MATEJ BEL UNIVERSITY, FACULTY OF ECONOMICS,
DEPARTMENT OF APPLIED INFORMATICS, BANSKÁ BYSTRICA, SLOVAKIA

THE UNIVERSITY OF ECONOMICS, FACULTY OF INFORMATICS AND STATISTICS,
DEPARTMENT OF STATISTICS AND PROBABILITY, PRAGUE, CZECH REPUBLIC

WROCLAW UNIVERSITY OF ECONOMICS, DEPARTMENT OF STATISTICS
AND ECONOMIC CYBERNETICS, WROCLAW, POLAND

12TH INTERNATIONAL SCIENTIFIC CONFERENCE APPLICATIONS OF MATHEMATICS AND STATISTICS IN ECONOMY

JAKUB FISCHER (ED.)

AUGUST 26 – 28, 2009, UHERSKÉ HRADIŠTĚ,
CZECH REPUBLIC

IN MEMORY OF FELIX KOSCHIN (1946 – 2009)

Editor
Jakub Fischer

Technical editor
Tomáš Kliegr

Cover design
Magdalena Martinovská

ISBN 978-80-245-1600-4

Jakub Fischer (Ed.)

Applications of Mathematics and Statistics in Economy: AMSE 2009

12th International conference on Mathematics and Statistics in Economy
Uherské Hradiště, Czech Republic, August 27-28, 2009
Proceedings

Preface

International conference *Mathematics and Statistics in Economy* devoted to actual problems of the application of mathematics and statistics in economy was held at Uherské Hradiště during 27nd and 28nd of August 2009. This conference was organized by the Department of Economic Statistics and the Department of Statistics and Probability of the University of Economics Prague. More than 60 specialists from the Czech Republic, Slovakia and Poland took part in this conference; they were representatives of University of Economics Prague, Matej Bel University in Banská Bystrica, Economic University in Wroclaw, Institute of Informatics of Czech Academy of Sciences, Technical University in Zvolen, Unicorn College, National Bank of Slovakia and other universities and institutions.

This conference took place for the 12th time. In 1998, during its first performance, the representatives of the departments of statistics from the Faculty of Informatics and Statistics of the University of Economics Prague and the Department of Applied Informatics of Matej Bel University agreed on the deepening of the cooperation of both institutions. Further, besides of personal special contacts among the members of these departments, the tradition of rotational organizing of international conferences with the same or similar topic has been created. In 2000, Polish colleagues were asked to join this conference too, because economical problems occurring in the Czech Republic, Slovakia and Poland are similar. The invitation of Polish colleagues from the Economical University in Wroclaw (departments of statistics of the University of Economics Prague have with this institution very good long-term relationships crowned by the awards of the title doctor honoris causa to Professor Z.Helwig in 1993) was only a natural consequence of the conception of the further development of topical orientation of these conferences, because the information exchange among specialists from central Europe countries seems to be very challenging. Furthermore, regular meetings of researchers from different universities will enable to get acquainted with the activities of other economically oriented faculties and contact not only universities but, above all, the researchers of individual departments interested in the same problems which are unavoidable with a view to narrow

specialization. The next (thirteenth) conference will take place in Slovakia.

The topics presented during this year´s conference were heterogeneous, from time-series problems over the use of different mathematical and statistical methods in banking, national accounting, insurance industry, business, quality control, marketing (or more generally in manager decision making) to the use of neural networks in data analysis. The time-table of this conference was organized so that all contributions were clustered according to their subject into less topical groups. In these proceedings, they are arranged in alphabetical order according to authors´ names. The conference AMSE 2009 was supported by the Ministry of Education, Youth and Sports of the Czech Republic, project n^{o} MSM6138439910.

November 2009

Prof. Ing. Richard Hindls, CSc.
Chair of Scientific Committee
Rector

Prof. Ing. Stanislava Hronová, CSc.
Chair of Organizing Committee
Vice Rector

University of Economics Prague

Organization

Scientific Program Committee

Chair
Richard Hindls

Members
Stanislava Hronová
Walenty Ostasiewicz
Rudolf Zimka

Organizing Committee

Chair
Stanislava Hronová

Members
Jakub Fischer
Savina Finardi
Věra Jeřábková
Tomáš Kliegr
Luboš Marek
Petr Mazouch
Jan Nemrava
Kristýna Vltavská
Michal Vrabec
Šárka Zívalová

Reviewers

Milan Bašta
Dagmar Blatná
Jan Čadil
Savina Finardi
Jakub Fischer
Stanislaw Heilpern
Stanislava Hronová
Eva Jarošová
Mária Kanderová

Kamil Kladívko
Pavol Král
Jitka Langhamrová
Ivana Malá
Luboš Marek
Petr Mazouch
František Mošna
Roman Pavelka
Iva Pecáková
Hana Řezanková
Jiří Trešl
Martin Zelený
Rudolf Zimka
Emília Zimková
Pavel Zimmermann

Contents

Nonparametric estimation of log volatility with wavelets

Milan Bašta[1]

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague 3, W. Churchill Sq. 4, 130 67
milan.basta@vse.cz

Abstract. Volatility plays a crucial role in financial markets. Favorite parametric models of volatility are ARCH/GARCH conditional models with conditional returns having for example normal distribution. Another class of models is a class of stochastic volatility models. We describe a nonparametric way to the estimation of the logarithm of volatility from the logarithm of squared returns. This approach is based on the discrete wavelet transform combined with the false discovery rate method of multiple testing. The procedure is applied to the estimation of the logarithm of volatility of stock prices of Citi-group. The results are discussed.

Keywords: Volatility, estimation, wavelets, denoising, false discovery rate

1 Introduction

Let P_t be the price of a financial asset and r_t be the return defined as

$$r_t \equiv \log\left(P_t \big/ P_{t-1}\right). \tag{1}$$

Assume that the conditional distribution of returns given information at time $t-1$ has the cumulative density function G and the conditional mean equal to m_t and the conditional variance equal to h_t, i.e.

$$r_t \mid I_{t-1} = G(m_t, h_t), \tag{2}$$

$$m_t \equiv E_{t-1}[r_t], \tag{3}$$

$$h_t \equiv E_{t-1}[(r_t - m_t)^2], \tag{4}$$

where E_{t-1} is the expectation operator given the information at time $t-1$ (i.e. I_{t-1}).

The conditional variance h_t is called *volatility*. We see that volatility describes the variability of conditional returns. The knowledge of both the volatility h_t and the

cumulative density function G may serve for the assessment of the risk that we undergo when holding a financial asset. Therefore, volatility plays a crucial role in financial markets, specifically in derivative pricing, hedging and portfolio management.

1.1 Properties of volatility

Volatility displays many well-known properties (see for example [4] and [8]). Some of these properties will be tackled shortly in this section. It is clear that volatility is varying with time. The analysis of the autocorrelation of squared returns r_t^2 shows that volatility is persistent with large (small) changes in the price today followed by large (small) changes in the price in the future (the clustering property of volatility). However, after some time volatility always seems to tend to its long-run mean (the mean reversion property of volatility). Innovations in the price of the asset have an asymmetric impact on volatility. Specifically, negative returns today usually imply higher future volatility than positive returns of the same size (the leverage effect).

1.2 Modeling volatility

For details see for example [4] and [8]. One of the most favorite models of volatility is the group of ARCH/GARCH conditional models. In the most basic model of this group, the so called GARCH(1,1) model, volatility h_t is assumed to behave according to the following process

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}, \tag{5}$$

where $\omega > 0$, $\alpha > 0$ and $\beta \geq 0$ are parameters. The current volatility h_t at time t is the function of the volatility h_{t-1} and squared returns r^2_{t-1} from the preceding period (time $t-1$). If the following condition

$$\alpha + \beta < 1 \tag{6}$$

is fulfilled, the process of returns r_t is weak stationary and volatility is mean reverting with the long-run level given by

$$\frac{\omega}{1-\alpha-\beta}. \tag{7}$$

The autocorrelation function of squared returns of the GARCH(1,1) model exhibits exponential decay. GARCH(1,1) does not capture the asymmetric impact of price innovations on volatility. Assuming that G in eq. (2) is normal, the parameters of the

model can be estimated with the maximum likelihood approach with normal errors. Many extensions of the GARCH models are available in literature.

A second class of volatility models is the class of stochastic volatility models. To illustrate, realize that

$$r_t = \sqrt{h_t}\,\varepsilon_t\,, \tag{8}$$

where volatility h_t is assumed to follow some latent stochastic process that need not be necessarily related to the process of returns and where

$$\varepsilon_t \sim G(m_t,1)\,. \tag{9}$$

If we define log volatility s_t as

$$s_t \equiv \log h_t\,, \tag{10}$$

where log stands for the natural logarithm, we can rewrite eq. (8) as

$$r_t = \exp\left(s_t/2\right)\varepsilon_t\,, \tag{11}$$

If we take the logarithm of the square of eq. (11) we get

$$\log r_t^{\,2} = s_t + \log(\varepsilon_t^2) = s_t + \eta_t + \gamma \tag{12}$$

with η_t and γ defined as

$$\eta_t \equiv \log\left(\varepsilon_t^2\right) - \gamma\,, \tag{13}$$

$$\gamma \equiv E[\log\left(\varepsilon_t^2\right)] = -1.27\,. \tag{14}$$

A specific parameterization may be given for s_t such as

$$s_t = \mu + \phi s_{t-1} + a_t \tag{15}$$

where μ and $|\phi| < 1$ are parameters and a_t is i.i.d.$(0,\sigma_a^2)$ and a_t and e_t are allowed to be correlated. A volatility model given by eq. (11) and eq. (15) is in fact the canonical model of stochastic volatility in discrete time.

1.3 Nonparametric estimation with wavelets

Log volatility s_t can be estimated nonparametrically from eq. (12) with the use of wavelets. Let us describe how the algorithm works. Rewrite equation (12) into the following form

$$\log r_t^2 - \gamma = s_t + \eta_t \,. \tag{16}$$

We see that the log volatility s_t cannot be observed directly. What we *can* observe is the log squared returns minus gamma (i.e. $\log r_t^2 - \gamma$) which is equal to the log volatility s_t plus the additive *noise* η_t. If we manage to remove the additive noise η_t from the time series of $\log r_t^2 - \gamma$ (the procedure of removing noise is called *denoising*) we would get the log volatility s_t. In Fig. 1 the time series of P_t, r_t, r_t^2, and $\log r_t^2 - \gamma$ are given for the shares of Citi-group, starting April 18, 2008 and ending April 24, 2009.

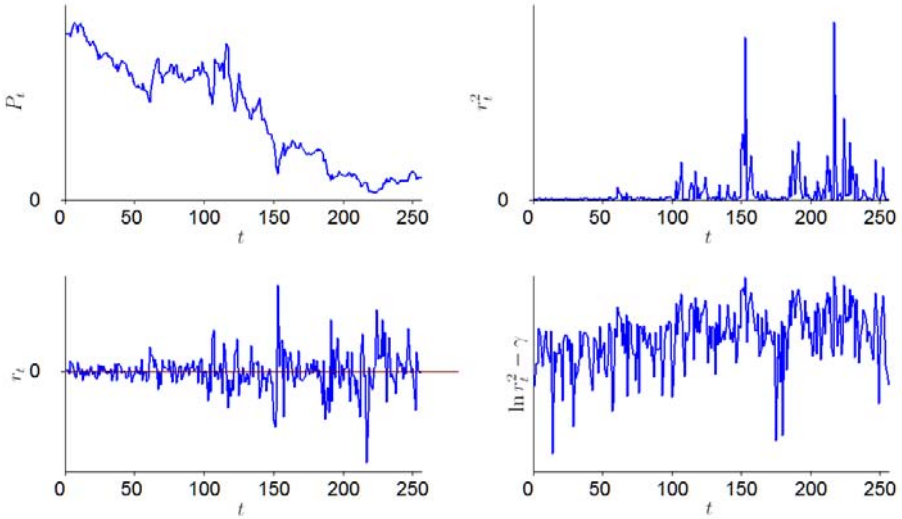


Fig. 1. The time series of P_t, r_t, r_t^2, and $\log r_t^2 - \gamma$ are given for the shares of Citi-group, starting April 18, 2008 and ending April 24, 2009. The effect of the financial crisis is clearly visible.

A question arises as how to remove the additive noise η_t. A typical approach in econometrics is to apply a moving average to the time series. However, the application of a moving average means rather smoothing than denoising. In this paper

we will present an approach to denoising based on wavelets and on the false discovery rate method of multiple testing. At first we will introduce the notion of wavelets and the discrete wavelet transform and further we will describe the false discovery rate method. In the end we will apply the combined procedure to denoising the time series of the log squared returns and illustrate the results on the time series of Citi-group of Fig. 1.

2 Wavelets

A rigorous way to introduce the concept of wavelets and the discrete wavelet transform is laborious and there is no place for it in this short paper. A detailed introduction to wavelets can be found in books such as [5], [7] and [9]. A short introduction can be found for example in [2]. In this text only a very short summary will be given.

Let us denote the input time series as $\{x_t : t = 0, ..., N-1\}$ and the coefficients of the discrete wavelet transform (DWT) as $\{w_k : k = 0, ..., N-1\}$. DWT is a linear transform with the *orthonormal* transform matrix O of size N, i.e. $O^T O = 1_N$, where O^T is a transpose of O and 1_N is the identity matrix of size N. The elements of O are given in a specific way (not specified here). If wavelet coefficients $\{w_k\}$ are written as a column vector $\mathbf{W}$ and the values of the time series $\{x_t\}$ as a column vector $\mathbf{X}$ then DWT and the backward synthesis of the original time series may be written as

$$\mathbf{W} = O\mathbf{X} \quad \text{a} \quad \mathbf{X} = O^T \mathbf{W}, \tag{17}$$

Even though DWT can be calculated according to eq. (17), the coefficients $\{w_k\}$ are usually calculated rather via the *pyramid algorithm* which is less time consuming. Pyramid algorithm can be thought of as a sequence of linear filtrations – more details can be found in [5], [7] and [9]. It is also important to stress that DWT can be applied only to a time series of the length $N = 2^J$, where J is a positive integer. There exists a variant of the DWT, called the maximal overlap discrete wavelet transform, MODWT, which does not require the time series to be of the length $N = 2^J$.

Now, let us group the sequence $\{w_k : k = 0, ..., N-1\}$ of the wavelet coefficients of length $N = 2^J$ into J subsequences $\{w_{j,m}\}$ such that

$$\left\{ w_{j,m} : m = 0,..., \frac{N}{2^j} - 1 \right\} \equiv \left\{ w_k : k = \frac{2^{j-1}-1}{2^{j-1}} N,..., \frac{2^j - 1}{2^j} N - 1 \right\}, , \tag{18}$$

where $j = 1, ..., J$. Moreover, let $v_{J,0} \equiv w_{N-1}$. The subsequence $\{w_{j,m}\}$ defined in eq. (18) is of the length $N/2^j$. The first $N/2$ values of $\{w_k\}$ are thus assigned to the subsequence $\{w_{1,m}\}$, the next $N/4$ values are assigned to the subsequence $\{w_{2,m}\}$ etc. until the last two values of the sequence $\{w_k\}$ are assigned to $w_{J,0}$ and $v_{J,0}$. It can be shown that the coefficients $\{w_{j,m}\}$ capture the *local* (subscript m) frequency content of

the time series in the interval of frequencies $[1/2^{j+1};1/2^{j}]$ and the coefficient $v_{J,0}$ captures the frequency content in the interval $[0;1/2^{J+1}]$.

3 Controlling the false discovery rate

In this paper, the denoising of the time series with DWT will be implemented through hypothesis testing (details will be given below). First, imagine a general problem of testing multiple hypotheses. Let T be the total number of hypotheses. In this situation it is crucial to choose a suitable significance level α, which will be the same for *each* test. If we choose the 'traditional' significance level $\alpha = 0.05$ then the probability of falsely rejecting *at least one* hypothesis will be extremely high, tending to 1 if T is large. Said informally, we allow many false rejections of null hypotheses in return for more correct rejections.

On the other hand, we could use the Bonferroni approach to hypotheses testing and set the significance level $\alpha = 0.05/T$. This approach guarantees that the probability of making at least one false rejection is not greater than 0.05. This means that we have lowered the chance of false rejection of any null hypothesis so much that it is difficult to reject the null hypothesis even if it is false.

In this paper we apply a method that is intermediate between the two just mentioned situations and at the same time is adaptive to the data – this method is called the *False Discovery Rate* method and details may be found for example in [3] or [6]. While testing T hypotheses simultaneously let us assume we reject the total number R of them. Out of these R rejected hypotheses let R_c be the number of correctly rejected (i.e. null is false) hypotheses and let R_f be the number of falsely rejected (i.e. the null is true) hypotheses,

$$R = R_c + R_f . \tag{19}$$

Now, for each from the total number of T hypotheses find its p-value and order the p-values so that

$$p_{(1)} \le p_{(2)} \le \ldots \le p_{(T)} . \tag{20}$$

Choose $0 \le \alpha \le 1$ and define

$$i \equiv \max\left(k \mid p_{(k)} < \frac{k}{T}\alpha \right). \tag{21}$$

Consequently reject all hypotheses (from the whole set of T hypotheses) whose p-values are less or equal to $p_{(i)}$. Define the *false discovery rate* Q as the fraction of rejected hypotheses that were falsely rejected i.e.

$$Q \equiv \frac{R_f}{R_f + R_c} . \tag{22}$$

For a selected α and for independent test statistics it can be shown that the *expected* value of Q will be lower or equal to α, i.e.

$$E(Q) \le \alpha . \tag{23}$$

4 Application to wavelet denoising

The idea beyond denoising the time series with wavelets is as follows. First let us return to the equation

$$\log r_t^2 - \gamma = s_t + \eta_t . \tag{24}$$

As DWT is a *linear* transform it must hold that (subscripting corresponds to eq. (18))

$$W_{j,m} = d_{j,m} + e_{j,m} , \tag{25}$$

where $W_{j,m}$ are the DWT coefficients of the log squared returns, $d_{j,m}$ the DWT coefficients of the log volatility and $e_{j,m}$ the DWT coefficients of the 'noise', i.e.

$$W_{j,m} \equiv DWT(\log r_t^2 - \gamma) \tag{26}$$

$$d_{j,m} \equiv DWT(s_t) \tag{27}$$

$$e_{j,m} \equiv DWT(\eta_t) \tag{28}$$

The approach to wavelet denoising of the time series of log squared returns will be carried out in a false discovery rate multiple testing framework. Let us assume the following sequence of null hypotheses, each of which claims that a wavelet coefficient $W_{j,m}$ corresponds to noise *only* (which means that $d_{j,m}$ is zero). The

alternative hypotheses claim that the wavelet coefficient contains signal (which means that $d_{j,m}$ is not zero).

$$H_{j,m}^{(0)} : d_{j,m} = 0 \qquad H_{j,m}^{(A)} : d_{j,m} \neq 0 \tag{29}$$

If null is true then the distribution of the wavelet coefficients $W_{j,m}$ has a cumulative density function denoted as F_j,

$$W_{j,m} = 0 + e_{j,m} \sim F_j \,. \tag{30}$$

Denote the elements of the DWT transform matrix as $(O)_{ki}$ with $k = 0,..., N-1$ subscripting the rows and $i = 0,..., N-1$ subscripting the columns and realize that

$$\sum_i (O)_{ki} = 0 \tag{31}$$

for all k except for $k = N-1$. Then we may write

$$e_{j,m} = \sum_i (O)_{ki} \eta_i = \sum_i (O)_{ki} \log(\varepsilon_i^2) \sim F_j \tag{32}$$

where k and j are related (see eq. (18)) as

$$k = \frac{2^{j-1} - 1}{2^{j-1}} N + m \tag{33}$$

If we assume that the distribution of ε_i is standard normal then a variable with cumulative density function F_j is a variable whose distribution is the same as a distribution of a weighted sum of log $\chi^2(1)$ random variables (where $\chi^2(1)$ stands for chi-square with one degree of freedom) where the weights are given be the elements in the rows of the transform matrix O. Now, let us denote the p-value corresponding to the hypothesis $H_{j,t}$ as $p_{j,t}$

$$p_{j,t} \equiv 2 \times \min\left(F_j(W_{j,t}), 1 - F_j(W_{j,t})\right) \tag{34}$$

Let $p_{(1)}, ..., p_{(T)}$ be the ordered p-values according to eq. (20) and define i according to eq. (21). Reject all hypotheses with p-values less or equal to $p_{(i)}$. Rejection of a hypothesis means that we 'claim' that the corresponding wavelet coefficient contains

signal and therefore the value of the coefficient must be retained. On the other hand, if we do not reject a hypothesis then we 'claim' the wavelet coefficient corresponds to noise only and its value must be set to zero. We can thus write

$$W_{j,t}^{(threshold)} \equiv \begin{cases} 0, & \text{pro } F_j^{-1}\left(p_{(i)}/2\right) < W_{j,t} < F_j^{-1}\left(1 - p_{(i)}/2\right) \\ W_{j,t}, & \text{otherwise} \end{cases} \tag{35}$$

Finally we can reconstruct the original time series (see eq. (17)) as

$$\mathbf{X}^{(denoised)} = O^T \mathbf{W}^{(threshold)}, \tag{36}$$

where $\mathbf{W}^{(threshold)}$ is a column vector which contains the values of $W_{j,t}^{(threshold)}$ and $\mathbf{X}^{(threshold)}$ is a column vector which contains the values of the denoised time series and O^T is the transpose of the DWT transform matrix.

4.1 Log volatility of Citi-group

In fig. 2 the procedure is applied to the time series of Citi-group of fig. 1. The smooth line denotes the denoised time series of log squared returns and is thus the estimate of the historical log volatility s_t.

The value of α was chosen to be $\alpha = 0.1$, the significance level for one hypothesis test was adaptively determined (by the false rate discovery method) to be 0.0016. The DWT transform matrix corresponds to wavelet *LA*8. Boundary wavelet coefficients were included into the analysis and treated as if they were not boundary. The result was averaged over circularly shifted time series (cycle spinning), which was implemented by the variant of the DWT called MODWT (for details see [7]).

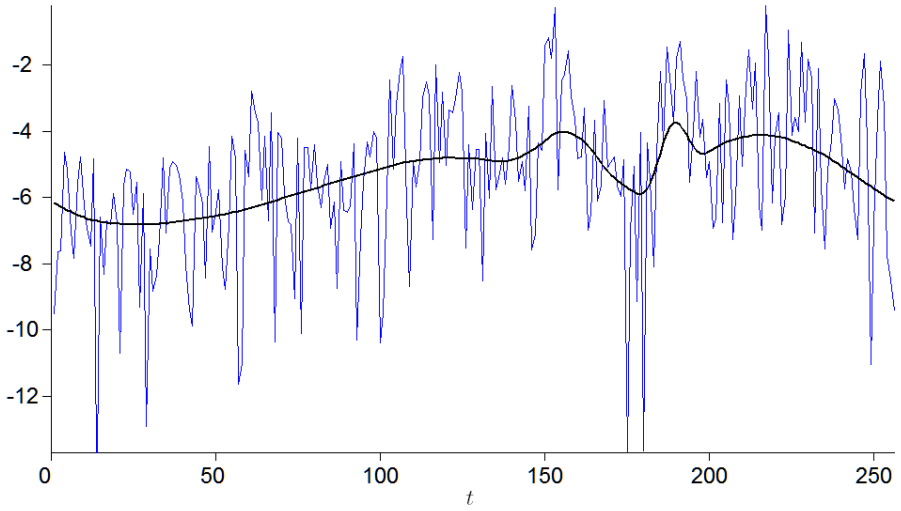


Fig. 2. The time series $\log r_t^2 - \gamma$ for the shares of Citi-group. The smooth line represents the estimate of log volatility obtained by DWT (cycle spinning variant) and false discovery rate method.

5 Conclusion and discussion

We have shown how to estimate log volatility from log squared returns with the use of wavelets and with the method of false discovery rate of multiple hypotheses testing. We have applied the procedure to the time series of shares of Citi-group.

Yet, several questions are to be answered. Is it beneficial for practical purposes to estimate the logarithm of volatility $s_t = \log h_t$ if only h_t is worked with in practical applications? Realize also that the exp(estimate(log volatility)) is "*not equal"* to the estimate(volatility) and volatility h_t cannot be directly estimated with our approach which involves *additive i.i.d.* noise.

Anyway, wavelets are an excellent tool for denoising time series. The denoising can be implemented through multiple hypothesis testing. The false discovery rate method seems to be a good data adaptive approach to multiple testing.

References

1. Abramovich, F., Benjamini, Y.: Adaptive Thresholding of Wavelet Coefficients. Computational Statistics and Data Analysis 22, 351--361 (1996)
2. Bašta, M.: Wavelety a jejich použití v analýze časových řad. Sborník prací účastníků vědeckého semináře doktorského studia, 107--117, FIS, VŠE, Praha (2008). ISBN 978-8024513614
3. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57, 289--300 (1995)

4. Engle, R. F., Patton, A. J.: What is a Good Volatility Model? Quantitative Finance 1, 237--245 (2001)
5. Gencay, R., Selcuk, F., Whitcher, B.: An Introduction to Wavelets and Other Filtering Methods in Finance and Economics. Academic Press (2002). ISBN 978-0122796708
6. Miller, C. J., Genovese, C., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, J., Moore, A.: Controlling the False Discovery Rate in Astrophysical Data Analysis. The Astronomical Journal 122, 3492--3505 (2001)
7. Percival, D., Walden, A.: Wavelet Methods for Time Series Analysis. Cambridge University Press (2000). ISBN 978-0521640688
8. Poon, S-H., Granger, C. W. J.: Forecasting Volatility in Financial Markets: A Review. Journal of Economic Literature XLI, 478--539 (2003)
9. Vidakovic, B.: Statistical Modeling by Wavelets. Wiley Series in Probability and Statistics, (1999). 978-0471293651

Application of Pareto Distribution in Modelling of Wage Distributions

Diana Bílková

Prague University of Economics, sq. W. Churchill 4, 130 67 Prague 3
bilkova@vse.cz

Abstract. Pareto distribution is usually used as a model of the distribution of the largest wages, not for the whole wage distribution. The Pareto distribution will be a good model of the wage distribution if the following ratios are equal: Ratio of the upper quartile to the median; ratio of the eighth decile to the sixth decile; ratio of the ninth decile to the eighth decile. This property can be used as one of the criterion to measure the quality of fit of Pareto distribution to some empirical wage distribution. If in a particular case the observed differences of the rates of the above mentioned quantiles are negligible, Pareto distribution will be an appropriate model of the considered wage distribution. In the case of differences are quite material, the approximation of the considered wage distribution with Pareto distribution will be more or less inappropriate.

Keywords: Pareto distribution, Pareto coefficient, ratio of two quantiles, wage distributions, differentiation of wages.

1 Pareto Distribution

Pareto distribution is usually used as a model of the distribution of the largest wages, in principle it cannot be used as a model for the whole wage distribution. In the statistical literature it is encouraged the application of this distribution as a model of wages, which are higher than median. Therefore in this article we will consider using the Pareto distribution to model wages higher than median.

