. Parameters of *CaViaR* models will be obtained by ALS method (3) and will be compared with LS method (1). Data set contains daily return in the period from 3 January 2005 to 19 February 2010 for *benchmark* as the WIG-Banks index, and additionally for the portfolio (we chosen some of banking companies) which minimizing the investment risk. By solving some optimization problem we obtained the weight for the components of portfolio (BPH, GETIN, MILLENNIUM) are: 55%, 29% and 16%. The significance level for *VaR* is 1%, 2% and 5%. We decided to implement three of described in previous part *CaViaR* models base on different number of sessions: 20, 60, 120 and 250: Model I: Symmetric Absolute Value (9) , Model II: Asymmetric slope (11) and Model III: Proportional Symmetric Adaptive (8).

The results obtained for the benchmark WIG-Banks index for chosen models we present in tables. The table contains the fraction of exceedances of the *VaR* and the *p-value* for the Kupiec test, which allows to accept or reject the hypothesis of equality between the number of exceedances and the significance level for *VaR*.

Table 1

The results of Kupiec test for *CaViaR-model I (WIG-Banks)*

CaViaR I	*VaR*	*Number of session*	$CaViaR_{1\%}$	$CaViaR_{2\%}$	$CaViaR_{5\%}$
LS	*VC*	20	1.91% (0.023)	2.80% (0.132)	5.34% (0.668)
		60	1.91% (0.023)	2.67% (0.202)	4.32% (0.371)
		120	1.65% (0.093)	2.16% (0.751)	4.96% (0.954)
		250	1.91% (0.023)	3.05% (0.051)	5.21% (0.789)
ALS	*VC*	20	1.40% (0.290)	2.54% (0.298)	5.34% (0.668)
		60	1.91% (0.023)	2.67% (0.202)	4.07% (0.215)
		120	1.65% (0.093)	2.03% (0.947)	4.96% (0.954)
		250	1.91% (0.023)	3.05% (0.051)	5.21% (0.789)

Table 2

The results of Kupiec test for *CaViaR-model II (WIG-Banks)*

CaViaR II	*VaR*	*Number of session*	$CaViaR_{1\%}$	$CaViaR_{2\%}$	$CaViaR_{5\%}$
LS	*VC*	20	1.78% (0.048)	2.92% (0.083)	5.46% (0.556)
		60	1.91% (0.023)	2.67% (0.202)	4.32% (0 371)
		120	1 65% (0.093)	2 16% (0.751)	4.96% (0.954)
		250	1.91% (0.023)	3.05% (0.051)	5.21% (0.789)
ALS	*VC*	20	1.27% (0.464)	2.54% (0.298)	5.34% (0.668)
		60	1.91% (0.023)	2.67% (0.202)	4.07% (0.215)
		120	1.65% (0.093)	2.03% (0.947)	4 96% (0.954)
		250	1.91% (0.023)	3.05% (0.051)	5.21% (0.789)

Table 3

The results of Kupiec test for *CaViaR-model III (WIG-Banks)*

CaViaR III	*VaR*	*Number of session*	$CaViaR_{1\%}$	$CaViaR_{2\%}$	$CaViaR_{5\%}$
LS	*VC*	20	1.91% (0.023)	3.05% (0.051)	5.59% (0.455)
		60	1.91% (0.023)	2.8% (0 132)	4 19% (0.286)
		120	1.65% (0.093)	2 16% (0.751)	4.96% (0.954)
		250	1.91% (0.023)	3 05% (0.051)	5.21% (0 789)
ALS	*VC*	20	1.4% (0.290)	2.29% (0.574)	5.21% (0.789)
		60	1.65% (0.093)	2.54% (0.298)	4.07% (0.215)
		120	1 52% (0.170)	2.03% (0.947)	4.83% (0 824)
		250	1.91% (0 023)	3.05% (0.051)	5 21% (0.789)

CaViaR-Model I for the level of significance of 1% estimates by the LS for each number of sessions beyond 120 returns the underestimated *VaR* (*p-value* < 0.05) – (Tables 1-3), while the same model estimated by the ALS approach gives very good results for 20 and 120 number of sessions. For the other levels of significance difference between the models is negligible, but in any case the ALS approach is better than the LS approach – the percentage of exceedances is closer to the level of significance, which was assumed. Exactly the same results

were obtained for *CaViaR-Model II* – the closer the number of exceedances of the prescribed level of significance was reached for the ALS approach. Better results were also obtained for the ALS approach for model III – more models estimate *VaR*, which gives percentage of exceedances closer to the assumed level of significance compared with the LS approach.

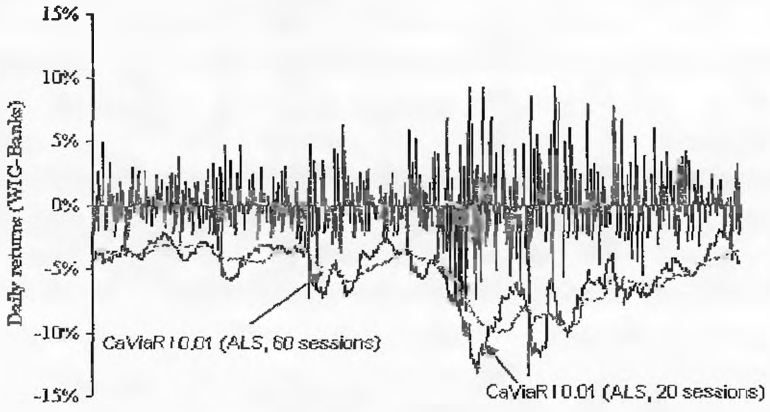


Figure 2. WIG-Banks daily returns and *CaViaR I*

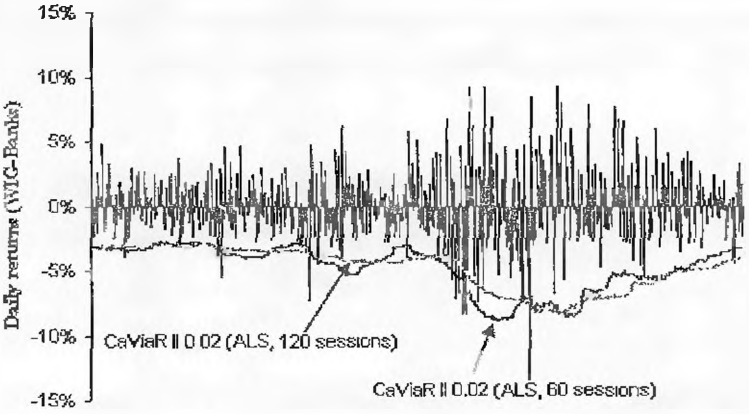


Figure 3. WIG-Banks daily returns and *CaViaR II*

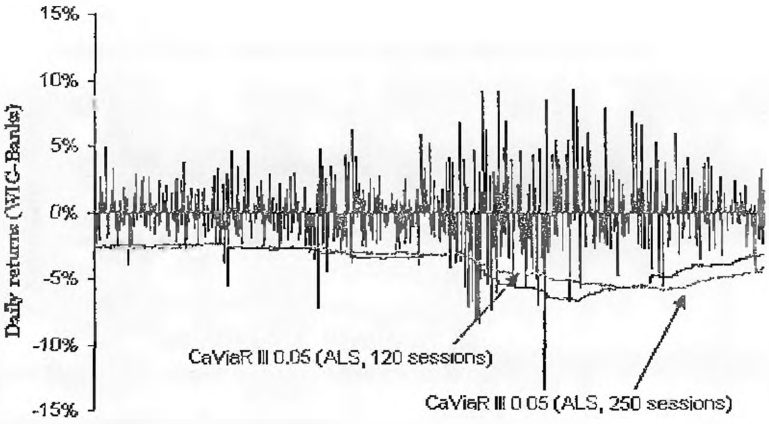


Figure 4. WIG-Banks daily returns and *CaViaR III*

Analogous analysis were made for the portfolio. The results are similar to the models for the WIG-Banks index (Tables 4-6). The difference between the Models I is evident for the 1% level of significance. The LS approach for each number of sessions gives underestimated *VaR*, but model I estimated by the ALS approach for 60 number of sessions gives the correct estimates. The difference for the other levels of significance is lower, but in any case the ALS approach gives estimates closer to the level of significance. Identical results were obtained for Model II and Model III – percentage of exceedances are closer to the level of significance, which was assumed.

Table 4

The results of Kupiec test for *CaViaR-model I (Portfolio)*

CaViaR I	*VaR*	*Number of session*	$CaViaR_{1\%}$	$CaViaR_{2\%}$	$CaViaR_{5\%}$
LS	*VC*	20	2.41% (0.001)	3.56% (0.005)	5.97% (0.224)
		60	1.91% (0.023)	2.67% (0.202)	5.21% (0.789)
		120	1.91% (0.023)	2.8% (0.132)	4.7% (0.698)
		250	2.03% (0.011)	3.18% (0.03)	5.72% (0.366)
ALS	*VC*	20	2.16% (0.005)	3.43% (0.009)	5.72% (0.366)
		60	1.52% (0.170)	2.54% (0.298)	5.08% (0.916)
		120	1.78% (0.048)	2.8% (0.132)	4.7% (0.698)
		250	1.91% (0.023)	3.18% (0.03)	5.72% (0.366)

Table 5

The results of Kupiec test for *CaViaR-model II (Portfolio)*

CaViaR II	*VaR*	*Number of session*	$CaViaR_{1\%}$	$CaViaR_{2\%}$	$CaViaR_{5\%}$
LS	*VC*	20	2.29% (0 002)	3.68% (0.002)	6.23% (0.128)
		60	1.78% (0.048)	2 67% (0.202)	5.21% (0.789)
		120	1.91% (0.023)	2.8% (0.132)	4.7% (0.698)
		250	2.03% (0.011)	3.18% (0.03)	5.72% (0.366)
ALS	*VC*	20	2.03% (0.011)	3.43% (0.009)	5.72% (0.366)
		60	1.52% (0.170)	2.54% (0.298)	5 21% (0.789)
		120	1.78% (0.048)	2.8% (0 132)	4.7% (0.698)
		250	1.91% (0.023)	3.18% (0.03)	5.72% (0.366)

Table 6

The results of Kupiec test for *CaViaR-model III (Portfolio)*

CaViaR III	*VaR*	*Number of session*	$CaViaR_{1\%}$	$CaViaR_{2\%}$	$CaViaR_{5\%}$
LS	*VC*	20	2.67% (0.000)	3.68% (0.002)	6.10% (0 171)
		60	2 03% (0.011)	2 67% (0.202)	5.34% (0.668)
		120	2 03% (0.011)	2 8% (0.132)	4.83% (0 824)
		250	2 03% (0.011)	3.18% (0.03)	5.84% (0.289)
ALS	*VC*	20	1.91% (0.023)	3.43% (0 009)	5 46% (0 556)
		60	1 52% (0.170)	2.67% (0.202)	4.96% (0.954)
		120	1.78% (0.048)	2 8% (0.132)	4.83% (0.824)
		250	1.91% (0.023)	3.18% (0.03)	5.59% (0.455)

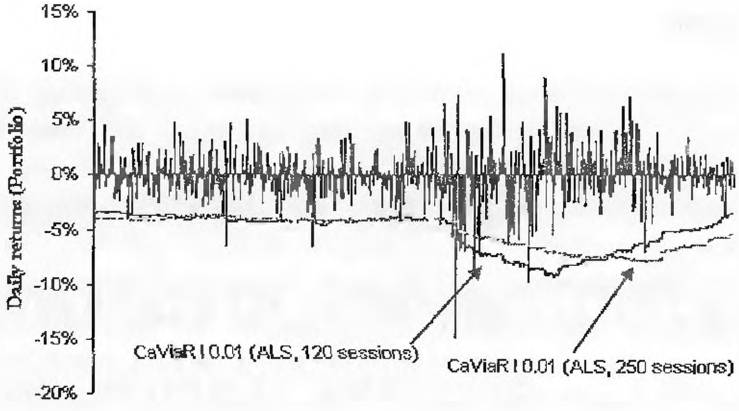


Figure 5. Portfolio daily returns and *CaViaR I*

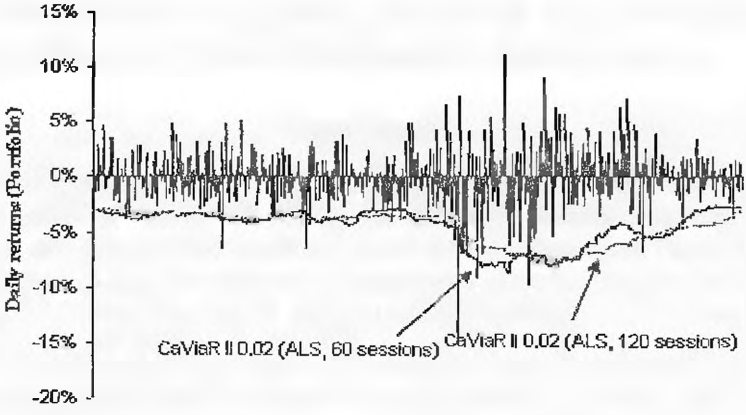


Figure 6. Portfolio daily returns and *CaViaR II*

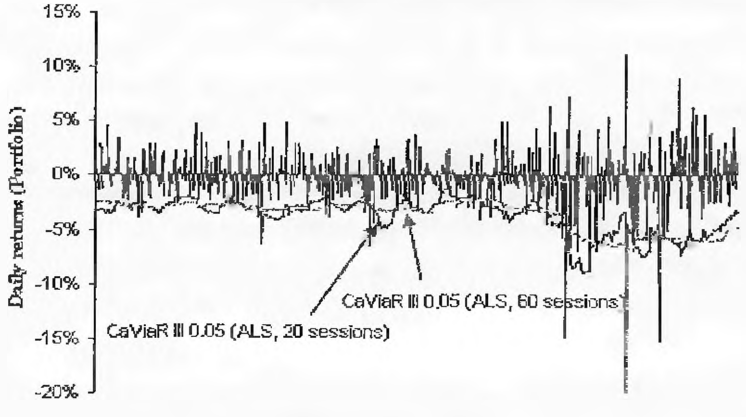


Figure 7. Portfolio daily returns and *CaViaR III*

Conclusion

In this paper, we have introduced a new approach to estimating conditional *VaR* using ALS regression. Asymmetric least squares method, which has been used for estimating parameters *CaViaR* models gives underestimated *VaR*. The main contribution of this paper is that we show the application chosen method of estimation on polish capital market. Using the ALS approach was not eliminated completely underestimating the *VaR*, but in comparison with the LS approach, the problem was a little limited. In some cases, the ALS approach was better in comparison with the LS approach, but there was no case in which the ALS was worse than the LS approach.

ASYMETRYCZNA METODA NAJMNIEJSZYCH KWADRATÓW W SZACOWANIU PARAMETRÓW MODELI CaViaR

Streszczenie

Modele *CaViaR* wykorzystywane do szacowania *Value at Risk* (*VaR*) posiadają zaletę w postaci braku założeń dla rozkładu stóp zwrotu. Prognozy *VaR* wyznaczane są na podstawie historycznych stóp zwrotu, wykorzystując własność autokorelacji. Autorzy zaprezentowali asymetryczną metodę najmniejszych kwadratów (*ALS*) do estymacji parametrów modeli *CaViaR*, która w kolejnym kroku została porównana z klasyczną metodą najmniejszych kwadratów. *ALS* jest metodą regresji, w której kwadraty reszt modelu posiadają różne wagi: reszty dodatnie otrzymują wagi równe rzędowi kwantyla, natomiast reszty ujemne otrzymują wagi równe dopełnieniu do jedności rzędu kwantyla. Ważenie reszt powoduje przesunięcie oszacowań w kierunku wartości skrajnych przy założeniu niskich rzędów kwantyla, podczas gdy w klasycznej metodzie najmniejszych kwadratów oszacowania opieraj się na warunkowych średnich.

Tomasz Żądło

ON PREDICTION OF LINEAR COMBINATION OF DOMAINS' TOTALS IN LONGITUDINAL ANALYSIS

Introduction

Let us introduce some notation presented earlier by Żądło[1]. In the paper longitudinal data for periods $t = 1,\ldots, M$ are considered. In the period t the population of size N_t is denoted by Ω_t. The population in the period t is divided into D disjoint domains (subpopulations) Ω_{dt} of size N_{dt}, where $d = 1,\ldots, D$. Let the set of population elements for which observations are available in the period t be denoted by s_t and its size by n_t. The set of domain elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}. Let: $\Omega_{rdt} = \Omega_{dt} - s_{dt}$, $N_{rdt} = N_{dt} - n_{dt}$.

Let M_{id} denotes the number of periods when the i'th population element may be potentially observed in the d'th domain (when the i'th population element belongs to the d'th domain). Let us denote the number of periods when the i'th population element (which belongs to the d'th domain) is observed by m_{id}. Let $m_{rid} = M_{id} - m_{id}$. We assume that the population may change in time and that one population element may change its domain membership in time (from technical point of view observations of some population element which change its domain membership are treated as observations of new population element). It means that i and t completely identify domain membership but additional subscript d will be needed as well. More about this assumptions will be written at the end of the next section.

---

[1] T. Żądło: *On Prediction of Domain Totals Based on Unbalanced Longitudinal Data*. In: *Survey Sampling in Economic and Social Research*. Eds. J. Wywiał, T. Żądło. Wydawnictwo Akademii Ekonomicznej, Katowice 2009, p. 97-111.

The set of elements which belong at least in one of periods $t = 1,\ldots, M$ to sets Ω_t is denoted by Ω and its size by N. Similarly, sets Ω_d, s, s_d, Ω_{rd} of sizes N_d, n, n_d, N_{rd} respectively are defined as sets of elements which belong at least in one of periods $t = 1,\ldots, M$ to sets Ω_{dt}, s_t, s_{dt}, Ω_{rdt} respectively. In the paper we consider the linear combination of two domains' totals – the d^*'th and the d''th domains. The d^*'th domain of interest in the period of interest t^* will be additionally denoted by a symbol * in the subscript i.e. $\Omega_{d^*t^*}$, and the set of elements which belong at least in one of periods $t = 1,\ldots, M$ to sets $\Omega_{d^*t^*}$, will be denoted by Ω_{d^*}. The $d^{\#}$'th domain of interest in the period of interest $t^{\#}$ will be additionally denoted by a symbol # in the subscript i.e. $\Omega_{d^{\#}t^{\#}}$, and the set of elements which belong at least in one of periods $t = 1,\ldots, M$ to sets $\Omega_{d^{\#}t^{\#}}$ will be denoted by $\Omega_{d^{\#}}$.

Values of the variable of interest are realizations of random variables Y_{tdj} for the i'th population element which belongs to the d'th domain in the period t_{ij}, where $i = 1,\ldots, N$, $j = 1,\ldots, M_{td}$, $d = 1,\ldots, D$. The vector of size $M_{td} \times 1$ of random variables Y_{tdj} for the i'th population element which belongs to the d'th domain will be denoted by $\mathbf{Y}_{id} = [Y_{tdj}]$, where $j = 1,\ldots, M_{td}$. Let us consider values of the variables of interest $Y_{i'd'j'}$ for the i'th population element which belongs to the d'-th domain observed in periods $t_{i'j'}$, where $i' = 1,\ldots, n$, $j' = 1,\ldots, m_{i'd'}$, $d' = 1,\ldots, D$. The vector of random variables $Y_{i'd'j'}$ (where $i' = 1,\ldots, n$, $j' = 1,\ldots, m_{i'd'}$, $d' = 1,\ldots, D$) of size $m_{i'd'} \times 1$ will be denoted by $\mathbf{Y}_{si'd'} = [Y_{i'd'j'}]$, where $j' = 1,\ldots, m_{i'd'}$. The vector of random variables $Y_{i''d''j''}$ of size $m_{ri''d''} \times 1$ for the i''-th population element which belongs to the d''-th domain for observations which are not available in the sample is denoted by $\mathbf{Y}_{rj''d''} = \left[Y_{i''d''j''}\right]$, where $j'' = 1,\ldots, m_{ri''d''}$.