The $100 \cdot P\%$ quantile of the wage distribution will be denoted by x_P, $0 < P < 1$. This value represents the upper bound of $100 \cdot P$ % lowest wages and also the lower bound of $100 \cdot (1 - P)$ % highest wages. A particular quantile (denoted as x_{P0}) which will be the lower bound of some small number of the highest wages is usually set to be the maximum wage. If the following formula (1) holds for any quantile x_P, the wage distribution is Pareto distribution

$$\frac{x_{P0}}{x_P} = \left(\frac{1 - P}{1 - P_0} \right)^b . \tag{1}$$

The parameter b of the Pareto distribution (1) is called the Pareto coefficient. It can be used as a characteristic of differentiation of 50 % highest wages.

We will now consider a pair of quantiles x_{P1} and x_{P2}, $P_1 < P_2$. It follows from (1) that

$$\frac{x_{P0}}{x_{P1}} = \left(\frac{1-P_1}{1-P_0}\right)^b \tag{2}$$

and

$$\frac{x_{P0}}{x_{P2}} = \left(\frac{1-P_2}{1-P_0}\right)^b. \tag{3}$$

From what we can derive for the rate of x_{P2} to x_{P1} that

$$\frac{x_{P2}}{x_{P1}} = \left(\frac{1-P_1}{1-P_2}\right)^b. \tag{4}$$

The rate

$$\frac{x_{P2}}{x_{P1}}$$

is an increasing function of the Pareto coefficient b. If the rate of quantiles increases, the relative differentiation of wages increases too. If only absolute differences between quantiles increase, only the absolute differentiation of wages increases.

It follows from the equation (1) that once the values x_{P0} and b are chosen we can determine the quantile x_P for any chosen P or the other way around for any value x_P we can find the corresponding value of P. In the first case it is advantageous to write the equation (1) as

$$x_P = \frac{x_{P0}}{\left(\frac{1-P}{1-P_0}\right)^b} \tag{5}$$

or after logarithmic transformation as

$$\log x_P = \log x_{P0} - b[\log(1 - P) - \log(1 - P_0)], \tag{6}$$

in the second case

$$1 - P = (1 - P_0)\sqrt[b]{\frac{x_{P0}}{x_P}} \tag{7}$$

or after logarithmic transformation as

$$\log(1 - P) = \log(1 - P_0) + \frac{1}{b}(\log x_{P0} - \log x_P). \tag{8}$$

The equations (2), (3) and (4) will after logarithmic transformation have the following form

$$b = \frac{\log \dfrac{x_{P0}}{x_{Pi}}}{\log \dfrac{1 - P_i}{1 - P_0}}, \quad \mathrm{i} = 1, 2, \tag{9}$$

$$b = \frac{\log \dfrac{x_{P2}}{x_{P1}}}{\log \dfrac{1 - P_1}{1 - P_2}}. \tag{10}$$

It follows from the equation (9) that instead of the Pareto coefficient b we can use any other quantile x_{P1} and x_{P2} of the Pareto distribution and it follows from the equation (10) that the Pareto coefficient b can be calculated using any known quantiles x_{P1} and x_{P2}. Then we can also determine the value x_{P0} using the formulas

$$x_{P0} = x_{P1}\left(\frac{1 - P_1}{1 - P_0}\right)^b, \tag{11}$$

$$x_{P0} = x_{P2}\left(\frac{1 - P_2}{1 - P_0}\right)^b. \tag{12}$$

The model characterized with the relationship (1) will be practically applicable if the following is known:

- The value of the quantile that characterizes the assumed wage maximum and the value of the Pareto coefficient b;
- The value of the quantile that characterizes the assumed wage maximum and the value of any other quantile;
- The values of any two quantiles of the Pareto distribution.

Any two quantiles can be written as x_P and x_{P+k}, where $0 < k < 1 - P$. Using the equation (4), we can derive for the rate of these two quantiles

$$\frac{x_{P+k}}{x_P} = \left(\frac{1-P}{1-P-k} \right)^b. \tag{13}$$

The rate (13) will be equal for such pairs of quantiles for which the following formula holds

$$\frac{1-P}{1-P-k} = c, \tag{14}$$

where c is a constant, i. e. the rate will be the same for all pairs of quantiles for which

$$k = \frac{c-1}{c}(1-P). \tag{15}$$

We will use the constant $c = 2$ in (15) and we will choose gradually $P = 0{,}5;\ 0{,}6;\ 0{,}8$. Then using the equation (13) we can show the equality of rates of some frequently used quantiles

$$\frac{x_{0,75}}{x_{0,5}} = \frac{x_{0,8}}{x_{0,6}} = \frac{x_{0,9}}{x_{0,8}}. \tag{16}$$

From the relationship (16) we can conclude that Pareto distribution assumes such a wage differentiation for which the rate of the upper quartile to median is the same as:

- The rate of the 8th to the 6th decile;
- And as the rate of the 9th to the 8th decile.

If in a particular case the observed differences of the rates of the above mentioned quantiles are negligible, Pareto distribution will be an appropriate model of the considered wage distribution. In the case the differences are quite material, the approximation of the considered wage distribution with Pareto distribution will be more or less inappropriate.

2 Parameter Estimates

If the Pareto distribution is chosen as a model for a particular distribution we have to keep in mind that this model is only an approximation. The wage distribution will be only approximated and the relations derived from the model will also hold for the "true distribution" only approximately. Which relations will hold more precisely and for which the precision will be lower will be mostly dependent on the method of parameter estimates.

There are many possibilities to choose from. In the following text the quantiles of Pareto distribution will be denoted as x_P and the quantiles of the observed wage distribution will be denoted as y_P.

First we need to decide which quantile to choose as x_{P0}. In this article we will assume that $x_{P0} = x_{0,99}$. From the equation (1) we can see that the considered Pareto distribution will be defined by the equation

$$\frac{x_{0,99}}{x_P} = \left(\frac{1-P}{0,01}\right)^b. \tag{17}$$

Then we need to determine the value $x_{0,99}$ and the value of the Pareto coefficient b. Because it is necessary to estimate the values of two parameters we need to choose two equations to estimate from.

A natural choice is the equation $x_{P0} = y_{P0}$; that is in our case $x_{0,99} = y_{0,99}$. As the other equation we set a quantile x_{P1} equal to the corresponding observed quantile, i.e. $x_{P1} = y_{P1}$. In this case, the parameters of the model will be

$$x_{P0} = y_{P0} \tag{18}$$

and using (9)

$$b = \frac{\log\frac{y_{P0}}{y_{P1}}}{\log\frac{1-P_1}{1-P_0}}. \tag{19}$$

We can get different modifications using different choice of the maximum wage and the second quantile. If we use equation $x_{0,99} = y_{0,99}$ and we use the median in the second equation, i.e. $x_{0,5} = y_{0,5}$ we get a model with parameters

$$x_{0,99} = y_{0,99}, \tag{20}$$

$$b = \frac{\log\frac{y_{0,99}}{y_{0,5}}}{\log\frac{0,5}{0,01}}. \tag{21}$$

Another possibility is setting any two quantiles of the model equal to the quantiles of the observed distribution

$$x_{P1} = y_{P1}, \tag{22}$$

$$x_{P2} = y_{P2}. \tag{23}$$

Using the formula (10) we get the following parameter estimates

$$b = \frac{\log \dfrac{y_{P_2}}{y_{P_1}}}{\log \dfrac{1-P_1}{1-P_2}} \tag{24}$$

and from (11) and (12) we get

$$x_{P0} = y_{P1}\left(\frac{1-P_1}{1-P_0}\right)^b = y_{P2}\left(\frac{1-P_2}{1-P_0}\right)^b. \tag{25}$$

With this alternative we can also get numerous modifications depending on the choice of quantiles y_{P1} and y_{P2} that are used.

The third possibility is based on the request that $x_{P0} = y_{P0}$ and that the rate of some other two quantiles of the Pareto distribution x_{P2}/x_{P1} is equal to the rate y_{P2}/y_{P1} of correspoding quantiles of the wage distribution observed. In this case we will estimate the parameters using (see (10))

$$x_{P0} = y_{P0}, \tag{26}$$

$$b = \frac{\log \dfrac{y_{P_2}}{y_{P_1}}}{\log \dfrac{1-P_1}{1-P_2}}. \tag{27}$$

In this case notwithstanding that $x_{P2}/x_{P1} = y_{P2}/y_{P1}$ hold, the equality of quantiles itself, $x_{P1} \neq y_{P1}$ and $x_{P2} \neq y_{P2}$, does not hold. In this case we can also arrive to numerous modifications depending on what maximum wage is chosen and what quantiles y_{P1} and y_{P2} are chosen.

For all of the above methods the equality of two characteristics of the model and the observed distribution was required. There are also different approaches to the parameter estimates.

The least squares method is frequently used for the Pareto distribution parameter estimates as well. We will consider the following quantiles of the observed wage distribution y_{P1}, y_{P2}, …, y_{Pk} and corresponding quantiles of the Pareto distribution x_{P1}, x_{P2}, …, x_{Pk}. The model distribution will be most precise when the sum of squared differences

$$\sum_{i=1}^{k} (y_{P_i} - x_{P_i})^2 \tag{28}$$

is minimized. In this case closed formula solution does not exist. Therefore sum of squared differences of logarithms of quantiles is often considered

$$\sum_{i=1}^{k}(\log y_{P_i} - \log x_{P_i})^2. \tag{29}$$

Minimizing the objective function (29), it is possible to derive the following estimates

$$b = \frac{k\sum_{i=1}^{k}\log y_{P_i}\ \log\frac{1-P_0}{1-P_i} - \sum_{i=1}^{k}\log y_{P_i}\ \sum_{i=1}^{k}\log\frac{1-P_0}{1-P_i}}{k\sum_{i=1}^{k}\log^2\frac{1-P_0}{1-P_i} - \left(\sum_{i=1}^{k}\log\frac{1-P_0}{1-P_i}\right)^2}, \tag{30}$$

$$\log x_{P_0} = \frac{\sum_{i=1}^{k}\log y_{P_i}}{k} - b\frac{\sum_{i=1}^{k}\log\frac{1-P_0}{1-P_i}}{k}. \tag{31}$$

In the case we use this estimating method, it is needed to keep in mind that the equality of model quantiles and observed quantiles is not guaranteed for any P. Again we can arrive to different results depending of what quantiles $y_{P1}, y_{P2}, \ldots, y_{Pk}$ are used for the calculations. Furthermore the parameter estimates also depend on the choice of the maximum wage.

3 Characteristics of the Appropriateness of Pareto Distribution

For the application of Pareto distribution as a model of the wage distribution, it is crucial that the model fits the observed distribution as close as possible. It is important that the observed relative frequencies in particular wage intervals are as close to the theoretical probabilities assigned to these intervals by the model as possible.

It is needed to note that the same parameter estimation method does not always lead to the best results. It is of particular importance in "what direction" is the observed wage distribution different from Pareto distribution. Pareto distribution assumes such wage differentiation that the relations (16) hold. With real data we can encounter many different situations

$$\frac{y_{0,75}}{y_{0,5}} < \frac{y_{0,8}}{y_{0,6}} < \frac{y_{0,9}}{y_{0,8}}, \tag{32}$$

$$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}, \tag{33}$$

$$\frac{y_{0,75}}{y_{0,5}} < \frac{y_{0,9}}{y_{0,8}} < \frac{y_{0,8}}{y_{0,6}}, \tag{34}$$

$$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,9}}{y_{0,8}} > \frac{y_{0,8}}{y_{0,6}}, \tag{35}$$

$$\frac{y_{0,8}}{y_{0,6}} < \frac{y_{0,75}}{y_{0,5}} < \frac{y_{0,9}}{y_{0,8}}, \tag{36}$$

$$\frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,9}}{y_{0,8}}. \tag{37}$$

It follows from (32) □ (37) that the observed distributions will more or less systematically differ from the Pareto distribution. In the case of (32) the differentiation of the observed wage distribution is higher; in the case of (33) the differentiation will be lower than in the case of Pareto distribution. Some bias occurs in cases (34), (35), (36) and (37) as well (but cannot be so specified). Systematical bias should be a signal for potential adjustment of the model which could be based for example on adding one or more parameters into the model. These adjustments usually lead to more complicated models. Therefore, the above mentioned bias is often neglected and simple models are preferred even though they lead to some bias.

4 Wage Distribution of Male and Female in Czech Republic in the Years 2001 – 2008

The data used in this article is the gross monthly wage of male and female in CZK in the Czech Republic in the years 2001 – 2008. Data were sorted in the table of interval distribution with opened lower and upper bound in the lowest and highest interval respectively. The source is the web page of the Czech statistical office. The following quantiles were calculated (see the table 1).

Table 1. Median $y_{0,50}$ (in CZK), 6th decile $y_{0,60}$ (in CZK), upper quartile $y_{0,75}$ (in CZK), 8th decile $y_{0,80}$ (in CZK), 9th decile $y_{0,90}$ (in CZK) and 99th percentile $y_{0,99}$ (in CZK) of gross monthly wages in Czech Republic in the years 2001 □ 2008, total and split up to male and female separated.

Total	**Year**	$y_{0,50}$	$y_{0,60}$	$y_{0,75}$	$y_{0,80}$	$y_{0,90}$	$y_{0,99}$
	2001	12 502	14 042	16 987	18 254	23 319	44 921
	2002	15 545	17 125	20 215	22 193	27 754	47 172
	2003	16 735	18 458	22 224	23 797	29 590	47 719
	2004	17 709	19 557	23 077	24 849	31 082	56 369
	2005	18 597	20 566	24 470	26 328	33 292	56 852
	2006	19 514	21 564	25 675	27 693	35 230	57 326
	2007	20 987	23 227	27 590	29 900	37 892	66 395
	2008	22 310	24 696	29 553	31 769	40 541	68 828

Male	Year	$y_{0,50}$	$y_{0,60}$	$y_{0,75}$	$y_{0,80}$	$y_{0,90}$	$y_{0,99}$
	2001	14 152	15 781	19 037	20 697	26 264	46 781
	2002	16 985	18 667	22 604	24 199	31 101	48 047
	2003	18 240	20 116	24 145	26 041	34 564	48 417
	2004	19 344	21 321	25 306	27 286	34 819	57 514
	2005	20 281	22 446	26 822	28 989	37 211	57 808
	2006	21 199	23 460	28 090	30 525	39 381	58 104
	2007	22 933	25 366	30 284	32 663	42 815	70 522
	2008	24 498	27 115	32 343	35 105	46 375	72 338

Female	Year	$y_{0,50}$	$y_{0,60}$	$y_{0,75}$	$y_{0,80}$	$y_{0,90}$	$y_{0,99}$
	2001	10 770	12 187	14 655	15 700	18 904	37 526
	2002	13 746	15 181	17 727	18 903	23 291	43 339
	2003	14 831	16 453	19 281	20 628	24 637	44 883
	2004	15 642	17 303	20 293	21 560	25 776	50 776
	2005	16 454	18 211	21 426	22 804	27 503	52 508
	2006	17 311	19 202	22 530	23 966	29 082	54 054
	2007	18 390	20 392	24 024	25 924	31 338	58 649
	2008	19 399	21 600	25 558	27 215	33 405	63 628

Table 2. The rates of quantiles $y_{0,75}/y_{0,50}$, $y_{0,80}/y_{0,60}$ and $y_{0,90}/y_{0,80}$ of the wage distributions in the years 2001 □ 2008 and its relations.

Total	Year	$\frac{y_{0,75}}{y_{0,50}}$	$\frac{y_{0,80}}{y_{0,60}}$	$\frac{y_{0,90}}{y_{0,80}}$	Relations between quantile rates
	2001	1,358 815	1,299 910	1,277 456	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)
	2002	1,300 422	1,295 897	1,250 612	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)
	2003	1,327 994	1,289 216	1,243 457	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)
	2004	1,303 112	1,270 608	1,250 812	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)
	2005	1,315 815	1,280 162	1,264 514	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)
	2006	1,315 734	1,284 213	1,272 161	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)
	2007	1,314 623	1,287 295	1,267 291	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)
	2008	1,324 653	1,286 403	1,276 118	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$ (33)

Male	Year	$\frac{y_{0,75}}{y_{0,50}}$	$\frac{y_{0,80}}{y_{0,60}}$	$\frac{y_{0,90}}{y_{0,80}}$	Relations between quantile rates	
	2001	1,345 148	1,311 556	1,268 936	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2002	1,330 847	1,296 386	1,285 203	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2003	1,323 680	1,294 561	1,327 273	$\frac{y_{0,8}}{y_{0,6}} < \frac{y_{0,75}}{y_{0,5}} < \frac{y_{0,9}}{y_{0,8}}$	(36)
	2004	1,308 222	1,279 734	1,276 084	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2005	1,322 543	1,291 532	1,283 632	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2006	1,325 086	1,301 135	1,290 146	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2007	1,320 542	1,287 669	1,310 810	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,9}}{y_{0,8}} > \frac{y_{0,8/}}{y_{0,6}}$	(35)
	2008	1,320 230	1,294 671	1,321 037	$\frac{y_{0,8}}{y_{0,6}} < \frac{y_{0,75}}{y_{0,5}} < \frac{y_{0,9}}{y_{0,8}}$	(36)

Female	Year	$\frac{y_{0,75}}{y_{0,50}}$	$\frac{y_{0,80}}{y_{0,60}}$	$\frac{y_{0,90}}{y_{0,80}}$	Relations between quantile rates	
	2001	1,360 723	1,288 227	1,204 113	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2002	1,289 624	1,245 137	1,232 163	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2003	1,300 019	1,253 747	1,194 319	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2004	1,297 375	1,246 052	1,195 526	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2005	1,302 189	1,252 237	1,206 076	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2006	1,301 488	1,248 134	1,213 470	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2007	1,306 362	1,271 283	1,208 841	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)
	2008	1,317 491	1,259 954	1,227 448	$\frac{y_{0,75}}{y_{0,5}} > \frac{y_{0,8}}{y_{0,6}} > \frac{y_{0,9}}{y_{0,8}}$	(33)

From the table 2 we can see that, with the exception of male in the year 2003, 2007 and 2008, all other wage distributions have lower differentiation than Pareto distribution. The systematical error occurred also in the case of male in the year 2003, 2007 and 2008. It follows from the empirical criterion (16) and from the table 2 that in all cases the differences between the rates of the considered quantiles are negligible and therefore Pareto distribution can be used as the model of the distribution.

Table 3. Estimated parameters of Pareto distribution for different choice of the estimation equations.

		Equations used					
		$x_{0,99} = y_{0,99}$, $x_{0,5} = y_{0,5}$		$x_{0,6} = y_{0,6}$, $x_{0,9} = y_{0,9}$		$x_{0,99} = y_{0,99}$, $\frac{x_{0,9}}{x_{0,6}} = \frac{y_{0,9}}{y_{0,6}}$	
		Parameter estimates		**Parameter estimates**		**Parameter estimates**	
Total	**Year**	x_{P0}	b	x_{P0}	b	x_{P0}	b
	2001	44 921	0,326 952	54 143	0,365 843	44 921	0,365 843
	2002	47 172	0,283 758	61 890	0,348 293	47 172	0,348 293
	2003	47 719	0,267 846	64 800	0,340 425	47 719	0,340 425
	2004	56 369	0,295 969	67 096	0,334 192	56 369	0,334 192
	2005	56 852	0,299 456	74 095	0,347 455	56 852	0,347 455
	2006	57 326	0,275 468	79 614	0,354 083	57 326	0,354 083
	2007	66 395	0,294 405	85 426	0,353 045	66 395	0,353 045
	2008	68 828	0,287 978	92 352	0,357 552	68 828	0,357 552

		Equations used					
		$x_{0,99} = y_{0,99}$, $x_{0,5} = y_{0,5}$		$x_{0,6} = y_{0,6}$, $x_{0,9} = y_{0,9}$		$x_{0,99} = y_{0,99}$, $\frac{x_{0,9}}{x_{0,6}} = \frac{y_{0,9}}{y_{0,6}}$	
		Parameter estimates		**Parameter estimates**		**Parameter estimates**	
Male	**Year**	x_{P0}	b	x_{P0}	b	x_{P0}	b
	2001	46 781	0,305 624	61 207	0,367 449	46 781	0,367 449
	2002	48 047	0,265 814	72 613	0,368 246	48 047	0,368 246
	2003	48 417	0,249 540	84 934	0,390 464	48 417	0,390 464
	2004	57 514	0,278 536	78 632	0,353 784	57 514	0,353 784
	2005	57 808	0,267 749	86 165	0,364 658	57 808	0,364 658
	2006	58 104	0,257 739	93 098	0,373 653	58 104	0,373 653
	2007	70 522	0,287 153	102 142	0,377 610	70 522	0,377 610
	2008	72 338	0,276 777	113 087	0,387 128	72 338	0,387 128

		Equations used					
		$x_{0,99} = y_{0,99}$, $x_{0,5} = y_{0,5}$		$x_{0,6} = y_{0,6}$, $x_{0,9} = y_{0,9}$		$x_{0,99} = y_{0,99}$, $\frac{x_{0,9}}{x_{0,6}} = \frac{y_{0,9}}{y_{0,6}}$	
		Parameter estimates		Parameter estimates		Parameter estimates	
Female	**Year**	x_{P0}	b	x_{P0}	b	x_{P0}	b
	2001	37 526	0,319 087	39 196	0,316 679	37 526	0,316 679
	2002	43 339	0,293 539	47 418	0,308 749	43 339	0,308 749
	2003	44 883	0,283 055	48 172	0,291 217	44 883	0,291 217
	2004	50 776	0,300 989	49 971	0,287 505	50 776	0,287 505
	2005	52 508	0,296 625	54 551	0,297 414	52 508	0,297 414
	2006	54 054	0,291 062	57 954	0,299 456	54 054	0,299 456
	2007	58 649	0,296 461	63 977	0,309 955	58 649	0,309 955
	2008	63 628	0,303 636	68 917	0,314 516	63 628	0,314 516

The 99th percentile will be considered as a characteristic of the maximum wage. The parameters of the Pareto distribution are estimated using the above described methods.

First we consider the conditions $x_{P0} = y_{P0}$ a $x_{P1} = y_{P1}$ and we chose median as the second quantile, i.e. $x_{0,99} = y_{0,99}$ a $x_{0,5} = y_{0,5}$. We estimate the parameter b using the formula (21). The summary of the parameter estimates is in the table 3.

Next we apply the conditions $x_{P1} = y_{P1}$ and $x_{P2} = y_{P2}$ and we choose 6th and 9th decile for y_{P1} and y_{P2}. We use the formulas (24) and (25) to estimate the parameters. The summary of the parameter estimates is in the table 3.

Parameters of the Pareto distribution can also be estimated using the equations $x_{P0} = y_{P0}$ and $x_{P2}/x_{P1} = y_{P2}/y_{P1}$. We choose the 9th and 6th decile in the rate y_{P2}/y_{P1}. In this case we use the relations (26) and (27) to estimate the parameters. The summary of the parameter estimates is also in the table 3.

In the end we also estimate the parameters of the Pareto distribution using the least squares method. We use the relations (30) and (31). In this method we choose 5th, 6th, 7th, 8th and 9th deciles of the observed wage distribution, i.e. $k = 5$. Parameters estimated using the least squares method are summarized in the table 4.

Table 4. Parameters estimated using the least squares method.

Year	Total		Male		Female	
	Parameter estimates		Parameter estimates		Parameter estimates	
	x_{P0}	b	x_{P0}	b	x_{P0}	b
2001	56 562	0,379 911	63 774	0,379 912	42 520	0,341 047
2002	64 026	0,358 469	73 770	0,372 825	49 188	0,320 682
2003	67 219	0,351 034	85 080	0,391 617	51 125	0,309 187
2004	69 311	0,344 615	80 310	0,360 986	52 763	0,303 849
2005	76 310	0,356 935	88 251	0,372 535	57 413	0,312 826
2006	81 721	0,362 626	95 225	0,381 012	60 917	0,315 022
2007	88 022	0,362 359	103 405	0,383 183	67 572	0,325 878
2008	94 849	0,366 387	114 131	0,391 293	72 463	0,330 659

Table 5. Sums of the absolute differences of the observed and theoretical frequencies.

Total	Year	Equations used			Least squares method
		$x_{0,99} = y_{0,99}$, $x_{0,5} = y_{0,5}$	$x_{0,6} = y_{0,6}$, $x_{0,9} = y_{0,9}$	$x_{0,99} = y_{0,99}$, $\frac{x_{0,9}}{x_{0,6}} = \frac{y_{0,9}}{y_{0,6}}$	
	2001	37 459	23 255	85 795	23 859
	2002	51 358	27 327	171 404	31 658
	2003	73 388	36 520	204 535	39 722
	2004	103 625	64 422	249 348	66 249
	2005	167 946	69 930	353 661	68 679
	2006	157 094	68 849	426 442	69 104
	2007	268 740	260 786	322 437	262 224
	2008	282 396	253 373	372 117	257 050

Male	Year	Equations used			Least squares method
		$x_{0,99} = y_{0,99}$, $x_{0,5} = y_{0,5}$	$x_{0,6} = y_{0,6}$, $x_{0,9} = y_{0,9}$	$x_{0,99} = y_{0,99}$, $\frac{x_{0,9}}{x_{0,6}} = \frac{y_{0,9}}{y_{0,6}}$	
	2001	20 603	10 089	56 291	9 959
	2002	33 576	19 711	111 796	20 298
	2003	47 909	23 576	96 863	23 747
	2004	60 241	32 457	178 858	33 076
	2005	81 505	35 349	220 276	36 321
	2006	96 789	37 737	250 764	37 653
	2007	140 965	138 678	202 143	139 428
	2008	135 960	133 953	173 262	135 089

		Equations used			
Female	**Year**	$x_{0,99} = y_{0,99}$, $x_{0,5} = y_{0,5}$	$x_{0,6} = y_{0,6}$, $x_{0,9} = y_{0,9}$	$x_{0,99} = y_{0,99}$, $\frac{x_{0,9}}{x_{0,6}} = \frac{y_{0,9}}{y_{0,6}}$	**Least squares method**
	2001	24 256	23 926	23 687	21 270
	2002	23 697	16 716	42 148	18 595
	2003	37 215	30 902	40 237	30 011
	2004	45 429	41 416	45 460	40 957
	2005	51 793	41 615	52 493	41 449
	2006	58 014	41 137	74 302	41 812
	2007	138 241	128 854	150 258	127 313
	2008	140 955	132 125	155 071	131 224

The values of the sum of absolute differences of observed and theoretical absolute frequencies of all intervals calculated for all cases considered wage distributions are in the table 5. In the case of the theoretical frequencies at first we determined theoretical probabilities using the formula (8). From these we determined theoretical absolute frequencies.

If we will prick the theoretical probabilities of separate wage intervals π_j, $j = 1, 2, \ldots, k$, we will obtained the theoretical absolute frequencies as $n \,.\, \pi_j$, where n is sample size. Then we can calculate the sum of absolute differences of observed absolute frequencies n_j and theoretical absolute frequencies $n \,.\, \pi_j$ of all intervals

$$S = \sum_{i=1}^{k} | n_j - n\,\pi_j |, \tag{38}$$

which are in the table 5. In the interest we can compare such results computed for lognormal distribution as a model of the whole wage distributions, see [1].

The question of suitability of a particular distribution using χ^2 test, when we work with such large datasets, was described for example in [1]. All calculated values of χ^2 criterion (high-order thousands) tends to reject every null hypothesis about supposed distribution in the case of such large sample size.

5 Conclusions

The appropriateness of particular modifications of the Pareto distribution can be evaluated comparing the theoretic and empirical frequencies. It is possible to compare both the absolute and relative differences between the theoretic and observed empirical distributions. In this article we used the absolute differences. The values of sums these differences are in the table 5. The values seem to be relatively high. The question of appropriateness of a given theoretic wage distribution in the case of large samples was described for example in [1] or [2]. Some more general conclusions can be made from the values of the absolute differences of observed and theoretic distributions.

With the exception of the wage distribution of women in the year 2001 the worst results are achieved using the equation $x_{0,99} = y_{0,99}$ and setting the ratio of other two quantiles of the Pareto distribution $x_{0,9}/x_{0,6}$ equal to the ratio $y_{0,9}/y_{0,6}$ of the corresponding empirical quantiles. This fact is less obvious for female distribution and most obvious for total distribution. This is also due to the larger sample size of the total sample (in comparison with the sample size of the sub groups of male and female). Again with the exception of the wage distribution of women in the year 2001 the second worst model is the estimate based on the equations $x_{0,99} = y_{0,99}$ and $x_{0,5} = y_{0,5}$. This fact is again less obvious for female distribution and most obvious for total distribution. In the case of the wage distribution of women in the year 2001 the worst estimate is based on the equations $x_{0,99} = y_{0,99}$ and $x_{0,5} = y_{0,5.}$ In the case of the total group is the third worst (second best) method the least squares method (with the exception of the year 2005). The best results are achieved with the method based on the equations $x_{0,6} = y_{0,6}$ and $x_{0,9} = y_{0,9}$. In the case of the total wage distribution in the year 2005 is the third worst method based on the equations $x_{0,6} = y_{0,6}$ and $x_{0,9} = y_{0,9}$ and the best method is the least squares method. In the case of the wage distribution of male (with the exception of the years 2001 and 2006) the third worst (second best) results are again achieved using the least squares method. The best results are achieved with the method based on the equations $x_{0,6} = y_{0,6}$ and $x_{0,9} = y_{0,9}$. In the years 2001 and 2006 (set of men) is the third worst method the method based on the equations $x_{0,6} = y_{0,6}$ a $x_{0,9} = y_{0,9}$ and the best is the least squares method. In the case of the female group (with the exception of the years 2001, 2002 and 2006) is the third worst (second best) method based on the equations $x_{0,6} = y_{0,6}$ and $x_{0,9} = y_{0,9}$ and the most precise results are achieved with the least squares method. In the years 2001, 2003, 2004 and 2005 was for the group of women the most precise the least squares method. The very best method for the group of male in the year 2001 was the least squares method. In this case other methods had much higher values of the above mentioned sum of absolute differences.

From the above described comparison, it is obvious that the simplest parameter estimating methods can be in the case of the Pareto distribution competing with more advanced methods.

References

1. Bílková, D.: Application of Lognormal Curves in Modeling of Wage Distributions. 7th International Conference APLIMAT (2008) − Applied Mathematics (Journal of Applied Mathematics), pp. 341 − 351 + CD. Bratislava (2008). ISBN 978-80-89313-02-0.
2. Bílková D.: Modelling of Wage Distributions in Czech Republic in the Years 2004 and 2005 Using Lognormal Curves and Curves of Pearson`s and Johnson`s system. Statistics 2/2008, pp. 149 – 166. Prague (2008). ISSN 0322-788X.
3. Novák, I.: Development of Wage Distributions in the National Economy and Trade of Czechoslovak Republic in the Years 1959 □ 1964 and the Possibilities of their Extrapolation (Inaugural Dissertation). Prague University of Economics, Prague (1965)
4. Novák, I.: Pareto Distribution As a Model of Wage Distributions (Development). Prague University of Economics, Prague (1966)

Approaches to Combining Forecasts

Dagmar Blatná

University of Economics Prague, W.Churchill sq.4
130 67 Prague 3, Czech Republic
blatna@vse.cz

Abstract. An extensive body of literature has shown that combining forecasts can improve forecast accuracy. Since pioneering work of Bates and Granger (1969), many techniques of combining forecasts has been developed. The main aim of the paper submitted is to describe key moments related to this subject, namely with respect to new approaches based on applications of robust methods..

Keywords: Combining forecasts technique, averaging, regression, robust regression methods

1 Introduction

The combining forecasts is a method that combines several different forecasting models in an appropriate way. The main purpose of combining forecasts is in better using of useful information provided by different forecasting models in order to improve the forecast accuracy. The combination of forecasts is a simple, pragmatic and sensible way to possibly produce better forecasts.

The systematic study of combining forecasting methods started in the late 1960s (Reid, 1969; Bates and Granger, 1969). Since then, the theories and applications of combining forecasts have become an important and interesting area in forecasting. Bates and Granger (1969) first introduced the idea of combining forecasts as a way of improving accuracy and since the various forecast combination techniques have been made to develop and improve the various forecast combination methods through empirical testing and simulations.

Generally, combining forecasting methods can be classified into two categories: the linear and the nonlinear ones. The linear combination methods are the simple and the weighted combination of forecasting models and the combination approaches are based on regression analysis. These methods are based on the assumption that there exists a linear relationship among the combined forecasting models. However, when some of the combined forecasting methods are derived from nonlinear models, or the conditional expectation, on which each individual forecasting model is based, is a nonlinear function of the information set, then a linear combination of individual forecasting methods is not the optimal method. The second category is composed of nonlinearly combined forecasting methods such as state-space methods that attempt to model non-stationarity in the combining weights. Currently, all types of artificial neural network methods are included in this category.

2 Averaging

The combination forecast is given by:

$$f_{c,t} = \sum_{i=1}^{n} w_i f_{i,t} \tag{1}$$

where $f_{i,t}$ is the *i*-th single forecast, *i*=1,…,*n*, $f_{c,t}$ is the combined forecast generated by the *n* single forecasts , and w_i is the combination weight assigned to f_i.

2.1 The Simple Average of Forecasts

In combining the forecasts generated by two or more models, it is important to decide the weights which will be assigned to each of the participating models. In the simple forecasting combination, the combination weight is assigned equally to each of the forecasts as follows:

$$w_i = \frac{1}{n} \tag{2}$$

The simple average method is a straightforward combination technique, empirical results show that this method can generate reliable forecasts in many situations. The simple combination forecasts compute the combination forecast without regard to the historical performance of the individual forecast and do not take into account the relative accuracy of the individual forecasting models that are combined.

The Robust Averages of Forecasts

It is well known that simple averages themselves are quite sensitive to extreme values, and forecasts can sometimes vary considerably. Because of this, some authors (e.g. Jose (2008) [11]) have suggested robust alternatives to use. Trimmed and Winsorized means provide forecasts which are slightly more accurate than the mean, and reduce the risk of high errors.

The α -trimmed mean is computed as

$$T_i(\alpha) = \frac{\sum_{i=g+1}^{n-g} f_{(i)} + (g - n\alpha)\left[f_{(g)} + f_{(n-g+1)}\right]}{n(1-2\alpha)} \tag{3}$$

where α = trimming, $g = [n\alpha] = \text{int}[n\alpha]$, $f_{(i)}$ is *i*-th order forecast.

The α-trimmed mean is easy computed and can be easily understood. For example, if $\alpha = 0.25$, them $T(0.25)$ is the average of the middle 50% of the order statistics, refered to as “the midmean” by Tukey.

The α-Winsorized mean is defined by

$$W_i(\alpha) = \frac{1}{n}\left[g.f_{(g)} + \sum_{i=g+1}^{n-g} f_{(i)} + g.f_{(n-g+1)} \right] \tag{4}$$

where $0 < \alpha < 1/2$ and $g = [n\alpha]$ is the largest integer k satisfying $k \leq n\alpha$.

2.2 Weighted Averages Combinining

Even though the use of equal weights for each of the individual forecasts offers the advantage of simplicity and also precludes the forecaster's own bias in the selection of weighting factors, there may be a good reason for weighting one individual forecast more than another. Many methods and approaches have been proposed in this area. Only those which frequently occur in applications are presented in this contribution.

Variance-Covariance method

When two individual forecasts are consistent over time then the "optimal" method proposed by Bates and Granger may be used. This method minimizes the variance of the forecasts errors over the time period covered. The weight assigned to the first forecast model is calculated in the following manner (the second forecast model would receive a weight of (1-*w*)):

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{5}$$

where σ_1^2, σ_2^2 are the variances of forecast errors for *i*-th model, ρ is the coefficient of correlation between the errors in the first set of forecasts and those in the second set.

Generally, the variance-covariance method determines the weight vector according to:

$$\boldsymbol{w} = \frac{\boldsymbol{u}\sum^{-1}}{\boldsymbol{u}\sum^{-1}u} \tag{6}$$

with the constraint $w_i \geq 0$. In equation (6) $\boldsymbol{w} = (w_1,...,w_n)'$, $\sum_{i=1}^{n} w_i = 1$; $\sum$ denotes the sample covariance matrix; and u is a conformable column vector of ones: $(1,...,1)'$. In practice we estimate $\boldsymbol{W}$ by replacing $\sum$ with an estimate

$$\hat{\sum}_{ij} = v^{-1} \sum_{t=T-v+1}^{T} e_{it}e_{jt} \tag{7}$$

where v means a number of the most recent observations, T is the total number of periods for which there is a history of forecasts errors.

This method calculates the weights by taking the historical performance of the individual forecasts into consideration and may be assigned as moving sample approach using all variance and covariance information and thus the method becomes an adaptive approach to combining forecasts. Granger and Ramanathan showed that this method is equivalent to a least squares regression in which the constant is suppressed and the weights are constrained to sum to one (see in 3.3).

Mowing Sample Approach Which Ignores Covariance Information

In this method weights are specified as follows

$$w_i = \frac{\left(\sum_{t=T-v+1}^{T} e_{i,t}^2 \right)^{-1}}{\left\{ \sum_{j=1}^{n} \left(\sum_{t=T-v+1}^{T} e_{j,t}^2 \right)^{-1} \right\}} \tag{8}$$

An 'Adaptive' Scheme Which Ignores Covariance Information

The weights are computed as

$$w_i = \alpha w_{i,T-1} + \left[(1-\alpha) \frac{\left(\sum_{t=T-v+1}^{T} e_{i,t}^2 \right)^{-1}}{\left\{ \sum_{j=1}^{n} \left(\sum_{t=T-v+1}^{T} e_{j,t}^2 \right)^{-1} \right\}} \right] \tag{9}$$

where $0 < \alpha < 1$. Smaller values of α imply that greater weights are given to recent observations.

Discount MSFE (Mean Square Forecast Error) Method

This approach weights recent observations more heavily than distant ones, and computes the combination forecast as a weighted average of the individual forecasts where the weights depend inversely on the historical performance of each individual forecasts. The weights are

$$w_i = \frac{\left(\sum_{t=1}^{T} \gamma^{T-t+1} e_{i,t}^2 \right)^{-1}}{\left\{ \sum_{j=1}^{n} \left(\sum_{t=1}^{T} \gamma^{T-t+1} e_{j,t}^2 \right)^{-1} \right\}} \tag{10}$$

where $0 < \gamma < 1$ is the discounting factor, *T* denote the observation lengths used to obtain the weights. For γ =1 this method corresponds to scheme (8).

The Approach Using All Variance and Covariance Information and Discounts Weights

In this approach the weights are defined as

$$\boldsymbol{W} = \frac{\sum^{\wedge -1} i}{\boldsymbol{i}\sum^{\wedge -1} i} \text{ where } \quad \left(\sum^{\wedge -1}\right)_{i,j} = \frac{\sum_{t=1}^{T} \gamma^t e_{it} e_{jt}}{\sum_{t=1}^{T} \gamma^t}, \quad \gamma \geq 1 \qquad (11)$$

Smaller values of γ imply that greater weights are given to recent observations.

The Bayesian Approach

There have been a number of attempts at using Bayesian analysis for forecast combination. Bunn [4] suggests a Bayesian methodology to combine forecasts based on the subjective probability. This approach develops the forecast combination as $f_c = p'f$ where *p'* is a simplex of probabilities which can be assessed and revised in a Bayesian manner. Each individual weight is interpreted as the probability that its respective forecast will perform the best (in the smallest absolute error sense) on the next occasion. Each probability is estimated as the fraction of occurrences in which its respective forecasting model has performed the best in the past. It is a robust nonparametric method of achieving differential weights with intuitive meaning which performs well when there is relatively little past data and/or when the decision maker wishes to incorporate expert judgement into the combining weights. Later a Bayesian model based on the magnitude of the relative performance of the forecasting models has been suggested. The Bayesian approach is not explained in detail in this contribution.

3 The Linear Regression Methods

Utilization of the framework of linear regression model, namely to consider forecasted process as the response variable and the individual forecasts as explanatory ones have been used already by Bates and Granger.

In most of papers, properties of the combinations which were created by means of linear regression model were investigated, and they tried to answer such questions as whether the intercept should be included or not, whether some constraints should be imposed on the coordinates of the estimator of coefficients of model, etc. There are three frequently applied basic regression models in combining forecast.

3.1 Unrestricted Regression Model

$$y_{t+h} = \beta_0 + \sum_{i=1}^{p} \beta_i f_{i,t} + \varepsilon_{t+h} \tag{12}$$

where y are the past actual values of interest, the independent variables $\{f_{i,t}\}$ are a sequence of forecasts of $\{y_{t+h}\}$ made at time t, h is the forecast horizon, ε_t is the error term following the OLS assumptions.

An unrestricted regression-based approach would appear to be the natural choice. However, many authors advice that some care should be taken. In case of including more forecast to be combined, the variable must be either stationary or made stationary. Another issue is that the forecast errors resulting from unrestricted least squares combining are likely to be serially correlated. The next problem with the unconstrained regression approach is multicollinearity. So, the unrestricted regression model is usually preferable only if the different forecasts are all unbiased.

3.2 Regression Model Without Intercept Term

In this method the constituent forecasts are used as regressors in an ordinary least squares (OLS) regression with the inclusion of a constant. This model may result in biased forecasts.

$$y_{t+h} = \sum_{i=1}^{p} \beta_i f_{i,t} + \varepsilon_{t+h} \tag{13}$$

3.3 Regression Model With Restricted Weights

Here the least squares regression is performed with the inclusion of a constant but the weights are constrained to sum to one.

$$y_{t+h} = \sum_{i=1}^{p} \beta_i f_{i,t} + \varepsilon_{t+h} \qquad \sum_{i'1}^{p} \beta_i = 1 \tag{14}$$

Under assumptions of unbiasedness of individual forecasts the most authors argue that this model is better than others. On occasion any of these weights may be negative, in which case interpretation is tenuous. So, many authors advice to use these methods only when weights are positive.

3.4 Regression Models for Integrated Processes

When the data used are realizations of integrated processes, such that their differences are I(0), then the following combination regressions are considered

$$y_{t+h} - y_t = \beta_0 + \sum_{i=1}^{p} \beta_i (f_{i,t} - y_t) + \varepsilon_{t+h} \tag{15}$$

$$y_{t+h} - y_t = \sum_{i=1}^{p} \beta_i (f_{i,t} - y_t) + \varepsilon_{t+h} \tag{16}$$

$$y_{t+1} - y_t = \sum_{i=1}^{p} \beta_i (f_{i,t} - y_t) + \varepsilon_{t+h} \qquad \sum_{i=1}^{p} \beta_i = 1 \tag{17}$$

Many generalizations of the combination regression considered here have been proposed, such as time-varying and other non-linear weighting and regression schemes, dynamic combining regressions, Bayesian shrinkage of combining weights toward equality, nonlinear combining regressions. The basic foundations of some of them are introduced.

3.5 Time –Varying Combining Weights

This approach was proposed by Diebold and Pauly) [7]. In the regression framework one may undertake weighted or rolling estimation of combining regressions, or one may estimate combining regressions with explicitly time varying parameters. The combining regression is based on the most recent v observations. For most applications it is adequate to assume that weighted matrix is diagonal, i.e.

$$\mathbf{W} = diag(w_{11}, ..., w_{TT}) \tag{18}$$

which means that the sum of squares is minimized

$$\sum_{i=1}^{T} w_{tt} \left(y_t - \sum_{i=0}^{n} \beta_i f_{it} \right)^2 \tag{19}$$

A simple method for ensuring that the influence of past observations declines with their distance from the present can is to specify

$$\mathbf{W} = diag(w_1, ..., w_T) \text{ ,where } w_t \geq w_{t-1}, \; t = 1, ..., T. \tag{20}$$

Diebold and Pauly [7] considered five basic weighting schemes:

- Equal weight (standard regression-based combining): $w_{tt} = 1$, for all t. (21)
- Linear: $w_{tt} = t$, for all t (22)

- Geometric: $w_{tt} = \lambda^{T-1}, \quad 0 < \lambda \leq 1,$ or $w_{it} = \lambda^{t}, \quad \lambda \geq 1$ (23)
- t^{λ} ('*t*-lambda'): $w_{tt} = t^{\lambda}, \quad \lambda \geq 0,$ (24)
- Box-Cox: $w_{tt}(\lambda) = \begin{cases} (t^{\lambda} - 1)/\lambda & if\ 0 < \lambda \leq 1 \\ \ln t & if\ \lambda = 0 \end{cases}$ (25)

3.6 Dynamic Combining Regression

Serial correlation is likely to appear in unrestricted regression-based forecast combining regressions. A combining regression with serially correlated disturbances is a special case of a combining regression that includes lagged dependent variables and lagged forecasts.

3.7 Bayesian Shrinkage Forecasts

In this method the weights can be viewed as a Bayesian estimator. The least-squares weights and the prior weights then emerge as polar cases for the posterior-mean combining weights. The actual posterior mean combining weights are a matrix weighted average of those for the two polar cases.

3.8 Robust Regression Methods

It is well known that the least squares regression is extremely sensitive to any error, or even atypical values in data set. The design matrix is created from the past forecasts, among which one or few values which are considerably far from others may sometimes appear, it is, some leverage points among the data may be expected. Similarly among the past values of the forecast process some outlying values can exist which are not necessarily wrong values, but the presence of which can negatively influence the results of combining forecasts. These circumstances imply that better results may be expected when applying robust methods instead of least squares. That is why one may give priority to some robust methods for estimating regression coefficients.
We consider the regression model

$$\mathbf{Y} = \mathbf{F}_t.\boldsymbol{\beta}_o + \boldsymbol{\varepsilon}_t \tag{25}$$

where $\mathbf{Y} = (Y_1,...,Y_t)^T$ represents the forecasted process (response variable), $\mathbf{F}_t = (f_{ij}), \quad i = 1,...,t,\ j = 1,...,k$ is the design matrix containing in the first column vector of ones, i.e. $f_{i1} = 1$ for i =1,…,t and the remaining k-1 columns are created

from (k-1) forecast, $\boldsymbol{\beta}_0 = (\beta_{01},...,\beta_{0k})^T$ is used to represent the vector of regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1,...,\varepsilon_t)^T$ is a vector containing i.i.d. random variables.

Some discussion concerning possibility of applying some robust methods (see [3],[7],[9],[18] have appeared. Employing the L_1-technique or trimmed least squares based on regression quantiles have been recommended as the first, later application of *M*-regression and at last stage robust regression methods with high breakdown point (especially *LTS* regression).

The L_1 -Norm Regression

The original proposal of Koenker and Bassett considered a linear combination of regression α-quantiles, 0< α <1, defined as any solution to the minimization problem:

$$\hat{\beta}_\alpha = \arg\min_{\beta \in R_p} \sum_{i=1}^{n} \rho_\alpha (Y_i - f_i'\beta) \tag{26}$$

where $\rho_\alpha(u) = \alpha|u|\mathrm{I}[u \geq 0] + (1-\alpha)|u|\mathrm{I}[u < 0]$

L_1 -norm estimator, i.e., the estimator minimizes $\sum |Y_i - f_i'\beta|$ is a special case for α =0.5. It is *LAD* (least absolute deviations) estimator or regression median $\hat{\beta}_{0.5}$.

The Trimmed Least Square (*TLS)* Regression ¨

TLS regression is based on Ruppert's and Carroll's generalization of the α-trimmed mean proposal. For some fixed $\alpha_1, \alpha_2, 0 < \alpha_1 < \alpha_2 < 1$

$$\beta^{TLS} = \arg\min_{\beta \in R_p} \sum_{i \in I_{\alpha_1,\alpha_2}} (Y_i - f_i'\beta)^2 \tag{27}$$

where $I_{\alpha_1,\alpha_2} = \left\{ i \middle| 1 \leq i \leq n, x_i'\hat{\beta}_{\alpha_1} < Y_i < x_i'\hat{\beta}_{\alpha_2} \right\}$.

The performance of both these methods may be seriously affected by the presence of leverage points among carries and hence it is necessary some care when using them. Both methods are able to cope with outliers (for the response variable) which in forecasts-combining-problem may be more frequently occuring case than appearance of leverage points.

The *M*-regression

The *M*-estimators of regression coefficients is defined by minimizing a sum of functions ρ of residuals

$$\text{minimize} \sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}}\right) \tag{28}$$

where residual $e_i = y_i - \beta f_i$, ρ is a smooth real almost everywhere differentiable function. (Considering $\rho(z) = z^2$ we obtain, as a special case of the *M*-estimators, the *LS* estimator).

Assuming that the derivate of function ρ *e*xists and $\psi = \rho'$, than we usually consider the *M*-estimator as solution to the system of equations

$$\sum_{i=1}^{n} \psi\left(\frac{e_i}{\hat{\sigma}}\right) f_i = 0 \tag{29}$$

with $\psi = \rho'$. If ρ is convex and ψ continous, the definitions in equations (28) and (29) are equivalent. Various ψ-functions lead to various *ME'*s.

Under assumptions of unbiasedness of individual forecasts in the regression combining the combination without intercept and with constraint imposed on the estimate of regression coefficients that they sum to one, is preferred (see chap. 3.3).

The conviction that robust methods have an important role to play in the formulation of the combination of forecasts has lead to explore the above topic when *M*-estimators of the regression coefficients are used in place of *LS*-estimator. Augliar, Rubio and Visek [3] proved then the constrained *M*-regression model produced forecasts that are asymptotically unbiased and more efficient than those from the unconstrained one.

The Least Trimmed Squares Regression

But some other robust technique can be more adequate than *M*-estimators when the individual forecasts include some „atypical" values - outliers are used as regressors, the problem of leverage points may appeal. In these cases technique with high breakdown point as the least trimmed squares or the least median of squares would be preferable. Visek [18] gives support to robust *TLS* method which may be effective in combining the forecasts.

The least trimmed squares (*LTS*) estimator (proposed by Rousseeuw [15]) are obtained by

$$\text{minimize} \sum_{i=1}^{h} e_{(i)}^2 , \tag{30}$$

where $e_{(i)}$ is the *i*-th order statistic among the squared residuals written in the ascending order, $h = \left[t/2\right] + \left[(t+1/2)\right]$ and [*x*] denotes the largest integer which is less or equal to *x*. The value of *h* implies the level of robustness of estimator, namely its breakdown point.

5 Conclusion

Many combining forecast methods in the forty years since the seminal work on combining have been suggested and can be found in literature. In this contribution, only some basic approaches were mentioned.

Despite a large literature on combining forecasts, there in no consensual agreement regarding the choice of which method or approach to use. Frequently, recommendations of various authors are contradictory. However, one result is common: combining forecasts is a well-established procedure for improving forecasting accuracy which takes advantage of the availability of both multiple information and computing resources for data-intensive forecasting and the research in this area has to be continued.

References

1. Anandalingam, G., Chen,L. Linear Combination of Forecasts: A General Bayesian Model. Journal of Forecasting, Vol.8 199-214 (1989), ISSN 0277-6693
2. Armstrong, J. S. Combining Forecasts: The End of the Beginning or the Beginning of the End? International Journal of Forecasting (1989), 5, 585-588, ISSN 0169-2070
3. Augilar,L.Z., Rubio,A.M.,Visek,J.A. Combining forecasts using constrained M-estimators. Bulletin 4/1996 of The Czech Econometric Society, pp.61-72, ISSN 1212-074x
4. Bunn, D. W., 'A Bayesian approach to the linear combination of forecasts', Op. Res. Q., 26 (1975), 325-9. ISSN 00303623.
5. De Menezes,L. M., Bunn,D.W., Taylor,J.W. Review of Guidelines for the Use of Combined Forecasts . European Journal of Operational Research, 2000, Vol. 120, pp. 190-204. ISSN 0377-2217.
6. Diebold, F. X., Pauly,P. Structural Change and the Combination of Forecasts. Journal of Forecasting, Vol. 6, 21-40 (1987), ISSN 0277-6693
7. Donaldson,R.G.,Kamstra,M. Forecasts Combining with Neural Networks. Journal of Forecasting, vol. 15 (1996), 49-61, ISSN 0277-6693
8. Granger,C.W.J. Invited Review Combining forecasts – Twenty Years Later. Journal of Forecasting, 8 (1989) 167-173. ISSN 0277-6693
9. Hallman,J., Kamstra,M . Combining Algorithms Based on Robust Estimation Techniques and Co-integrating Restrictions .Journal of Forecasting, Vol. 8, 189-198 (1989), ISSN 0277-6693
10. Hampel,F.R.,Ronchetti,E.M.,Rousseeuw,P.J., Stahel, W.A. Robust Statistics. The Approach Based on Influence Functions. J.Wiley, N.York 1986. ISBN 0-471-82921-8.
11. Jose, V.R.R., Winkler, R.L.. Simple robust averages of forecasts: Some empiricical results. International Journal of Forecasting 24 (2008), 163-169, ISSN 0169-2070.
12 .Maronna, R.A.,Martin,R.D.,Yohai,V.J.. *Robust Statistics*. Theory and Methods. J Wiley, London, 2006, ISBN-13 978-0-470-01092-11
13. Newbold, P., Granger, C.W. J**.** Experience with Forecasting Univariate Time Series and the Combination of Forecasts. J.R. Statist. Soc., A (1974), 137, Part 2, pp.131-164,ISSN 0035-9254
14. Rapach,D.E., Strauss, J.K. Forecasting US Employtment Growth Using Forecast Combining Methods. Journal of Forecasting 27 (2008), pp. 75-93., ISSN 0277-6693
15. Rousseeuw,P.J., Leroy,A.M. Robust Regression and Outlier Detection. J.Wiley, New Jersey

2003. ISBN 0-471-48855-0.
16. Stock, J.H , Watson, M.W. Combination forecasts of Output Growth in a Seven-Country Data Set. Journal of Forecasting 23 (2004), 405-430, ISSN 0277-6693
17. Swanson, N.R., Zeng, T. Choosing Among Competing Econometric Forecasts: Regression-based Forecast Combination Using Model Selection. Journal of Forecasting 20(2001), 425-440, ISSN 0277-6693
18. Víšek, J.A. Combining forecasts using the least trimmed squares. Kybernetika 37 (2001), pp.193-204, ISSN 0023-5954.
19. Visek,J.A. Robust constrained combinations of forecasts. Bulletin vol.5 no 8/1998 The Czech Econometric Society, 53-80. ISSN 1212-074x
20. Winkler, R. L.,Makridakis**, S. :** The Combination of Forecasts'. Journal of the Royal Statist. Soc., (1983), A, 146, 150-157, ISSN 0035- 9254.