The proposed approach may be used to predict the linear combination of two domains' totals for any (past, current and future) periods.

1. Superpopulation model

We consider superpopulation models (studied earlier by Żądło[2]) used for longitudinal data[3] which are – what is important for further considerations – special cases of the General Linear Model (GLM) and the General Linear Mixed Model (GLMM). The following two-stage model is assumed. Firstly:

[2] Ibidem.

[3] Compare: G. Verbeke, G. Molenberghs: *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York 2000; D. Hedeker, R.D. Gibbons: *Longitudinal Data Analysis*. John Wiley & Sons, New Jersey 2006.

$$\mathbf{Y}_{id} = \mathbf{Z}_{id}\boldsymbol{\beta}_{id} + \mathbf{e}_{id}, \tag{1}$$

where $i = 1,\ldots, N$; $d = 1,\ldots, D$, $\mathbf{Y}_{id}$ is a random vector of size $M_{id} \times 1$, $\mathbf{Z}_{id}$ is known matrix of size $M_{id} \times q$, $\boldsymbol{\beta}_{id}$ is a vector of unknown parameters of size $q \times 1$, $\mathbf{e}_{id}$ is a random component vector of size $M_{id} \times 1$. Vectors $\mathbf{e}_{id}$ ($i = 1,\ldots, N$; $d = 1,\ldots, D$) are independent with $\mathbf{0}$ vector of expected values and variance-covariance matrix $\mathbf{R}_{id}$. Although $\mathbf{R}_{id}$ may depend on i it is often assumed that $\mathbf{R}_{id} = \sigma_e^2 \mathbf{I}_{M_{id}}$ where $\mathbf{I}_{M_{id}}$ is the identity matrix of rank M_{id}. Secondly, we assume that:

$$\boldsymbol{\beta}_{id} = \mathbf{K}_{id}\boldsymbol{\beta} + \mathbf{v}_{id}, \tag{2}$$

where $i = 1,\ldots, N$; $d = 1,\ldots, D$, $\mathbf{K}_{id}$ is known matrix of size $q \times p$, $\boldsymbol{\beta}$ is a vector of unknown parameters of size $p \times 1$, $\mathbf{v}_{id}$ is a vector of random components of size $q \times 1$. It is assumed that vectors $\mathbf{v}_{id}$ ($i = 1,\ldots, N$; $d = 1,\ldots, D$) are independent with $\mathbf{0}$ vector of expected values and variance-covariance matrix $\mathbf{G}_{id} = \mathbf{H}$ what means that $\mathbf{G}_{id}$ does not depend on i.

Similar assumptions to (1) and (2) are presented by Verbeke, Molenberghs[4] but there are two differences. Firstly, in the book assumptions are made for profiles defined by elements. In this paper assumptions are made for profiles defined by elements and domain membership i.e. $\mathbf{Y}_{id}$ (of size $M_{id} \times 1$) what allows to take the possibility of population changes in time into account. Secondly, in the book the assumptions are made only for the sampled elements (i.e. $i = 1,\ldots, n$). In this paper they are made for all of the population elements ($i = 1,\ldots, N$). Based on (1) and (2) it is obtained that:

$$\mathbf{Y}_{id} = \mathbf{X}_{id}\boldsymbol{\beta} + \mathbf{Z}_{id}\mathbf{v}_{id} + \mathbf{e}_{id}, \tag{3}$$

where $i = 1,\ldots, N$; $d = 1,\ldots, D$, $\mathbf{X}_{id} = \mathbf{Z}_{id}\mathbf{K}_{id}$ is known matrix of size $M_{id} \times p$. Let $\mathbf{V}_{id} = D_{\xi}^2(\mathbf{Y}_{id})$. Hence, $\mathbf{V}_{id} = \mathbf{Z}_{id}\mathbf{H}\mathbf{Z}_{id}^T + \mathbf{R}_{id}$.

Let $\mathbf{A}_d$ be a column vector and $col_{1\le d\le D}(\mathbf{A}_d) = \left[\mathbf{A}_1^T \quad \ldots \quad \mathbf{A}_d^T \quad \ldots \quad \mathbf{A}_D^T\right]^T$ be a column vector obtained by stacking $\mathbf{A}_d$ vectors. Note that by stacking $\mathbf{Y}_{id}$ vectors (i.e. $\mathbf{Y} = col_{1\le d\le D}(col_{1\le i\le N_d}(\mathbf{Y}_{id}))$ from (3) we obtain the formula of the GLMM.

[4] G. Verbeke, G. Molenberghs: Op. cit., p. 20.

Let us consider the following special case of (3) – (similar model is considered by Verbeke, Molenberghs[5]:

$$Y_{idj} = (\beta_d + v_{id})t_{ij} + e_{idj} = \beta_d t_{ij} + v_{id} t_{ij} + e_{idj}, \tag{4}$$

where $i = 1,\ldots, N$; $d = 1,\ldots,D$, $j = 1,\ldots, M_{id}$. The model (4) is the simple (without constant) trend model with random coefficient. Note that if t_{ij} is not the number of period but it is its some (e.g. logarithmic) transformation, the model (4) will be a nonlinear model.

A special case of (4), where:

$$\forall_d \ \beta_d = \beta \tag{5}$$

will be also considered.

In the considered model we assume that (Verbeke and Molenberghs[6]) $\mathbf{R}_{id} = \sigma_e^2 \mathbf{I}_{M_{id}}$. What is more, $\mathbf{H} = \sigma_v^2$. Hence,

$$Cov_\xi(Y_{idj}, Y_{i'd'j'}) = \begin{cases} 0 & for \quad i \neq i' \vee d \neq d' \\ \sigma_e^2 + t_{ij}^2 \sigma_v^2 & for \quad i = i' \wedge j = j' \\ t_{ij} t_{i'j'} \sigma_v^2 & for \quad i = i' \wedge d = d' \wedge j \neq j' \end{cases} \tag{6}$$

and $\mathbf{V}_{id} = D_\xi^2(\mathbf{Y}_{id}) = \sigma_e^2 \mathbf{I}_{M_{id}} + \sigma_v^2 \mathbf{t}_{M_{id}} \mathbf{t}_{M_{id}}^T$, where $\mathbf{t}_{M_{id}}$ is a vector of size $M_{id} \times 1$ of t_{ij} for the i'th population element which belongs to the d'th domain. The model (4) is a special case of (3), where $\boldsymbol{\beta} = \left[\beta_1 \ \ldots \ \beta_d \ \ldots \ \beta_D\right]^T$, $\mathbf{X} = col_{1 \le d \le D}(col_{1 \le i \le N_d}(\mathbf{X}_{id}))$, $\mathbf{X}_{id} = \mathbf{Z}_{id}\mathbf{K}_{id}$, $\mathbf{Z}_{id} = \left[t_{i1} \ \ldots \ t_{ij} \ \ldots \ t_{iM_{id}}\right]^T$ and matrix $\mathbf{K}_{id}$ is of size $1 \times D$ with the d'th element which equals 1 and other elements equal 0. Hence, matrices $\mathbf{X}_{id}$ and $\mathbf{X}$ are of sizes $M_{id} \times D$ and $\left(\sum_{d=1}^{D}\sum_{i=1}^{N} M_{id}\right) \times D$ respectively.

[5] Ibidem, p. 25.
[6] Ibidem.

We have assumed that the population may change in time and that one population element may change its domain membership in time. Observations of new element of the population or observations of the population element after the change of the domain membership are treated as realizations of new profile (3). Hence, because of covariance structure (6) where nonzero covariances are only within profiles, we assume independence of observations for some population element before and after changing domain membership.

2. EBLUP of domain total

In the paper the BLUP proposed by Royall[7] and its empirical version are studied. Henderson's[8] BLUP and EBLUP, what was shown by Żądło[9], cannot be used in this case.

Till the end of this section considerations are based on the assumption (4) – equations under assumptions (4) and (5) will be presented in the next section. Based on the Royall's[10] results Żądło[11] shows that the BLUP of the d^{*}'th domain total for the t^{*}'th period and its MSE under the model (4) are given by:

$$\hat{\theta}_{BLU^{*}} = \sum_{i \in s_{d^{*}t^{*}}} Y_{it^{*}} + t^{*} N_{rd^{*}t^{*}} \hat{\beta}_{d^{*}} + \sigma_v^2 t^{*} \sum_{i=1}^{N_{rd^{*}t^{*}}} \sum_{j=1}^{m_{id^{*}}} b_{id^{*}}^{-1} t_{ij} \left(Y_{id^{*}j} - t_{ij} \hat{\beta}_{d^{*}} \right), \qquad (7)$$

where $\hat{\beta}_{d^{*}} = \left(\sum_{i=1}^{n_{d^{*}}} b_{id^{*}}^{-1} \sum_{j=1}^{m_{id^{*}}} t_{ij}^2 \right)^{-1} \left(\sum_{i=1}^{n_{d^{*}}} b_{id^{*}}^{-1} \sum_{j=1}^{m_{id^{*}}} Y_{idj} t_{ij} \right)^{-1}$, $b_{id^{*}} = \sigma_e^2 + \sigma_v^2 \sum_{j=1}^{m_{id^{*}}} t_{ij}^2$.

$$MSE_{\xi}(\hat{\theta}_{BLU^{*}}) = g_{1^{*}}(\delta) + g_{2^{*}}(\delta), \qquad (8)$$

where

$$g_{1^{*}}(\delta) = N_{rd^{*}t^{*}} (\sigma_e^2 + t^{*2} \sigma_v^2) - \sigma_v^4 t^{*2} \sum_{i=1}^{N_{rd^{*}t^{*}}} b_{id^{*}}^{-1} \sum_{j=1}^{m_{id^{*}}} t_{ij}^2, \qquad (9)$$

[7] R.M. Royall: *The Linear Least Squares Prediction Approach to Two-Stage Sampling*. "Journal of the American Statistical Association" 1973, Vol. 71, No. 355, p. 657-664.

[8] C.R. Henderson: *Estimation of Genetic Parameters*. "Annals of Mathematical Statistics" 1950, Vol. 21, p. 309-310.

[9] T. Żądło: *On prediction of...*, op. cit.

[10] R.M. Royall: Op. cit.

[11] T. Żądło: *On prediction of...*, op. cit.

$$g_{2*}(\delta) = \left(t^* N_{rd^*t^*} - \sigma_v^2 t^* \sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} t_{ij}^2 \right)^2 \left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} t_{ij}^2 \right)^{-1} \qquad (10)$$

If the unknown parameters in the formula of the BLUP are replaced by any even, translation–invariant estimators, random components are symmetrically distributed around zero and the expectation of the EBLUP is finite, then the EBLUP remains unbiased (the proof for empirical version of the Royall's BLUP is presented by Żądło[12]). Kackar and Harville[13] show that standard estimating procedures including maximum likelihood (ML) and restricted maximum likelihood (REML) under normality yield even, translation–invariant estimators. The formula of the MSE of the Royall's[14] EBLUP and MSE estimators will be used based on the results presented by Żądło[15] – the proof may be treated as the generalization of results of Datta and Lahiri[16] for the EBLUP proposed by Henderson[17]. Other MSE estimators (but for the EBLUP derived by Henderson[18]) are proposed by Das, Jiang, Rao[19] and Prasad, Rao[20].

Żądło[21] shows that the MSE of the EBLUP for ML and REML variance estimators in the considered case is given by:

$$MSE_{\xi}(\hat{\theta}_{EBLU^*}(\hat{\delta})) = g_{1*}(\delta) + g_{2*}(\delta) + g_{3*}^*(\delta) + o(D^{-1}), \qquad (11)$$

where $g_{1*}(\delta)$ and $g_{2*}(\delta)$ are given by (9) and (10) and

$$g_{3*}^*(\delta) = t^{*2} \sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-3} \sum_{j=1}^{m_{id^*}} t_{ij}^2 \left(I_{vv}^{(-1)} \sigma_v^4 - 2 I_{ve}^{(-1)} \sigma_e^2 \sigma_v^2 + I_{ee}^{(-1)} \sigma_e^4 \right) \qquad (12)$$

---

[12] Idem: *On Unbiasedness of Some EBLU Predictor*. In: *Proceedings in Computational Statistics 2004*. Ed. J. Antoch. Physica-Verlag, Heidelberg-New York 2004, p. 2019-2026.

[13] R.N. Kackar, D.A. Harville: *Unbiasedness of Two-stage Estimation and Prediction Procedures for Mixed Linear Models* "Communications in Statistics Series" 1981, Vol. 10, p. 1249-1261.

[14] R.M. Royall: Op. cit.

[15] T. Żądło: *On MSE of EBLUP*. "Statistical Papers" 2009, Vol. 50, No. 1, p. 101-118.

[16] G.S. Datta, P. Lahiri: *A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems*. "Statistica Sinica" 2000, Vol. 10, No. 2, p. 613-627.

[17] C.R. Henderson: Op. cit.

[18] Ibidem.

[19] K. Das, J. Jiang, J.N.K. Rao: *Mean Squared Error of Empirical Predictor*. "The Annals of Statistics" 2004, Vol. 32, No. 2, p. 818-840.

[20] N.G.N. Prasad, J.N.K. Rao: *The Estimation of Mean the Mean Squared Error of Small Area Estimators*. "Journal of the American Statistical Association" 1990, Vol. 85, No. 409, p. 163-171.

[21] T. Żądło: *On prediction of...*, op. cit.

(additional asterisk in the superscript of $g_{3*}^{*}(\delta)$ is according to the notation of Datta and Lahiri[22]), where:

$$I_{vv}^{(-1)} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} t_{ij}^2\right)^2, \quad I_{ve}^{(-1)} = -2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} t_{ij}^2\right),$$

$$I_{ee}^{(-1)} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right),$$

$$b =$$

$$= \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right)\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} t_{ij}^2\right)^2\right) - \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} t_{ij}^2\right)\right)^2.$$

The estimator of MSE is given by (Żądło[23]):

$$\hat{MSE}_{\xi}(\hat{\theta}_{EBLU^*}(\hat{\delta})) = g_{1*}(\hat{\delta}) + g_{2*}(\hat{\delta}) + 2g_{3*}^{*}(\hat{\delta}) - \mathbf{B}_{\hat{\delta}}^{T}(\hat{\delta})\frac{\partial g_{1*}(\hat{\delta})}{\partial \delta}, \qquad (13)$$

where $g_{1*}(\hat{\delta})$, $g_{2*}(\hat{\delta})$, $g_{3*}^{*}(\hat{\delta})$ are given by (9), (10) and (12) respectively, where δ is replaced by $\hat{\delta}$, $\frac{\partial g_{1*}(\hat{\delta})}{\partial \delta}$ is given by:

$$\frac{\partial g_{1*}(\delta)}{\partial \delta} = \left[\frac{\partial g_{1*}(\delta)}{\partial \sigma_e^2} \quad \frac{\partial g_{1*}(\delta)}{\partial \sigma_v^2}\right]^T =$$

$$= \begin{bmatrix} N_{rd^*t^*} - \sigma_v^4 t^{*2}\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-2}\sum_{j=1}^{m_{id^*}} t_{ij}^2 \\ t^{*2}N_{rd^*t^*} + 2\sigma_v^4 t^{*2}\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}} t_{ij}^2 - \sigma_v^4 t^{*2}\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-2}\left(\sum_{j=1}^{m_{id^*}} t_{ij}^2\right)^2 \end{bmatrix}, \qquad (14)$$

[22] G.S. Datta, P. Lahiri: Op. cit.

[23] T. Żądło: *On prediction of...*, op. cit.

where δ is replaced by $\hat{\delta}$. $\mathbf{B}_{\hat{\delta}}(\hat{\delta})$ is given by $\mathbf{B}_{\hat{\delta}}(\delta)$ (the bias of $\hat{\delta}$), where δ is replaced by $\hat{\delta}$. If δ is estimated using ML method, the bias of $\hat{\delta}$ is given by:

$$\mathbf{B}_{\hat{\delta}^{ML}}(\delta) = \frac{1}{2}\mathbf{I}_{\delta}^{-1}(\delta)\, col_{1\le k\le q} tr\left[\mathbf{I}_{\beta}^{-1}(\delta)\frac{\partial}{\partial \delta_k}\mathbf{I}_{\beta}(\delta)\right] + o(D^{-1}). \tag{15}$$

where:

$$\mathbf{I}_{\delta}^{-1} = \begin{bmatrix} I_{vv}^{(-1)} & I_{ve}^{(-1)} \\ I_{ve}^{(-1)} & I_{ee}^{(-1)} \end{bmatrix},$$

$$col_{1\le k\le q} tr\left[\mathbf{I}_{\beta}^{-1}(\delta)\frac{\partial}{\partial \delta_k}\mathbf{I}_{\beta}(\delta)\right] = -\begin{bmatrix} \sum_{d=1}^{D}\left(\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} t_{ij}^2\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2}\sum_{j=1}^{m_{id}} t_{ij}^2\right) \\ \sum_{d=1}^{D}\left(\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} t_{ij}^2\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} t_{ij}^2\right)^2\right) \end{bmatrix}$$

and δ is replaced by $\hat{\delta}$. The bias of REML estimators $\hat{\delta}$ may be omitted (and hence the whole forth term on the right side of (13)) because:

$$\mathbf{B}_{\hat{\delta}^{REML}}(\delta) = o(D^{-1}). \tag{16}$$

The MSE estimator (13) is approximately unbiased in the sense that

$$E_{\xi}\left(\hat{MSE}_{\xi}(\hat{\theta}_{EBLU*}(\hat{\delta}))\right) = MSE_{\xi}(\hat{\theta}_{EBLU*}(\hat{\delta})) + o(D^{-1}). \tag{17}$$

3. Special case of EBLUP of domain total

In this section we present special cases of the equations presented in the previous section assuming (4) and (5). The BLUP (7) simplifies to:

$$\hat{\theta}_{BLU*} = \sum_{i\in s_{d^*i^*}} Y_i + t^* N_{rd^*i^*}\hat{\beta} + \sigma_v^2 t^* \sum_{i=1}^{N_{rd^*i^*}}\sum_{j=1}^{m_i} b_{id^*}^{-1} t_{ij}(Y_{id^*j} - t_{ij}\hat{\beta}), \tag{18}$$

where

$$\hat{\beta} = \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_i} t_{ij}^2\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_i} Y_{idj}t_{ij}\right)^{-1}. \tag{19}$$

Its MSE is given by (8), where $g_{1^*}(\delta)$ is given by (9) and

$$g_{2^*}(\delta)=\left(t^* N_{id^*t^*}-\sigma_v^2 t^* \sum_{i=1}^{N_{id^*t^*}} b_{id^*}^{-1} \sum_{j=1}^{m_i} t_{ij}^2\right)^2 \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1} \sum_{j=1}^{m_i} t_{ij}^2\right)^{-1}. \qquad (20)$$

The MSE of the EBLUP is given by (11), where $g_{1^*}(\delta)$ is given by (9), $g_{2^*}(\delta)$ by (20) and $g_{3^*}(\delta)$ by (12). The estimator of the MSE is given by (13), where $g_{1^*}(\hat{\delta})$, $g_{2^*}(\hat{\delta})$, $g_{3^*}^*(\hat{\delta})$ are given by (9), (20) and (12) respectively, where δ is replaced by $\hat{\delta}$, $\frac{\partial g_{1^*}(\delta)}{\partial \delta}$ is given by (14), where δ is replaced by $\hat{\delta}$, $\mathbf{B}_{\hat{\delta}^{REML}}(\delta)$ is given by (16), and $\mathbf{B}_{\hat{\delta}^{ML}}(\delta)$ is given by (15), where:

$$col_{1\le k\le q} tr\left[\mathbf{I}_\beta^{-1}(\delta)\frac{\partial}{\partial \delta_k}\mathbf{I}_\beta(\delta)\right]=$$

$$=-\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} t_{ij}^2\right)^{-1}\left[\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\sum_{j=1}^{m_{id}} t_{ij}^2\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} t_{ij}^2\right)^2\right)\right]^T \qquad (21)$$

These equations have been introduced because it will be shown in section 5 that MSE and MSE estimator of EBLUP of linear combination of two domains'total under (4) and simplifying assumption (5) include additional term comparing with equations in more general assumption (4).