Testing the Validity of the Black Version of the Capital Asset Pricing Model

Martin Boďa, Mária Kanderová

Univerzita Mateja Bela v Banskej Bystrici, Ekonomická fakulta,
Tajovského 40, 975 90 Banská Bystrica
martin.boda@umb.sk, maria.kanderova@umb.sk

Abstract. The Capital Asset Pricing Model (CAPM) is a fundamental vehicle by which contemporary financial theory approach the explication of risky assets price determination. The article focuses on empirical verification of the Black version of the CAPM - in theory and practice - and discusses the shortcomings of statistical verification of the CAPM.

Keywords: The Capital Asset Pricing Model (CAPM), the Black version of the CAPM, three implications of the CAPM, likelihood ratio statistics, F statistics, risk premium.

1 The Introduction

Although the Capital Asset Pricing Model (CAPM) is very popular in the sphere of theoretical finance; over the recent 50 years enough evidence has been gathered to disprove its elusive validity. For its validity is tied with glittering assumptions of investing at financial markets, it still remains of appeal to many authors attempting to prove its empirical validity. The failure of statistical tests to recognise the empirical validity of the CAPM motivated several modifications of the CAPM in the belief that they would reflect better the rationality of financial markets, still in consistence with the general theory of investing at financial markets. One of them is the famous Black version of the CAPM. This article focuses upon the derivation of an appropriate statistical framework for testing its empirical validity and on its full implementation on a sample of U. S. data, and is a natural continuation of our earlier work ([2], [1]).

2 The Formula of the CAPM

The foundations of entire investment theory were laid by Harry Markowitz (1952) who in his pioneer work established that the decision-making of an investor goes under the mean-variance (M-V) rule, i. e. an investor aims to build up a portfolio of the highest expected return (mean, M) at the given level of risk that he is willing to accept (variance, V). These theoretical results were further extended and set up the CAPM – by William F. Sharpe (1964) and John Lintner (1965) into the so-called

Sharpe-Lintner version and by Fisher Black (1972) into the respective Black version. These authors were interested in the process of the determination of prices of risky assets at financial markets and founded upon rationality of investors and efficiency of financial markets. It is the Harry Markowitz's assumption of mean-variance efficiency of financial markets that is key to the validity of the CAPM. Should thus the CAPM be found invalid facing the real milieu of financial markets, it would directly imply that financial markets are not efficient in the sense of Harry Markowitz and would subsequently discard all stylized emblems of contemporary financial management.

It is not the aim of the article to investigate into the legitimacy of the model or enlighten its interpretation insomuch as its full derivation or the interpretation is readily available in literature. The reader may familiarize oneself with the theoretical derivation of the CAPM in the aforesaid works of William F. Sharpe, John Lintner and Fisher Black, or e. g. in [4], or in [6].

It must be said that there are many versions of the model and numerous modifications stemming from attempts to best reflect the reality of financial markets, or even from the efforts to apply the CAPM to human capital markets or property markets. Nonetheless, taking the CAPM traditionally in the domain of financial markets, it may appear either in the Sharpe and Lintner variant, or in the Black version. The distinction consists in their attitude towards the existence of lending and borrowing facilities at a riskless rate of interest. Whilst Sharpe and Lintner admit that an investor is able to borrow or lend at a riskless rate of interest, the Black version operates in the environment with the absence of riskless assets. The riskless asset is an asset which yields the given return with probability 1, and its return is identified with the riskless rate of interest. In practical applications, a proxy for the riskless rate of interest such as a government zero coupon rate or a government bond rate is employed instead.

2.1 The Sharpe-Lintner Version of the CAPM

The Sharpe and Lintner formula of the CAPM reads for the expected return of the *i*-th asset at a given time in this way:

$$\mathrm{E}R_i = R_f + \beta_i^{SL}\left(\mathrm{E}R_m - R_f\right), \tag{1}$$

$$\beta_i^{SL} = \frac{\mathrm{cov}\left(R_i, R_m\right)}{\mathrm{D}R_m}, \tag{2}$$

where R_i stands for the return of risky asset i, R_m denotes the return on the market portfolio and R_f is the return on the riskfree asset (i. e. the riskless rate of interest). Here needs be remarked that the riskless rate R_f may be deemed as nonstochastic and constant over time, or it may be considered as a random variable (which is compatible with empirical observations). The expression in the brackets of formula (1) $\mathrm{E}R_m - R_f$ is called the risk premium and measures the expected excess return that an investor can gain by investing into the market portfolio in place of investing into a riskless asset.

2.2 The Black Version of the CAPM

The formula of Black is more general as it operates with the return on the zero-beta portfolio rather than with the riskless rate of interest. Because no financial asset can be unreservedly found riskless as was defined above, the Black model does not resort to the existence of a riskless asset. It only assumes the existence of a portfolio uncorrelated with the market portfolio, independent of market moves. If such a portfolio exists, the beta coefficient of the CAPM is zero, which thus motivates the name the zero-beta portfolio. Generality rests upon the fact that if a riskless asset does exist in the financial market, its beta coefficient is zero and it itself is a zero-beta portfolio. The same holds for portfolios of riskless assets, they, too, are zero-beta portfolios. In the environment where there is no riskless asset, the expected return on the *i*-th asset is supposed to be linearly related to its beta by virtue of the excess of the return on the zero-beta portfolio. The Black version assumes that the expected return of the *i*-th asset at a certain time is governed by the formula

$$\mathrm{E}R_i = \mathrm{E}R_{om} + \beta_i^B \left(\mathrm{E}R_m - \mathrm{E}R_{om}\right) , \tag{3}$$

$$\beta_i^B = \frac{\mathrm{cov}\left(R_i, R_m\right)}{\mathrm{D}R_m} , \tag{4}$$

where R_{om} represents the return on the zero-beta portfolio. The bracketed expression in (3) $\mathrm{E}R_m - \mathrm{E}R_{om}$ is also termed the risk premium and quantifies the expected excess return that an investor can gain by investing into the market portfolio instead of investing into a zero-beta portfolio.

The econometric analysis of this model is comparatively complicated as the zero-beta portfolio is not observable and its return is an unobserved quantity.

2.3 The Testable Implications of the CAPM

It is of interest to check the validity of the CAPM by rigorous statistical testing. Equations (1) and (3) bear three testable implications as stated in [5]:

C1. The relationship between the expected return on a risky asset and its risk expressed by β_i^{SL} or by β_i^B is linear.

C2. β_i^{SL} or β_i^B is the complete measure of risk of a risky asset, and no other measure of risk appears in (1) or (3).

C3. The risk premium $\mathrm{E}R_m - R_f$ or $\mathrm{E}R_m - \mathrm{E}R_{om}$ is positive, which implies that in the milieu of risk-averse investors higher risk is associated with higher expected returns.

Should it be possible to find all three validity implications satisfied, it would be suggestive of the empirical trueness of the CAPM. In contrast, statistical evidence against at least one of them would be at odds with the theoretical development and empirical validity of the model.

2.4 The Econometric Model of the Black Version of the CAPM

The article focuses upon the Black version of the CAPM. The following explication presents, in line with the above-declared aim, an econometric model of this version and develops suitable statistical tests for its empirical verification.

For this purpose the following model of linear regression for risky asset i is taken under advisement:

$$R_i - \eta = \alpha_i + \beta_i \left(R_m - \eta \right) + \varepsilon_i \ , \tag{5}$$

in which the expected return on the zero-beta portfolio is re-denoted into η, the other expectations are removed and in comparison with (3) the parameter α_i and the disturbance term ε_i are added. Having formulated the model for a single risky asset in this fashion, the usual ordinary least squares estimate or the maximum likelihood estimate for β_i coincide with its theoretical counterpart in (4).

It is necessary to test for each asset i in the market

C1. Whether the intercept α_i can be though of as zero.
C2. Whether the coefficient β_i exhausts the entire variation of asset excess returns.
C3. Whether the expected market premium $\mathrm{E}R_m - \eta$ is positive.

Parameter estimation in (5) and testing C1 – C3 must be done not individually for each asset available but for all assets in the market, though.

For the CAPM to be valid, it need be true at an arbitrary time and for all assets and portfolios of them available in the market. Accordingly, N financial assets over T time periods are considered and the Black model reads at any time t

$$\boldsymbol{R}_t - \eta \mathbf{1} = \boldsymbol{\alpha} + \boldsymbol{\beta} \, (R_{mt} - \eta) + \boldsymbol{\varepsilon}_t \ , \tag{6}$$

whereas it is required

$$\begin{gathered} \mathrm{E}\boldsymbol{\varepsilon}_t = \mathbf{0} \ , \quad \mathrm{cov}\boldsymbol{\varepsilon}_t = \boldsymbol{\Sigma} \ (\text{p. d.}) \ , \\ \mathrm{E}R_{mt} = \mu_m \ , \quad \mathrm{D}R_{mt} = \sigma_m^{\,2} \ , \\ \mathrm{cov}(\boldsymbol{\varepsilon}_t \ , R_{mt}) = \mathbf{0} \ . \end{gathered} \tag{7}$$

In (6) and (7) $\boldsymbol{R}_t$ is an $(N \times 1)$ vector of returns for N assets (or portfolios of assets) at time t, $\mathbf{1}$ is the $(N \times 1)$ vector of ones, $\boldsymbol{\alpha}$ denotes the $(N \times 1)$ vector of intercepts and $\boldsymbol{\beta}$ the $(N \times 1)$ vector of beta coefficients of the regression, R_{mt} is a scalar to represent the aforesaid stochastic market return at time t, η is a scalar to represent a constant zero-beta portfolio return, and finally $\boldsymbol{\varepsilon}_t$ stands for the $(N \times 1)$ vector of disturbances at time t.

As the further development bases upon the method of maximum likelihood and model (6) and (7) suffers from the fact that R_{mt} is a random variable, it need be assumed that the asset returns $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_T$ conditional on market returns $R_{m1}, \ldots, R_{mT}$ are independent identically distributed multivariate (N-dimensional) normal.

3 The Methodological Set-up

For a smooth theoretical derivation of the statistical tests several aspects of maximum likelihood theory of point estimation as well one result on the multiavariate statistical theory will be needful.

Let $\xi_1, ..., \xi_n$ be a sequence of independent random vectors with the same distribution associated with probability density function $p(\mathbf{x}, \cdot)$ with respect to some σ-finite measure μ. Let the probability density function $p(\mathbf{x}, \cdot)$ depend upon a set of real vector parameters $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ with dimensions $m_1, ..., m_k$ respectively, all at least equal one. It is required that the density function $p(\mathbf{x}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$ is well-behaved in correspondence with usual regularity conditions. The joint probability density function $L(\mathbf{x}_1, ..., \mathbf{x}_n, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k) \equiv L(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$ of $\xi_1, ..., \xi_n$ is called the likelihood function and its logarithm $l(\mathbf{x}_1, ..., \mathbf{x}_n, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k) \equiv l(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$ the log-likelihood function. The maximum likelihood estimates for $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ may be defined by a system of equations

$$\frac{\partial}{\partial \boldsymbol{\theta}_j} l(\xi_1, \ldots, \xi_n, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \Big|_{\substack{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1 \\ \vdots \\ \boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k}} \equiv \mathbf{0} \tag{8}$$

for all $j \in \{1, ..., k\}$. The covariance matrix of the maximum likelihood estimates for $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ can by obtained by the inverse of the Fisher information matrix $\mathfrak{I}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$ established by the expression

$$\mathfrak{I}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) = \left(-E \frac{1}{n} \frac{\partial^2 l}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \right)_{\substack{i \in \{1, \ldots, k\} \\ j \in \{1, \ldots, k\}}}, \tag{9}$$

so that it holds for the covariance matrix of the maximum likelihood estimate for any parameter $\boldsymbol{\theta}_j$ that

$$\operatorname{cov} \hat{\boldsymbol{\theta}}_j = \left\{ \mathfrak{I}^{-1}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \right\}_{\boldsymbol{\theta}_j \boldsymbol{\theta}_j} \tag{10}$$

for all $j \in \{1, ..., k\}$.

Let us assume that the true value of parameter $\boldsymbol{\theta}_j$ for some $j \in \{1, ..., k\}$ is $\boldsymbol{\theta}_{j0}$ and that the rest $k - 1$ parameters must be estimated by the maximum likelihood method. The log-likelihood function l evaluated at the maximum likelihood estimates for $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ when the true value of parameter $\boldsymbol{\theta}_j$ for some $j \in \{1, ..., k\}$ is not known is termed unconstrained (similarwise, the maximum likelihood estimates are called unconstrained). The log-likelihood function l^* evaluated at the maximum likelihood estimates for $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, ..., \boldsymbol{\theta}_k$ and the true value of parameter $\boldsymbol{\theta}_j$ for some $j \in \{1, ..., k\}$ $\boldsymbol{\theta}_{j0}$ is called constrained (and, correspondingly, the maximum likelihood estimates for the unknown parameters $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, ..., \boldsymbol{\theta}_k$ are called constrained). Then, considering the referred regularity conditions, the likelihood ratio statistics

$$LR = 2\left(l - l^*\right) = \\ = 2\left(l(\xi_1,\ldots,\xi_n,\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_n) - l^*(\xi_1,\ldots,\xi_n,\hat{\boldsymbol{\theta}}^*_1,\ldots,\hat{\boldsymbol{\theta}}^*_{j-1},\boldsymbol{\theta}_{j0},\hat{\boldsymbol{\theta}}^*_{j+1},\ldots,\hat{\boldsymbol{\theta}}^*_k)\right) \tag{11}$$

is distributed asymptotically $\chi^2(m_j)$.

The following theorem is to be found and proven in e. g. [7] and [9]:

Let $\boldsymbol{\xi}_{m\times 1}$ be distributed $N_m(\mathbf{0}, \mathbf{V})$ and $\mathbf{A}_{m\times \mathrm{m}}$ be distributed $\mathrm{W}_m(n, \mathbf{V})$, whilst $n \geq m$ and $\boldsymbol{\xi}$ and $\mathbf{A}$ independent, then

$$F = \frac{n-m+1}{m}\boldsymbol{\xi}'\mathbf{A}^{-1}\boldsymbol{\xi} \tag{12}$$

follows an (exact) $F(m, n - m + 1)$ distribution.

4 The Derivation of the Tests

In this section tests for validating the three implications of the Black version of the CAPM are described. In consistence with model (6), implication 1 requires $\boldsymbol{\alpha} = \mathbf{0}$, implication 2 is that the covariance matrix $\boldsymbol{\Sigma}$ is formed only by virtue of $\boldsymbol{\beta}$ only and implication 3 forces the expected market premium $\mathrm{E}R_m - \eta$ to be positive.

4.1 Testing Implication 1 of the Black Version of the CAPM

The unconstrained model set by equation (6) allows any values for the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Given the assumptions of normality and conditionality on market returns, the probability density of any $\boldsymbol{R}_t$ conditional on market return R_{mt} is

$$p(\mathbf{R}_t \,|\, R_{mt};\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{N}{2}}\left(\det\boldsymbol{\Sigma}\right)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{R}_t-\eta\mathbf{1}-\boldsymbol{\alpha}-\boldsymbol{\beta}(R_{mt}-\eta))'\boldsymbol{\Sigma}^{-1}(\mathbf{R}_t-\eta\mathbf{1}-\boldsymbol{\alpha}-\boldsymbol{\beta}(R_{mt}-\eta))}, \tag{13}$$

and combining it with the assumption of *iid* property, the joint probability density of $\boldsymbol{R}_1$, ..., $\boldsymbol{R}_T$ conditional on market returns R_{m1}, ..., R_{mT}, or the unconstrained likelihood function, runs in the form

$$p(\mathbf{R}_1,\ldots,\mathbf{R}_T \,|\, R_{m1},\ldots,R_{mT};\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}) = L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}) = \\ = (2\pi)^{-\frac{NT}{2}}\left(\det\boldsymbol{\Sigma}\right)^{-\frac{T}{2}} e^{-\frac{1}{2}\sum_{t=1}^{t=T}(\mathbf{R}_t-\eta\mathbf{1}-\boldsymbol{\alpha}-\boldsymbol{\beta}(R_{mt}-\eta))'\boldsymbol{\Sigma}^{-1}(\mathbf{R}_t-\eta\mathbf{1}-\boldsymbol{\alpha}-\boldsymbol{\beta}(R_{mt}-\eta))}. \tag{14}$$

Consequently, the unconstrained log-likelihood function is

$$l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}) = -\tfrac{NT}{2}\log(2\pi) - \tfrac{T}{2}\log\left(\det\boldsymbol{\Sigma}\right) - \\ -\tfrac{1}{2}\sum_{t=1}^{t=T}\left(\mathbf{R}_t-\eta\mathbf{1}-\boldsymbol{\alpha}-\boldsymbol{\beta}(R_{mt}-\eta)\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{R}_t-\eta\mathbf{1}-\boldsymbol{\alpha}-\boldsymbol{\beta}(R_{mt}-\eta)\right). \tag{15}$$

and the respective unconstrained maximum likelihood estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are determined by the solution to the following system of equations

$$\left.\frac{\partial l}{\partial \boldsymbol{\alpha}}\right|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}\\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}\\ \boldsymbol{\Sigma}=\hat{\boldsymbol{\Sigma}}}} = \hat{\boldsymbol{\Sigma}}^{-1}\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \eta\mathbf{1} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}(R_{mt} - \eta)\right) \equiv \mathbf{0},$$

$$\left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}\\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}\\ \boldsymbol{\Sigma}=\hat{\boldsymbol{\Sigma}}}} = \hat{\boldsymbol{\Sigma}}^{-1}\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \eta\mathbf{1} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}(R_{mt} - \eta)\right)\left(R_{mt} - \eta\right) \equiv \mathbf{0}, \tag{16}$$

$$\left.\frac{\partial l}{\partial \boldsymbol{\Sigma}}\right|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}\\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}\\ \boldsymbol{\Sigma}=\hat{\boldsymbol{\Sigma}}}} = -\tfrac{T}{2}\hat{\boldsymbol{\Sigma}}^{-1} + \tfrac{1}{2}\hat{\boldsymbol{\Sigma}}^{-1}\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \eta\mathbf{1} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}(R_{mt} - \eta)\right)\left(\mathbf{R}_t - \eta\mathbf{1} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}(R_{mt} - \eta)\right)'\hat{\boldsymbol{\Sigma}}^{-1} \equiv \mathbf{0}.$$

The resulting maximum likelihood estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}(\eta) &= \bar{\mathbf{R}} - \eta\mathbf{1} - \hat{\boldsymbol{\beta}}(\bar{R}_m - \eta),\\
\hat{\boldsymbol{\beta}} &= \frac{\frac{1}{T}\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \bar{\mathbf{R}}\right)\left(R_{mt} - \bar{R}_m\right)}{s_m^2},\\
\hat{\boldsymbol{\Sigma}} &= \tfrac{1}{T}\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \bar{\mathbf{R}} - \hat{\boldsymbol{\beta}}(R_{mt} - \bar{R}_m)\right)\left(\mathbf{R}_t - \bar{\mathbf{R}} - \hat{\boldsymbol{\beta}}(R_{mt} - \bar{R}_m)\right)',
\end{aligned} \tag{17}$$

whilst $\bar{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{t=T}\mathbf{R}_t$, $\bar{R}_m = \frac{1}{T}\sum_{t=1}^{t=T}R_{mt}$ and $s_m^2 = \frac{1}{T}\sum_{t=1}^{t=T}\left(R_{mt} - \bar{R}_m\right)^2$. It is worth noticing that only the maximum likelihood estimate for $\boldsymbol{\alpha}$ depends on the value of η, which is indicated by being a function of η.

The results for the constrained model are obtained in the same manner. The constrained model is also governed by equation (6), however, with $\boldsymbol{\alpha} = \mathbf{0}$ and with arbitrary values for the remaining parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. The constrained log-likelihood function reads

$$\begin{aligned}
l^*(\mathbf{0}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = &-\tfrac{NT}{2}\log\left(2\pi\right) - \tfrac{T}{2}\log\left(\det\boldsymbol{\Sigma}\right) -\\
&-\tfrac{1}{2}\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \eta\mathbf{1} - \boldsymbol{\beta}(R_{mt} - \eta)\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{R}_t - \eta\mathbf{1} - \boldsymbol{\beta}(R_{mt} - \eta)\right)
\end{aligned}. \tag{18}$$

and produces the maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in the form

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^*(\eta) &= \frac{\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \eta\mathbf{1}\right)\left(R_{mt} - \eta\right)}{\sum_{t=1}^{t=T}\left(R_{mt} - \eta\right)^2},\\
\hat{\boldsymbol{\Sigma}}^*(\eta) &= \tfrac{1}{T}\sum_{t=1}^{t=T}\left(\mathbf{R}_t - \eta\mathbf{1} - \hat{\boldsymbol{\beta}}(R_{mt} - \eta)\right)\left(\mathbf{R}_t - \eta\mathbf{1} - \hat{\boldsymbol{\beta}}(R_{mt} - \eta)\right)',
\end{aligned} \tag{19}$$

both of which are dependent upon the value of η.

For it holds that

$$\begin{aligned}
&\sum_{t=1}^{t=T}\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{\beta}}(R_{mt}-\eta)\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{\beta}}(R_{mt}-\eta)\right)=\\
&=\mathrm{Tr}\sum_{t=1}^{t=T}\left(\mathbf{R}_t-\bar{\mathbf{R}}-\hat{\boldsymbol{\beta}}(R_{mt}-\bar{R}_m)\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{R}_t-\bar{\mathbf{R}}-\hat{\boldsymbol{\beta}}(R_{mt}-\bar{R}_m)\right)=\\
&=\sum_{t=1}^{t=T}\mathrm{Tr}\left(\mathbf{R}_t-\bar{\mathbf{R}}-\hat{\boldsymbol{\beta}}(R_{mt}-\bar{R}_m)\right)\left(\mathbf{R}_t-\bar{\mathbf{R}}-\hat{\boldsymbol{\beta}}(R_{mt}-\bar{R}_m)\right)'\hat{\boldsymbol{\Sigma}}^{-1}=\\
&=\sum_{t=1}^{t=T}\mathrm{Tr}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Sigma}}^{-1}=\sum_{t=1}^{t=T}\mathrm{Tr}\,\mathbf{I}_N=NT\,,
\end{aligned}\tag{20}$$

and, similarly,

$$\begin{aligned}
&\sum_{t=1}^{t=T}\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\beta}}^*(R_{mt}-\eta)\right)'\hat{\boldsymbol{\Sigma}}^{*-1}\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\beta}}^*(R_{mt}-\eta)\right)=\\
&=\mathrm{Tr}\sum_{t=1}^{t=T}\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\beta}}^*(R_{mt}-\eta)\right)'\hat{\boldsymbol{\Sigma}}^{*-1}\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\beta}}^*(R_{mt}-\eta)\right)=\\
&=\sum_{t=1}^{t=T}\mathrm{Tr}\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\beta}}^*(R_{mt}-\eta)\right)\left(\mathbf{R}_t-\eta\mathbf{1}-\hat{\boldsymbol{\beta}}^*(R_{mt}-\eta)\right)'\hat{\boldsymbol{\Sigma}}^{*-1}=\\
&=\sum_{t=1}^{t=T}\mathrm{Tr}\hat{\boldsymbol{\Sigma}}^*\hat{\boldsymbol{\Sigma}}^{*-1}=\sum_{t=1}^{t=T}\mathrm{Tr}\,\mathbf{I}_N=NT\,,
\end{aligned}\tag{21}$$

evaluating the unconstrained log-likelihood function at the respective unconstrained maximum likelihood estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ it is arrived at

$$l(\hat{\boldsymbol{\alpha}}(\eta),\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}})=-\tfrac{NT}{2}\log(2\pi)-\tfrac{T}{2}\log\left(\det\hat{\boldsymbol{\Sigma}}\right)-\tfrac{NT}{2}\,,\tag{22}$$

and evaluating the constrained log-likelihood function at $\boldsymbol{\alpha}=\mathbf{0}$ and at the constrained maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ it is obtained that

$$l^*(\mathbf{0},\hat{\boldsymbol{\beta}}^*(\eta),\hat{\boldsymbol{\Sigma}}^*(\eta))=-\tfrac{NT}{2}\log(2\pi)-\tfrac{T}{2}\log\left(\det\hat{\boldsymbol{\Sigma}}^*(\eta)\right)-\tfrac{NT}{2}\;;\tag{23}$$

and the likelihood ratio statistics in (11) goes into the form

$$\begin{aligned}
LR(\eta)&=2\left(l(\hat{\boldsymbol{\alpha}}(\eta),\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}})-l^*(\mathbf{0},\hat{\boldsymbol{\beta}}^*(\eta),\hat{\boldsymbol{\Sigma}}^*(\eta))\right)=\\
&=T\left[\log\left(\det\hat{\boldsymbol{\Sigma}}^*(\eta)\right)-\log\left(\det\hat{\boldsymbol{\Sigma}}\right)\right].
\end{aligned}\tag{24}$$

The likelihood ratio statistics (24) does depend on η and under the null hypothesis H_0: $\boldsymbol{\alpha}=\mathbf{0}$, if the value of η is known, follows an asymptotic $\chi^2(N)$ distribution. However, the value of η is unknown and must be established, which can be done by virtue of (24). It is apparent that the value of η minimizing the likelihood ratio (24) is the maximum likelihood estimate for η. Having thus defined the maximum likelihood estimate for η in rather an unusual way as

$$\hat{\eta}^*=\arg\min_{\eta\in\Re}LR(\eta)\,,\tag{25}$$

and substituting it into (24), the result is that the likelihood ratio statistics

$$LR(\hat{\eta}^*)=T\left[\log\left(\det\hat{\boldsymbol{\Sigma}}^*(\hat{\eta}^*)\right)-\log\left(\det\hat{\boldsymbol{\Sigma}}\right)\right].\tag{26}$$

loses one degree of freedom and has an asymptotic $\chi^2(N-1)$ distribution instead.

The test grounded in the likelihood statistics (26) is, however, a large sample test. To construct a finite sample test, the F statistics (12) may be employed. To this end, the covariance matrix of the unconstrained maximum likelihood for $\boldsymbol{\alpha}$ need be set-up, which may be effected through the inverse of the Fisher information matrix. The elements of the Fisher information matrix are successively

$$\begin{aligned}
\mathfrak{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} &= -\mathrm{E}\,\tfrac{1}{T}\partial^2\ell/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}\,,\\
\mathfrak{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\mathrm{E}\,\tfrac{1}{T}\partial^2\ell/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta} = \left(s_m^2 + (\bar{R}_m - \eta)^2\right)\boldsymbol{\Sigma}^{-1}\,,\\
\mathfrak{I}_{\boldsymbol{\Sigma}\boldsymbol{\Sigma}} &= -\mathrm{E}\,\tfrac{1}{T}\partial^2\ell/\partial\boldsymbol{\Sigma}\partial\boldsymbol{\Sigma} = \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\,,\\
\mathfrak{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}} &= -\mathrm{E}\,\tfrac{1}{T}\partial^2\ell/\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta} = (\bar{R}_m - \eta)\boldsymbol{\Sigma}^{-1} = \mathfrak{I}_{\boldsymbol{\beta}\boldsymbol{\alpha}}\,,\\
\mathfrak{I}_{\boldsymbol{\alpha}\boldsymbol{\Sigma}} &= -\mathrm{E}\,\tfrac{1}{T}\partial^2\ell/\partial\boldsymbol{\alpha}\partial\boldsymbol{\Sigma} = \mathbf{0} = \mathfrak{I}_{\boldsymbol{\Sigma}\boldsymbol{\alpha}}\,,\\
\mathfrak{I}_{\boldsymbol{\beta}\boldsymbol{\Sigma}} &= -\mathrm{E}\,\tfrac{1}{T}\partial^2\ell/\partial\boldsymbol{\beta}\partial\boldsymbol{\Sigma} = \mathbf{0} = \mathfrak{I}_{\boldsymbol{\Sigma}\boldsymbol{\beta}}\,,
\end{aligned} \tag{27}$$

put in the Fisher information matrix itself as

$$\mathfrak{I}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}) = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & (\bar{R}_m - \eta)\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ (\bar{R}_m - \eta)\boldsymbol{\Sigma}^{-1} & \left(s_m^2 + (\bar{R}_m - \eta)^2\right)\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1} \end{pmatrix}, \tag{28}$$

and its inverse reads

$$\mathfrak{I}^{-1}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}) = \begin{pmatrix} \left(1 + \frac{(\bar{R}_m - \eta)^2}{s_m^2}\right)\boldsymbol{\Sigma} & -\frac{\bar{R}_m - \eta}{s_m^2}\boldsymbol{\Sigma} & \mathbf{0} \\ -\frac{\bar{R}_m - \eta}{s_m^2}\boldsymbol{\Sigma} & \frac{1}{s_m^2}\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 2\boldsymbol{\Sigma}\boldsymbol{\Sigma} \end{pmatrix}. \tag{29}$$

As it is easy to see that the unconstrained maximum likelihood estimates for $\boldsymbol{\alpha}$ is unbiased and N-variate normal – being just a linear combination of N-variate normal $\boldsymbol{R}_1, ..., \boldsymbol{R}_T$ and of conditionally constant market returns $R_{m1}, ..., R_{mT}$ – with the specification completed by $\boldsymbol{\alpha}\boldsymbol{\alpha}$ element of $\mathfrak{I}^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$

$$\hat{\boldsymbol{\alpha}} \sim N_N\left(\boldsymbol{\alpha}, \left(1 + \tfrac{(\bar{R}_m - \eta)^2}{s_m^2}\right)\boldsymbol{\Sigma}\right), \cdot \tag{30}$$

so it is tedious (and for the sake of this article omitted) to show that

$$T\hat{\boldsymbol{\Sigma}} \sim W_N(T-2, \boldsymbol{\Sigma})\ . \tag{31}$$

On the basis of (30) and (31), it is possible to fructify the theorem stated in Section 3. It suffices to put only

$$\boldsymbol{\xi} \equiv \sqrt{T}\left(1 + \tfrac{(\bar{R}_m - \eta)^2}{s_m^2}\right)^{-\frac{1}{2}}, \quad \mathbf{A} \equiv T\hat{\boldsymbol{\Sigma}}, \quad m \equiv N, \quad m \equiv T - 2, \tag{32}$$

and the F statistics (12) takes the form

$$F(\eta) = \frac{T-N-1}{N} \frac{s_m^2}{s_m^2 + (\bar{R}_m - \eta)^2} \hat{\boldsymbol{\alpha}}'(\eta) \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\alpha}}(\eta) \tag{33}$$

with an $F(N, T-N-1)$ distribution under null hypothesis H_0: $\boldsymbol{\alpha} = \mathbf{0}$ and is dependent on η. As η is unknown, the test using statistics (33) cannot be applied directly. In (33) η must be replaced by its maximum likelihood estimate. Because the maximum likelihood ration for η minimizes the $LR(\eta)$, it also minimizes $F(\eta)$. If η_0 is the true value of η, it always holds that

$$F(\hat{\eta}^*) \leq F(\eta_0) \ , \tag{34}$$

which results in the fact that test based on $F(\eta)$ with the substitution of the maximum likelihood estimate for η does not reject H_0 too often. If the null hypothesis is rejected using the maximum likelihood estimate for η, it is rejected for any value of η_0. However, it is stated in [3] that this testing approach provides a useful check as the asymptotic likelihood ratio test based on (26) has been found to reject the null too often.

4.2 Testing Implication 2 of the Black Version of the CAPM

For testing implication 2 it need be made clear what is meant by exhausting the entire variation of asset excess returns. To determine this, it suffices to derive the covariance of any $\boldsymbol{R}_t$ in N-asset model (6), allowing for restrictions (7); one obtains for any time t

$$\operatorname{cov} \mathbf{R}_t = \operatorname{cov}\left(\eta \mathbf{1} + \boldsymbol{\alpha} + \boldsymbol{\beta}(R_{mt} - \eta) + \boldsymbol{\varepsilon}_t\right) = \boldsymbol{\beta} \mathrm{D} R_{mt} \boldsymbol{\beta}' + \operatorname{cov} \boldsymbol{\varepsilon}_t = \sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma} \tag{35}$$

Thus, if $\boldsymbol{\beta}$ is to exhaust the entire variation of the $\boldsymbol{R}_t$'s there cannot be a significant difference between $\operatorname{cov}\boldsymbol{R}_t$ and $\sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}'$. Only this permits $\boldsymbol{\beta}$ to explain the entire variation of the R_t's.

This ascertainment compels the framework which consists in comparing covariance matrices. To test implication 2 it suffices to test whether the actual covariance matrix $\operatorname{cov}\boldsymbol{R}_t$ equals $\sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}'$, or rather whether there cannot be established a statistically significant difference between $\operatorname{cov}\boldsymbol{R}_t$ equals $\sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}'$. This cannot be effected in the ideal fashion for two reasons. First, the available statistical procedures for testing the null hypothesis H_0: $\operatorname{cov}\boldsymbol{R}_t = \sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}'$ require that the null specification $\sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}'$ be full-ranked with non-zero determinant, which is not case, apparently. It is therefore not feasible to test that all elements in the matrices in the null hypothesis comply. Second, the specification of $\sigma_m^2 \boldsymbol{\beta}\boldsymbol{\beta}'$ in the null hypothesis is not known and need be estimated.

Under this approach to verification of the Black version of the CAPM, implication 2 cannot be tested.

4.3 Testing Implication 3 of the Black Version of the CAPM

Of $R_{m1}, \ldots, R_{mT}$ has so far been assumed only that they are a random drawing from a distribution with a finite expectation and a finite dispersion. However, on the

distribution no assumption has been made. Insomuch as testing the positiveness of $ER_m - \eta$ is needed, with η treated as a constant estimable by the maximum likelihood method, this can be accomplished under no particular distributional assumptions by a nonparametric procedure. On the other hand, it may be assumed that $R_{m1}, ..., R_{mT}$ are a drawing from a certain distribution and the testing procedure can build upon this assumption.

In the article two approaches are presented: either testing by the Wilcoxon signed-rank test free of distributional assumptions, or testing on the footing of the standard one-sample T test when normality is assumed.

For these procedures are notorious and well-described in literature, the formal description of testing is left-out. It is needful to say only that for the Wilcoxon signed-rank test the normal approximation was employed and that prior to using the one-sample T test normality testing of market returns was conducted. The reader may consult e. g. [11].

5 The Empirical Verification

For an illustration of the aforesaid statistical issues for implication 1 only, the estimation procedure and the tests were conducted on a set U. S. data. Available were monthly stock prices of 17 U. S. companies represented in S&P 500 Index: *Exxon Mobil Corporation, Microsoft Corporation, Johnson & Johnson, Procter & Gamble Company, International Business Machines Corporation, J P Morgan Chase & Company, AT&T Incorporated, Apple Incorporated, Chevron Corporation, General Electric Company, Wal-Mart Stores Incorporated, Intel Corporation, Bank of America Corporation, Pfizer Incorporated, The Coca-Cola Company, Hewlett-Packard Company, Pepsico Incorporated.* Out of the monthly stock prices logarithmic returns were constructed and so was done with the monthly values of S&P 500 Index. However, it is needful to note that logarithmic returns of S&P 500 Index (or of any stock index whatsoever) are chosen merely as a proxy to market returns, which are unobservable. To ensure comparability, the stock returns and index returns were annualized. The series of returns encompassed 240 monthly observations over 20 calendar years from July 1989 up to June 2009. The series were downloaded from <*http://finance.yahoo.com*>. The computations were performed in the environment of Microsoft ® Excel 2003 and the testing of implication 3 was accomplished in NCSS 2004.

The methodological procedure ran in this fashion: The series of 20 years of observations was divided into four non-overlapping 5-year sub-periods, each counting 60 monthly observations, to which two non-overlapping 10-year sub-periods corresponded, each counting 120 monthly observations. The estimation of α's and β's and the application of the described tests were first effected on the 5-year sub-periods (M7/1989 – M6/1994, M7/1994 – M6/1999, M7/1999 – M6/2004, M7/2004 – M7/2009), further on the 10-year sub-periods (M7/1989 – M6/1999, M7/1999 – M6/2009), and eventually on the entire 20-year period (M7/1989 – M6/2009).

The results of the tests are reported in Table 1 as for implication 1 and in Table 2 as for implication 3.

Table 1. Results on Testing Implication 1

The period	No. of observations	Eta	Likelihood ratio statistics	Significance	F statistics	Significance
5-year sub-periods						
M7/1989 - M6/1994	60	-0.0049	8.9734	0.9145	0.3986	0.9785
M7/1994 - M6/1999	60	0.0606	13.4832	0.6372	0.6225	0.8546
M7/1999 - M6/2004	60	0.0841	10.8287	0.8199	0.4887	0.9443
M7/2004 - M6/2009	60	0.1640	22.8874	0.1168	1.1474	0.3461
10-year sub-periods						
M7/1989 - M6/1999	120	0.0916	14.4065	0.5685	0.7653	0.7279
M7/1999 - M6/2009	120	0.0606	9.2646	0.9021	0.4816	0.9562
20-year period						
M7/1989 - M6/2009	240	-0.0083	14.9727	0.5266	0.8406	0.6449

Table 2. Results on Testing Implication 2

The period	No. of observations	The Wilcoxon statistics (approx.)	Signifiacance	D'Agostino omnibus statistics	Significance	T statistics	Significance
5-year sub-periods							
M7/1989 - M6/1994	60	1.3398	0.0902	3.8091	0.1489	1.1254	0.1325
M7/1994 - M6/1999	60	2.2158	0.0134	29.4186	0.0000	1.5669	0.0612
M7/1999 - M6/2004	60	-1.4944	0.9325	0.6506	0.7223	-1.6225	0.9450
M7/2004 - M6/2009	60	-2.3631	0.9909	25.5015	0.0000	-2.6840	0.9953
10-year sub-periods							
M7/1989 - M6/1999	120	1.6237	0.0522	23.4089	0.0000	1.0225	0.1543
M7/1999 - M6/2009	120	-1.1287	0.8705	15.3915	0.0005	-1.8135	0.9639
20-year period							
M7/1989 - M6/2009	240	2.8644	0.0021	37.1388	0.0000	1.7555	0.0402

Before the testing of implication 1 itself, the asset returns were inspected as to the presence of multivariate normality. (On this, the results are not reported.) The Mardina's tests of multivariate normality of 1970, which are closer described e. g. in [8], were not suggestive of its presence in the asset returns $\boldsymbol{R}_1, ..., \boldsymbol{R}_T$. The test founded on the multivariate kurtosis rejected normality at all levels of significance in the entire 20-year period and in all the sub-periods considered, whereas although the test based on the multivariate skewness rejected normality at all levels of significance in the entire 20-year period, it was indicative that normality in sub-periods M7/1994 – M6/1999, M7/1999 – M6/2004, M7/2004 – M7/2009 could be characteristic of the data. The summary result yet is that the assumption of normality has not been substantiated and that further tests to conduct may be therefore invalidated. This manifestation, of possible symmetry and leptokurtosis, is typical of financial returns and is consistent with other empirical studies.

In the face of the failure of the normality assumptions, the results on testing implication 1 are favourable as in the entire 20-year period and all the sub-periods considered the null hypothesis of $\boldsymbol{\alpha} = \boldsymbol{0}$ cannot be rejected at all custom levels of significance. The Black version of the CAPM as proposed by (6) has been found to hold.

This, however, cannot be judged on meeting implication 3. The results can be interpreted as relatively inconclusive, or rather, relatively unfavourable. The Wilcoxon signed-rank test tested the null hypothesis of the median of risk premium to be zero against the alternative hypothesis of the median of risk premium to be positive. The one-sample T test performed the same testing, though, about the expectation. The fact is that the interpretation value crucially depends upon the assumption of normality of market returns. This assumption is invalidated by D'Agostino omnibus test of normality of 1990 in the entire period and all the sub-periods with the exception of M7/1989 – M6/1994 and M7/1999 – M6/2004. The positiveness of the expected risk premium at the significance level of 0.05 can only be detected in M7/1994 – M6/1999 and the entire 20-year period (as far as the Wilcoxon signed-rank test results are taken into consideration). The results are also determined by the credibility of the maximum likelihood estimate for η, which substantially varied over the periods considered.

6 The conclusion

Over the late three-four decades enough empirical evidence has been obtained against validity of the Capital Asset Pricing Model. None the less, the model still enjoys great popularity for its intuitive and understandable construction resulting from a set of assumptions peculiar to the environment of developed capital markets. Much literature is devoted to the promotion of the model; yet scarcely is its validity questioned. The idea that its advocating cannot go without investigating its validity incites this article. Thus motivated, the article describes the framework for the estimation of the parameters of the CAPM after Black, and acquaints the reader with statistical tests for judging its validity. Three implications of the Black version of the CAPM need be tested; however, under the outline of the approach presented in the article only two are testable. The empirical part contains the results of empirical verification of the Black version of the CAPM. The results building upon the failed assumption of normality are surprisingly uniform and are suggestive that the CAPM over the entire 20-year period from 1989 to 2009 is valid with respect to the chosen portfolio of seventeen assets, as for implication 1. These findings are consistence with the theoretic representation of the CAPM. Contrariwise, implication 3 has not been found met and the positiveness of the risk premium may be challenged. Yet, this ascertainment is not at odds with the theoretic formula of the CAPM.

The article was prepared under the grant scheme VEGA No. 1/4633/07 Construction and analysis of dynamic macro-economic models of open economies and VEGA No. 1/4634/07 Variant methods of prediction of financial situation of small and medium enterprises after the introduction of common European currency in the Slovak Republic.

References

1. BOĎA, M. The Sharpe-Lintner Version of the CAPM: Framework for Empirical Verification. In: Firma a konkurenční prostředí 2009 – 1. část. MSD, Brno, pp. 196--202. (2009). ISBN 978-80-7392-084-5.
2. BOĎA, M., KANDEROVÁ, M. The Capital Asset Pricing Model Overviewed. The contribution held at Applications of Mathematics and Statistics in Economy 2008 in Wisla. Publishing under way. (2008)
3. CAMPBELL, J. Y., LO, A. W., MACKINLAY, A. C. The Econometrics of Financial Markets. 2nd printing, with corrections. Princeton University Press, Princeton, 611 pp. (1997). ISBN 0-691-04301-9.
4. ELTON, Edwin J., GRUBER, Martin J. et al. Modern Portfolio Theory and Investment Analysis. 7th ed. Wiley, Hoboken, 728 pp. (2007). ISBN 0-470-05082-9.
5. FAMA, E. F., MACBETH, J: D. Risk, Return and Equilibrium: Empirical Tests. In: The Journal of Political Economy, vol. 81., no. 3, pp. 607--636. (1973)
6. MLYNAROVIČ, V. Finančné investovanie. Teórie a aplikácie. Iura Edition, Bratislava, 293 pp. (2001). ISBN 80-89047-16-5.
7. MUIRHEAD, R. J. Aspects of Multivariate Statistical Theory. 1982 ed. Wiley, New York, 673 pp. (2005). ISBN 0-471-76985-1.
8. POTOCKÝ, R., LAMOŠ, F. Pravdepodobnosť a matematická štatistika. Alfa, Bratislava, 344 pp. (1989). ISBN 80-7184-767-4.
9. RAO, C. R. Linear Statistical Inference and Its Applications. 1973 2nd ed. Wiley, New York, 625 pp. (2002). ISBN 0-471-21875-8.
10. RENCHER, A. C. Methods of Multivariate Analysis. 2nd ed. Wiley, New York, 708 pp. (2002). ISBN 0-471-41889-7.
11. SHESKIN, D. Handbook of Parametric and Nonparametric Statistical Procedures. 3rd ed. Chapman & Hall, Boca Raton, 1184 pp. (2004). ISBN 1-58488-440-1.

The Linear Regression Model of Education Expenditure in the EU

Jana Borůvková[1], Bohumil Minařík[2]

[1] College of Polytechnics Jihlava
Tolsteho 16, Jihlava, The Czech Republic
e-mail: boruvkova@vspj.cz

[2] Mendel University of Agriculture and Forestry in Brno
Zemedelska 1, Brno, The Czech Republic
e-mail: minarik@mendelu.cz

Abstract. The most prevailing indicator of public expenditure for education is their proportion in GDP. Expenditures for one student per year or the proportion of students per 100 inhabitants rank among other important indicators. This paper deals with the analysis of public expenditures, which were directed at the sphere of education using the latest available data of EUROSTAT. Information from 26 EU countries and Norway was used. A linear regression model was compiled with three predictors and the interpreted variable of expenditure for a student. This model originated on the basis of the wider range of indicators using step regression and the rate of its determination amounted to about 90%. Through the correction of non-linearity introduced by some countries substantially better results were achieved without necessity to exclude these countries. The resulting model with five parameters leaves 1.4% variability of the interpreted variable unexplained. The model correctness was also examined using regression diagnostics.

Keywords: Education expenditure, Linear regression modeling, Model Building, GDP per inhabitant, Proportion of educational expenditure to GDP

1 Introduction

The education, research, development and innovation are interconnected and represent the most important long-term source of economic development and prosperity of countries and their regions.

As stressed in the document *Key data on higher education in Europe 2007* [1], education represents great benefit both for individuals – it helps to achieve aims in life, profession and position and for all society – degree of education is one of basic criteria for current and future competitive strength of the economy. Thus education expenditures predetermine, to a certain extent, the competitive strength of the

economy. At the same time, amount of costs currently invested is appreciated in a considerable lapse of time.

Majority of developed EU countries finance the substantial part of educational system from public funds, which distinguishes them e.g. from the USA that have, besides state school system, a wide network of private educational institutions with long tradition.

2 Material and methodology

The statistical monitoring of public education expenditure is methodically controlled by a common specialized group of UNESCO, OECD and EUROSTAT organizations. The proportion of education expenditure in GDP is the most common indicator of public expenditure on education. Other important indicators are e.g.: student costs per year, GDP per capita or a number of students per 100 inhabitants.

This paper deals with the description and analysis of costs directed to school systems of selected European countries [2]. The data, published by EUROSTAT for the year 2005[1] [3], helped to create a linear regressive model with three explaining variables – GDP proportion in percents directed at educational system, GDP per capita and a number of students per 100 inhabitants and the explained variable is cost per student.

The described model was created on the basis of originally wider range of indicators using the technique of creating regressive model by means of step regression. A lot of multi-factor regressive models have been created which differed from the authors´ model in including other predicators and their interactions. No other model, however, could be designated as more convenient either from the logical point of view or formally statistical point of view (determination coefficient and residual characteristics).

To create a multiple linear regressive model, official statistical information on 26 EU countries, extended by Norway, has been used (except Luxemburg that did not have complete processed indicators). In all, the model has been created based on data from 27 European countries.

The designed model is trying to answer the question: "How do expenditure per student (estimation of explained variable y') depend on:

- proportion of education expenditure in GDP (explaining variable, predictor x_1),
- GDP per capita (explaining variable, predicator x_2) and
- a number of students per 100 inhabitants (explaining variable, predictor x_3)."

While analyzing individual variables and their binary relations, outliers in the direction of explained variable axis and outliers in direction of both axes were present. It is caused by different economic level of involved countries and by the

[1] While the paper was created no later data on education expenditure were available.

difference in financing schools from public funds in particular countries. To prevent the negative influence of outliers without elimination of this monitoring (it introduces into the model a statistically insignificant yet not negligible nonlinearity), we have introduced a auxiliary variable, predictor x_4 taking values –1, 0 and 1.

The aim of regression modeling is to measure partial effects which predictor variability shows on variable y (in this case variables x_1, x_2, x_3 and x_4) by means of the simple linear relation $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \varepsilon = y' + \varepsilon$ (ε is the difference between the observed value and the corresponding value that is predicted by the model and thus it represents the variance that is not explained by the model and residuum is its estimation).

If all variables before the analysis are standardized, it is possible according to standardized regression coefficients to evaluate the relative contribution of predictor x_j to prediction of variable y.

Linear relations between predictors and explained variable are condition for creating a correct linear model. On the contrary particular predictors should be correlated minimally.

The determination coefficient R^2, which is a proportion of variability explained by the model and the total variability, equals a part of variability in variable y that is explained by predictors x_1, x_2, x_3 and x_4.

To evaluate the regression model quality, an analysis of residuals has been used:

- Residuals have a normal classification with zero mean value.
- The dependence of raw residuals on predicted values shows no systematic relations.
- Detection of possible influential points (outliers).

3 RESULTS AND DISCUSSION

3.1 Basic data analysis

To create the basic ideas of the data there was made a preliminary data analysis and summary characteristics were calculated for variables x_1, x_2 and x_3), which are listed in *Tab. 1*.

Tab. 1 Summary characteristics of variables.

	Valid *n*	Mean	Median	Min.	Max.	Std.Dev
Proportion of education expenditure to GDP (in %)	28	5.32	5.30	3.48	8.28	1.15
GDP per capita (in €)	28	21882.1	19450.0	2800.00	65000.0	15281.5
Number of students per 100 inhabitants	28	23.06	22.69	18.46	29.38	2.92
Expenditure per student (in €)	27	5287.44	5907.90	1454.20	9132.90	2084.63

The illustrated values of variable, which have been identified in different countries, are graphically depicted in *Fig. 1*. For the classification the natural breaks method was used, which creates a class of its own similar elements – in this case states that have similar annual expenditure per student.

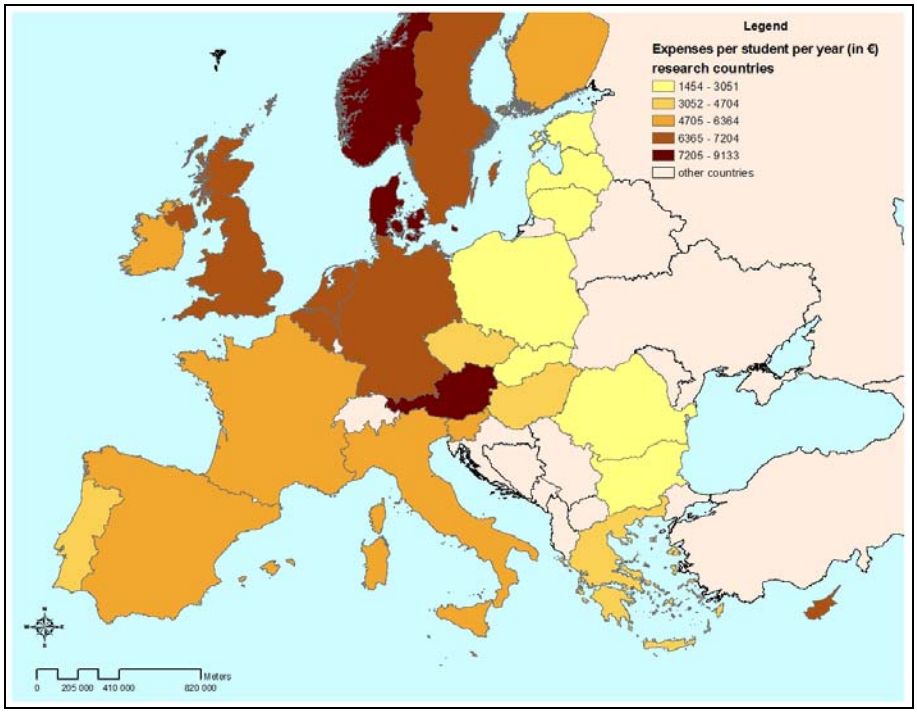


Fig. 1 Expenditure per student in €.

3.2 Characteristics of the individual countries and the relationship between predictors

The Figures 2 to 4 in particular show that the predictors are of a low correlation, which implies the distribution of countries in all four quadrant of each image. The first quadrant includes countries that showed above-average value in the two controlled variables, the third quadrant on the contrary, countries with average value of both variables. The second and fourth quadrant includes countries in which always one of the two variables achieved above average and the second average value.

In *Fig. 2* we see that group of countries BE, FI, SE, DK and NO shows the average high value of GDP per capita (NO in this group the highest) and the above average proportion of expenditures on education of GDP (from this group the highest DK). The antithesis is a group of countries, not reaching even the below average in both variables (12, mostly relatively less developed countries, including CZ). LU has a completely unique position, which at the largest GDP per capita shows one of the lowest share of expenditure on education in GDP.

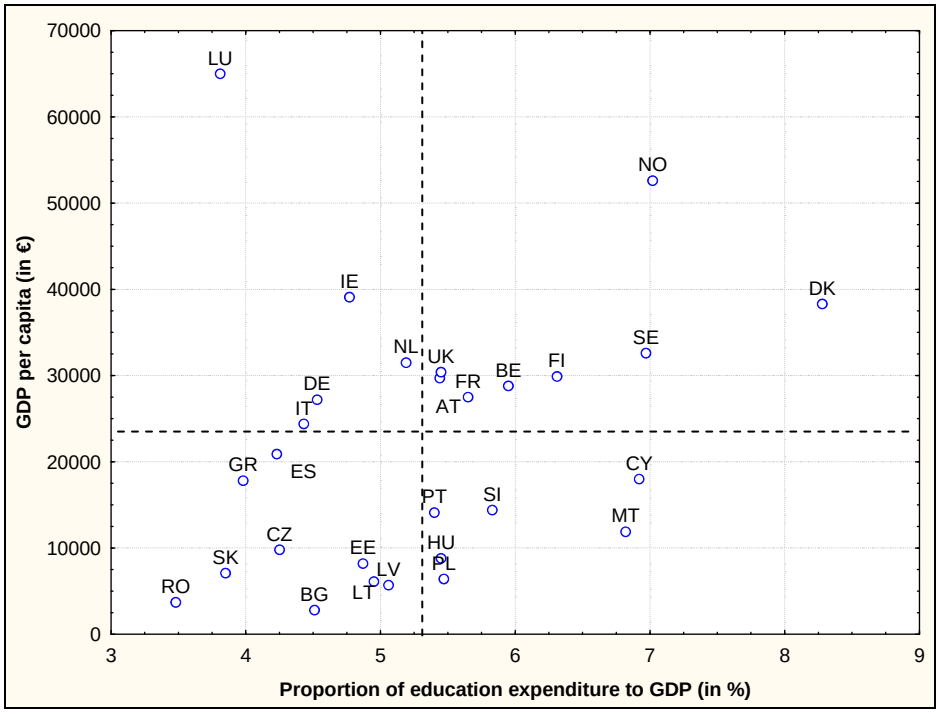


Fig. 2 Proportion of educational expenditure to GDP (in %) vs. GDP per inhabitant (in €). Countries are identified by the official international two-letter code.