4. EBLUP of linear combination of two domains' totals

Let us consider linear combination (LC) of two domain totals – the d^*'th domain total in the period t^* and the $d^\#$'th domain total in the period $t^\#$ denoted by $\theta_{(LC)}=a\theta_{*}+b\theta_{\#}$. Special cases of $\theta_{(LC)}$ are inter alia: sum of domains' totals ($a=1$, $b=1$), difference of domain totals ($a=1$, $b=-1$), sum of domains' means ($a=N_{d^*}^{-1}, b=N_{d^\#}^{-1}$) and difference of domains' means ($a=N_{d^*}^{-1}, b=-N_{d^\#}^{-1}$). The BLUP, the EBLUP and MSE components will be denoted by $\hat{\theta}_{BLU(LC)}$, $\hat{\theta}_{EBLU(LC)}$, $g_{1(LC)}(\delta)$, $g_{2(LC)}(\delta)$, $g_{3(LC)}^*(\delta)$. In this section considerations are based on the assumption (4) (equations under assumptions (4) and (5) will be presented in the next section).

Firstly, we consider the formulae of the predictor. Based on the Royall's[24] theorem we obtain:

$$\hat{\theta}_{BLU(LC)} = a\hat{\theta}_{BLU^*} + b\hat{\theta}_{BLU\#}, \tag{22}$$

$$\hat{\theta}_{EBLU(LC)} = a\hat{\theta}_{EBLU^*} + b\hat{\theta}_{EBLU\#}, \tag{23}$$

Secondly, we analyze the form of the MSE and the estimator of the MSE assuming (4). We obtain:

$$g_{1(LC)}(\delta) = a^2 g_{1^*}(\delta) + b^2 g_{1\#}(\delta), \tag{24}$$

$$g_{2(LC)}(\delta) = a^2 g_{2^*}(\delta) + b^2 g_{2\#}(\delta), \tag{25}$$

$$g_{3(LC)}(\delta) = a^2 g_{3^*}(\delta) + b^2 g_{3\#}(\delta). \tag{26}$$

Hence,

$$MSE_\xi(\hat{\theta}_{BLU(LC)}) = a^2 MSE_\xi(\hat{\theta}_{BLU^*}) + b^2 MSE_\xi(\hat{\theta}_{BLU\#}), \tag{27}$$

$$MSE_\xi(\hat{\theta}_{EBLU(LC)}) = a^2 MSE_\xi(\hat{\theta}_{EBLU^*}) + b^2 MSE_\xi(\hat{\theta}_{EBLU\#}). \tag{28}$$

Moreover, based on (4) we obtain:

$$\mathbf{B}_{\hat{\delta}}^T(\delta)\frac{\partial g_{1(LC)}(\delta)}{\partial \delta} = a^2 \mathbf{B}_{\hat{\delta}}^T(\delta)\frac{\partial g_{1^*}(\delta)}{\partial \delta} + b^2 \mathbf{B}_{\hat{\delta}}^T(\delta)\frac{\partial g_{1\#}(\delta)}{\partial \delta} \tag{29}$$

and then

$$\hat{MSE}_\xi(\hat{\theta}_{EBLU(LC)}) = a^2 \hat{MSE}_\xi(\hat{\theta}_{EBLU^*}) + b^2 \hat{MSE}_\xi(\hat{\theta}_{EBLU\#}). \tag{30}$$

5. Special case of EBLUP of linear combination of two domains' totals

Firstly, we consider the formulae of the predictor of the linear combination of domains' totals assuming (4) and (5). In this case formulae (22) and (23) remain true.

[24] R.M. Royall: Op. cit.

Secondly, we analyze the form of the MSE and the estimator of the MSE assuming (4) and (5). In this case results (24), (26), (29) remain true but instead of (25) we obtain:

$$g_{2(LC)}(\delta) = a^2 g_{2*}(\delta) + b^2 g_{2\#}(\delta) + 2ab g_{2*\#}(\delta), \tag{31}$$

where:

$$\begin{aligned} g_{2*\#}(\delta) = &\left(t^* N_{rd^*t^*} - \sigma_v^2 t^* \sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1} \sum_{j=1}^{m_i} t_{ij}^2 \right) \times \\ &\times \left(t^\# N_{rd^\#t^\#} - \sigma_v^2 t^\# \sum_{i=1}^{N_{rd^\#t^\#}} b_{id^\#}^{-1} \sum_{j=1}^{m_i} t_{ij}^2 \right) \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-1} \sum_{j=1}^{m_i} t_{ij}^2 \right)^{-1} \end{aligned} \tag{32}$$

Hence, assuming (4) and (5) we obtain:

$$MSE_\xi(\hat{\theta}_{BLU(LC)}) = a^2 MSE_\xi(\hat{\theta}_{BLU*}) + b^2 MSE_\xi(\hat{\theta}_{BLU\#}) + 2ab g_{2*\#}(\delta), \tag{33}$$

$$MSE_\xi(\hat{\theta}_{EBLU(LC)}) = a^2 MSE_\xi(\hat{\theta}_{EBLU*}) + b^2 MSE_\xi(\hat{\theta}_{EBLU\#}) + 2ab g_{2*\#}(\delta), \tag{34}$$

and

$$\hat{MSE}_\xi(\hat{\theta}_{EBLU(LC)}) = a^2 \hat{MSE}_\xi(\hat{\theta}_{EBLU*}) + b^2 \hat{MSE}_\xi(\hat{\theta}_{EBLU\#}) + 2ab g_{2*\#}(\hat{\delta}). \tag{35}$$

6. Simulation analysis

The Monte Carlo simulation analysis is prepared to check the accuracy of the EBLUP in the cases of model misspecification (comparing with the case of correct model specification) and biases of the proposed MSE estimators (including the cases of model misspecification). It is based on the artificial data and was prepared using R software[25]. Survey is conducted in 9 periods and its purpose is to predict the sum of domain totals in the period 10. The population consists of 2500 units and does not change in time. It is divided into 20 domains of sizes

[25] R Development Care Team: *A Language and Environment for Statistical Computing.*

from 50 to 200 units (details are presented in the Table 1). The rotating scheme 2-(2)-2 of the sample is considered what means that each elementary sample (out of 4) after some start-up phase is surveyed in 2 periods, then it is not surveyed during 2 periods and finally it is surveyed in 2 periods again.

Table 1

Sample sizes in domains in 9 periods

d	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	sum
N_d	50	50	50	50	50	100	100	100	100	100	150	150	150	150	150	200	200	200	200	200	2500
n_{d1}	2	1	2	2	1	0	3	0	2	2	2	2	2	3	5	2	2	2	2	3	40
n_{d2}	1	0	2	3	2	2	3	0	1	2	1	1	2	2	3	5	2	4	1	3	40
n_{d3}	2	1	0	3	1	2	1	1	1	2	2	0	4	4	2	4	2	5	2	1	40
n_{d4}	2	1	0	2	0	0	1	1	2	2	5	1	5	3	4	1	1	4	2	3	40
n_{d5}	1	1	0	1	0	0	3	0	3	2	4	3	4	2	4	3	1	3	3	2	40
n_{d6}	2	2	0	1	0	3	3	0	1	2	2	2	3	2	2	6	1	3	4	1	40
n_{d7}	3	1	0	1	0	3	1	1	0	1	3	2	3	2	2	5	1	5	3	3	40
n_{d8}	1	0	0	0	1	0	2	1	2	2	3	3	5	2	2	2	2	4	3	5	40
n_{d9}	1	1	0	1	1	0	2	0	3	2	3	2	4	3	2	2	1	4	5	3	40
sum	15	8	4	14	6	10	19	4	15	17	25	16	32	23	26	30	13	34	25	24	360

This rotating scheme is used in the Population Economic Activity Survey conducted by Polish Central Statistical Office. In the considered simulation study elementary samples include 10 elements each, hence, according to the used rotating scheme 40 population elements from 4 elementary samples (each of size 10) were observed in each period.

Data are generated from the superpopulation model (4), assuming that:

$$\forall_d \beta_d = \beta \text{ and } \forall_{i,j} t_{ij} = 1. \tag{36}$$

The number of iterations (generated populations) is 50 000. Special cases of the equations presented in the previous sections based on (36) are used.

In the simulation study the problem of prediction of the sum of two domains' totals – first domain in the future period 10 and other domains (out of 20) in the future period 10 is studied. Hence, the purpose of the survey is prediction of 19 characteristics (19 sums of domain totals) for domains: 1st and 2nd, 1st and 3rd,..., 1st and 20th. The following predictors are used: BLUP – the BLUP (assuming that σ_e^2 and σ_v^2 are known); EBLUP ML – the EBLUP, where σ_e^2 and σ_v^2 are estimated using ML; EBLUP REML – the EBLUP, where σ_e^2 and σ_v^2 are estimated using REML. To estimate model parameters using ML and REML, R function *lme* was used with default arguments.

In the simulation the following, arbitrarily chosen values of parameters are considered: $\beta = 10$, $\sigma_e^2 = 1$, $\sigma_v^2 = 3$. We consider 9 cases of different distributions of random components (with values of variances presented above): NN – normal distribution of v_{id}, normal distribution of e_{idj}, NU – normal distribution of v_{id}, uniform distribution of e_{idj}, NE – normal distribution of v_{id}, shifted exponential distribution of e_{idj}, UN – uniform distribution of v_{id}, normal distribution of e_{idj}, EN – shifted exponential distribution of v_{id}, normal distribution of e_{idj}, UU – uniform distribution of v_{id}, uniform distribution of e_{idj}, EE – shifted exponential distribution of both v_{id} and e_{idj}, UE – uniform distribution of v_{id}, shifted exponential distribution of e_{idj}, EU – shifted exponential distribution of v_{id}, uniform distribution of e_{idj}. Random components are generated using *rnorm*, *runif*, *rexp* (after transformation) R functions.

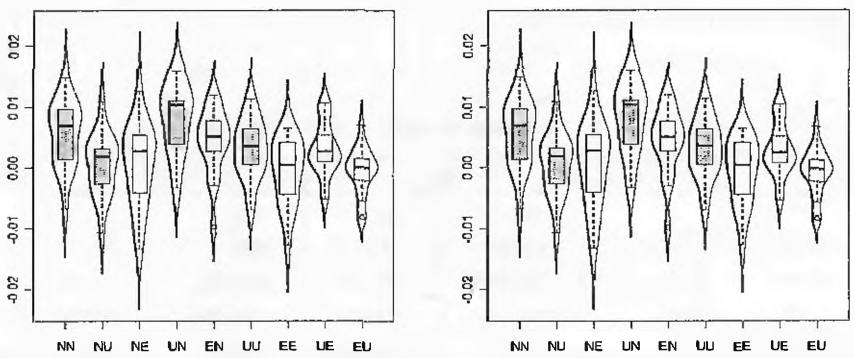
Table 2

Results for 19 sums of domain totals for NN case

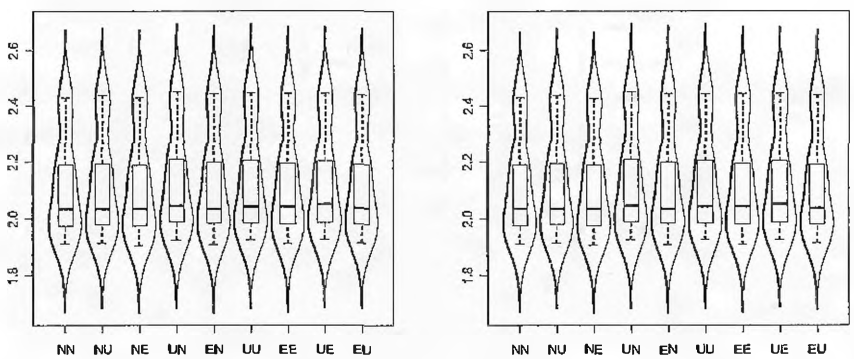
case		NN		EE	
predictor		EBLUP REML	EBLUP ML	EBLUP REML	EBLUP ML
predictor's relative bias (in %)	minimum	-0 0068*	-0.0068*	-0.0128	-0,0128
	Q1	0.0014*	0,0014*	-0 0046	-0 0046
	Me	0.0069*	0.0069*	0 0004	0 0004
	mean	0.0055*	0.0055*	-0.0007	-0.0007
	Q3	0.0097*	0,0097*	0.0040	0.0040
	maximum	0.0148*	0,0148*	0 0065	0 0065
relative RMSE (in %)	minimum	1 9121	1 9121	1 9104	1 910427
	Q1	1.9745	1 9745	1 9795	1.979505
	Me	2 0335	2 0335	2.0421	2.042082
	mean	2.1177	2.1177	2.1223	2.122336
	Q3	2.1932	2.1932	2.1951	2.195104
	maximum	2.4297	2.4297	2.4442	2.444231
MSE estimator's relative bias (in %)	minimum	-0.1676	-0.1675	-1.1917	-1.3874
	Q1	0.3596	0 3593	-0.4036	-0.5430
	Me	0.5719	0 5713	0.0216	-0.2152
	mean	0.5766	0.5761	0.0449	-0.1427
	Q3	0.8511	0.8507	0.4600	0.2482
	maximum	1.3139	1.3131	1.1807	0.9456

* model unbiased predictor – real values equal 0

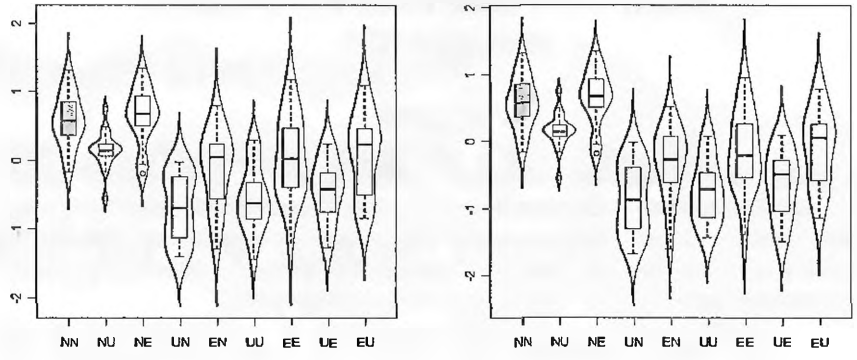
The ratio of the MSE of the EBLUP and the MSE of the BLUP is less than 1.001 for all of models in all of domains. Hence, the accuracy of the predictor decreases due to the estimation of unknown variance parameters only by less than 0.1%. In the simulation study the following statistics are computed: relative biases of predictors in %, relative RMSE of predictors in %, relative biases of MSE estimators in % for different distributions of random components for all of 19 sums of domain totals. In the Table 2 detailed results for NN and EE cases are presented. Results for all of cases of different distributions of random components are summarized on Graphs 1-3 using both boxplots and violinplots (the function *simple.violinplot* in *UsingR* package was used).



Graph 1. Relative biases of EBLUP REML (on the left) and EBLUP ML (on the right) in %



Graph 2. Relative RMSE of EBLUP REML (on the left) and EBLUP ML (on the right) in %



Graph 3. Relative biases of MSE estimators of EBLUP REML (on the left) and EBLUP ML (on the right) in %

It is known[26], that if (inter alia) random components are symmetrically distributed (normal distribution in not necessary) the EBLUP remains unbiased. Hence in NN, NU, UN, UU cases (grey boxplots) on the Graph 1 the simulation bias is presented (real values equal 0). Note that absolute values of relative biases for the considered predictors in all of cases are not high – they do not exceed 0.02%.

For all of cases similar values of relative RMSEs are obtained – they are between 1.9% and 2.9% for 19 characteristics (Graph 2).

The absolute values of relative biases of MSE estimators (see Graph 3) for the assumed in derivations NN case (grey boxplot) are less then 1.32%. For other (nonnormal) cases (when derived equations are not appropriate) they are also small – they do not exceed 1.7% and although it is higher than 1.32% it is still acceptable. Biases of MSE estimators for non NN cases are especially small when random components v_{ul} are normal (NU and NE cases). Despite small biases of the MSE estimator other MSE estimators are worth considering in the future studies[27].

[26] T. Żądło: *On Unbiasedness of...*, op. cit.

[27] J. Jiang, P. Lahiri, S.-M. Wan: *A Unified Jackknife Theory for Empirical Best Prediction with M-estimation*. "The Annals of Statistics" 2002, Vol. 30, No. 6, p. 1782-1810.

O PREDYKCJI LINIOWEJ KOMBINACJI WARTOŚCI GLOBALNYCH W DOMENACH W BADANIACH OKRESOWYCH

Streszczenie

W artykule zaproponowano empiryczny najlepszy liniowy nieobciążony predyktor liniowej kombinacji wartości globalnych w domenach, zakładając model nadpopulacji ze składnikami losowymi specyficznymi dla profili wielookresowych wraz z błędem średniokwardatowym i jego estymatorem. Rozważania wzbogacono o badania symulacyjne, z uwzględnieniem problemu złej specyfikacji modelu nadpopulacji.

Agnieszka Orwat-Acedańska

THE CLASSIFICATION OF POLISH MUTUAL BALANCED FUNDS ACCORDING TO THE MANAGEMENT STYLE USING ANDREWS ESTIMATORS

Introduction

The funds' managers invest in assets with different risk characteristics. Since investments efficiency differs among sectors it is not easy to distinguish between returns that result from investment decisions on sectors and investment decisions on particular assets within a given sector. A style analysis introduced by William Sharpe in 1992 allows to attribute the results of portfolio management to the decision on investments in different asset classes. Non-negativity of the Sharpe style model coefficients implies that the exact distribution of OLS estimators of the coefficients is not known. The knowledge of the estimators' distribution is essential for correct interval estimation as well as hypothesis testing and hence for meaningful inference on the impact of the decisions on asset classes for the funds' performance. A method proposed by Donald Andrews allows to build correctly confidence intervals for the parameters.

The article's aim is to estimate the confidence intervals for the style coefficients for the mutual balanced funds operating on the Polish financial market. These results are then used to classify the funds.

1. Sharpe style analysis model

The style describes the way a fund invests its money that allows to reach a predetermined investment results. The style analysis aims at attributing a fund's rate of returns in a given period to investment in different asset classes. The relationship between a fund's rate of return and rates of return from indexes representing different asset classes in period t, $t = 1,2,\ldots,T$ is given by:

$$R_t = \beta_1 F_{t1} + \beta_2 F_{t2} + \ldots + \beta_k F_{tk} + \varepsilon_t, \tag{1}$$

where R_t – fund's rate of return in period t; F_{ti}, $i = 1,2,\ldots,k$ – a rate of return from the index i in period t; β_i, $i = 1,2,\ldots,k$ – the i'th parameter of the model describing sensitivity of R_t to F_{ti} (i'th style share); ε_t – error term. $\mathbf{F} = (F_{t1}, F_{t2},\ldots,F_{tk})'$

Let $\mathbf{R} = (R_1, R_2,\ldots,R_T)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_T)'$, $\mathbf{F} = (\mathbf{F}_1, \mathbf{F}_2,\ldots,\mathbf{F}_T)'$, where $\mathbf{F} = (F_{t1}, F_{t2},\ldots,F_{tk})'$ means a random vector of rates of returns from the indexes in period t. A vector of unknown parameters $\beta = (\beta_1, \beta_2,\ldots,\beta_k)'$ represents the set of style shares defining a portfolio of k asset classes. A product $\mathbf{F}_t'\beta$ is known as "*return on a style portfolio*" [1].