Fig. 3 shows the exceptional position of UK, which recorded the highest number of students per 100 inhabitants in slightly below average shares of spending on education in GDP. In the third quadrant again we see a group of few less developed countries (the lowest ever number of students per 100 inhabitants has a BG), but this time somewhat unusually in LU, DE, AT and NL.

In *Fig. 4*, the latter group of countries is located in the fourth quadrant, which means a combination of high GDP per capita and a low number of students per 100 inhabitants. The group is opposed to the Baltic countries and PL in the second quadrant.

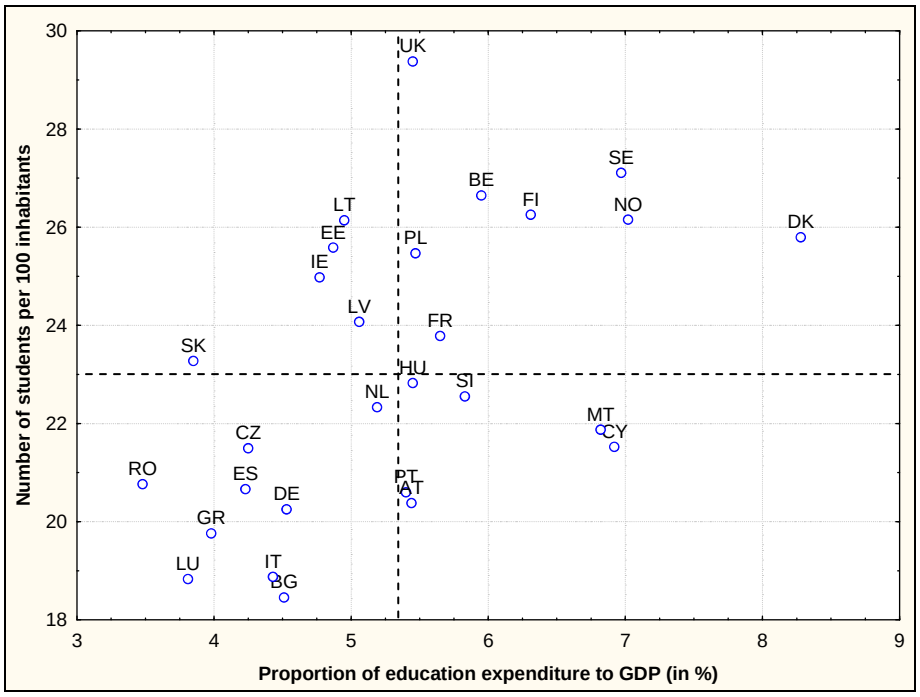


Fig. 3 Proportion of educational expenditure to GDP (in %) vs. number of students per 100 inhabitants.

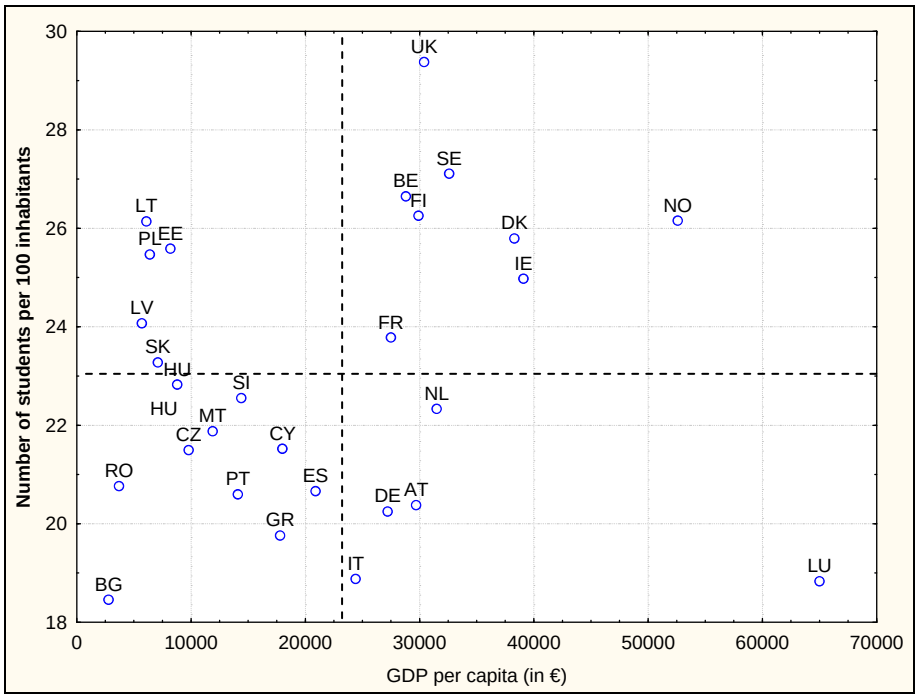


Fig. 4 GDP per inhabitant (in €) vs. number of students per 100 inhabitants.

Tab. 2 lists Pearson´s correlation coefficients and *Tab. 3* partial correlation coefficients, including the level of significance of variables x_1, x_2 and x_3. The values of the correlation coefficients are not mostly statistically significant. Only in one case, Pearson´s correlation coefficient is statistically significant, but it is relatively low.

Tab. 2 Pearson´s pair correlation coefficients and their significance.
Highlighted values are considered as statistically significant.

	Proportion of education expenditure to GDP (in %)	GDP per capita (in €)	Number of students per 100 inhabitants
Proportion of education expenditure to GDP (in %)	1.0000	.2719	.5183
	p= ---	p=.162	p=.005
GDP per capita (in €)	.2719	1.0000	.1303
	p=.162	p= ---	p=.509
Number of students per 100 inhabitants	.5183	.1303	1.0000
	p=.005	p=.509	p= ---

Objective information on the partial correlation relations of predictors and explanatory variables provide partial correlation coefficients in *Tab. 3*, which are generally higher than pair coefficients. In particular, it appears reasonably expected negative relationship between the predictor variable x_3 and explained variable.

Tab. 3 Pearson´s partial correlation coefficients and their significance.
Highlighted values are considered as statistically significant.

	Proportion of education expenditure to GDP (in %)	GDP per capita (in €)	Number of students per 100 inhabitants
Expenditure per student (in €)	.6592	.9040	-.5155
	p=.000	p=.000	p=.008

3.3 Linear regression model

The original regression equation to estimate the theoretical values of $y´$ with predictors x_1, x_2 and x_3 is in the shape of $y´ = 2978 + 648\,x_1 + 0.128\,x_2 – 162.5\,x_3$.

This regression equation explains nearly 90% of variability of explained variable. Diagnostics of residues (see *Fig. 5*, which results in parabolic course residues) showed that this model incorrectly explained variable for some countries, especially those with extremely low or extremely high value of GDP per capita. The countries that have issued per student significantly less than the model shows, are both

countries with the highest GDP per capita (IE, DK, NO), as well as countries with the lowest GDP per capita (BG, RO). In addition, the model has not proved useful for countries with extreme value of the explained variable (UK, AT, SI, ES).

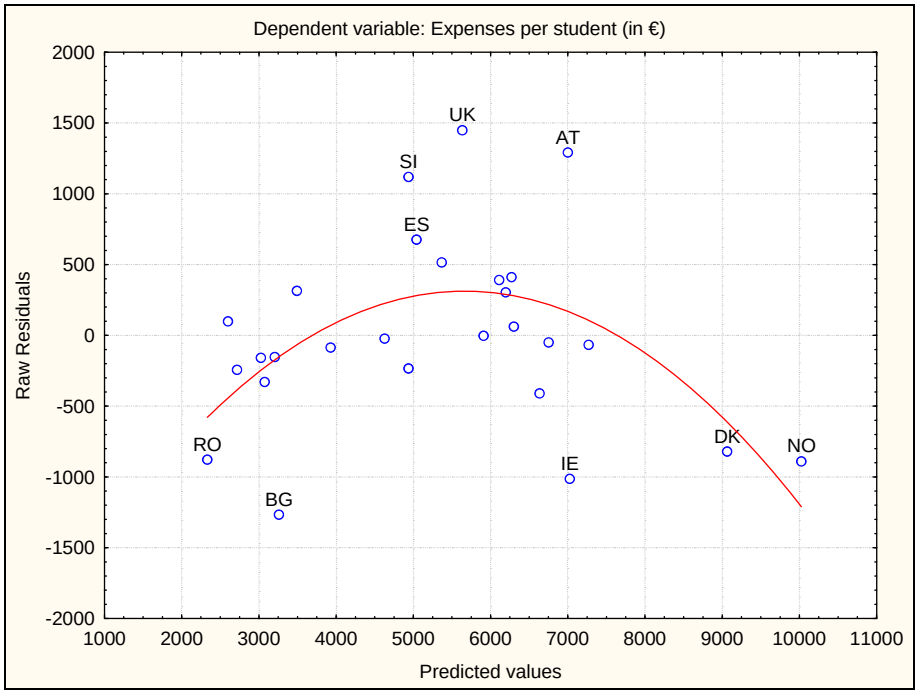


Fig. 5 Predicted vs. residual values.

It was the original model of extended auxiliary variable, and the construction of model correctness was verified again through forward and backward step regression. Regression equation of the extended model takes the form

$$y' = 3047 + 699.7\, x_1 + 0.132\, x_2 - 179.3\, x_3 + 1069.2\, x_4$$

Adjusted index of determination is equal to 0.9864, which corresponds to about 1.4% unsolved variability illustrated variables. Detail the parameters of the model contained in *Tab. 4*.

Tab. 4 Parameters of regression model wih auxiliary variable.
Highlighted values are considered as statistically significant.

	Exp. per student (in €) Param.	Exp. per student (in €) Std.Err	Exp. per student (in €) t	Exp. per student (in €) p
Intercept	3046.620	441.1792	6.90563	0.000001
Auxiliary variable	1069.159	89.1550	11.99214	0.000000
Proportion of education expenditure to GDP (in %)	699.704	57.5537	12.15743	0.000000
GDP per capita (in €)	0.132	0.0047	28.02263	0.000000
Number of students per 100 inhabitants	-179.345	21.0256	-8.52983	0.000000

	Exp. per student (in €) Param.	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept	3046.620	2131.670	3961.569
Auxiliary variable	1069.159	884.263	1254.055
Proportion of education expenditure to GDP (in %)	699.704	580.345	819.063
GDP per capita (in €)	0.132	0.123	0.142
Number of students per 100 inhabitants	-179.345	-222.950	-135.741

A positive difference in the proportion of expenditure on education to GDP by one percentage point corresponds with an increase in expenditure per student of € 699.7, positive difference in GDP per capita of € 1 corresponds to an increase in expenditure per student of 0.13 €. In contrast, a positive difference in the number of students per 100 inhabitants of the unit corresponds with the decline in spending per student for 180 €.

To assess the relative contribution of predictors we use the coefficients β, which are listed in *Tab. 5*. These coefficients have been calculated so that before the creation of model the data are were standardized. From the values of the coefficient it is clear that the greatest impact on the explained variable *y´* has a variable "GDP per capita" and the least impact has a variable "Number of students per 100 inhabitants".

Tab. 5 Values of the standardized regression coefficients.
Highlighted values are considered as statistically significant.

Expenditure per student (in €)	Beta (ß)	Std.Err. ß	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Auxiliary variable	0.301129	0.025111	0.249053	0.353205
Proportion of education expenditure to GDP (in %)	0.380463	0.031295	0.315562	0.445365
GDP per capita (in €)	0.824109	0.029409	0.763119	0.885099
Number of students per 100 inhabitants	-0.245318	0.028760	-0.304962	-0.185673

In *Tab. 6* there is given a coefficient of determination R^2, which reaches 0.986, which can be interpreted in such a way that the created model explains nearly 99% of variability illustrated variables. The appropriateness of the model also shows the *F*-statistic and its *p*-value. In this case we get a statistically significant value, which also shows the suitability of the model.

Tab. 6 Coefficient of determination R^2 and the significance testing of model.
Highlighted values are considered as statistically significant.

	Multiple *R*	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model
Expenditure per student (in €)	0.993168	0.986383	0.983907	111449305	4	27862326
		SS Rezid.	**df Rezid.**	**MS Rezid.**	***F***	***p***
Expenditure per student (in €)		1538563	22	69934.67	398.4051	0.00

Let's ask a question, what lies behind the approximately 1.4% of unexplained variability. These may be errors in the data, individual differences in the content of the indicators and methodology for their determination in different countries, differences in the financing of education systems in different countries, other predictors that were not included in the model and possibly other effects. Despite all these influences included model explained about 98.6% of all variability illustrated

variables, which can be attained under the conditions previously considered difficult anticipative success.

3.4 Analysis of residues of the extended model

One of the assumptions, the appropriateness of using the model created is normality of residues. *Fig. 6* is a *p*-graph residue, which indicates the normality of residues. Neither the Kolmogorov-Smirnov normality test does not allow to reject the hypothesis of normal distribution of residue levels ($p > 0.2$).

In *Fig. 7* there is a graph of calculated values against residuals. This graph does not show any systematic relations and thus confirms model suitability.

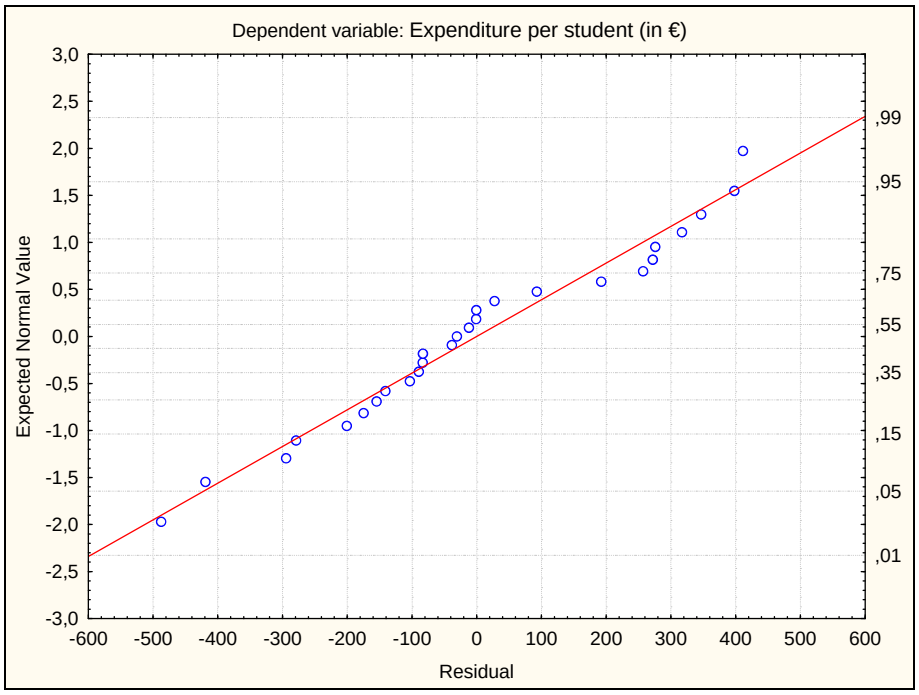


Fig. 6 Normal probability plot; Raw residuals.

The Absolute Percentage Error (A.P.E.) indicates model suitability for each chosen country, calculated and shown in *Fig. 8*. The natural breaks classification, which identifies grouping of similar data elements – in this case countries having the same A.P.E., was used.

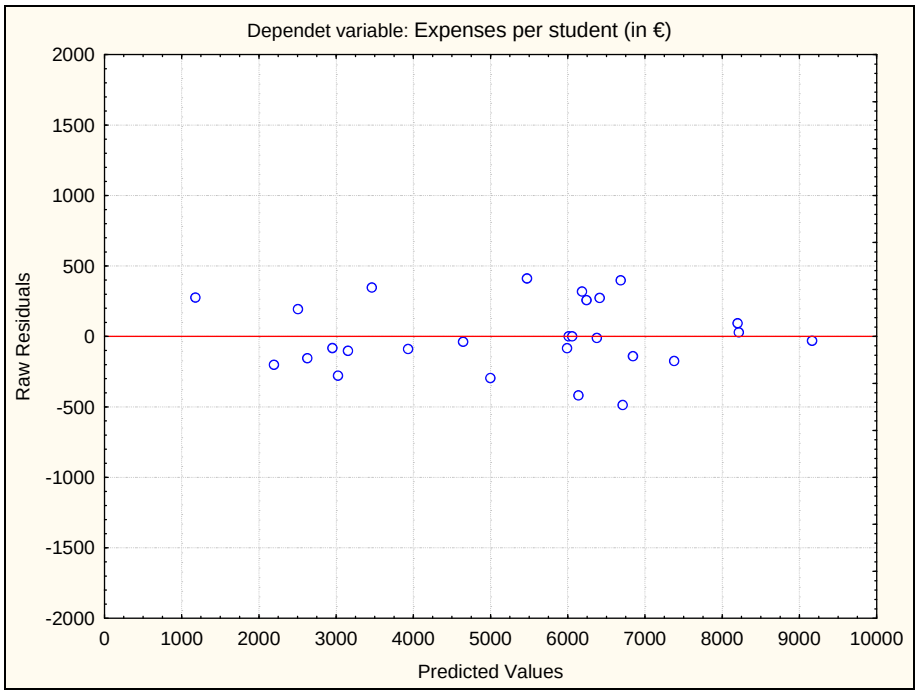


Fig. 7 Predicted vs. residual values.

The average extended model value of A.P.E. is 4.7%, while it was almost 12% for the original model. The picture 8 shows that the model is the least suiTab. for RO (A.P.E. almost 20%), less suiTab. for CZ, LV and BG (A.P.E. up to 10%). On the contrary, almost precise values of student expenditure were calculated for eight countries: FR, IT, IE, DK, NO, GR, AT, SI.

The model enabled to set the value of explained variable for LU (data not available in another way). The calculated value 9856 € per a student is the highest among all countries included (the second NO shows 9133 €. and corresponds to unofficial estimates stating 10,000 €.

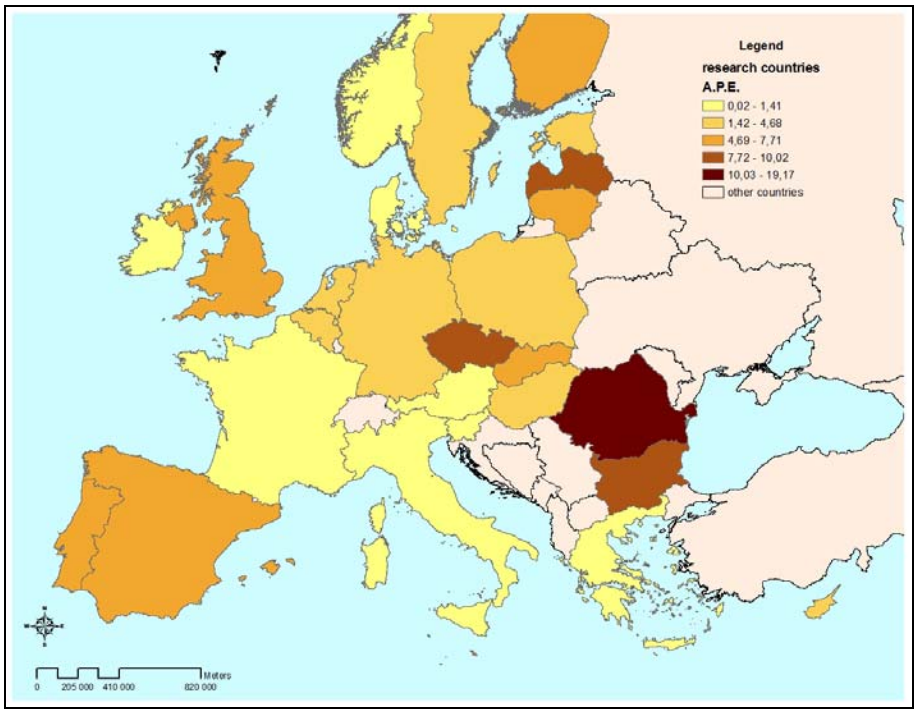


Fig. 8 Model suitability for each individual country expressed in A.P.E. characteristics.

4 Conclusion

This paper tries to analyze a starting position of public education financing in each individual European countries, in most cases members of the EU.

With regard to the authors´ profession, statistical methods were applied to official data obtained from public electronic resources that were at disposal. Besides general descriptions of individual, mostly financial indicators and their binary and partial relations, it was important to obtain a multifactorial regression model which captures the relation between public expenditure per student (in Euro) and explaining predictors, which, under the condition that statistical expectations are met, clarify maximum variability of explained variable. This kind of model was found by choosing from a wide range of indicators and many model equations and by using forward stepwise regression.

Three found predictors, namely education expenditure as a proportion of GDP (%), GPD per capita (Euro) and a number of students per 100 inhabitants, were unable to explain more than 90% of expenditure variability per a student. The presence of outliers, which reduced predicative ability of the created regression model, was the reason of this fact. On the basis of experience obtained from similar situations, the authors used categorical variable, the values of which eliminated the given

unfavorable consequences without the need to exclude some countries from the analysis and consequently reduce their number. The extended model with five highly important and statistical parameters explained 98.6% of variability of explained variable.

However, the model proved to be more suiTab. for some of the countries, while less for others. The A.P.E. characteristic was used to explain individual predictive value of each country. With regard to a low number of included countries, as expected, confidential intervals of individual parameters and the model as a whole are wide enough to provide opportunities for inductive consideration on the importance of individual variables. The GDP predictor per capita was classified uniquely the most important since provides information on economic development of the country. The rest of the predictors inform about national traditions, usages, ways of public education financing in each country. Values of each individual country vary very much (e.g. a very low education expenditure proportion of GDP in extremely rich LU, the same way as a number of students per 100 inhabitants which can be compared to e.g. BG and lower than in RO). The regression equation enabled to estimate the explained variable for LU, the value of which was not available in statistics. The calculated value almost equals to expert estimates which state student expenditure at the amount of approximately 10,000 €.

The paper was written within the framework of the research program MSM 6215648904 at Faculty of Economics, Mendel University of Agriculture and Forestry in Brno titled *"Czech Economy in the Process of Integration and Globalization, and Development of Agrarian Sector and the Sector of Services under the New Conditions of Integrated Agrarian Market*", topical direction no 5 "*Social and Economic Relations of Sustainable and Multifunctional Agriculture and Measures of Agrarian and Regional Policies"* .

Reference

1. **Key Data on Higher Education in Europe**. Brussel – Luxembourg: ECSC-EC-EAEC (2007). ISBN 978-92-79-0610-2
2. OECD: **Education at a Glance: OECD Indicators – 2007 Edition**, http://213.253.134.43/oecd/pdfs/browseit/9607051E.PDF yearbook OECD [Accessed 2008-03-27] (2008)
3. EUROSTAT: **Economy and finance**, http://epp.eurostat.ec.europa.eu/portal/page?_pageid=0,1136173,0_45570455&_dad=portal&_schema=PORTAL [Accessed 2009-02-07] (2009)

Risk-Neutral Option Pricing[1]

Martin Cícha

Department of Statistics and Probability Calculus,
University of Economics, Prague
cicham@vse.cz

Abstract. Risk-neutral approach is one of two main approaches to derivatives pricing. This paper introduces a derivatives pricing tool derived from the binomial tree process of underlying asset price under assumption of constant volatility. This article focuses on stocks, but proposed approach can be generalized and applied for another underlying assets as well. The assumption of constant volatility leads to a lognormal distribution of underlying asset price. The volatility observed in the market is often a function of option strike and time to maturity rather than a constant. Hence, using lognormal distribution in pricing formulas leads to incorrect prices. I provide evidence of nonlognormality of CEZ stock using American-style call warrants in this paper.

Keywords: risk-neutral pricing, implied density function, volatility smile

1 Introduction

Optional contracts are large group of financial instruments and unique approach to their correct pricing has not been introduced yet. The replication strategy and risk-neutral approach are two main approaches which lead to the identical results. Under assumption of constant volatility we derive lognormal distribution of an underlying asset. Using lognormal distribution in pricing formulas we get theoretical price of the option. It is often the case that traders use volatility as function of option strike price and option time to maturity rather than as a constant value, which leads to volatility surface. Lognormal distribution of underlying asset can not be used for pricing any more. It is possible to use the information that is embedded in options prices to derive implied distribution of underlying asset. These density functions estimated from cross-section of observed option prices are gaining increasing attention. They are used to price complex derivatives and derivatives not traded in the market. Using implied distribution of underlying asset we can also find the mispriced derivatives and earn excess returns above the risk-free rate of interest by buying options that are undervalued by the market and selling options that are overvalued by the market. A number of authors have used implied density functions as indicator of market sentiment to examine whether options markets anticipated major economic events. Central banks, in particular, have been interested in using implied density functions to

[1] This paper is supported by IGA VSE # 26/08

assess market participant's expectations of future change in interest rates, stock prices and exchange rates. There can be found many research papers concerning this issue in research bulletins of central banks. See for example, FED and Bank of England (Bliss and Panigirtzoglou, 2002), Bank of England (Bahra, 1997), ECB (Schneider and Glatzer, 2003). Methods for estimating implied density function fall into five groups: stochastic process methods, implied binomial trees, finite-difference methods, implied volatility smoothing methods and density function approximating methods. Stochastic process methods begin by assuming a model for stochastic process driving the prices of the underlying asset. This approach can be used in absence of options prices. For instance, Hoerdahl applied the Longstaff-Swchartz model to Swedish interest rates in (Hoerdahl, 2000). The implied binomial tree method was developed by (Rubinstein, 1994). The smoothed implied volatility method was originally developed by Shimko in (Shimko, 1993). This is an approximating function method to the volatility smile rather than to the density function. Approximating function methods are based on minimizing the difference between observed options prices and fitted prices produced by particular functional form, chosen to allow for a variety of possible shapes. For instance, Melick and Thomas used mixtures of lognormals in (Melick and Thomas, 1994), Mandan and Milne used Hermite polynomials in (Mandan and Milne, 1994) and alternatively Ait-Sahalia used non-parametric kernel estimator in (Ait-Sahalia and Lo, 1998).

The paper is organized as follows. Section 2 introduces binomial tree model in the real world and shows the transition from the real to the risk-neutral world. Further, Section 2 discusses real and risk-neutral distribution of stock price. Section 3 introduces the risk-neutral option pricing approach and applies it to the real market options. Section 4 discusses violence in constant volatility assumption. Section 4 also shows that warrants on CEZ stock does not have constant volatility, hence CEZ stock price is not lognormally distributed at any future time. Furthermore, Section 4 introduces deriving volatility smile and volatility term structure from American-style call warrants on CEZ. Section 4 also discusses two-lognormal mixture implied probability distribution of CEZ stock price and shows estimation of parameters. Using these statistics we can make assumption about future CEZ stock price. Section 5 concludes.

2 The binomial tree model and expectation approach

A derivative is a function of underlying asset, say stock, at specific time in the future. Let us assume that the price of the stock follows process of binomial tree. This fact causes that only two things can happen to the stock at each time *t*: an 'up' move or 'down' move. We start at time $t = 0$ and continue onwards to the time $t = T$ using time step δt. We will assign probabilities to these moves: probability p to move up, and thus $1-p$ to move down. We also need a cash bond B to represent the time-value of money. There will be some continuously compounded interest rate r that will hold for the period δt and will cause that 1 $ today will be worth $e^{r\delta t}$ $ a tick later.

We can use two approaches to the derivative pricing at this moment. The first one is

so called replication strategy. The replication strategy is based on the fact that any derivative of a stock, derivative pay off particularly, can be constructed in advance from an appropriate portfolio of bond and stock. At the end of the tick-period, of course at time $t = T$ as well, the value of the derivative would exactly cancel the value of the portfolio whatever the stock price was, see (Joshi, 2005) for details. Therefore, if one knows the value of this portfolio one also knows the value of the derivative. The other derivatives pricing approach is based on expectation and change of the measure. We can imagine the value of the derivative as a expectation of discounted claim under martingale measure. I will deal only with the expectation approach in this paper.

Let us assume that the stock price process $\{S_t, t \geq 0\}$ has a constant growth rate μ and constant noise σ. The stock jumps from the value S_t to the values at time $t + \delta t$. The value $S_{t+\delta t}$ after up jump determines (1). The value $S_{t+\delta t}$ after down jump determines (2).

$$S_{t+\delta t}^{up} = S_t \exp\left(\mu \delta t + \sigma \sqrt{\delta t}\right). \tag{1}$$

$$S_{t+\delta t}^{down} = S_t \exp\left(\mu \delta t - \sigma \sqrt{\delta t}\right). \tag{2}$$

All jumps are equally likely in the real world, i.e. under a $\mathbb{P}$ measure. The real world measure $\mathbb{P}$ under which S_t occurs in the real world is irrelevant for correct pricing (Baxter, 1999). We have to find a measure $\mathbb{Q}$ under which the process of discounted stock $\left\{B_t^{-1} S_t, t \geq 0\right\}$ is a martingale. Let us define the variable

$$q_t = \frac{S_t \exp\left(r \delta t - S_{t+\delta t}^{down}\right)}{S_{t+\delta t}^{up} - S_{t+\delta t}^{down}}. \tag{3}$$

If the condition of rational market is fulfilled, following relation holds $S_{t+\delta t}^{down} < S_t e^{r \delta t} < S_{t+\delta t}^{up}$ which force q_t into $(0,1)$. We demand the same constraint for probability. It can be shown that q_t is the martingale measure probability (Baxter, 1999). We can calculate q_t approximately equal to

$$q_t = \frac{1}{2}\left(1 - \sqrt{\delta t}\left(\frac{\mu + \frac{1}{2}\sigma^2 - r}{\sigma}\right)\right). \tag{4}$$

Consequently, q_t depends no more on time t but only on time step δt. Hence, the whole tree has only one unique probability q. The binomial tree model is parameterized by the time step δt. As that quantity gets smaller, the model should ever more closely approximate a real world. For a fixed time t, we get stock value under $\mathbb{P}$

$$S_t = S_0 e^{\left(\mu + \sigma\sqrt{t}\left(\frac{2X_n - n}{\sqrt{n}}\right)\right)}, \tag{5}$$

where n is number of time steps till time t and X_n is the total number of the n separate jumps which were up-jumps. X_n is binomially distributed random variable with mean $\frac{n}{2}$ and variance $\frac{n}{4}$, so that $\frac{2X_n - n}{\sqrt{n}}$ has zero mean and variance equals 1. Applying central limit theorem, this distribution converges to the standard normal distribution. As δt gets smaller, the distribution of S_t becomes log-normal with mean $e^{1/2\sigma^2 t} S_0 e^{\mu t}$ and variance $S_0^2\left(e^{2t\left(\mu+\sigma^2\right)} - e^{t\left(\sigma^2+2\mu\right)}\right)$. Distribution of S_t is useless for correct pricing under real world $\mathbb{P}$ measure and thus we apply the same procedure under martingale measure $\mathbb{Q}$. Using up-move probability (??), X_n is still binomially distributed with mean nq and variance $nq(1-q)$. Hence, $\frac{2X_n - n}{\sqrt{n}}$ has mean $-\sqrt{t}\,\frac{\mu + \frac{1}{2}\sigma^2 - r}{\sigma}$ and variance asymptotically approaching one. As δt gets smaller central limit theorem tells us that this formula converges to a normal variable with the same mean and variance exactly one. Then transforming this normal random variable we get the distribution of S_t which is log-normal just like under $\mathbb{P}$ measure, but with mean $S_0 e^{tr}$ and variance $S_0^2\left(e^{tr}\right)^2\left(e^{\sigma^2 t} - 1\right)$. Thus, distribution of S_t under $\mathbb{Q}$ martingale measure is fully determined by time t and a noise σ (so-called volatility) and does not depend on growth rate μ.

Since we have found the marginal distribution of S_t under $\mathbb{Q}$, we can start deal with the derivatives pricing. We need to know the stock distribution, claim, time of exercising the claim and the riskless interest rate r. The price of the claim X (i.e. derivative price) is then expected value of discounted claim under $\mathbb{Q}$:

$$V_0 = E_{\mathbb{Q}}\left(B_T^{-1} X\right), \tag{6}$$

where T is time of exercising the claim and B_T^{-1} is discounted zero coupon bond which is defined as $B_0 e^{-rT}$.

3 Risk neutral derivatives pricing

In order to compute the derivative price (6), we have to find the distribution of discounted claim under $\mathbb{Q}$ using the transformation of random variable technique. To be able to do that, it is easier using normal distribution preferably $N(0,1)$ instead of log-normal distribution of S_t. Distribution of S_t is defined by:

$$S_t = S_0 e^{\left(\left(r-\frac{1}{2}\sigma^2\right)t+\sigma\sqrt{t}Z\right)}, \tag{7}$$

where Z is N(0, 1) under $\mathbb{Q}$ (Joshi, 2005).

3.1 Pricing a call option

Call option (European) gives the holder the right, but not the obligation, to buy a stock at specified price (strike) at time $t = T$. Rational investor will exercise the option only if the stock value is greater than the strike price k at time T. Thus, the claim X is the function of stock $X = (S_T - k)^+$ at time $t = T$. Primarily, we have to determine the moments of the claim or of the discounted claim under $\mathbb{Q}$. Using (7) we get density (8) and distribution function (9) of discounted claim $Y = B_T^{-1} X$.

$$f(y) = \frac{\sqrt{\frac{2}{\pi}}}{2\sigma\sqrt{t}\,\mathrm{e}^{\frac{\left(\frac{t\sigma^2}{2}+\log\left(\frac{y+k\mathrm{e}^{-rt}}{S_0}\right)\right)^2}{2\sigma^2 t}}\left(y+k\mathrm{e}^{-rt}\right)} \tag{8}$$

$$F(y) = \frac{1}{2}erf\left(\frac{\sqrt{2}\left(\frac{t\sigma^2}{2}+\log\left(\frac{y+k\mathrm{e}^{-rt}}{S_0}\right)\right)}{2\sigma\sqrt{t}}\right)+\frac{1}{2} \tag{9}$$

Having appropriate density function we compute the expected value according to (6). We have to keep in mind that the claim X and discounted claim Y can reach only positive values. Hence, the price V of call option is

$$V(S_0T) = \int_0^\infty y\,f(y)dy, \tag{10}$$

Suppose a stock with constant volatility σ of 20% and constant drift μ of 5% with continuously compounded interest rate r constant at 3%. What is the price of an option to buy the stock for \$8 in 5 years time, given a current stock price of \$5?
The density (8) and the distribution function (9) are displayed in Figure 1. We solve equation (10) numerically with the outcome of 0.35999. Thus, the price of call option is \$0.36 at time $t = 0$. All other prices would lead to arbitrary risk-free profits through constructing appropriate replication portfolio of a bond and a stock (Joshi, 2005).

Let us take a look at Black-Scholes formula (Hull, 2006). Fischer Black and Myron Scholes constructed self financing replication portfolio of bond and stock in continuous world. Thus, they had to solve a Partial Differential Equation (PDE) to get the correct option price. The PDE has an explicit solution therefore their formula (11) became so popular. Applying (11) the price of call option is \$0.36 which is exactly the same price as using risk neutral approach. The explicit solution to B-S for call options is given by:

$$V(S_0,T) = S_0\Phi\left(\frac{\log\frac{S_0}{k}+\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right)- \tag{11}$$

$$-ke^{-rT}\Phi\left(\frac{\log\frac{S_0}{k}+\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right),$$

where the notation $\Phi(x)$ denotes the distribution function of $N(0,1)$.

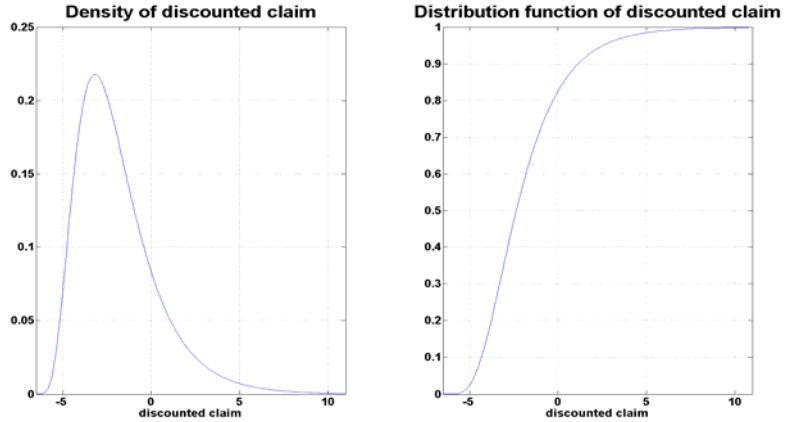


Figure 1: Density and distribution function of discounted call option claim with $\sigma = 0.2$, $k = 8$, $S_0 = 5$, $T = 5$, $r = 0.03$.

3.2 Pricing a put option

Put option (European) gives the holder the right, but not the obligation, to sell a stock at specified price (strike) at time $t = T$. Rational investor will exercise the option only if the stock value is less than the strike price at time T. Thus, the claim X is the function of stock $X = (k - S_T)^+$ at time $t = T$. Applying the transformation of random variable technique, we get the density (12) and the distribution function (13) of discounted claim Y.

$$f(y) = \frac{e^{rt}\sqrt{\frac{2}{\pi}}}{2\sigma\sqrt{t}\,e^{\frac{\left(\log\left(\frac{k-ye^{rt}}{S_0}\right)-t\left(r-\frac{\sigma^2}{2}\right)\right)^2}{2\sigma^2 t}}\left(k-ye^{rt}\right)}. \tag{12}$$

$$F(y)=\frac{1}{2}-\frac{1}{2}erf\left(\frac{\sqrt{2}\left(\log\left(\frac{k-ye^{rt}}{S_0}\right)-t\left(r-\frac{\sigma^2}{2}\right)\right)}{2\sigma\sqrt{t}}\right). \tag{13}$$

Having appropriate density function, we compute the expected value (10) where $t=0$ for $y<0$ and $t=y$ for $y>0$.

Density and distribution functions of discounted claim under $\mathbb{Q}$ martingale measure for $\sigma=0.2$, $k=8$, $S_0=5$, $T=5$, $r=0.03$ are shown in figure 2. Again, we solve the integral (10) numerically. The price of put option is \$2.2457. We get exactly the same price applying Black-Scholes formula for put option. The Black-Scholes formula for put option is given by explicit PDE solution (14) , see (Hull, 2006) for details.

$$\begin{aligned} V(S_0,T)= \quad & ke^{-rT}\Phi\left(-\frac{log\frac{S_0}{k}+\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right)- \\ & -S_0\Phi\left(\frac{\log\frac{S_0}{k}+\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right), \end{aligned} \tag{14}$$

where the notation $\Phi(x)$ denotes the distribution function of $N(0,1)$.

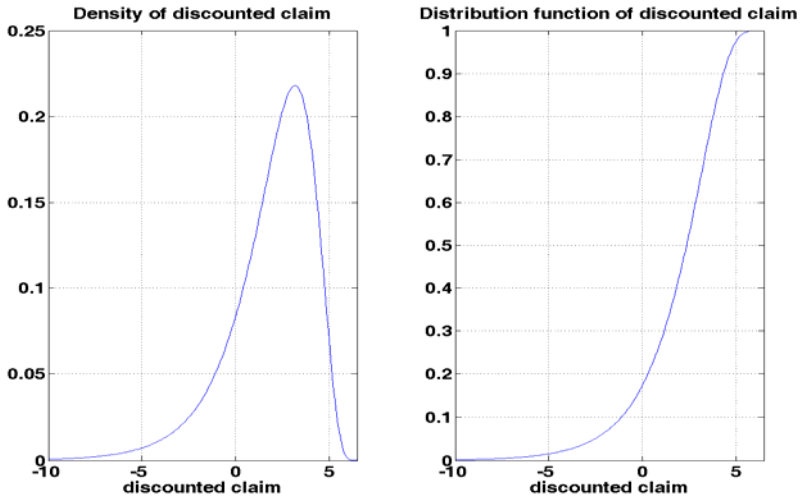


Figure 2: Density and distribution function of discounted put option claim with $\sigma = 0.2$, $k = 8$, $S_0 = 5$, $T = 5$, $r = 0.03$.

3.3 Pricing a arbitrary derivative

Let us suppose a contract paying off some arbitrary function of stock price at time $t = T$. The contract also has guaranteed minimum payout and maximum payout according to gains of the stock. More precisely, it is a five-year contract which pays out 90% of the ratio of the terminal and initial values of the stock. Or it pays out 130% of initial stock price if otherwise it would be less, or 180% of initial stock price if otherwise it would be more. How much is this contract worth at time $t = 0$? Thus, the claim X is

$$X = min\left\{ max\left\{ 1.3S_0, 0.9\frac{S_T}{S_0} \right\}, 1.8S_0 \right\}, \tag{15}$$

To price this contract, we have to compute the expected value of discounted claim under $\mathbb{Q}$ martingale measure (6) in the some way as pricing call and put options. The best way to do that is separating the claim into three partial claims. Then, we determine the value of each partial claim using a corresponding interval of stock price S at time $t = T$. At the end, we sum all values of partial claims and we get total value of the derivative (15). The partial claims and corresponding intervals are shown in (16)-(18).

$$S_T \in \left(0, \frac{1.3}{0.9} S_0^2\right) X_1 = 1.3 S_0. \tag{16}$$

$$S_T \in \left(\frac{1.3}{0.9} S_0^2, 2S_0^2\right) X_2 = 0.9 \frac{S_T}{S_0}. \tag{17}$$

$$S_T \in \left(2S_0^2, \infty\right) X_3 = 1.8 S_0. \tag{18}$$

We assume stock drift $\mu = 7\%$, stock volatility $\sigma = 15\%$, initial stock value $S_0 = 1$ and interest rate $r = 6.5\%$. We compute the value of partial claim (16) at first. To do that, we have to find the probability $P\left(S_T \in \left(0, \frac{1.3}{0.9} S_0^2\right)\right)$ under $\mathbb{Q}$ which is a probability of exercising the first claim. The distribution of S_T is given by (7). The density and distribution functions of S_T are shown in Figure 3. The resulting probability is

$$P_{\mathbb{Q}}\left(S_T \in \left(0, \frac{1.3}{0.9} S_0^2\right)\right) = 0.616.$$

Hence, using the derivative pricing formula (6) multiplied by the corresponding probability, we get the value of first partial claim.

$$V_0^1 = E_{\mathbb{Q}}^1\left(B_T^{-1} X_1\right) P_{\mathbb{Q}}\left(S_T \in \left(0, \frac{1.3}{0.9} S_0^2\right)\right) = e^{-rT} X_1 P_{\mathbb{Q}}\left(S_T \in \left(0, \frac{1.3}{0.9} S_0^2\right)\right).$$

Partial claim X_1 has value of 0.57863.

Applying the same procedure, we find value of X_3 (18). Having

$$V_0^3 = E_{\mathbb{Q}}^3\left(B_T^{-1} X_3\right) P_{\mathbb{Q}}\left(S_T \in \left(2S_0^2, \infty\right)\right) = e^{-rT} X_3 P_{\mathbb{Q}}\left(S_T \in \left(2S_0^2, \infty\right)\right).$$

where

$$P_{\mathbb{Q}}\left(S_T \in \left(2S_0^2, \infty\right)\right) = 0.1029.$$

Partial claim X_3 has value of 0.1338.

Now we are to price the remaining partial claim X_2. Again we apply (6)

$$V_0^2 = E_{\mathbb{Q}}^2\left(B_T^{-1} X_2\right) = 0.9e^{-rT} \int_{-\infty}^{\infty} t\, f(y)dy,$$

where $t = y$ for $y \in (\frac{1.3}{0.9} S_0^2, 2S_0^2)$ and $t = 0$ otherwise, $f(y)$ denotes density function of S_T under $\mathbb{Q}$ martingale measure. The value of partial claim X_2 is 0.30588.

The value of the derivative is 1.01831835 which is sum of values of the partial claims. Constructing the replication portfolio to fully hedge the claim (15) and solving corresponding PDE, we get value of 1.01831835 which is exactly the same price as applying risk-neutral approach (Baxter, 1999).

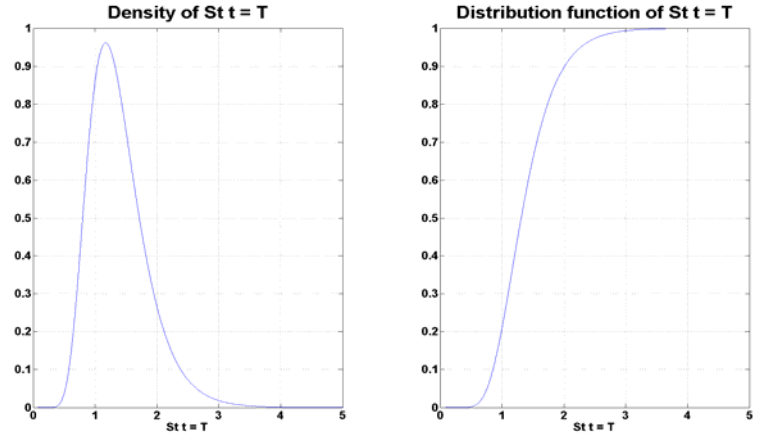


Figure 3: Density and distribution function of S_T with $\sigma = 0.15$, $S_0 = 1$, $T = 5$, $r = 0.065$.

4 Market practice

Let me take a look at real market options prices. How close are the market prices of options to those predicted by Black-Scholes and risk-neutral approach using lognormal distribution? Are the probability distributions of the asset prices really lognormal at any future time? Traders use the Black-Scholes model, but not exactly in the way that Black and Scholes originally intended. The risk-neutral approach is general enough to produce correct prices under changed conditions in the real market. Many information about underlying, including probability distribution, can be derived from options prices. I analyze the distribution of CEZ stock price using CEZ warrants prices.

4.1 Data description

I chose the options on CEZ stock due to their relatively market activity comparing to the options on the other Czech stocks. The only optional instruments on CEZ stock are warrants denominated in EUR and traded on Boerse Stuttgart. I used all priced call warrants on CEZ traded on Boerse Stuttgart as at 23 April 2009. Market mid prices were taken from Boerse Stuttgart via Bloomberg. Warrants are much like call options, and will often confer the same rights as an equity option and can even be traded in secondary markets. However, they have several key differences. Firstly, warrants are issued by private parties, typically the corporation on which a warrant is based and by investment banks, rather than a public options exchange. Secondly, warrants are not standardized like exchange-listed options. While investors can write stock options on the CEZ share, they are not permitted to do so with CEZ-listed warrants. As all CEZ warrants traded on Boerse Stuttgart are written by banks, they are not dilutive[2], hence we can handle the warrants as options, respecting the ratio naturally. The ratio determines how many warrants are needed in order to acquire the right to buy or sell one unit of the underlying instrument. Because all considered warrants are American-style calls on dividend paying stock, we are forced to modify our pricing formula for European call (10). Unlike European exercising style, American style means that the right to exercise the warrant can be invoked on any trading day during the life of the warrant. It can be shown that it is never optimal to exercise an American call option on non-dividend paying stock before the expiration date. When underlying stock pays off dividends, it is optimal to exercise the option either at a time immediately before the stock goes ex-dividend or at a time of expiration (Hull, 2006). CEZ pays off the dividend once a year, more importantly, once during the life of all considered options. Using Black's approximation, we price two Europeans options with exercise dates at maturity and at time immediately before dividend. The value of American-style option is then the greater of the values (Hull, 2006).

4.2 Volatility smile

Volatility is an unobservable parameter and must be estimated. If there are no tradable options in the market we can estimate the volatility from a history of the stock price. We face the problem of the length of time series in this case. For tradable options, we determine the volatility from option's market price and the pricing formula. Hence, this kind of volatility is implied. Our pricing formulas assume that the probability density function (PDF) of the underlying asset is lognormal with constant volatility at any given future time. This assumption is not the one made by traders. According to empirical studies, market assumes the probability distribution of an equity price with heavier left tail and less heavy right tail than the lognormal distribution. The reason is that traders consider that the probability of a larger downward movement in the stock price is higher than that predicted by the lognormal

[2]Warrants issued by the company itself are dilutive. When the warrant issued by the company is exercised, the company issues new shares of stock, hence the number of outstanding shares increases.

probability distribution. For standard options, traders use volatility smiles to allow for nonlognormality. They adjust the volatility parameter according to volatility smile and then they use lognormal distribution. The volatility smile defines the relationship between the implied volatility of an option and its strike price. For equity options, the volatility smile tends to be downward sloping. This fact means that call options with strike price below the market price of the stock (in-the-money) tend to have higher implied volatility than call options with strike price above the market price of the stock (out-of-the-money). Hence, in-the-money call options are relatively more expensive than out-of-the-money calls. We can also define the implied volatility as a function of strike and option's time to maturity which is standard practice in the market. We call this relationship volatility surface.

Having an option market price, we need pricing formula to derive the implied volatility of particular option. Carr has further developed and generalized the Geske-Johnson approach and has introduced general approach to the valuation of the American-style option, see (Carr, 1995) for details. Since CEZ pays off dividend once during the life of all considered options, it is optimal to exercise the option either immediately before the stock goes ex-dividend or at option's maturity. For that reason, we consider $n = 2$ for $\omega_1(\delta_k)$ and for $\omega_2(\delta_d)$ in pricing formula which leads to using bivariate normal distribution. Using the Carr risk-neutral approach, the value of American call is

$$V_0 = k\omega_1(\delta_k) - S_0\omega_2(\delta_d), \tag{19}$$

where $\omega_1(\delta_k)$ and $\omega_2(\delta_d)$ are defined in see (Carr, 1995).

Solving the equation (19) for σ, we obtain option's implied volatility. I solved the equation for all options with any market activity maturing on 17 June 2009 for all trading days since 1 January 2009 till 23 April 2009. The resulting implied volatility smile time series is displayed in Figure 4. We can observe that the volatility tends to increase with the increasing option's time to maturity and tends to decrease with increasing strike price. This outcome is in accordance with theory. Traders often call this downward sloping shape as a volatility skew. The warrant with strike value of CZK 900 seems to have constantly lower implied volatility, hence it seems to be constantly undervalued. To explain this discrepancy between the theory and the market, we have to take a look at market activity. The trading volume is far higher than by the others warrants and hence bid-offer spread is much smaller for all analyzed days. Due to higher liquidity on this warrant, its price represents the market opinion better than prices of the others warrants. I will use weights on the warrants subject to their liquidity to express market price significance later in the paper.

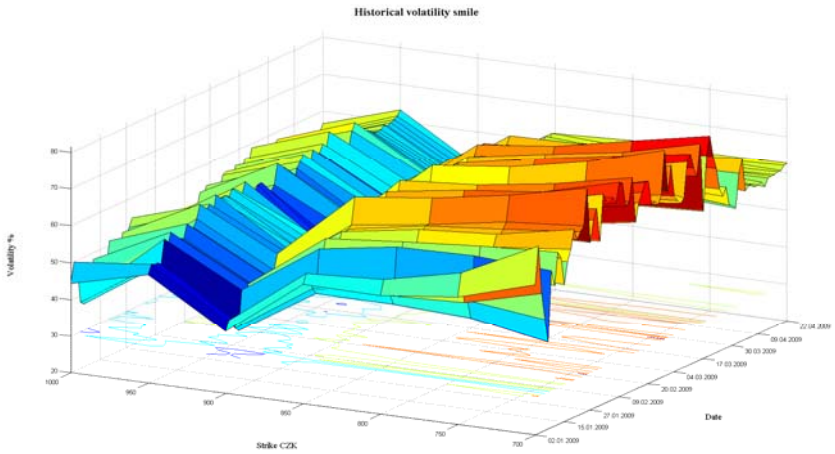


Figure 4: Volatility skew for American call options on CEZ stock

4.3 Implied distribution

The variance that is implied by an option's price is the market ex ante estimate of the underlying asset's return volatility over the remaining life of an options. More interestingly, it is possible to derive the higher moments of future asset values from the market options prices. These can be extracted in the form of an ex ante risk-neutral probability distribution of the underlying asset at particular future date. Rather then specifying the underlying asset price dynamics to infer the risk-neutral density function, it is possible to make assumption about the functional form of density function itself and to recover its parameters by minimizing the distance between the observed prices and those that are generated by the assumed probability distribution.
Since all considered warrants on CEZ stock are American-style calls with one dividend payout during warrants life, it is important to know whether it is optimal to exercise the warrant immediately before ex-dividend date or at maturity. To identify this, I used Black's approximation approach. Black's approximation takes account of early exercise in call options. This involves calculating the price of European options that mature at option's maturity date T and at time immediately before the stock goes ex-dividend (t_1), and then setting the American price equal to the greater of the two. We start considering the possibility of early exercise just prior to the ex-dividend date (i.e., at time t_1). If the warrant is exercised at time t_1, the investor receives $S_{t_1} - K$.

If the warrant is not exercised, the stock price drops to $S_{t_1} - D_1$, where D_1 is the dividend. The dividend on CEZ is determined in amount of CZK 50 per stock for 2009. The ex-dividend date is 13 May 2009. Because lower bound for call options is $S_{t_1} - D_1 - ke^{-r(T-t_1)}$ (Hull, 2006), it follows that if

$$S_{t_1} - D_1 - ke^{-r(T-t_1)} \geq S_{t_1} - D_1,$$

that is,

$$D_1 \leq k\left(1 - e^{-r(T-t_1)}\right),$$

it cannot be optimal to exercise at time t_1. On the other hand, if

$$D_1 > k\left(1 - e^{-r(T-t_1)}\right), \tag{20}$$

for any reasonable assumption about the stochastic process followed by the stock price, it can be shown that it is always optimal to exercise at time t_1. The inequality in (20) will tend to be satisfied when ex-dividend date t_1 is fairly close to the maturity of the option T and the dividend is large. I used forward mid rate between PRIBOR and PRIBID as risk-free interest rate. Applying this method to all considered warrants, I got the information whether it is optimal to exercise the warrant at time t_1 or at time T. Using this information, we can derive the implied distribution of CEZ equity price.