The model (1) main assumptions are as follows:

a) $T > k$;
b) the terror terms ε_t are independent and identically distributed random variables such that $E(\varepsilon_t) = 0$, $D^2(\varepsilon_t) < \infty$, $t = 1,2,\ldots,T$;
c) the vectors $(R_t, \mathbf{F}_t')$ for every $t = 1,2,\ldots,T$ are independent and identically distributed;
d) the vector $\mathbf{F}_t$ and the terror term ε_t are uncorrelated for every $t = 1,2,\ldots,T$.

In the model R_t and $\mathbf{F}$ are treated as random variables. For the estimation purposes their realizations are used, so to simplify the notation they will also be noted as R_t and $\mathbf{F}$. Using a matrix representation:

$$\mathbf{R} = \mathbf{F}\beta + \boldsymbol{\varepsilon}, \tag{2}$$

further assumptions read as follows:

e) $rz(\mathbf{F}) = k$;

[1] W.F. Sharpe: *Asset Allocation Management Style and Performance Measurement.* "Journal of Portfolio Management" 1992, No. 18(2), p. 7.

f) the vector β of the Sharpe style shares, known also as the Sharpe style weights, satisfies:

$$1'\beta = 1, \text{ where } 1 \text{ is k-dimensional vector of ones,} \tag{3}$$

$$\beta \geq 0. \tag{4}$$

The product of estimated parameters and the rates of return from the indexes represent an ideal market portfolio, also known as a passive portfolio. Rates of return from the passive part correspond to rates of return from a portfolio style, whereas rates of return from the active part correspond to the model's random errors. An investor's ability to select assets with returns higher than the market return in a given period is called a selection effect. Since this effect represents part of the return that exceeds return from a portfolio of randomly selected assets it could be attributed to active portfolio management strategy. On the other hand an allocation effect represents passive management strategy. In this context it could be said that passive managers can provide their customers only with an "investment style", whereas active ones provide both, style and selection[2]. Determination coefficients from the style equation (2) measures the style effect. One minus determination coefficient assesses magnitude of the selection effect.

2. Chassis estimation of the Sharpe style weights

An OLS estimator of the Sharpe style weights $\hat{\beta}_{SMNK}$ is a solution to the following optimization problem:

$$\min_{\beta \in \Theta} \frac{1}{T}(R - F\beta)'(R - F\beta), \tag{5}$$

s. t.

$$\Theta = \{\beta: \ 1'\beta = 1, \ \beta \geq 0\}. \tag{6}$$

[2] Ibidem.

If only the summation constrain (3) is concerned then the OLS estimator of β is asymptotically distributed as normal with mean β and known covariance[3]. If also the non-negativity constrain (4) is added the exact distribution of the OLS estimator in not known. The robust method that allows to build the correct confidence intervals in this situation is the method proposed by Andrews.

3. Andrews estimation in the style analysis

Andrews proved[4] that asymptotic distribution of $\hat{\beta}_{MNK}$ can be approximated by a random vector $\hat{\lambda}$:

$$\sqrt{T}(\hat{\beta}_{SMNK} - \beta) \xrightarrow[T\to\infty]{dystr.} \hat{\lambda}, \tag{7}$$

where $\hat{\lambda}$ is a solution to the following problem:

$$\min_{\lambda}(\lambda - Z)'M(\lambda - Z), \tag{8}$$

s. t.

$$1'\lambda = 0,\ Q\lambda \le 0, \tag{9}$$

where $M = E(F_t F_t')$ – $(k\times k)$-dimensional matrix satisfying det(M) > 0; Z – k-dimensional random vector such that $Z = M^{-1}G$, G is k-dimensional random normal vector $G \sim N_k(0,V)$ with a covariance matrix $V = E(\varepsilon_t^2 F_t F_t')$, where det(V) > 0; $Q = [q_{ji}]_{j=1,\ldots,l;\, i=1,\ldots,k}$, – a matrix such that $q_{ji} = -1$, when $\beta_i = 0$ and $q_{ji} = 0$ otherwise; $j = 1,\ldots,l$; l – number of zero coefficients. The condition $1'\lambda = 0$ results from the fact that the asymptotic distribution of the random vector $\hat{\lambda}$ is close to the distribution of the statistic $\hat{\beta}_{SMNK} - \beta$. which satisfies $1'\beta = 1$, $1'\hat{\beta}_{SMNK} = 1$ and hence $1'(\hat{\beta}_{SMNK} - \beta) = 0$. The constrain $Q\lambda \le 0$ represents the fact that $\lambda_i \ge 0$ for $\beta_i = 0$, $i = 1,2,\ldots,k$, which comes from the very definition of $\hat{\lambda}$.

[3] W.H. Greene: *Econometric Analysis*. Prentice Hall, New Jersey 2002, p. 194.

[4] D.K.W. Andrews: *Estimation When a Parameter Is on Boundary*. "Econometrica" 1999, No. 67, p. 1341-1383.

The estimation of asymptotic confidence intervals for β is based on estimation of the distribution of the random vector $\hat{\lambda}$[5]. Here we assume that the conditions a) – f) of the Sharpe style analysis model are satisfied.

In the first step we calculate an estimator $\hat{\mathbf{M}}$ of the matrix M and an estimator $\hat{\mathbf{V}}$ of the matrix V according to: $\hat{\mathbf{M}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{F}_t'$, $\hat{\mathbf{V}} = \frac{1}{T}\sum_{t=1}^{T}e_t^2\mathbf{F}_t\mathbf{F}_t'$, where $e_t = R_t - \mathbf{F}_t'\hat{\beta}_{SMNK}$.

In the next step we build the matrix Q. For this purpose we test the null hypotheses $H_0 : \beta_i = 0$, $i = 1,2,\ldots,k$ against the alternatives $H_0 : \beta_i > 0$ at the significance level α. The vector $\tilde{\beta}$ of the OLS estimators of β satisfying the summation condition (3) is given by:

$$\tilde{\beta} = \ddot{\beta} - \hat{\mathbf{M}}^{-1}\mathbf{1}(\mathbf{1}'\hat{\mathbf{M}}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\ddot{\beta} - 1), \tag{10}$$

where $\ddot{\beta}$ is k-dimensional of OLS estimates without restrictions:

$$\ddot{\beta} = \hat{\mathbf{M}}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{R}_t\right)$$

From the assumption b) it follows that the estimator $\tilde{\beta}$ given by (10) is asymptotically normal distributed[6]. If the null is true then the statistic z_i for every $i = 1,2,\ldots,k$:

$$z_i = \tilde{\beta}_i / s(\tilde{\beta}_i) \tag{11}$$

where $\tilde{\beta}_i$ is i-th element of the vector of OLS estimates $\tilde{\beta}$ of β has asymptotically Student t distribution[7].

The standard error of $\tilde{\beta}_i$ is given by:

$$s(\tilde{\beta}_i) = \sqrt{\tilde{c}_{ii}},\ i = 1,2,\ldots,k, \tag{12}$$

---

[5] T. Kim, H. White, D. Stone: *Asymptotic and Bayesian Confidence Intervals for Sharpe-Style Weights*. "Journal of Financial Econometric" 2005, No. 3(3), p. 319.

[6] W.H. Greene: *Econometric Analysis...*, op. cit., s. 68.

[7] Ibidem. For big samples the Andrews method does not need the terror term normality assumption. For small samples the normality condition is necessary for the limiting distribution of z_i to be Student t.

where $\tilde{c}_{ii}$ is the i-th diagonal element of the covariance matrix $\tilde{\mathbf{C}}$ of the estimator $\tilde{\boldsymbol{\beta}}$:

$$\tilde{\mathbf{C}} = \frac{1}{T}(\tilde{\mathbf{D}} + \frac{1}{m^2}\hat{\mathbf{M}}^{-1}\mathbf{1}\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1}\mathbf{1}'\hat{\mathbf{M}}^{-1} - \frac{1}{m}\tilde{\mathbf{D}}\mathbf{1}\mathbf{1}'\hat{\mathbf{M}}^{-1} - \frac{1}{m}\hat{\mathbf{M}}^{-1}\mathbf{1}\mathbf{1}'\tilde{\mathbf{D}}), \quad (13)$$

where: $\tilde{\mathbf{D}} = \hat{\mathbf{M}}^{-1}\tilde{\mathbf{V}}\hat{\mathbf{M}}^{-1}$, $\tilde{\mathbf{V}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{e}_t^2\mathbf{F}_t\mathbf{F}_t'$, $\tilde{e}_t = R_t - \mathbf{F}_t'\tilde{\boldsymbol{\beta}}$, $m = \mathbf{1}'\hat{\mathbf{M}}^{-1}\mathbf{1}$. If for a given significance level α an cquality $z_t \geq z_\alpha$ holds, where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha$, Φ – being the normal cumulative distribution, then the null is rejected in favor of H_1. Otherwise the null is no rejected. The results from these tests are used to construct the matrix Q.

In the next step Monte Carlo realizations of a random vector $\hat{\mathbf{G}}_h$, $h = 1, 2,\ldots, N$ (N – number of realizations) such that $\hat{\mathbf{G}}_h \sim N_k(\mathbf{0}, \hat{\mathbf{V}})$ are generated. For every $\hat{\mathbf{G}}_h$ a vector $\hat{\boldsymbol{\lambda}}_h$ is defined as a solution to the following problem:

$$\min_{\boldsymbol{\lambda}} (\boldsymbol{\lambda} - \hat{\mathbf{M}}^{-1}\hat{\mathbf{G}}_h)\hat{\mathbf{M}}(\boldsymbol{\lambda} - \hat{\mathbf{M}}^{-1}\hat{\mathbf{G}}_h), \quad (14)$$

s.t.

$$\mathbf{1}'\boldsymbol{\lambda} = 0, \ \mathbf{Q}\boldsymbol{\lambda} \leq 0 \quad (15)$$

From the set of solutions $\hat{\boldsymbol{\lambda}}_h$ $\alpha/2$-quantile z_L and $(1 - \alpha/2)$-quantile z_U are calculated. Approximately the following equality holds: $P(z_L \leq \hat{\boldsymbol{\lambda}} \leq z_U) \approx 1 - \alpha$.

Andrews asymptotic confidence intervals for the Sharpe style weights vector β at the confidence level $1 - \alpha$ is given by:

$$(\hat{\beta}_{SMNK} - z_U T^{-1/2}, \hat{\beta}_{SMNK} - z_L T^{-1/2}), \quad (16)$$

such that $P(\hat{\beta}_{SMNK} - z_U T^{-1/2} \leq \beta \leq \hat{\beta}_{SMNK} - z_L T^{-1/2}) \approx 1 - \alpha$.

For the parameters for which the null is not rejected z_L are close to 0 and z_U are positive. Therefore the confidence intervals are approximations to the interval (16) given by $(\hat{\beta}_{SMNK} - z_U T^{-1/2}, \hat{\beta}_{SMNK})$. Due to finite sample lower bounds for the Andrews asymptotic intervals might be slightly negative, whereas upper bounds might be close to 0 for the parameters which do not significantly differ from 0.

4. Robust estimation of the covariance matrix V

If the assumption c) does not hold, for example when R_t or $\mathbf{F}_t'$ are heteroscedastic or serially correlated, then the standard estimator of $\mathbf{V} = \mathrm{E}(\varepsilon_t^2 \mathbf{F}_t \mathbf{F}_t')$ is biased. A robust Newey–West heteroscedacsity and autocorrelation consistent (HAC) estimator can be used to obtain consistent and unbiased estimates of V. The HAC estimator is given by[8]:

$$\hat{\mathrm{avar}}_{\mathrm{HAC}}(\mathbf{G}) = \frac{1}{T}\sum_{t=1}^{T} e_t^2 \mathbf{F}_t \mathbf{F}_t' + \frac{1}{T}\sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-l}(\mathbf{F}_t \mathbf{F}_{t-l}' + \mathbf{F}_{t-l}\mathbf{F}_t'), \qquad (17)$$

where l – lag truncation, $w_l = 1 - \dfrac{1}{L+1}$. There are several ways to set the lag l. One of them is to assume[9]: $l \approx T^{0,25}$.

5. Hierarchical classification methods

In the set theory classification is defined as a nonempty family of subsets K_i, $i = 1,2,\ldots,k$ over a set of objects $\mathbf{K}$ that satisfies the conditions:[10]

$$\mathbf{K}_i \cap \mathbf{K}_j = \phi, \quad \bigcup_{i=1}^{k} \mathbf{K}_i = \mathbf{K}, \; i \neq j, \quad i = 1,2,\ldots,k, \qquad (18)$$

where ϕ is an empty set. Hence classification is treated as a set of classes taken from the set of classified objects[11].

Hierarchical grouping procedures can be described with the following scheme: given a distance matrix for the set of objects it is initially assumed that every object forms a separate class. Then a pair of class is found for which the distance between them is the shortest. They are merged and form one new class. Then the new distance matrix is calculated. The procedure continues until only one class is left. Differences between methods come from different ways of calculating the distance between the classes. The most popular methods are: single

[8] Ibidem, p. 200.
[9] Ibidem.
[10] E. Gatnar, M. Walesiak: *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydawnictwo Akademii Ekonomicznej, Wroclaw 2004.
[11] Ibidem.

linkage, complete single, unweighted pair-group average, weighted pair-group average, unweighted pair-group centroid, weighted pair-group centroid and Ward's method[12].

6. Results of empirical analysis

Management style of every of 13 open balanced mutual funds operating on the market during the whole period 02.01.2002–30.06.2008 was analysed. These were: Pioneer Zrównoważony FIO, ING FIO Zrównoważony, OFI Union Investment Zrównoważony, Arka BZ WBK Zrównoważony FIO, Legg Mason Zrównoważony Środkowoeuropejski FIO, Skarbiec FIO Zrównoważony, DWS Polska FIO Zrównowazony, PKO Zrównoważony FIO, Aviva Investors FIO subfundusz Aviva Investors Zrównoważony, KBC Beta SFIO, Novo FIO Subfundusz Novo Zrównoważonego Wzrostu, Millennium FIO Subfundusz Zrównoważony, BPH FIO Parasolowy Subfundusz BPH Aktywnego Zarządzania[13].

In every model logarithms of monthly rates of return from a fund's participation unit prices are treated as dependent variables. The set of independent variables is the same for all models and consists of logarithms of monthly rates of return from the indexes representing a fund's investments in different classes of stocks and bonds. Among the independent variables there are the rates of return from sector stock subindices: WIG-banks, WIG-construction, WIG-informatics, WIG-food industry and WIG-telecommunication; as well as rates of return from the following bond accounting prices: 2-year zero coupon bonds (OK), 5-year fixed interest bonds (PS) and 10-year fixed interest bonds (DS)[14].

The rates of return from the subindices and from the bonds are not highly correlated. Nonetheless correlation coefficients between rates of return from the stock indices are significantly different from 0 at the 0.05 significance level. This is also the case for correlation between the bonds PS and the bonds DS (see Table 1).

[12] Ibidem.

[13] In what follows short forms of these names are used.

[14] The monthly accounting prices of bonds were calculated from daily prices, which were average prices for all bonds of a given type quoted in that day.

Table 1

Correlation matrix of independent variables in the style model

	WIG-banks	WIG-construction	WIG-informatics	WIG-food industry	WIG-telecom	PS	DS	OK
WIG-banks	1 00	**0.63**	**0.64**	**0.58**	**0.60**	0 19	**0.26**	0 03
WIG-construction	**0.63**	1 00	**0.65**	**0.65**	**0.38**	0 01	0 08	0 02
WIG-informatics	**0.64**	**0.65**	1 00	**0.57**	**0.51**	0.08	0 14	0 08
WIG-food industry	**0.58**	**0.65**	**0.57**	1 00	**0.34**	0 10	0 08	-0 05
WIG-telecommunication	**0.60**	**0.38**	**0.51**	**0.34**	1 00	-0 03	0 13	0 19
PS	0.19	0 01	0.08	0 10	-0 03	1 00	**0.50**	0 04
DS	**0.26**	0.08	0.14	0 08	0 13	**0.50**	1 00	0.08
OK	0 03	0 02	0 08	-0.05	0.19	0.04	0 08	1 00

Coefficients different from 0 at the significance level 0 05 are bolded.

Implementation of the Andrews method was preceded by statistical verification of the model's assumptions. The error term independence assumption was examined by Ljung–Box test for autocorrelation and Koenker–Basset test for homoscedasticity. Additionally Lagrange multiplier ARCH effect test was employed. In both tests, LB and ARCH, the lag parameters of 1 and 4 were assumed. The results of these tests are presented in Table 2. For every model homoscedasticity hypothesis should not be rejected[15]. This is also the case for lack of autocorrelation and ARCH effect hypotheses[16].

Table 2

Results of hypotheses testing of the models' random errors

Fund	Koenkera–Basset test		Ljung–Box test				Lagrange multiplier ARCH test			
			$q=1$		$q=4$		$q=1$		$q=4$	
	test statistics	p-value	test statistics	p-value	test statistics	p-value	test statistics	p-value	test statistics	p-value
Pioneer	0 00	0 99	0 26	0 92	4 16	0.57	0.55	0.83	5 99	0.32
ING	0.87	0.73	1 75	0 46	4.90	0.46	0 41	0 87	2.47	0 81
UniKorona	0 02	0 99	2 98	0 20	5.61	0.37	0.57	0.82	5 14	0.43
Arka	2.63	0.26	1 86	0.43	1 86	0.43	3.01	0 20	6 87	0 23
Legg Mason	2.07	0 38	0 04	0 99	6 12	0 31	0 76	0.76	3 52	0.66
Skarbiec	0 02	0.99	0.50	0.85	6 89	0 23	0 23	0 93	1.32	0 93
DWS	1 59	0 50	2.73	0.24	6 81	0 24	9 45	0 01	5 25	0 42
PKO	1.72	0 47	7 15	0 06	10 91	0.03	0 86	0 73	1.20	0 94
Aviva	6.09	0 01	3 14	0 18	17 90	0.05	3.36	0 15	11 34	0 02
KBC	2.56	0 27	2 74	0.24	3 01	0 74	3 99	0 09	1 46	0 92
Novo	0 00	0 99	1 62	0 49	6 22	0 30	9 73	0.01	11.42	0 02
Millenium	0 97	0.69	1 48	0.53	2 31	0 83	9 91	0 01	8 11	0 14
BPH	5 66	0 02	0 19	0 94	0 46	0 98	5 26	0 03	10 01	0 05

[15] For each model p-values of the Koenker–Basset test were higher than 0.01.

[16] For each model p-values of the Ljung–Box and ARCH effect tests were higher than 0.01

The above mentioned tests were also utilized to verify the independence of the vectors $(R_t, \mathbf{F}_t')$ for every $t = 1,2,...,T$. One dimensional autocorrelation and homoscedasticity tests were carried separately for each dimension of this 9-dimensional vector. The results shown in Table 3 suggest that the homoscedasticity hypothesis should be rejected at the 0.05 significance level for such variables as WIG-informatics, WIG-telecommunication, PS, DS as well as Novo, Millenium and BPH[17].

Table 3

Results of hypotheses testing of the vectors $(R_t^j, \mathbf{F}_t')$

Independent variables	Koenker–Basset test		Ljung–Box test			
			$q = 1$		$q = 4$	
	test statistics	p-value	test statistics	p-value	test statistics	p-value
WIG-banks	0.02	0 99	0 54	0.83	2.37	0 82
WIG-construction	0 33	0 90	4.55	0 06	7 38	0 19
WIG-informatics	5.00	0.04	0 07	0.98	6.04	0 32
WIG-food industry	0 22	0 94	0 05	0 99	3 36	0.69
WIG-telecom	8.28	0.00	2 22	0 34	6.80	0 24
PS	7.56	0.00	2 97	0.20	11 12	0.03
DS	11 23	0.00	0 80	0.75	14 86	0.01
OK	0 45	0 86	0 08	0 98	5 12	0 43
Dependent variables						
Pioneer	0 99	0 68	0 05	0.98	1.31	0 93
ING	1 38	0 56	0 08	0 98	1 37	0 93
UniKorona	0 11	0.97	0 52	0 84	1 91	0 87
Arka	0 40	0 88	0 02	0 99	2 13	0 85
Le[illegible] Mason	0 46	0 86	0 12	0 96	1.27	0 93
Skarbiec	0.02	0 99	0 00	0 99	1 25	0 94
DWS	3 33	0 15	0.00	0 99	2 64	0.79
PKO	0 16	0 95	0 89	0 72	1 46	0 92
Aviva	2 30	0.32	0 11	0 97	4.12	0 58
KBC	1 91	0 42	0.15	0 96	0 66	0 97
Novo	7 69	0.00	0.29	0.91	4 54	0 52
Millenium	8 79	0.00	0.16	0 95	2 79	0.77
BPH	7 51	0.00	0.03	0.99	2.75	0.77

Table 4

Results of the model's random errors normality tests

Fund	Shapiro–Wilk test		Anderson–Darling test	
	test statistics	p-value	test statistics	p-value
1	2	3	4	5
Pioneer	0 893	0 992	0 155	0 957
ING	0 437	0 985	0 181	0 915
UniKorona	0 460	0.986	0 341	0 496

[17] For these variables p-values are less than significance level 0.05.

cont. Table 4

1	2	3	4	5
Arka	0.128	0 976	0 381	0 402
Legg Mason	0.441	0.984	0.410	0.344
Skarbiec	0.809	0.990	0 221	0 833
DWS	0.007	0.952	0.560	0.148
PKO	0 217	0.979	0.396	0.370
Aviva	0 187	0.978	0.616	0 109
KBC	0 000	0 922	1 162	0 005
Novo	0.005	0 949	0 834	0 032
Millenium	0.000	0 913	1 009	0 012
BPH	0 004	0.947	0 905	0 021

Due to relatively small sample size (78 observations) the normality assumption of the terror terms was also examined using Shapiro–Wilk and Andreson–Darling tests. The results from Table 4 suggest that there are no grounds to reject the normality hypothesis for each model.

For the point estimates $\hat{\beta}_{SMNK_i}$, $i = 1,2,\ldots,8$ obtained as solutions to the problem (5)-(6) the confidence intervals were calculated using Andrews method for each fund. A confidence level of 0.95 was assumed. Since, for several indices and funds the serial independence and homoscedasticity assumptions were rejected, a robust HAC covariance estimator of V was employed. Number of Monte Carlo draws of the random vector $\hat{G}_h$ was set to 1000. The results of the interval estimation are given in Table 5. The parameters which do not differ from 0 at the significance level 0.01 are bolded[18]. The intervals for other parameters do not cover 0 with probability 0.95, so the coefficients do differ from 0 at the significance level 0.05. The funds were arranged according to the magnitude of passiveness (value of determination coefficients).

In the next stage of the research procedure the models were respectified. Only those variables were left that in the previous stage significantly differ from 0. Re-estimation was again preceded by verification of models' assumptions. Similar to the earlier stage the robust HAC estimator of the covariance matrix V was utilized. The number of Monte Carlo draws was again set at 1000. The results of the second estimation are given in Table 6.

[18] The p-values are higher than 0.01 and the corresponding confidence intervals cover 0.

Table 5

Andrews confidence intervals for the Sharpe style analysis models

Dependent variable	Independent variable	Value of estimator $\hat{\beta}_{SMNK_j}$	p-value	z_l	z_u	lower bound	upper bound	R^2
1	2	3	4	5	6	7	8	9
Pioneer	WIG-banks	0.279	0.00	-0.732	0.668	0.203	0.362	0.870
	WIG-construction	0.049	0.06	-0.525	0.553	-0.014	0.109	
	WIG-informatics	0.065	0.00	-0.409	0.385	0.022	0.112	
	WIG-food industry	0.049	0.06	-0.456	0.483	-0.006	0.100	
	WIG-telecom	0.056	0.01	-0.463	0.444	0.006	0.109	
	PS	0.234	0.00	-1.511	1.135	0.106	0.405	
	DS	0.177	0.00	-1.698	1.013	0.062	0.369	
	OK	0.091	0.21	0.000	1.754	-0.108	0.091	
ING	WIG-banks	0.235	0.00	-0.818	0.529	0.175	0.327	0.868
	WIG-construction	0.066	0.01	-0.461	0.524	0.006	0.118	
	WIG-informatics	0.076	0.00	-0.408	0.396	0.031	0.122	
	WIG-food industry	0.075	0.00	-0.494	0.417	0.028	0.131	
	WIG-telecom	0.038	0.11	0.000	0.519	-0.021	0.038	
	PS	0.264	0.00	-1.519	0.277	0.233	0.436	
	DS	0.105	0.05	0.000	1.070	-0.016	0.105	
	OK	0.142	0.05	0.000	1.410	-0.018	0.142	
UniKorona	WIG-banks	0.206	0.00	-0.652	0.594	0.139	0.280	0.832
	WIG-construction	0.079	0.00	-0.523	0.475	0.026	0.139	
	WIG-informatics	0.066	0.00	-0.353	0.376	0.024	0.106	
	WIG-food industry	0.088	0.01	-0.619	0.653	0.014	0.158	
	WIG-telecom	0.052	0.06	-0.601	0.558	-0.011	0.120	
	PS	0.317	0.00	-2.009	1.672	0.128	0.545	
	DS	0.191	0.01	-1.693	1.360	0.037	0.383	
	OK	0.000	0.51	0.000	1.733	-0.196	0.000	
Arka	WIG-banks	0.286	0.00	-1.025	0.823	0.193	0.402	0.813
	WIG-construction	0.107	0.00	-0.581	0.508	0.049	0.173	
	WIG-informatics	0.062	0.02	-0.578	0.527	0.002	0.127	
	WIG-food industry	0.042	0.08	0.000	0.583	-0.024	0.042	
	WIG-telecom	0.080	0.05	-0.606	0.723	-0.002	0.149	
	PS	0.344	0.00	-1.937	0.347	0.301	0.563	
	DS	0.080	0.12	0.000	1.262	-0.063	0.080	
	OK	0.010	0.75	0.000	1.402	-0.159	0.000	
Legg Mason	WIG-banks	0.185	0.00	-0.794	0.497	0.129	0.275	0.766
	WIG-construction	0.042	0.05	0.000	0.383	-0.001	0.042	
	WIG-informatics	0.033	0.07	0.000	0.397	-0.012	0.033	
	WIG-food industry	0.053	0.01	-0.494	0.365	0.012	0.109	
	WIG-telecom	0.082	0.00	-0.521	0.480	0.028	0.141	
	PS	0.312	0.00	-1.599	1.074	0.190	0.493	
	DS	0.247	0.00	-1.505	1.217	0.109	0.417	
	OK	0.045	0.36	0.000	1.889	-0.169	0.045	
Skarbiec	WIG-banks	0.139	0.00	-0.796	0.741	0.055	0.229	0.767
	WIG-construction	0.051	0.02	-0.473	0.432	0.002	0.105	
	WIG-informatics	0.095	0.00	-0.433	0.383	0.052	0.144	
	WIG-food industry	0.057	0.05	-0.541	0.575	-0.008	0.119	
	WIG-telecom	0.073	0.00	-0.445	0.473	0.020	0.124	
	PS	0.248	0.01	-2.143	1.675	0.059	0.491	
	DS	0.226	0.00	-1.834	1.393	0.068	0.434	
	OK	0.109	0.17	0.000	2.092	-0.128	0.109	

cont. Table 5

1	2	3	4	5	6	7	8	9
DWS	WIG-banks	0 204	0 00	-0 927	0 653	0 130	0 309	0.750
	WIG-cosntruction	0 074	0 02	-0 553	0 575	0 009	0 137	
	WIG-informatics	0.056	0.06	-0 597	0 541	-0.006	0.123	
	WIG-food industry	0 070	0 00	-0 506	0 488	0 015	0 127	
	WIG-telecom	0.038	0.12	0.000	0 566	-0.026	0.038	
	PS	0 353	0 00	-1 836	0.460	0 301	0 561	
	DS	0 157	0.02	0 000	1 166	0 025	0 157	
	OK	0.048	0.26	0.000	1 594	-0.133	0.048	
PKO	WIG-banks	0 135	0 00	-0 732	0 668	0 086	0 198	0 870
	WIG-cosntruction	0 056	0 02	-0 525	0 553	0 005	0 107	
	WIG-informatics	0.023	0.16	-0 409	0,385	-0.021	0.023	
	WIG-food industry	0 071	0 00	-0 456	0 483	0 035	0.110	
	WIG-telecom	0.035	0.06	-0 463	0 444	0.000	0.071	
	PS	0 320	0 00	-1 511	1 135	0 171	0 511	
	DS	0 230	0 00	-1 698	1 013	0 108	0 394	
	OK	0.129	0.10	0 000	1 754	-0.064	0.129	
Aviva	WIG-banks	0 100	0 00	-0 606	0 539	0 039	0.169	0 732
	WIG-cosntruction	0 059	0 00	-0.375	0.345	0.020	0.102	
	WIG-informatics	0.027	0.16	0.000	0 416	-0.020	0.027	
	WIG-food industry	0 074	0 00	-0 401	0 359	0 034	0 120	
	WIG-telecom	0 055	0 01	-0 436	0 390	0 011	0 104	
	PS	0 331	0 00	-1.833	1 426	0.169	0 538	
	DS	0.304	0 00	-1.610	1.250	0.163	0 487	
	OK	0.050	0.33	0 000	1 660	-0.138	0.050	
KBC	WIG-banks	0 243	0 00	-1.496	1.209	0.106	0.412	0.723
	WIG-cosntruction	0.113	0 01	-0.854	0 754	0 027	0 209	
	WIG-informatics	0.022	0.34	0.000	0 978	-0.089	0.022	
	WIG-food industry	0 101	0 01	-0 849	0 734	0.018	0.197	
	WIG-telecom	0 129	0 00	-0 763	0.672	0 053	0.215	
	PS	0.344	0.00	-1 908	0 679	0.267	0.560	
	DS	0.000	1.00	0 000	1 544	-0.175	0.000	
	OK	0.048	0.27	0 000	1.949	-0.172	0.048	
Novo	WIG-banks	0 172	0 00	-1.066	0.588	0 106	0 293	0 664
	WIG-cosntruction	0.079	0.00	-0 546	0 492	0 023	0 141	
	WIG-informatics	0.055	0 01	-0.513	0 415	0 008	0 113	
	WIG-food industry	0 055	0.15	0 000	0 832	-0.039	0.055	
	WIG-telecom	0.020	0 36	0 000	0.812	-0.072	0.020	
	PS	0 392	0 00	-2.396	0 548	0 330	0.663	
	DS	0.138	0.03	0.000	1.274	-0.006	0.138	
	OK	0.089	0.18	0.000	1.780	-0.112	0.089	
Millenium	WIG-banks	0.129	0 00	-1.134	0 780	0.040	0.257	0 653
	WIG-cosntruction	0.059	0.04	0.000	0 577	-0.006	0.059	
	WIG-informatics	0.051	0.03	0.000	0 479	-0.003	0.051	
	WIG-food industry	0.076	0 00	-0 586	0 397	0.031	0 142	
	WIG-telecom	0.056	0.04	-0.555	0.557	-0.007	0.118	
	PS	0 274	0 00	-1 876	1 363	0 120	0 487	
	DS	0 253	0.00	-1 512	1 246	0 112	0.424	
	OK	0.103	0.22	0 000	2 143	-0.140	0.103	

cont. Table 5

1	2	3	4	5	6	7	8	9
BPH	WIG-banks	0.159	0.00	-1.370	0.785	0.070	0.314	0.596
	WIG-construction	0.088	0.00	-0.624	0.583	0.022	0.159	
	WIG-informatics	0.000	0.86	0.000	0.711	-0.080	0.000	
	WIG-food industry	0.091	0.00	-0.566	0.501	0.034	0.155	
	WIG-telecom	0.008	0.39	0.000	0.660	-0.067	0.008	
	PS	0.256	0.00	-1.998	1.314	0.107	0.482	
	DS	0.322	0.00	-2.175	1.795	0.118	0.568	
	OK	0.076	0.29	0.000	2.492	-0.206	0.076	

Table 6

Andrews confidence intervals after the models' respecification

Dependent variable	Independent variable	Value of estimator $\hat{\beta}_{SMNK_i}$	*p*-value	z_d	z_u	lower bound	upper bound	R^2
1	2	3	4	5	6	7	8	9
Pioneer	WIG-banks	0.332	0.00	-0.622	0.642	0.259	0.402	0.854
	WIG-informatics	0.099	0.00	-0.395	0.396	0.055	0.144	
	WIG-telecom	0.048	0.04	-0.476	0.493	-0.007	0.102	
	PS	0.289	0.00	-1.372	1.557	0.112	0.444	
	DS	0.232	0.00	-1.568	1.51	0.061	0.409	
ING	WIG-banks	0.255	0.00	-0.495	0.549	0.218	0.336	0.853
	WIG-construction	0.117	0.02	-0.44	0.466	0.001	0.104	
	WIG-informatics	0.106	0.00	-0.513	0.479	0.024	0.137	
	WIG-food industry	0.131	0.00	-0.487	0.502	0.027	0.139	
	PS	0.391	0.00	-0.481	0.465	0.451	0.558	
UniKorona	WIG-banks	0.238	0.00	-0.552	0.501	0.181	0.301	0.828
	WIG-construction	0.074	0.01	-0.468	0.49	0.018	0.127	
	WIG-informatics	0.077	0.00	-0.418	0.441	0.027	0.124	
	WIG-food industry	0.087	0.00	-0.524	0.551	0.025	0.146	
	PS	0.366	0.00	-1.429	1.298	0.219	0.528	
	DS	0.158	0.02	-1.272	1.325	0.008	0.301	
Arka	WIG-banks	0.347	0.00	-0.644	0.654	0.273	0.422	0.806
	WIG-construction	0.112	0.00	-0.427	0.409	0.066	0.161	
	WIG-informatics	0.083	0.01	-0.592	0.582	0.017	0.15	
	PS	0.458	0.00	-0.484	0.469	0.405	0.512	
Legg Mason	WIG-banks	0.226	0.00	-0.626	0.572	0.161	0.297	0.759
	WIG-food industry	0.085	0.00	-0.443	0.457	0.033	0.135	
	WIG-telecom	0.086	0.00	-0.482	0.476	0.032	0.140	
	PS	0.355	0.00	-1.175	1.182	0.221	0.488	
	DS	0.249	0.00	-1.274	1.331	0.098	0.393	
Skarbiec	WIG-banks	0.154	0.00	-0.793	0.805	0.063	0.244	0.758
	WIG-construction	0.071	0.00	-0.372	0.39	0.027	0.113	
	WIG-informatics	0.104	0.00	-0.475	0.5	0.047	0.158	
	WIG-telecom	0.074	0.00	-0.475	0.466	0.021	0.128	
	PS	0.287	0.00	-1.812	1.852	0.078	0.493	
	DS	0.310	0.00	-1.747	1.756	0.111	0.507	
DWS	WIG-banks	0.257	0.00	-0.685	0.726	0.175	0.335	0.757
	WIG-construction	0.085	0.00	-0.445	0.449	0.034	0.135	
	WIG-food industry	0.081	0.00	-0.506	0.51	0.023	0.138	
	PS	0.431	0.00	-1.313	1.304	0.284	0.580	
	DS	0.146	0.03	-1.361	1.371	-0.009	0.300	

cont. Table 6

1	2	3	4	5	6	7	8	9
PKO	WIG-banks	0.176	0.00	-0.433	0.423	0.128	0.225	0.749
	WIG-construction	0.058	0.01	-0.449	0.446	0.008	0.109	
	WIG-food industry	0.075	0.00	-0.345	0.352	0.035	0.114	
	PS	0.416	0.00	-1.616	1.586	0.236	0.599	
	DS	0.275	0.00	-1.597	1.608	0.092	0.455	
Avva	WIG-banks	0.110	0.00	-0.562	0.544	0.049	0.174	0.733
	WIG-construction	0.066	0.00	-0.342	0.355	0.025	0.104	
	WIG-food industry	0.079	0.00	-0.458	0.385	0.035	0.130	
	WIG-telecom	0.060	0.01	-0.442	0.434	0.011	0.110	
	PS	0.355	0.00	-1.468	1.527	0.182	0.521	
	DS	0.330	0.00	-1.382	1.53	0.157	0.486	
KBC	WIG-banks	0.255	0.00	-1.211	1.192	0.120	0.392	0.732
	WIG-construction	0.117	0.00	-0.697	0.738	0.033	0.196	
	WIG-food industry	0.106	0.01	-0.782	0.712	0.025	0.195	
	WIG-telecom	0.131	0.00	-0.679	0.742	0.047	0.208	
	PS	0.391	0.00	-1.139	1.287	0.246	0.520	
Novo	WIG-banks	0.219	0.00	-0.726	0.743	0.135	0.302	0.647
	WIG-construction	0.093	0.00	-0.498	0.535	0.032	0.149	
	WIG-informatics	0.063	0.03	-0.564	0.572	-0.002	0.127	
	PS	0.625	0.00	-0.711	0.72	0.543	0.706	
Millenium	WIG-banks	0.230	0.00	-0.688	0.706	0.150	0.308	0.618
	WIG-food industry	0.120	0.00	-0.477	0.447	0.070	0.174	
	PS	0.412	0.00	-1.541	1.595	0.232	0.587	
	DS	0.237	0.00	-1.557	1.488	0.069	0.413	
BPH	WIG-banks	0.168	0.00	-1.008	0.996	0.056	0.283	0.611
	WIG-construction	0.087	0.00	-0.552	0.534	0.026	0.149	
	WIG-food industry	0.091	0.00	-0.529	0.495	0.034	0.150	
	PS	0.295	0.00	-1.621	1.582	0.116	0.479	
	DS	0.359	0.00	-1.885	1.889	0.145	0.573	

Every variable in each model is different from 0 at the significance level of 0.05. Parameter values reflect average impact of investments in different asset classes on rates of return from the fund's participation units. These results were further used to classify the funds. Figure 1 contains a hierarchical tree depicting the results of hierarchical classification of the funds according to the Sharpe style weights after respecification of the models. Euclidean distance was used to construct the distance matrix. As a result of the procedure two main groups of funds emerge. The first one consists of Arka BZ WBK Zrównoważony FIO, Novo FIO Subfundusz Novo Zrównoważonego Wzrostu, ING FIO Zrównoważony oraz KBC Beta SFIO. For these funds investments in bonds are represented only by one type of bond – 5-year fixed interest (PS). The other funds form the second group. It is characterized by the fact that investment performance of the funds from this group is highly influenced by rates of return from the WIG-banks subindex. Pioneer is the fund within the group that differs the most from other funds. The similar results were obtained using unweighted pair-group average method and other non-Euclidean distance measures.

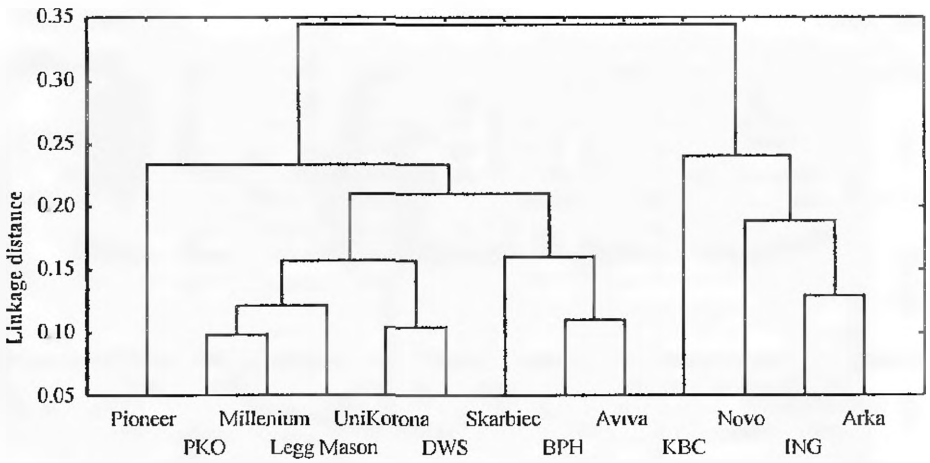


Figure 1. Classification of the balanced mutual funds according to estimated style weights

Conclusion

Investment decisions taken by mutual funds are highly complicated. Therefore performance measurement process should take into account as many aspects of investment policy as possible. Attribute method based on the multifactor Sharpe model can be seen as an important supplement for traditional single factor efficiency measures.

The point estimates of the Sharpe style weights mirror structure of mutual balanced fund portfolios according to stock and bond shares. The funds management styles are usually very different. The main virtue of the presented Andrews method in comparison with other non-classic methods (for example Bayesian method) is that it could be used when the true values of parameters lie on the boundary of a parameter space as well as when they are inside the space. It must be emphasized however that the conclusion presented in the article is valid only for a certain set of market indices used for the Sharpe style analysis. The correct choice of these indices, treated as exogenous variables, is very important for the discussed method. More precise conclusions could be drawn from dynamic analysis of Sharpe style models.

KLASYFIKACJA POLSKICH OTWARTYCH FUNDUSZY INWESTYCYJNYCH ZRÓWNOWAŻONYCH ZE WZGLĘDU NA STYL ZARZĄDZANIA W OPARCIU O ESTYMACJĘ ANDREWSA

Streszczenie

Zarządzający funduszami inwestycyjnymi zrównoważonymi inwestują w aktywa o różnych stopniach ryzyka. Efektywność inwestycji w poszczególnych sektorach jest niejednorodna, zatem trudno jest odróżnić zyski funduszy, wynikające z wyboru klas aktywów między sektorami od zysków, wynikających z wyboru konkretnych aktywów w ramach danego sektora. Istotnym elementem oceny wyników zarządzania portfelem jest przypisywanie wyników (*performance atribution*), które może być realizowane za pomocą statystycznej metody analizy stylu. Podejście to zostało zapoczątkowane przez Williama Sharpe'a w 1992 roku. Założenie nieumiejętności współczynników modeluy zaproponowanego przez Sharpe'a, powoduje, że rozkład wektora estymatorów MNK parametrów modelu nie jest znany. Znajomość rozkładu estymatorów parametrów modelu ma kluczowe znaczenie dla estymacji przedziałowej oraz weryfikacji hipotezy o istotności parametrów modelu. Jest ona więc istotna dla poprawnego formułowania wniosków dotyczących wpływu analizowanych czynników na osiągnięte przez fundusz stopy zwrotu.

Celem pracy jest aplikacja metody Andrewsa w estymacji przedziałowej wag modeli analizy stylu zarządzania otwartymi funduszami inwestycyjnymi (OFI) zrównoważonymi, działającymi na polskim rynku finansowym oraz klasyfikacja tych funduszy w oparciu o uzyskane wyniki.

Ewa Witek

THE USE OF FINITE MIXTURE MODELS IN THE CLASSIFICATION OF THE EU MEMBER STATES

Introduction

The introduction of the euro ten years ago was a major step in European integration. It has also been one of its major successes: around 329 million EU citizens use the euro as their currency and enjoy its benefits. When the euro was launched on 1 January 1999, it became the new official currency of 11 Member States, replacing the old national currencies in two stages. First, it was introduced as a virtual currency for cash less transactions and accounting purposes, while the old currencies continued to be used for cash payments and considered as sub-units of the euro. Then it was launched in a physical form, as banknotes and coins, on 1 January 2002.

The euro is not the currency of all the EU Member States. Two countries (Denmark and the United Kingdom) agreed on an opt-out clause in the Treaty exempting them from participation, while the remainder (many of the newest EU members plus Sweden) have yet to meet the conditions for adopting the single currency. Nowadays 16 out of the 27 EU countries have adopted the euro and 11 are waiting to fulfill the convergence criteria. The euro zone entry date of the non-euro countries may be shifted due to the uncertain global situation. The recent world economic crisis has made the currency swap more difficult for many countries. One of the biggest problems is estimating price developments because of the fluctuations in the exchange rate. The complaints are heard first of all in many of the newest EU countries in Eastern Europe, of which only two – Slovakia and Slovenia – are members of the euro zone.

An accession country that plans to join the Monetary Union must align many aspects of its society – social, economic and political – with those of the EU Member States. Adopting the euro also demands extensive preparations; in particular it requires economic and legal convergence. The economic convergence criteria are designed to ensure that a member state's economy is sufficiently prepared for adoption of the single currency and can integrate smoothly into the monetary regime of the euro zone. Legal convergence requires that national legislation, in particular the national central bank and monetary issues, is compatible with the Treaty.

The aim of this paper is to determine the effect of different covariates on the number of years in the euro zone. The countries will be characterized by the selected economic indicators and convergence criteria. Poisson regression has been recognized as an important tool for analyzing this kind of data. However, observed response variable analyzed under such models often exhibit overdispersion i.e., the variance of the observation is greater than its mean (this may be reflected in over-large residual deviance and adjusted residuals which have a variance > 1). By using finite mixture models to explain overdispersion we adopt the viewpoint that there are a finite number of clusters which may be characterized by different values of generalized linear models coefficients. Finite mixture models are given by a convex combination of u different components. For each component it is assumed that it follows a parametric distribution or is given by a generalized linear model (GLM). Mixture models have shown promise in a number of practical applications, including tissue segmentation, character recognition, minefield and seismic fault detection and classification of astronomical data. The article presents an application of the mixture models in economic analysis.

Exactly ten years ago the euro was introduced, first as an electronic means of payment (the banknotes and coins came into circulation in 2002). We would like to detect inhomogeneities of the euro and non-euro countries characterized by the selected economic indicators and convergence criteria.

1. Mixture model

Finite mixture models are a popular technique for modeling unobserved heterogeneity or approximating general distribution functions. They are used in a lot of different areas such as astronomy, biology, economic, marketing or medi-

cine. An overview of mixture models is given in Titterington et al.[1] or McLachlan and Peel[2]. The mixture is assumed to consist of s components where each component follows a parametric distribution. Each component has a weight assigned which indicates the a priori probability for an observation to come from this component and the mixture distribution is given by the weighted sum over the u components. The mixture model is given by:

$$f(y_i|\mathbf{x}_i^j, \mathbf{x}_i^{\varpi}, \Theta) = \sum_{s=1}^{u} \tau_s(\mathbf{x}_i^{\varpi}, \alpha) f_s(y_i|\mathbf{x}_i^j, \Theta_s), \tag{1}$$

where:

f_s – density function of component s,

y_i – observed value of Y,

$\mathbf{x}_i = (\mathbf{x}_i^j, \mathbf{x}_i^{\varpi})$ – a covariate vector in which $\mathbf{x}_i^j$ and $\mathbf{x}_i^{\varpi}$ are m_1 and m_2 – covariate vectors corresponding to the regression part and mixing part of the model respectively,

Θ_s – the component specific parameter vector for the density function f_s,

Θ – the vector of all parameters for the mixture density function, $\Theta = (\tau_s, \alpha_s, \Theta_s)$,

τ_s – the prior probability of component s; $(\tau_s \geq 0 \wedge \sum_{s=1}^{u} \tau_s = 1), \Theta_s \neq \Theta_l \forall s \neq l$.

We assume that the component specific densities are from the same parametric families. If f_s is from the exponential family of distributions and for each component a generalized linear model is fitted (GLM's, McCullagh and Nelder[3]) these models are also called GLIMMIX (Wedel and DeSarbo[4]).

The posterior probability that observation $(\mathbf{x}_i, y_i)$ belongs to class r is given by:

$$p(r|\mathbf{x}_i, y_i, \Theta) = \frac{\tau_r(\mathbf{x}_i^{\varpi}, \alpha) f(y_i|\mathbf{x}_i^j, \Theta_r)}{\sum_{s=1}^{u} \tau_s(\mathbf{x}_i^{\varpi}, \alpha) f(y_i|\mathbf{x}_i^j, \Theta_r)}. \tag{2}$$

[1] D. M. Titterington, A. F. Smith. U. E. Makov: *Statistical Analysis of Finite Mixture Distribution*. John Wiley & Sons, San Diego 1985.

[2] G. J. McLachlan, D. Peel: *Finite Mixture Models*. Wiley, New York 2000, p. 81-116.

[3] P. McCullagh, J. Nelder: *Generalized linear models*. Chapman and Hall, New York 1989.

[4] M. Wedel, W. S. DeSarbo: *A Mixture Likelihood Approach for Generalized Linear Models*. "Journal of Classification" 1995, No. 12, p. 21-55.

2. Parameter estimation and model selection

The parameters of the model are usually estimated by maximum likelihood using the Expectation-Maximization (EM) algorithm[5]. The log-likelihood of a sample of n observations $\{(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_n, y_n)\}$ is given by:

$$\log L = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \Theta) = \sum_{i=1}^{n} \log(\sum_{s=1}^{u} \tau_s f_s(y_i|\mathbf{x}_i, \Theta_s)), \qquad (3)$$

and can usually not be maximized directly. Each EM iteration consists of two steps – an E-step and an M-step:

– E-step – estimation of the posterior class probabilities for each observation in the k'th iteration:

$$\hat{p}_{is} = p(s|\mathbf{x}_i, y_i, \Theta^{(k)}), \qquad (4)$$

using equation (2) and derivation of the prior class probabilities:

$$\hat{\tau}_s = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{is} \qquad (5)$$

– M-step – given the estimates for the a posteriori probabilities, obtain new estimates $\theta^{(k+1)}$ (in the $k + 1$'th iteration) of the parameters by maximizing

$$Q(\theta^{(k+1)} \mid \theta^{(k)}) = Q_1(\theta_s^{(k+1)} \mid \theta^{(k)}) + Q_2(\alpha^{(k+1)} \mid \theta^{(k)}) \qquad (6)$$

where:

$$Q_1(\theta_s^{(k+1)} \mid \theta^{(k)}) = \sum_{i=1}^{n}\sum_{s=1}^{u} \hat{p}_{is} \log(f_s(y_i \mid \mathbf{x}_i^J, \Theta_s^{(k+1)})), \qquad (7)$$

$$Q_2(\alpha^{(k+1)} \mid \theta^{(k)} = \sum_{i=1}^{n}\sum_{s=1}^{u} \hat{p}_{is} \log(\hat{\tau}_s(\mathbf{x}_i^{\varpi}, \alpha^{(k+1)}). \qquad (8)$$

[5] A. P. Dempster, N. M. Laird, D. B. Rubin: *Maxiumum Likelihood for Incomplete Data via the EM Algorithm (with discussion)*. "Journal of the Royal Statistical Society" 1977, Series B, No. 39, p. 1-38.

We consider in the empirical part of the article:

$$f_s(y_i \mid \mathbf{x}_i^J, \beta_s) = Poisson(y_i|\lambda_{is}) = \frac{\lambda_{is}^{y_i}}{y_i!}\exp^{-\lambda_{is}}, \tag{9}$$

$$\lambda_{is} \equiv \lambda_s(\mathbf{x}_i^J, \beta_s) = \exp(\beta'_s \mathbf{x}_i^J). \tag{10}$$

The E and M steps are repeated until the likelihood improvement falls under a pre-specified threshold or a maximum number of iterations is reached (see Wang[6] for more details).

In order to select the optimal clustering model several measures have been proposed (see i.e. McLachlan and Peel[7]). Three information criteria are available in `flexmix` package of R: BIC (*Bayesian Information Criterion*), AIC (*Akaike Information Criterion*) and ICL (*Integrated Completed Likelihood*). The performance of some of these criteria were compared by Biernacki et al.[8] In general, BIC was found to be consistent under correct specification of the component densities (Kass and Raftery[9], Keribin[10]) and has given good results in a range of applications (i.e. Fraley and Raftery[11], Stanford and Raftery[12]). For this reason BIC criterion will be used in further analysis. The BIC criterion is defined:

$$BIC_s = -2\log p(\mathbf{x}_i, y_i|\hat{\Theta}_s, M_s) + v_s \log(n), \tag{11}$$

where:

$\log p(\mathbf{x}_i, y_i|\hat{\Theta}_s, M_s)$ – is the maximized loglikelihood for the model M_s, v_s is the number of parameters to be estimated in that model, n is the number of observations in the data.

---

[6] P. Wang: *Mixed Regression Models for Discrete Data*. PhD thesis, University of British Columbia, Vancouver 1994.

[7] G. J. McLachlan, D. Peel: Op. cit.

[8] C. Biernacki, G. Celeux, G. Govaert: *Choosing Models in Model-based Clustering and Discriminant Analysis*. "Journal of Statistical Computation and Simulation" 1999, No. 64, p. 49-71.

[9] R. E. Kass, A. E. Raftery: *Bayes Factors*. "Journal of the American Statistical Association" 1995, No. 90 , p. 928-934.

[10] C. Keribin: *Consistent Estimation of the Order of Mixture Models*. "Sankhya Indian Journal Statistics" 2000, No. 62, p 49-66.

[11] C. Fraley, A. E. Raftery. *Model-based Clustering, Discriminant Analysis, and Density Estimation*. "Journal of the American Statistical Association" 2002, No. 97, p. 611-631.

[12] D. Stanford, A. E. Raftery: *Principal Curve Clustering with Noise*. "IEEE Transactions on Pattern Analysis and Machine Intelligence" 2000, No. 22, p. 601-609.

The first term in BIC criterion measures the goodness-of-fit, whereas the second term penalizes model complexity. One selects the model that minimizes BIC value.

3. Example

All computations and graphics in this paper have been done in `flexmix`[13] and `clusterSim`[14] packages of R. The data was sourced from AMECO database[15].

The following variables (different economic variables i.e., two of the convergence criteria) in the year 2008 were used in the analysis: y – period of time in the euro zone, x_1 – average share of imports and exports of goods in world trade including intra EU trade, x_2 – unemployment rate, x_3 – private final consumption expenditure at current prices per head of population, x_4 – annual rates of inflation, x_5 – gross domestic product per head of population, x_6 – gross fixed capital formation at current prices: total economy, x_7 – real long-term interest rates, x_8 – balance on current transactions with the rest of the world[16].

At the very beginning of our analysis we checked variable's evidence for being useful for clustering using the HINoV method of variable selection (Carmone, Kara and Maxwell[17]). The HINoV method was applied in `clusterSim` packgage of R, the results are presented in Figure 1.

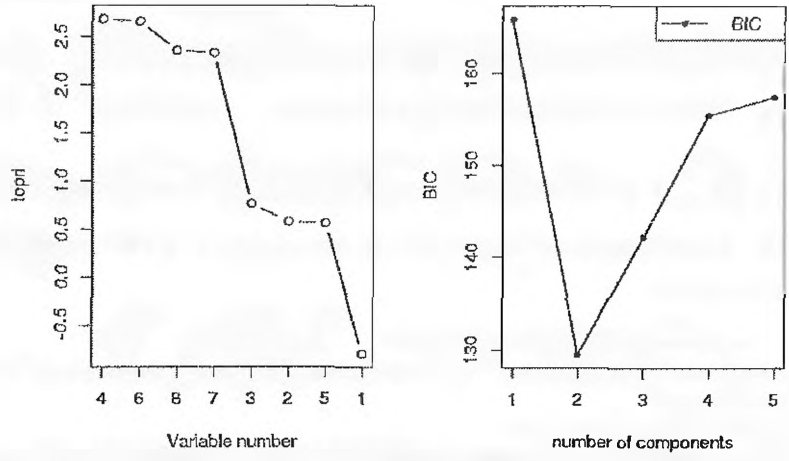


Figure 1. Left: The results of HINoV method of variable selecion; Right: The results of BIC criterion}

[13] B. Grün, F. Leisch, `flexmix`, http://www.R-project.org, 2004.

[14] M. Walesiak, A. Dudek, `clusterSim`, http://www.R-project.org, 2008.

[15] http://ec.europa.eu/economy_finance/db_indicators/db_indicators8646_en.htm

[16] We consider only variables corresponding to the regression part of the model.

[17] F. J. Carmone, A. Kara, S. Maxwell: HINoV: *A New Method to Improve Market Segment Definition by Identifying Noisy Variables*. "Journal of Marketing Research" 1999, No. 36, p. 501-509.

The variables selected by the variable selection method were (in order of selection) x_4, x_6, x_8, x_7.

The Poisson regression model has been widely used for analyzing count data in which each observation consist of discrete response variable and a vector of covariates. For instance, the response variable may represent the number of failures of piece of equipment, the number of purchases of particular commodity, the number of reported misconduct incidents.

We assume that the response variable – the number of years in the euro zone (for each country) follows a Poisson distribution either. As a special case of the generalized linear model, the Poisson regression is applied. We used as covariates x_4, x_6, x_8, x_7 (the results of HINoV variable selection method). In practise, however, the model sometimes fits poorly, suggesting the need for alternative models. In this case, it is not uncommon that observed data are overdispersed i.e., the variance of the observation is greater than its mean. Score tests for three alternative specifications of overdispersion were developed by Dean[18]. Dean's test statistics yields values $P(a) = 77.74$, $P(b) = 77.85$ and $P(c) = 80.52$ respectively. Because all three statistics have standard null distributions, they provide strong evidence that the data is overdispersed and that a more complex model, such as a mixture of generalized models, is needed. Then, the mixture of the generalized linear models was applied. The optimal number of the mixture components was chosen using BIC criterion (see e.g., Blum et al.[19], Murtagh and Starck[20], Murtagh and Starck[21], Wang et al.[22], Wang et al. [23]). Figure 1 shows that the optimal number of components is 2.

We estimated parameters of two components using EM algorithm. In further analysis we ran the test for significance of regression coefficients. For x_8 the coefficients of both components were not significantly different from 0. The significant parameter estimates of both components, the estimated prior probabilities τ_s and n_s – the number of observations assigned to the corresponding clusters are presented in Table 1.

---

[18] D. B. Dean: *Testing for Overdispersion in Poisson and Binomial Regression Models*. "Journal of the American Statistical Association" 1992, No. 87, p. 451-457.

[19] F. S. Blum, Y. Zhang, B. M. Sadler, R. J. Kozick: *On the Approximation of Correlated Non-Gaussian Noise PDFs Using Gaussian Mixture Models*. Conference on the Applications of Heavy Tailed Distributions in Economics. "Engineering and Statistics" 1999, American University DC.

[20] F. Murtagh, J. L. Starck: *Bayes Factors for Edge Detection from Wavelet Product Spaces*. "Optical Engineering" 2003, No. 4, p. 1375-1382.

[21] F. Murtagh, J. L. Starck: *Quantization from Bayes Factors with Application to Multilevel Yhresholding*. "Pattern Recognition Letters" 2003, No. 24.

[22] P. Wang, M. L. Puterman, I. Cockburn, N. Le: *Mixed Poisson Regression Models with Covariate Dependent Rates*. "Biometrics" 1996, No. 52, p. 381-400.

[23] P. Wang, I. Cockburn, M. L. Puterman: *Analysis of Patent Data-A Mixed-Poisson-Regression-Model Approach*. "Journal of Business & Economic Statistics" 1998, No. 16, p. 27-41.

Table 1

The significant parameter estimates of two components in the mixture

Cluster	τ_s	n_s	GLM model
I	0.503	10	$\lambda_1 = \exp(30.8692 - 5.101x_4 + 0.00046\,x_6 - 5.6297x_7)$
II	0.497	17	$\lambda_2 = \exp(1.5519 - 0.15381x_4 + 0.00052\,x_6 - 0.5059x_7)$

We refer to the two components of the model as "State 1" and "State 2", respectively. For instance, $\hat{\beta}_{14} = 5.10$ is the estimated effect the inflation has on the countries being in underlying state one, while $\hat{\beta}_{24} = 0.15$ is the negative effect when the country is in state two. Compared to the second mixture component, the coefficient of x_4 is much larger in cluster one. Apparently, inflation and annual interest rates are major determinants of belonging to this cluster. In the second class the effect of x_7 (annual interest rates) is positive, but considerably smaller than in the first one. The coefficients for x_6 (gross fixed capital formation at current prices) in both components have a positive, but not a major effect on the years in the euro zone.

Conclusions

We have shown the use of the mixture models in the classification of EU countries. Mixture models analysis yields two groups of countries. The second class comprises countries such as Belgium, Ireland, Greece, France, Italy, Luxemburg, Portugal, Slovenia, Sweden and Austria. With the exception of Sweden these are all countries of the euro zone, characterized by average annual inflation rates of 4.1%, gross fixed capital formation at current prices level of 108.95 billion Euro and the real long-term interest rates at the average level of 1.3%. As far as the first class is concerned the average values are: $x_4 = 6.0\%$, $x_6 = 94.53$ billion Euro, $x_7 = 0.44\%$. This class comprises the rest of EU countries, not belonging to the euro zone yet, with the exception of 4 old euro area countries – Germany, Spain, Finland, Netherlands – and 2 new euro area countries – Cyprus and Malta. We have analyzed the relationship between the number of years in the euro zone and the current economic situation in these two groups of countries. We have considered different economic indicators and convergence criteria. The mixture model analysis has confirmed that the largest coefficients and a major effect on the belonging to the second mixture component (mostly euro zone countries), convergence criteria variables such as annual inflation rates and real long-term interest rates have.

In the future the analysis should be extended to include the outliers i.e. opt-out countries such as Denmark and the UK. The other idea is to use all of the convergence criteria indicators and discuss relationships between the results of the analysis and UE requirements.

ZASTOSOWANIE MODELI MIESZANEK DO KLASYFIKACJI KRAJÓW UNII EUROPEJSKIEJ

Streszczenie

Modele mieszanek rozkładów różnego typu są stosowane bardzo często wtedy, gdy zbiór obserwacji jest wysoce niejednorodny. Pozwalają one po pierwsze: na określenie optymalnej liczby klas (podzbiorów), po drugie na: zidentyfikowanie wielowymiarowych rozkładów zmiennych, charakteryzujących obiekty należące do poszczególnych klas.

Celem artykułu jest weryfikacja i ocena ważności kryteriów wejścia poszczególnych państw do strefy euro. Na podstawie danych dotyczących 27 krajów Europy zbudowano model mieszanek rozkładów Poissona, wykorzystując pakiet `flexmix` w programie R. Określono optymalną liczbę klas, opierając się na kryteriach informacyjnych (np. BIC) oraz zinterpretowano parametry oszacowanego modelu.

Janusz L. Wywiał

TEST-ESTIMATOR AND DOUBLE TEST

Introduction

The two problems are considered. The first one is estimating a mean value on the basis of two samples. Firstly the hypothesis on equality of means in the both samples is tested. If the hypothesis is not rejected the mean value is estimated on the basis of the combined samples. When the hypothesis is rejected the mean value is estimated only on the basis of one sample. The parameters of such test-estimator are derived.

The next problem is testing the hypothesis H_0 on a mean value on the basis of two samples. Previously, we test the hypothesis H'_0 on equality of means in the both samples. If H'_0 is not rejected the both samples are combined to testing the hypothesis H_0. In the case when the hypothesis H'_0 is rejected the hypothesis H_0 is tested only on the basis of one sample. The significance level of such double test is considered.

Properties of preliminary test predictors (test-estimators) are considered e.g. by Bancroft[1], Bancroft and Han[2], Giles, Lieberman, Giles[3], J. Giles and D. Giles[4], Korsell[5], Sclove, Morris and Radhakrishnan[6]. Some practical applications of preliminary test estimators are considered e.g. by Asano[7], Bock, Yancey, Judge[8].

---

[1] T.A. Bancroft: *On Biases in Estimation Die to the Use of Preliminary Tests of Sigificance.* „The Annals of Mathematical Statistics" 1944, Vol. 15, No. 2, p. 190-204. Idem: *Testimating, Testipredicting and Testitesting as Aids in Using Snedecor and Cochran's „Statistical Methods"* „Biometrics" 1975, Vol. 31, No. 2, p. 319-323.

[2] T.A. Bancroft, C.P. Han: *Inference Based on Conditional Apecification: A Note and a Bibliography.* „International Statistical Review" 1977, Vol. 45, p. 117-127.

[3] D.E.A. Giles, O. Lieberman, J. Giles: *The Opitmal Size of a Preliminary Test of Linear Restrictions in a Misspecified Regression Model.* „Journal of the American Statistical Association" 1992, Vol. 87, No. 420, p. 1153-1157.

[4] J.A. Giles, D.E.A. Giles: *Pre-test Estimation and Testing in Econometrics: Recent Developments.* „Journal of Economic Surveys" 1993, Vol. 7, No. 2, p. 146-197.

[5] N. Korsell: *Statistical Properties of Preliminary Test Estimators.* Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 17. „Acta Universitatis Upsaliensis", Uppsala 2006.

1. Test-estimator

Let $X = [X_1, X_2,..., X_n]$ be a simple sample from a probability density $f_{\theta,\omega}$, where $\theta \in \Theta$, $\omega \in \Omega = \Omega_0 \cup \Omega_1$, $\Omega_0 \cap \Omega_1 = \phi$. The value (observation) of the simple sample will be denoted by $x = [x_1, x_2,..., x_n]$. The tested hypothesis and the alternative one are denoted by H_0: $\omega \in \Omega_0$ and H_1: $\omega \in \Omega_1$, respectively. Let K be a subset of the sample space X called the critical region. The test statistic is:

$$\varphi(x) = \begin{cases} 1 & if \quad x \in K \\ 0 & if \quad x \in K^c = R - K \end{cases} \tag{1}$$

The power function:

$$\beta(\omega) = P_\omega(X \in K) = \int_K f_{\theta,\omega}(x)dx = E(\varphi(X)) \tag{2}$$

The size of the test: $\alpha = \sup_{\omega \in \Omega_0}(\beta(\omega))$.

Let us suppose that our aim is the estimation of the parameter θ. The estimator $T_0 = T_0(X)$ is preferred, when $\omega \in \Omega_0$ and $T_1 = T_1(X)$, when $\omega \in \Omega_1$. We will use the estimator T_1, when the hypothesis H_0 is rejected and T_0 if the hypothesis H_0 is not rejected. So, the estimator can be written in the following way.

$$T(X) = \varphi(X)T_1(X) + (1-\varphi(X))T_0(X) \tag{3}$$

The statistic $T(X)$ can be called the *test-estimator* (see Bancroft's[9] definition of testimating). The expected value of the test-estimator is[10]:

$$E(T(X),\omega) = \int_{R^n} T(x) f_{\theta,\omega}(x)dx = \int_K T_1(x) f_{\theta,\omega}(x)dx + \int_{K^c} T_0(x) f_{\theta,\omega}(x)dx \tag{4}$$

---

[6] L. Sclove, C. Morris, R. Radhakrishnan: *Non Optimality of Preliminary Test Estimators for the Mean of a Multivariate Normal Distribution*. „The Annals of Mathematical Statistics" 1972, Vol. 43, No. 5, p. 1481-1490.

[7] C. Asano: *Estimation After Preliminary Test of Significance and Their Applications to Biometrical Researches*. „Bulletin Mathematical Statistics" 1960, Vol. 9, p. 1-24.

[8] M. Bock, T. Yancey, G. Judge. *The Statistical Consequences of Preliminary Test Estimators in Regression*. „Journal of the American Statistical Association" 1973, Vol. 87, nr 420, p. 1153-1157.

[9] T A. Bancroft: *Testimating...*, op. cit.

[10] J. L Wywiał: *Prediction of domain total using preliminary test*. „Sankhya" (in print).

Let $f_{\theta,\omega}(x \mid K) = \frac{f_{\theta,\omega}(x)}{P_\omega(X \in K)}$ for $x \in K$ and $f_{\theta,\omega}(x \mid K^c) = \frac{f_{\theta,\omega}(x)}{P_\omega(X \in K^c)}$ for $x \in K^c$.

So, the expression (4) can be rewritten in the following way:

$$E(T(X),\omega)=\beta(\omega)E(T_1(X)|K,\ \omega) + (1-\beta(\omega))E(T_0(X)|K^c,\ \omega) \tag{5}$$

where:

$$E\left(T_1(X) \mid K,\omega\right) = \int_K T_1(x) f_{\theta,\omega}(x \mid K)dx,$$

$$E\left(T_1(X) \mid K^c,\omega\right) = \int_{K^c} T_0(x) f_{\theta,\omega}(x \mid K^c)dx.$$

The variance of the test-estimator[11]:

$$D^2(T(X),\omega) = \beta(\omega)\ D^2(T_1(X)|K,\ \omega) + (1-\beta(\omega))\ D^2(T_0(X)|K^c,\ \omega)+$$

$$+\beta(\omega)(1-\beta(\omega))(E(T_1(X)|K,\ \omega) - E(T_0(X)|K^c,\ \omega))^2 \tag{6}$$

Example 1. Let $Y_i \sim N(\mu_i,\delta_i)$, $i = 1, 2$. Our purpose is estimation of the mean μ_1. The estimation is supported by testing the hypothesis H_0: $\omega = 0$ where $\omega = \mu_1-\mu_2$, against H_1: $\omega \neq 0$. If the hypothesis H_0 is rejected the expected value μ_1 is estimated by means of Y_1 otherwise when H_0 is not rejected the expected value μ_1 is estimated by means of $\frac{Y_1+Y_2}{2}$. The hypothesis H_0 is tested on the basis of the statistic $Z = \frac{Y_1 - Y_2}{\sqrt{\delta_1+\delta_2}} \sim N(\kappa,1)$, where $\kappa = \frac{\omega}{\sqrt{\delta_1+\delta_2}}$. Let $P(|Z| \geq z_\alpha|H_0) = \alpha$ where α is the significance level. So, $\varphi(y_1,y_2) = 1$ if and only if $y \in K = \left\{(y_1,y_2): y_2 \geq y_1 + z_\alpha\sqrt{\delta_1+\delta_2}\ \text{ or } y_2 \leq y_1 - z_\alpha\sqrt{\delta_1+\delta_2}\right\}$. $\varphi(y_1,y_2) = 0$ if and only if $y \in K^c = \left\{(y_1,y_2): y_1 - z_\alpha\sqrt{\delta_1+\delta_2} \leq y_2 \leq y_1 + z_\alpha\sqrt{\delta_1+\delta_2}\right\}$. So, in order to estimate the parameter μ_1 we can use the following test estimator.

$$T_1(Y) = \varphi(Y)Y_1 + (1 - \varphi(Y))(Y_1 + Y_2)/2,$$

[11] Ibidem.

where Y = [Y_1 Y_2]. When the hypothesis H_0 is true the conditional (truncated) distributions $f_0(y|K)$ and $f_1(y|K^c)$ are symmetric around the point $(y_1,y_2) = (\mu_1,\mu_1)$. So, each marginal distribution of the truncated distributions is symmetric around μ_1. Hence, $E(Y_1|K,0) = \mu_1$, $E(Y_1|K^c,0) = \mu_1$, $E(Y_2|K,0) = \mu_1$, and $E\left(\frac{Y_1+Y_2}{2}\middle|K^c,0\right) = \mu_1$. So, the expression (5) leads to the equation $E(T_1(Y),0) = \mu_1$. Hence, when the hypothesis H_0 is true the test-estimator $T(Y)$ is unbiased. Particularly, let us note that if $\delta_1 + \delta_2 = 1$ then $K = \{(y_1,y_2): y_2 \geq y_1 + z_\alpha \text{ or } y_2 \leq y_1 - z_\alpha\}$ and $K^c = \{(y_1,y_2): y_1 - z_\alpha < y_2 < y_1 + z_\alpha\}$.

Example 2. Let $\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, $i = 1, 2$, are means from the following two independent samples $X_i = [X_{i1} \ldots X_{in_i}]$ and $X_{ij} \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$. Similarly, like in the Example 1 we test hypothesis H_0: $\omega = 0$, where $\omega = \mu_1 - \mu_2$ against the alternative one H_1: $\omega \neq 0$. If the hypothesis H_0 is rejected under the significance level α the expected value μ_1 is estimated by means of $\overline{X}_1$ otherwise when H_0 is not rejected the expected value μ_1 is estimated by means of $(\overline{X}_1 + \overline{X}_1)/2$. When in the Example 1 we assume that $Y_i = \overline{X}_i$, $\delta_i = \sigma_i^2/n_i$, $i = 1, 2$ then the following result can be evaluated under the assumption that H_0 is true. The statistic $T_2(X) = \varphi(X)\overline{X}_1 + (1 - \varphi(X))(\overline{X}_1 + \overline{X}_2)/2$, is unbiased test-estimator for the mean μ_1. Similarly, the statistic $T_3(X) = \varphi(X)\overline{X}_1 + (1-\varphi(X))(n_1\overline{X}_1 + n_2\overline{X}_2)/n$, $n = n_1 + n_2$ is unbiased test-estimator for the mean μ_1, too.

Example 3. Let us consider the problem defined in the Example 2 but in addition we assume that $\sigma_1 = \sigma_2 = \sigma$ and σ is not known. The hypothesis H_0: $\omega = 0$, where $\omega = \mu_1 - \mu_2$ against the alternative one H_1: $\omega \neq 0$ is now tested by means of well known Fisher's statistic:

$$F = (\overline{X}_1 - \overline{X}_2)^2 \left(\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)^{-1}$$

$$S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2, \; i = 1, 2.$$

with one and $n_1 + n_2 - 2$ degrees of freedom. Under the significance level $\alpha = P(F \geq f_\alpha)$ the rejection region of the test is the set:

$$K = \left\{ x : \left(\overline{X}_1 - \overline{X}_2\right)^2 \geq f_\alpha \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}$$

Basic properties of multidimensional geometry lead to conclusion that when the hypothesis H_0 is true the set K is a cone symmetric around the point $\{x_1 = \mu_1, \ldots, x_{n_1+n_2} = \mu_1\}$. Hence, under the considered case the above defined statistics T_2 and T_3 are unbiased test-estimators for the mean μ_1 too.

Example 4. Let us generalize the problem considered in the Example 3. Let $X_i = \left[X_{i1} \ldots X_{in_i}\right]$ be independent sample drawn from the distributions of the random variable $X_i \sim N\left(\mu_i, \sigma^2\right)$, $i = 1,\ldots,k$, and σ and μ_i, $i = 1,\ldots,k \geq 2$ are not known. The sample means and variances are denoted by $\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$. $\hat{S}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2$, i=1,...,k $\hat{S}^2 = \frac{1}{k} \sum_{i=1}^{k_i} \hat{S}_i^2$. Let $U_i = \overline{X}_i - \overline{X}_1$ for $i = 2,\ldots,k$. So, $U^T = [U_2 \ldots U_k] \sim N(\mu; \sigma^2 A)$, where:

$$A = \frac{1}{n_1} \begin{bmatrix} 1 + \frac{n_1}{n_2} & 1 & 1 & \ldots & 1 \\ 1 & 1 + \frac{n_1}{n_3} & 1 & \ldots & 1 \\ 1 & 1 & 1 + \frac{n_1}{n_4} & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & 1 & \ldots & 1 + \frac{n_1}{n_k} \end{bmatrix}$$

It is well known that the distribution vector U does not depend on distribution of the variance $\hat{S}^2$. So, this leads to the conclusion that the ratio $F_1 = \frac{U^T A^{-1} U}{\hat{S}^2}$ has noncentral F distribution with $k-1$ and $n-k$ degrees of freedom, where $n = \sum_{i=1}^{k} n_i$ and

$$A^{-1}=\frac{1}{n}\begin{bmatrix}(n-n_2)n_2 & -n_2n_3 & -n_2n_4 & \dots & -n_2n_k\\ -n_3n_2 & (n-n_3)n_3 & -n_3n_4 & \dots & -n_3n_k\\ -n_4n_2 & -n_4n_3 & (n-n_4)n_4 & \dots & -n_4n_k\\ \dots & \dots & \dots & \dots & \dots\\ -n_kn_2 & -n_kn_3 & -n_kn_4 & \dots & (n-n_k)n_k\end{bmatrix}.$$

We consider the hypothesis H_0: $\omega_i = 0$, where $\omega_i = \mu_1-\mu_i$, $i = 2,\dots,k$ against the alternative one H_1: $\omega_i \neq 0$ for at least one index $i = 2,\dots,k$. When hypothesis H_0 is true, the statistic F_1 has central F distribution with $k-1$ and $n-k$ degrees of freedom.

Under the significance level $\alpha = P(F \geq f_\alpha)$ the rejection region of the test is the set:

$$K=\left\{x: U^T A^{-1}U \geq f_\alpha \hat{S}^2\right\}$$

Basic properties of multidimensional geometry lead to conclusion that, when the hypothesis H_0 is true the set K is a cone symmetric around the point:

$$\left\{x_{11}=\mu_1,\dots,x_{1n_1}=\mu_1,\dots,x_{n-n_{k+1}}=\mu_1,x_n=\mu_1\right\}.$$

Hence, the statistic $T_4(X)=\varphi(X)\overline{X}_1+(1-\varphi(X))\hat{X}_1$, where $\hat{X}_1=\frac{1}{k}\sum_{i=1}^{k_i}\overline{X}_i$ is unbiased test-estimator for the mean μ_1. Similarly, the statistic $T_5(X)=\varphi(X)\overline{X}_1+(1-\varphi(X))\hat{X}_2$, where $\hat{X}_2=\frac{1}{n}\sum_{i=1}^{k_i}\overline{X}_i n_i$ is unbiased test-estimator for the mean μ_1, too.

Example 5. Let us consider the problem formulated in the Example 4 again. But now the following hypotheses are tested simultaneously H_{0i}: $\omega_i = 0$, where $\omega_i=\mu_1-\mu_i$, against the alternative one H_{1i}: $\omega_i \neq 0$ for $i = 2, \dots, k$, respectively. The i-th hypothesis is verified by means of the following test.

$$\varphi_i(x_i)=\begin{cases}1 & if \quad x_i \in K_i\\ 0 & if \quad x_i \in K_i^c\end{cases},$$

where:

$$K_i = \left\{x : |\bar{x}_i - \bar{x}_1| \geq t_{\alpha/2}\sqrt{Q_i}\right\},$$

$$K_i^c = R^n - K_i = \left\{x : |\bar{x}_i - \bar{x}_1| < t_{\alpha/2}\sqrt{Q_i}\right\},$$

$$Q_i = \frac{(n_1 - 1)\hat{S}_1^2 + (n_i - 1)\hat{S}_i^2}{n_1 + n_i - 2}\left(\frac{1}{n_1} + \frac{1}{n_i}\right)$$

$i = 2,\ldots,k$ and $P(|T_i| \geq t_{\alpha/2}) = \alpha$ and T_i has Student distribution with $n_1 + n_i - 2$ degrees of freedom. More formally, the hypotheses can be redefined as follows:

$$H_0 \Leftrightarrow (H_{02} \text{ and } H_{03} \text{ and} \ldots \text{and } H_{0k}), \quad H_1 \Leftrightarrow (H_{12} \text{ or } H_{13} \text{ or} \ldots \text{or } H_{1k}).$$

The hypothesis H_0 against the alternative one H_1 can be verified by means of the following test.

$$\varphi(x) = \prod_{i=2}^{k} \varphi_i(x_i) = \begin{cases} 1 & \text{if} \quad x \in K = \bigcup_{i=2}^{k} K_i \\ 0 & \text{if} \quad x \in K^c = \bigcap_{i=2}^{k} K_i^c \end{cases}$$

Hence, the hypothesis H_0 is true if the all hypotheses H_{0i}, $i = 2,\ldots,k$, are true. The alternative hypothesis H_1 is true if at least one of the hypotheses H_{1i}, $i = 2, \ldots, k$, is true. Using the well known Bonferroni's inequality (see e.g. Miller[12], the hypothesis H_0 is rejected under the significance level $\alpha_* \leq (k-1)\alpha$ if at least one of the test $\varphi_i(x_i)$, $i = 2,\ldots,k$, rejects the hypothesis H_{0i} under the significance level α. The considered simultaneous method of testing hypotheses let us identify the set of not rejected hypotheses from the sequence (H_{02} and H_{03} and...and H_{0k}). So, these sample means for which $\varphi_i(x_i) = 0$ can be used to construction the following test-estimator of the mean μ_1.

$$T_6(X) = \varphi(X)\bar{X}_1 + (1 - \varphi(X))\bar{X}_*,$$

where:

$$\bar{X}_* = \frac{\bar{X}_1 n_1 + \sum_{i=2}^{k_i} (1 - \varphi(X_i))\bar{X}_i n_i}{n_1 + \sum_{i=2}^{k_i} (1 - \varphi(X_i))n_i}$$

[12] R.G. Miller: *Simultaneous Statistical Inference*. Springer-Verlag, New York 1981.

So, if $\varphi(x) = 1$ then all the hypotheses H_{0i}, $i = 2,\ldots, k$, are rejected and the expected value μ_1 is estimated on the basis of the sample mean $\overline{X}_1$. In the case when $\varphi(x) = 0$ some of the hypotheses H_{0i}, $i = 2,\ldots, k$, are not rejected and the parameter μ_1 is estimated by means of the sample average $\overline{X}_*$. The sample X has multidimensional normal distribution. Its density function $f(x)$ as well as the truncated densities $f(x|K)$ and $f(x|K^c)$ are symmetric around the point $x = E(X)$ if the hypothesis H_0 is true. This leads, similarly as in the Example 4, to the conclusion that the statistic T_6 is unbiased estimator of mean value μ_1

2. Double test

Let $\varphi(x)$, given by expression (1), be the test of the hypothesis H_0: $\omega \in \Omega_0$ and H_1: $\omega \in \Omega_1$, see the previous chapter. Now our purpose is testing the hypothesis $H_0^{'}: \theta \in \Theta_0$ against the alternative one $H_1^{'}: \theta \in \Theta_1$. But firstly, the hypothesis H_0 is tested in order to chose an appropriate test for the hypothesis $H_1^{'}$. Let us assume that we prefer the test $\varphi_0(x)$ when $x \in K$ and $\varphi_1(x)$ if $x \in K^c$, where:

$$\varphi_0(x) = \begin{cases} 1, & if \quad x \in K_0, \\ 0 & if \quad x \in K_0^c = R - K_0. \end{cases}$$

$$\varphi_1(x) = \begin{cases} 1, & if \quad x \in K_1, \\ 0 & if \quad x \in K_1^c = R - K_1. \end{cases} \tag{7}$$

Their power functions are:

$$\beta_i(\omega) = P_\omega(X \in K_i) = \int_{K_i} f_{\theta,\omega}(x)dx = E(\varphi_i(X)),\ i = 0, 1 \tag{8}$$

The sizes of the tests: $\alpha = sup_{\omega \in \Omega_0} \beta_i(\omega)$.

The test of the hypothesis $H_0^{'}$ against the alternative one $H_1^{'}$ is:

$$\varphi_{01}(x) = \varphi(x)\varphi_0(x) + (1 - \varphi(x))\varphi_1(x) \tag{9}$$

$$\varphi_{01}(x) = \begin{cases} 1, & if \quad x \in K_{01} \\ 0 & if \quad x \in K_{01}^c = R - K_0 \end{cases}, \tag{10}$$

where:

$$K_{01}=K\cap K_0\cup K^c\cap K_1.$$

The power function:

$$\beta_{01}(\omega)=P_\omega(X\in K_{01})=P_\omega(X\in K\cap K_0\cup X\in K^c\cap K_1)=P_\omega(X\in K\cap K_0)$$
$$+P_\omega(X\in K^c\cap K_1)\le P_\omega(X\in K_0)+P_\omega(X\in K_1)=\beta_0(\omega)+\beta_1(\omega), \quad (11)$$

$$\beta_{01}(\omega)=\int_{K\cap K_0}f_{\theta,\omega}(x)dx+\int_{K^c\cap K_1}f_{\theta,\omega}(x)dx=\int_{R^n}\varphi(x)\varphi_0(x)f_{\theta,\omega}(x)dx+$$
$$+\int_{R^n}(1-\varphi(x))\varphi_1(x)f_{\theta,\omega}(x)dx=E(\varphi_{01}(x)). \quad (12)$$

This lets us write the power function in the following way:

$$\beta_{01}(\omega)=\beta_{0/1}(\omega)\beta(\omega)+\beta_{1/0}(\omega)(1-\beta(\omega))=\beta_{1/0}(\omega)+(\beta_{0/1}(\omega)-\beta_{1/0}(\omega))\beta(\omega) \quad (13)$$

where:

$$\begin{cases}\beta_{0/1}(\omega)=\dfrac{P_\omega(X\in K_0\cap K)}{P_\omega(X\in K)}=\dfrac{P_\omega(X\in K_0\cap K)}{\beta(\omega)}\\ \beta_{1/0}(\omega)=\dfrac{P_\omega(X\in K_1\cap K^c)}{P_\omega(X\in K^c)}=\dfrac{P_\omega(X\in K_1\cap K^c)}{1-\beta(\omega)}\end{cases}. \quad (14)$$

So, $\beta_{0/1}(\omega)$ and $\beta_{1/0}$ are power functions of the defined conditional tests.

Let $\alpha_{01}=\beta_{01}(\omega)$, $\alpha_{0/1}=\beta_{0/1}(\omega)$, $\alpha_{1/0}=\beta_{1/0}$ be significance levels of the tests. The equation (13) leads to the following

$$\alpha_{01}(\omega)=\alpha_{0/1}\alpha+\alpha_{1/0}(1-\alpha)=\alpha_{1/0}+(\alpha_{0/1}-\alpha_{1/0})\alpha. \quad (15)$$

On the basis of the equation (11) we have:

$$\alpha_{01}\le 2\alpha. \quad (16)$$

Example 6. Let $Y_i\sim N(\mu_i,\delta_i)$, $i=1,2$. Our purpose is testing the hypothesis H_0': $\mu_1=0$ against the alternative H_1': $\mu_1=\mu_0>0$. In order to do it, firstly the hypothesis H_0: $\omega=0$ where $\omega=\mu_2-\mu_1$ against the alternative one H_1: $\omega>0$, is tested. The hypothesis H_0 is tested on the basis of the statistic

$Z = \frac{Y_2 - Y_1}{\sqrt{\delta_1 + \delta_2}} \sim N(\omega, \delta_1 + \delta_2)$. Let $P(Z \geq z_\alpha | H_0) = \alpha$, where α is the significance level. So, $\varphi(y_1,y_2) = 1$ if and only if $(y_1,y_2) \in K = \{(y_1,y_2): y_2 \geq y_1 + z_\alpha \sqrt{\delta_1 + \delta_2}\}$. $\varphi(y_1,y_2) = 0$ if and only if $(y_1,y_2) \in K^c = \{(y_1,y_2): y_2 < y_1 + z_\alpha \sqrt{\delta_1 + \delta_2}\}$ If the hypothesis H_0 is rejected the hypothesis $H_0^{'}$ is tested by means of the statistic $Z_0 = \frac{Y_1}{\sqrt{\delta_1}}$ otherwise when H_0 is not rejected the hypothesis $H_0^{'}$ is tested by means of the statistic $Z_1 = \frac{Y_2 + Y_1}{\sqrt{\delta_1 + \delta_2}}$. Let $P(Z_i \geq z_\alpha | H_0) = \alpha$, $i = 0, 1$. Hence, $\varphi_0(y_1,y_2) = 1$ if and only if $(y_1,y_2) \in K_0 = \{(y_1,y_2): y_1 \geq z_\alpha \sqrt{\delta_1}\}$ and $\varphi_1(y_1,y_2) = 1$ if and only if $(y_1,y_2) \in K_1 = \{(y_1,y_2): y_2 \geq -y_1 + z_\alpha \sqrt{\delta_1 + \delta_2}\}$. Finally, $\varphi_{01}(y_1,y_2) = 1$ if and only if $(y_1,y_2) \in K_{01} = \{(y_1,y_2): K \cap K_0 \cup K^c \cap K_1\}$. The significance level of the test $\varphi_{01}(y_1,y_2)$ is not greater than 2α. We can show that if $\mu_1 = \mu_0$ and $\mu_2 = \mu_0$ than the power of the test is greater than its significance level. So, the test $\varphi_{01}(y_1,y_2)$ is unbiased. Moreover, if μ_0 increases then the power of the test increases, too.

Particularly, let us note that if $\delta_1 = \delta_2 = 0.5$, then:

$$K = \{(y_1,y_2) : y_1 \geq y_2 + z_\alpha\}, \quad K^c = \{(y_1,y_2) : y_1 < y_2 + z_\alpha\},$$

$$K_0 = \{(y_1,y_2): y_1 \geq z_\alpha / \sqrt{2}, -\infty < y_2 < \infty\},$$

$$K_{01} = \{(y_1,y_2): (y_2 \geq y_1 + z_\alpha, y_1 \geq z_\alpha / \sqrt{2}) \text{ or } (y_2 < y_1 + z_\alpha, y_1 \geq y_2 < -y_1 + z_\alpha)\}$$

Moreover, similarly like in the Example 2 we can consider case, when:

$$Y_i = \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \; X_{ij} \sim N(\mu_i, \delta_i), \; \delta_i = \frac{\sigma_i^2}{n_i}, \; i = 1, 2.$$

Appendix

The derivation of the expected value shown by the expression (5) is as follows[13].

$$E(T(X),\omega)=\int_{R^n}\left(\varphi(x)T_1(x)+(1-\varphi(x))T_0(x)\right)f_{\omega,\theta}(x)dx=$$

$$=\int_K T_1(x)f_{\omega,\theta}(x)dx \quad \int_{+K^c} T_0(x)f_{\omega,\theta}(x)dx=\beta(\omega)\int_K T_1(x)f_{\omega,0}(x\mid K)dx+$$

$$+\left(1-\beta(\omega)\right)\int_{K^c}T_0(x)f_{\omega,\theta}(x\mid K^c)dx=\beta(\omega)E(T_1(X)\mid K,\omega)+$$

$$\left(1-\beta(\omega)\right)E(T_1(X)\mid K^c,\omega).$$

Derivation of the variance given by the expression (6) is[14]:

$$D^2(T(X),\omega)=\int_{R^n}\left(T(X)-E(T(X),\omega)\right)^2 f_{\omega,\theta}(x)dx=$$

$$=\int_{R^n}\left(\varphi(x)T_1(x)+(1-\varphi(x))T_0(x)-E(T(X),\omega)\right)^2 f_{\omega,\theta}(x)dx=$$

$$=\int_{R^n}\left(\varphi^2(x)\left(T_1(x)-E(T(X),\omega)\right)^2+\right.$$

$$\varphi(x)(1-\varphi(x))\left(T_1(x)-E(T(X),\omega)\right)\left(T_0(x)-E(T(X),\omega)\right)+$$

$$\left.+(1-\varphi(x))^2\left(T_0(x)-E(T(X),\omega)\right)^2 f_{\omega,\theta}(x)dx\right.=$$

$$=\int_{R^n}\varphi(x)\left(T_1(x)-E(T(X),\omega)\right)^2 f_{\omega,\theta}(x)dx+$$

$$+\int_{R^n}\left(1-\varphi(x)\right)\left(T_0(x)-E(T(X),\omega)\right)^2 f_{\omega,\theta}(x)dx=$$

$$=\int_K\left(T_1(x)-E(T(X),\omega)\right)^2 f_{\omega,0}(x)dx+$$

[13] J.L. Wywial: Op. cit.
[14] Ibidem.

$$\int_{K^c}\left(T_0(x)-E(T(X),\omega)\right)^2 f_{\omega,\theta}(x)dx=$$

$$=\int_{K}\left(\left(T_1(x)-E(T_1(X),K,\omega)\right)+\left(E(T_1(X),K,\omega)-E(T(X),\omega)\right)\right)^2 f_{\omega,0}(x)dx+$$

$$=\int_{K}\left(\left(T_1(x)-E(T_1(X),K,\omega)\right)+\left(E(T_1(X),K,\omega)-E(T(X),\omega)\right)\right)^2 f_{\omega,\theta}(x)dx+$$

$$+\int_{K}\left(\left(T_0(x)-E(T_0(X),K^c,\omega)\right)+\left(E(T_0(X),K^c,\omega)-E(T(X),\omega)\right)\right)^2 f_{\omega,0}(x)dx=$$

$$=\beta(\omega)\int_{K}\left(T_1(x)-E(T_1(X),K,\omega)\right)^2 f_{\omega,\theta}(x\mid K)dx+$$

$$+\beta(\omega)\left(E(T_1(X),K,\omega)-E(T(X),\omega)\right)^2\int_{K} f_{\omega,\theta}(x\mid K)dx+$$

$$+2\beta(\omega)\left(E(T_1(X),K,\omega)-E(T(X),K,\omega)\right)\cdot$$

$$\int_{K}\left(T_1(x)-E(T_1(X),K,\omega)\right)^2 f_{\omega,\theta}(x\mid K)dx+$$

$$+\left(1-\beta(\omega)\right)\int_{K^c}\left(T_0(x)-E(T_0(X),K^c,\omega)\right)^2 f_{\omega,\theta}(x\mid K^c)dx+$$

$$+\left(1-\beta(\omega)\right)\left(E(T_0(X),K^c,\omega)-E(T(X),\omega)\right)^2\int_{K^c} f_{\omega,\theta}(x\mid K^c)dx+$$

$$+2\left(1-\beta(\omega)\right)\left(E(T_0(X),K^c,\omega)-E(T(X),K,\omega)\right)\cdot$$

$$\int_{K}\left(T_0(x)-E(T_0(X),K^c,\omega)\right)^2 f_{\omega,\theta}(x\mid K^c)dx=$$

$$=\beta(\omega)D^2(T_1(X)\mid K,\omega)+\beta(\omega)\left(E(T_1(X)\mid K,\omega)-E(T(X),\omega)\right)^2+$$

$$+\left(1-\beta(\omega)\right)D^2(T_0(X)\mid K^c,\omega)+\left(1-\beta(\omega)\right)\left(E(T_0(X)\mid K^c,\omega)-E(T(X),\omega)\right)^2$$

This and the expression (5) lead to the the following result.

$$D^2(T(X),\omega)=\beta(\omega)D^2(T_1(X)\mid K,\omega)+\left(1-\beta(\omega)\right)D^2(T_0(X)\mid K^c,\omega)+$$

$$+\beta(\omega)\left(E(T_1(X)\mid K,\omega)-\beta(\omega)E(T_1(X)\mid K,\omega)-\left(1-\beta(\omega)\right)E(T_0(X)\mid K^c,\omega)\right)^2+$$

$$+\left(1-\beta(\omega)\right)\left(E(T_0(X)\mid K^c,\omega)-\beta(\omega)E(T_1(X)\mid K,\omega)-\left(1-\beta(\omega)\right)E(T_0(X)\mid K^c,\omega)\right)^2=$$

$$= \beta(\omega)D^2(T_1(X) \mid K, \omega) + (1-\beta(\omega))D^2(T_0(X) \mid K^c, \omega) +$$

$$+ \beta(\omega)(1-\beta(\omega))^2\left(E(T_1(X) \mid K, \omega) - E(T_0(X) \mid K^c, \omega)\right)^2 +$$

$$+ \beta^2(\omega)(1-\beta(\omega))\left(E(T_1(X) \mid K, \omega) - E(T_0(X) \mid K^c, \omega)\right)^2 +$$

This result leads to the expression (6).

TEST-ESTYMATOR I TEST PODWÓJNY

Streszczenie

W wielu sytuacjach wybiera się metodę wnioskowania statystycznego, w zależności od tego czy spełnione są założenia warunkujące możliwość stosowania danego estymatora albo testu statystycznego. W praktyce weryfikuje się hipotezy o tym czy te założenia są spełnione. Wybór odpowiedniego estymatora zależy od wyniku testowania. Oznacza to, że gdy odrzucimy hipotezę sprawdzaną, to stosujemy estymator inny od tego, który będzie użyty w przypadku, gdy tej hipotezy nie odrzucimy. Wybrany tą drogą estymator (czyli poprzedzony testowaniem odpowiedniej hipotezy statystycznej) nazywamy test-estymatorem. W literaturze zamiast „test-estymator" stosuje się również pojęcie „wstępny test-estymator", które jest dosłownym tłumaczeniem nazwy *preliminary test estimator*. Podobnie ma się rzecz, gdy wybór testu dla określonej hipotezy statystycznej, który jest poprzedzony weryfikacją innej hipotezy o założeniach stosowalności testów dla hipotezy głównej (pierwotnej). Ta procedura weryfikacji jest nazwana testem podwójnym. Z kolei w literaturze anglosaskiej spotyka się nazwę *testitest* dla tego typu procedury wnioskowania statystycznego.

W pracy przedstawiono ogólne wyrażenia określające test estymator i test podwójny. Analizowano proste przykłady konstrukcji test estymatorów, służących ocenie wartości przeciętnej zmiennej losowej. Przedstawiono również test podwójny, służący weryfikacji hipotezy o ustalonej wartości przeciętnej. Wstępne testy statystyczne dotyczyły weryfikacji jednorodności dwóch prób. W zależności od tego czy przyjęto, czy odrzucono hipotezę o tej jednorodności, dalsze wnioskowanie prowadzono albo na podstawie jednej próby, albo na podstawie połączonych obu prób.

Informacja o Katedrze Statystyki

W strukturze Katedry z ystyki, Statystycznej Analizy Jak Staty-stycznych. Obecnie w Kat r zwy-czajny, 3 profesorów uczelni

Zainteresowania naukow ają się wokół zagadnień metody reprezentacyjnej, statystyki małych obszarów, statystycznej kontroli jakości, wnioskowania statystycznego w audycie finansowym, metody klasyfikacji analizy danych oraz wykorzystywania metody symulacji komputerowej w analizach statystycznych.

Pracownicy Katedry współpracują z Vilniaus Universitetas, Université de Neuchâtel, University of Florida.

Pracownicy Katedry prowadzą zajęcia na kierunkach Informatyka i ekonometria, Logistyka, Zarządzanie, Gospodarka turystyczna, Finanse i rachunkowość oraz Międzynarodowe stosunki gospodarcze.

ISBN 978-83-7246-652-5