The price of European call option is given by (6) at time t where the claim is $X = (S_T - k)^+$. Unlike density function used in chapter "Pricing a call option", I assume that the risk-neutral distribution of stock is the mixture of two lognormals rather than lognormal at any future time. We can use arbitrary density function which fulfills the criteria of probability density function and has finite variance. The problem with using other than the Gaussian density functions is that the underlying price distribution changes as the holding period changes. Under assumption of lognormally distributed daily returns must be arbitrary length holding period price distributions also lognormal. No other finite variance distribution is similarly stable (Bahra, 1997). Moreover, the functional form assumed for the density function should be relatively flexible. In particular, it should be able to capture the main contributions to the volatility smile, namely the skewness and the kurtosis of the underlying distributions. A weighted sum of independent lognormal density functions fits these criteria. The probability distribution of any future stock price is given by

$$f(S_T) = \theta L\left(S_T \mid \mu_1, \sigma_1, S_0\right) + \left(1-\theta\right) L\left(S_T \mid \mu_2, \sigma_2, S_0\right), \tag{21}$$

where $L(S_T)$ is lognormal density function given by (22) and θ is probability weight satisfying the condition that $\theta \in [0,1]$. Thus the fitted value for a call price, given parameters $\{\mu_1, \sigma_1, \mu_2, \sigma_2, \theta\}$, is given by

$$\hat{C}_t\left(k, \tau \mid \mu_1, \sigma_1, \mu_2, \sigma_2, \theta\right) = e^{-r\tau}\left(\theta \int_0^\infty x f(x \mid \mu_1, \sigma_1) dx + \left(1-\theta\right) \int_0^\infty x f(x \mid \mu_2, \sigma_2) dx\right).$$

Having fitted values of calls warrants, we can minimize the sum of squared errors, with respect to the five distributional parameters and risk-free rate r, between the warrant prices generated by mixture distribution model and those actually observed in the market Figure 5 shows the two-lognormal mixture implied density function of CEZ stock with its weighted component lognormal density functions on 17 June 2009. Further, Figure 5 shows the lognormal density function with the same mean and standard deviation as the implied distribution. Figure 5 displays the expected value of implied distribution and forward price as well. We can observe that the expected value of CEZ stock derived from implied distribution and forward price of CEZ stock are practically identical which was one of the criteria using by minimizing the objective function. We can observe that the resulting implied distribution of CEZ stock price has greater kurtosis than lognormal distribution and gives higher price to in-the-money warrants and lower price to out-of-the-money warrants than initially considered lognormal distribution.

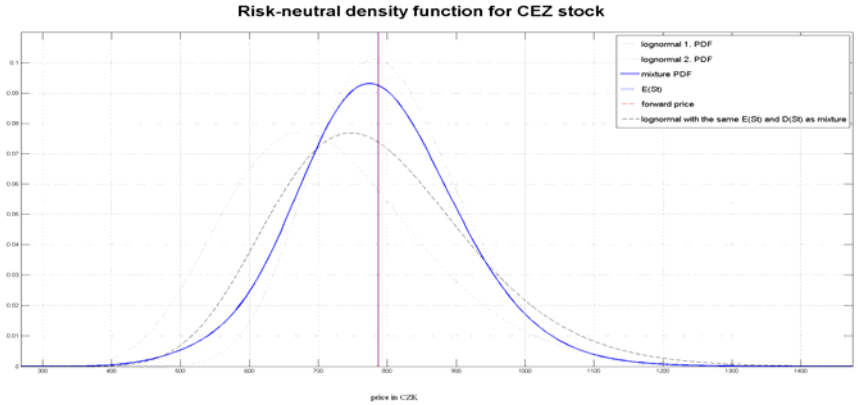


Figure 5: Two-lognormal mixture implied density function of CEZ stock on 17 June 2009 with its weighted component lognormal density functions, lognormal density function with the same mean and standard deviation as the implied distribution, expected value of implied distribution, forward price of CEZ stock

Conclusions

In this paper, I have presented risk-neutral option pricing approach and have shown its application in pricing real market derivatives. I have derived probability distribution of an underlying asset price, stock in particular, from binomial tree process of the stock price in real world. Since real world is irrelevant for pricing, I changed the measure and have derived risk-neutral distribution of the stock price. The discrete trees are only an approximation to the way that prices actually move. In practice, a price can change at any instant, rather than just at some fixed tick-times. Using central limit theorem, I came over to continuous time. Under assumption of constant volatility is the distribution of stock price lognormal. I have priced three different derivatives and showed no bias between prices produced by risk-neutral approach and those produced by replication strategy approach. The replication

strategy approach leads to solving partial differential equations PDE (heavily used Black-Scholes formula for instance). Presented risk-neutral approach is suitable for pricing any European-style claim function of underlying asset. I see its main advantage in its relative simplicity comparing to continuous stochastic process approach which leads to solving stochastic differential equations SDE.

The results in this paper provide evidence of nonlognormality of CEZ stock price. I used American-style call warrants on CEZ with dividend pay off during the life of warrant to construct volatility smile time series. As CEZ pay off the dividend, we have to take into account the possibility of early exercise. We can observe that volatility is function of strike price rather than a constant value. Furthermore, I have derived the implied distribution of CEZ stock price from warrants prices observed in the market. I used two-lognormal mixture density function which is, in my opinion, flexible enough to capture the main contributions to the volatility smile, namely the skewness and the kurtosis of the underlying distribution. Minimizing the objective function subject to highly nonlinear constraint, I have found parameters estimates of the implied density function. Not all warrants are equally liquid and thus their prices are not equally reliable. Hence, I used traded volume on each warrant as weights.

The resulting implied distribution of CEZ stock price has greater kurtosis than lognormal distribution and gives higher price to in-the-money warrants and lower price to out-of-the-money warrants than initially considered lognormal distribution. Using implied density function, we can price not traded derivatives or search for mispriced ones. Implied distribution provides us information about future in behaviour of underlying asset in terms of probabilities as well. I have derived higher moments of CEZ stock price like skewness and kurtosis. These statistics provide a useful way of tracking the behaviour of implied density functions over the life of a single contract and making comparisons across contracts. Observing day-to-day changes in these statistic, we may also analyze market reactions and future expectations to company earnings releases, money market operations, government bond auctions, etc.

References

1. Ait-Sahalia Y and Lo A W (1998): Nonparametric estimation of state-price densities implied in financial asset prices.*Journal of Finance*, 53(2): 499-547
2. Bahra B (1997): Implied risk-neutral probability density functions from options prices.*Bank of England working Paper*,(66)
3. Baxter M and Rennie A (1999)*Financial Calculus*. Cambridge University Press, Cambridge.
4. Bliss R R and Panigirtzoglou N (2002): Testing the stability of implied probability density functions.*Journal of Banking and Finance*,26(3)
5. Carr P (1995) The Valuation of American Exchange Options with Application to Real Options. In Trigeorgis L:*Real Options in Capital Investment: New Contributions*.Praeger Publishers, Westport CT, 109-120.
6. Hull J C (2006) *Options, Futures, and Other Derivatives*. 6, Prentice Hall,
7. Joshi M (2005) *The Concepts and Practice of Mathematical Finance*. Cambridge University Press, Cambridge
8. Rubinstein M (1994): Implied Binomial Trees. *Journal of Finance*, 69 (3), 771-818

Fibonacci and his Sequence

Jan Coufal

University of Economics and Management,
Department of Informatics and Quantitative Methods,
Nárožní 2600/9a, 158 00 Prague 5, Czech republic
E-mail: jan.coufal@vsem.cz, jan.n.coufal@seznam.cz

Abstract. Leonardo of Pisa (* c. 1170 – † c. 1250), also known as Leonardo Pisano, Leonardo Bonacci, Leonardo Fibonacci, or, most commonly, simply Fibonacci, was an Italian mathematician, considered by some "*the most talented mathematician of the Middle Ages*"
Fibonacci is best known to the modern world for:
– The spreading of the Hindu-Arabic numeral system in Europe, primarily through the publication in the early 13th century of his Book of Calculation, the *Liber Abaci*.
– A number sequence named after him known as the Fibonacci numbers, which he did not discover but used as an example in the Liber Abaci. In the Fibonacci sequence of numbers, each number is the sum of the previous two numbers, starting with 0 and 1. Thus the sequence begins 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610 etc.

Keywords: Leonardo Pisano, Fibonacci, Fibonacci sequence, Fibonacci numbers, Fibonacci Roulette System

Fibonacci

Leonardo Pisano Fibonacci (* c. 1170 probably in Pisa, now in Italy – † c. 1250 possibly in Pisa), also known as Leonardo of Pisa, Leonardo Di Pisa, Leonardo Pisano, Leonardo Bonacci, Leonardo Fibonacci, or, most commonly, simply Fibonacci (derived from filius Bonacci, meaning son of Bonaccio), was an Italian mathematician, considered by some "*the most talented mathematician of the Middle Ages*".

Leonardo was born in Pisa (now in Italy) about 1170. His father Guglielmo was nicknamed Bonaccio ("good natured" or "simple"). Leonardo's mother, Alessandra, died when he was nine years old. Fibonacci himself sometimes used the name Bigollo, which may mean good-for-nothing or a traveller. As stated in [1]:

Did his countrymen wish to express by this epithet their disdain for a man who concerned himself with questions of no practical value, or does the word in the Tuscan dialect mean a much-travelled man, which he was?

Fibonacci was born in Italy but was educated in North Africa where his father, Guilielmo, held a diplomatic post. His father's job was to represent the merchants of the Republic of Pisa who were trading in Bugia, later called Bougie and now called

Bejaia. Bejaia is a Mediterranean port in northeastern Algeria. The town lies at the mouth of the Wadi Soummam near Mount Gouraya and Cape Carbon. Fibonacci was taught mathematics in Bugia and travelled widely with his father and recognized the enormous advantages of the mathematical systems used in the countries they visited. Fibonacci writes in his famous book *Liber abaci* (Book of Calculation , 1202):

After my father's appointment by his homeland as state official in the customs house of Bugia for the Pisan merchants who thronged to it, he took charge; and in view of its future usefulness and convenience, had me in my boyhood come to him and there wanted me to devote myself to and be instructed in the study of calculation for some days.

There, following my introduction, as a consequence of marvelous instruction in the art, to the nine digits of the Hindus, the knowledge of the art very much appealed to me before all others, and for it I realized that all its aspects were studied in Egypt, Syria, Greece, Sicily, and Provence, with their varying methods; and at these places thereafter, while on business.

I pursued my study in depth and learned the give-and-take of disputation. But all this even, and the algorism, as well as the art of Pythagoras, I considered as almost a mistake in respect to the method of the Hindus. (Modus Indorum). Therefore, embracing more stringently that method of the Hindus, and taking stricter pains in its study, while adding certain things from my own understanding and inserting also certain things from the niceties of Euclid's geometric art, I have striven to compose this book in its entirety as understandably as I could, dividing it into fifteen chapters.

Almost everything which I have introduced I have displayed with exact proof, in order that those further seeking this knowledge, with its pre-eminent method, might be instructed, and further, in order that the Latin people might not be discovered to be without it, as they have been up to now. If I have perchance omitted anything more or less proper or necessary, I beg indulgence, since there is no one who is blameless and utterly provident in all things.

The nine Indian figures are:

9 8 7 6 5 4 3 2 1

With these nine figures, and with the sign 0 *... any number may be written.*

Fibonacci ended his travels around the year 1200 and at that time he returned to Pisa. There he wrote a number of important texts which played an important role in reviving ancient mathematical skills and he made significant contributions of his own. Fibonacci lived in the days before printing, so his books were hand written and the only way to have a copy of one of his books was to have another hand-written copy made. Of his books we still have copies of *Liber abaci* (1202), *Practica geometriae* (1220), *Flos* (1225), and *Liber quadratorum* (published in 1225). Given that relatively few hand-made copies would ever have been produced, we are fortunate to have access to his writing in these works. However, we know that he wrote some other texts which, unfortunately, are lost. His book on commercial arithmetic *Di minor guisa* is lost as is his *Commentary on Book X of Euclid's Elements* which contained a numerical treatment of irrational numbers which Euclid had approached from a geometric point of view.

One might have thought that at a time when Europe was little interested in scholarship, Fibonacci would have been largely ignored. This, however, is not so and widespread interest in his work undoubtedly contributed strongly to his importance.

Fibonacci was a contemporary of Jordanus but he was a far more sophisticated mathematician and his achievements were clearly recognized, although it was the practical applications rather than the abstract theorems that made him famous to his contemporaries.

The Holy Roman emperor was Frederick II. He had been crowned king of Germany in 1212 and then crowned Holy Roman emperor by the Pope in St Peter's Church in Rome in November 1220. Frederick II supported Pisa in its conflicts with Genoa at sea and with Lucca and Florence on land, and he spent the years up to 1227 consolidating his power in Italy. State control was introduced on trade and manufacture, and civil servants to oversee this monopoly were trained at the University of Naples which Frederick founded for this purpose in 1224.

Frederick became aware of Fibonacci's work through the scholars at his court who had corresponded with Fibonacci since his return to Pisa around 1200. These scholars included Michael Scotus who was the court astrologer, Theodorus Physicus the court philosopher and Dominicus Hispanus who suggested to Frederick that he meet Fibonacci when Frederick's court met in Pisa around 1225.

Johannes of Palermo, another member of Frederick II's court, presented a number of problems as challenges to the great mathematician Fibonacci. Three of these problems were solved by Fibonacci and he gives solutions in Flos which he sent to Frederick II. We give some details of one of these problems below.

After 1228 there is only one known document which refers to Fibonacci. This is a decree made by the Republic of Pisa in 1240 in which a salary is awarded to:

... the serious and learned Master Leonardo Bigollo ....

This salary was given to Fibonacci in recognition for the services that he had given to the city, advising on matters of accounting and teaching the citizens.

Liber abaci, published in 1202 after Fibonacci's return to Italy, was dedicated to Scotus. The book was based on the arithmetic and algebra that Fibonacci had accumulated during his travels. The book, which went on to be widely copied and imitated, introduced the Hindu-Arabic place-valued decimal system and the use of Arabic numerals into Europe. Indeed, although mainly a book about the use of Arab numerals, which became known as algorism, simultaneous linear equations are also studied in this work. Certainly many of the problems that Fibonacci considers in Liber abaci were similar to those appearing in Arab sources.

The second section of Liber abaci contains a large collection of problems aimed at merchants. They relate to the price of goods, how to calculate profit on transactions, how to convert between the various currencies in use in Mediterranean countries, and problems which had originated in China.

A problem in the third section of Liber abaci led to the introduction of the Fibonacci numbers and the Fibonacci sequence for which Fibonacci is best remembered today:

A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on becomes productive?

The resulting sequence is $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610$ etc. (Fibonacci omitted the first term in Liber abaci). This sequence, in which each number is the sum of the two preceding numbers, has proved

extremely fruitful and appears in many different areas of mathematics and science. The Fibonacci Quarterly is a modern journal devoted to studying mathematics related to this sequence.

Many other problems are given in this third section, including these types, and many many more:

A spider climbs so many feet up a wall each day and slips back a fixed number each night, how many days does it take him to climb the wall.

A hound whose speed increases arithmetically chases a hare whose speed also increases arithmetically, how far do they travel before the hound catches the hare.

Calculate the amount of money two people have after a certain amount changes hands and the proportional increase and decrease are given.

There are also problems involving perfect numbers, problems involving the Chinese remainder theorem and problems involving summing arithmetic and geometric series.

Fibonacci treats numbers such as $\sqrt{10}$ in the fourth section, both with rational approximations and with geometric constructions.

A second edition of Liber abaci was produced by Fibonacci in 1228 with a preface, typical of so many second editions of books, stating that:

... new material has been added [to the book] from which superfluous had been removed...

Another of Fibonacci's books is Practica geometriae written in 1220 which is dedicated to Dominicus Hispanus whom we mentioned above. It contains a large collection of geometry problems arranged into eight chapters with theorems based on Euclid's Elements and Euclid's On Divisions. In addition to geometrical theorems with precise proofs, the book includes practical information for surveyors, including a chapter on how to calculate the height of tall objects using similar triangles. The final chapter presents what Fibonacci called geometrical subtleties [1]:

Among those included is the calculation of the sides of the pentagon and the decagon from the diameter of circumscribed and inscribed circles; the inverse calculation is also given, as well as that of the sides from the surfaces. ... to complete the section on equilateral triangles, a rectangle and a square are inscribed in such a triangle and their sides are algebraically calculated ...

In Flos Fibonacci gives an accurate approximation to a root of $10x + 2x^2 + x^3 = 20$, one of the problems that he was challenged to solve by Johannes of Palermo. This problem was not made up by Johannes of Palermo, rather he took it from Omar Khayyam's algebra book where it is solved by means of the intersection of a circle and a hyperbola. Fibonacci proves that the root of the equation is neither an integer nor a fraction, nor the square root of a fraction. He then continues:

And because it was not possible to solve this equation in any other of the above ways, I worked to reduce the solution to an approximation.

Without explaining his methods, Fibonacci then gives the approximate solution in sexadecimal notation as $1.22.7.42.33.4.40$ (this is written to base 60, so it is $1 + \frac{22}{60} + \frac{7}{60^2} + \frac{42}{60^3} + \frac{33}{60^4} + \frac{4}{60^5} + \frac{40}{60^6}$). This converts to the decimal 1.3688081075 which is correct to nine decimal places, a remarkable achievement.

Liber quadratorum, written in 1225, is Fibonacci's most impressive piece of work, although not the work for which he is most famous. The book's name means the book of squares and it is a number theory book which, among other things, examines methods to find Pythagorean triples. Fibonacci first notes that square numbers can be constructed as sums of odd numbers, essentially describing an inductive construction using the formula $n^2 + (2n+1) = (n+1)^2$. Fibonacci writes:

I thought about the origin of all square numbers and discovered that they arose from the regular ascent of odd numbers. For unity is a square and from it is produced the first square, namely 1*; adding* 3 *to this makes the second square, namely* 4*, whose root is* 2*; if to this sum is added a third odd number, namely* 5*, the third square will be produced, namely* 9*, whose root is* 3*; and so the sequence and series of square numbers always rise through the regular addition of odd numbers.*

To construct the Pythagorean triples, Fibonacci proceeds as follows:

Thus when I wish to find two square numbers whose addition produces a square number, I take any odd square number as one of the two square numbers and I find the other square number by the addition of all the odd numbers from unity up to but excluding the odd square number. For example, I take 9 *as one of the two squares mentioned; the remaining square will be obtained by the addition of all the odd numbers below* 9*, namely* 1*,* 3*,* 5*,* 7*, whose sum is* 16*, a square number, which when added to* 9 *gives* 25*, a square number.*

Fibonacci also proves many interesting number theory results such as:

there is no x, y such that $x^2 + y^2$ and $x^2 - y^2$ are both squares, and $x^4 - y^4$ cannot be a square.

He defined the concept of a congruum, a number of the form $ab(a+b)(a-b)$, if $a+b$ is even, and 4 times this if $a+b$ is odd. Fibonacci proved that a congruum must be divisible by 24 and he also showed that for x, c such that $x^2 + c$ and $x^2 - c$ are both squares, then c is a congruum. He also proved that a square cannot be a congruum.

As stated in [2]:

... the Liber quadratorum alone ranks Fibonacci as the major contributor to number theory between Diophantus and the 17th-century French mathematician Pierre de Fermat.

Fibonacci's influence was more limited than one might have hoped and apart from his role in spreading the use of the Hindu-Arabic numerals and his rabbit problem, Fibonacci's contribution to mathematics has been largely overlooked. As explained in [1]:

Direct influence was exerted only by those portions of the "Liber abaci" and of the "Practica" that served to introduce Indian-Arabic numerals and methods and contributed to the mastering of the problems of daily life. Here Fibonacci became the teacher of the masters of computation and of the surveyors, as one learns from the "Summa" of Luca Pacioli ... Fibonacci was also the teacher of the "Cossists", who took their name from the word 'causa' which was first used in the West by Fibonacci

in place of 'res' or 'radix'. His alphabetic designation for the general number or coefficient was first improved by Viète ...

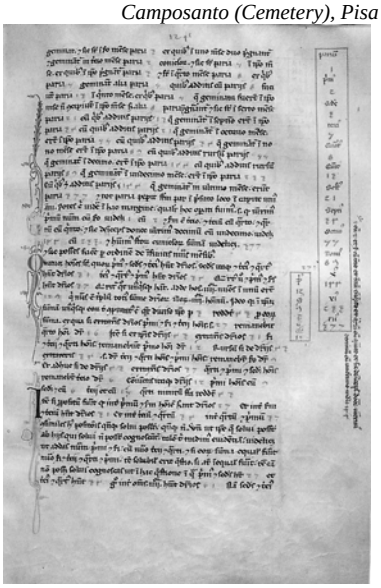
Fibonacci's work in number theory was almost wholly ignored and virtually unknown during the Middle ages. Three hundred years later we find the same results appearing in the work of Maurolico.

The portrait above is from a modern engraving and is believed to not be based on authentic sources.

Portrait of Fibonacci

Monument dedicated to Fibonacci, Camposanto (Cemetery), Pisa

Detail of monument in Composanto



A page of the Liber Abaci from the Biblioteca Nazionale di Firenze

Fibonacci sequence and Fibonacci numbers

In the Fibonacci sequence of numbers, each number is the sum of the previous two numbers, starting with 0 and 1.

The Fibonacci sequence was well known in ancient India, where it was applied to the metrical sciences (prosody), long before it was known in Europe. Developments have been attributed to Pingala (200 BC), Virahanka (6th century AD), Gopāla (c.1135 AD), and Hemachandra (c.1150 AD).

The Fibonacci sequence is formed by adding S to a pattern of length $n+1$, or L to a pattern of length n; and the prosodicists showed that the number of patterns of length n is the sum of the two previous numbers in the sequence. Fibonacci sequence are used in a polyphase version of the merge sort algorithm in which an unsorted list is divided into two lists whose lengths correspond to sequential Fibonacci numbers – by dividing the list so that the two parts have lengths in the approximate proportion $\frac{1+\sqrt{5}}{2}$. Donald Knuth reviews this work in The Art of Computer Programming. Here a tape-drive implementation of the polyphase merge sort was described. Fibonacci sequence arise in the analysis of the Fibonacci heap data structure.

The Fibonacci sequence and principle is also used in the financial markets. It is used in trading algorithms, applications and strategies. Some typical forms include: the Fibonacci Fans, the Fibonacci Arcs, the Fibonacci Retracements, the Fibonacci Expansion, the Fibonacci Channels and the Fibonacci Time Extensions (or the Fibonacci Time Zones).

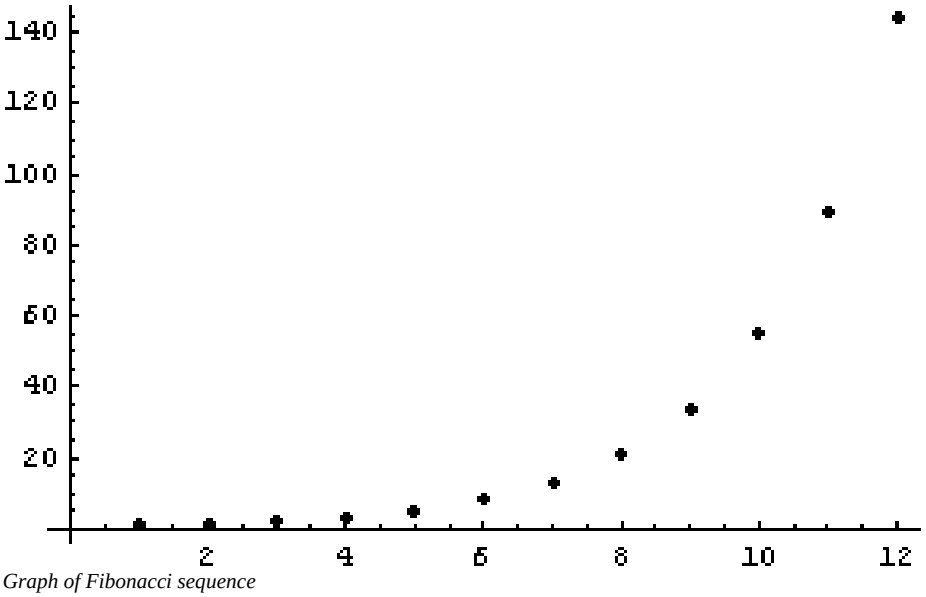
In the West, the sequence was studied by Leonardo of Pisa, known as Fibonacci, in his Liber Abaci (1202). He considers the growth of an idealized (biologically unrealistic) rabbit population, assuming that:

* In the "zeroth" month, there is one pair of rabbits (additional pairs of rabbits is 0).
* In the first month, the first pair begets another pair (additional pairs of rabbits is 1).
* In the second month, both pairs of rabbits have another pair, and the first pair dies (additional pairs of rabbits 1).
* In the third month, the second pair and the new two pairs have a total of three new pairs, and the older second pair dies (additional pairs of rabbits 2).

The laws of this are that each pair of rabbits has 2 pairs in its lifetime, and dies.

Let the population at month $n+2$ be $F(n+2)$. At this time, only rabbits who were alive at month n are fertile and produce offspring, so $F(n)$ pairs are added to the current population of $F(n+1)$. Thus the total is $F(n+2)=F(n+1)+F(n)$, $F(0)=0$ and $F(1)=1$. That is recurrence formula of Fibonacci sequence. Members of Fibonacci sequence are Fibonacci numbers.

If you have more than one surname, please make sure that the Volume Editor knows how you are to be listed in the author index.



Graph of Fibonacci sequence

Fibonacci numbers also appear in the description of the reproduction of a population of idealized bees, according to the following rules:

* If an egg is laid by an unmated female, it hatches a male.
* If, however, an egg was fertilized by a male, it hatches a female.

Thus, a male bee will always have one parent, and a female bee will have two.

If one traces the ancestry of any male bee (1 bee), he has 1 female parent (1 bee). This female had 2 parents, a male and a female (2 bees). The female had two parents, a male and a female, and the male had one female (3 bees). Those two females each had two parents, and the male had one (5 bees). This sequence of numbers of parents is the Fibonacci sequence. This is an idealization that does not describe *actual* bee ancestries. In reality, some ancestors of a particular bee will always be sisters or brothers, thus breaking the lineage of distinct parents.

Closed form expression of Fibonacci sequence

The Fibonacci recursion $F(n+2)-F(n+1)-F(n)=0$, $F(0)=0$ and $F(1)=1$ is the linear difference equation. Characteristic equation is $\lambda^2-\lambda-1=0$.

Roots are $\lambda_1=\frac{1}{2}+\frac{\sqrt{5}}{2}=\frac{1+\sqrt{5}}{2}$ and $\lambda_2=\frac{1}{2}-\frac{\sqrt{5}}{2}=\frac{1-\sqrt{5}}{2}=1-\lambda_1=-\frac{1}{\lambda_1}$. Number $\frac{1}{\lambda_1}=\frac{2}{1+\sqrt{5}}\approx 0.618034$ is golden ratio.

General solution is

$$F(n)=C_1\lambda_1^n+C_2\lambda_2^n=C_1\left(\tfrac{1}{2}+\tfrac{\sqrt{5}}{2}\right)^n+C_2\left(\tfrac{1}{2}-\tfrac{\sqrt{5}}{2}\right)^n=$$
$$=C_1\left(\tfrac{1+\sqrt{5}}{2}\right)^n+C_2\left(\tfrac{1-\sqrt{5}}{2}\right)^n=C_1\lambda_1^n+C_2(1-\lambda_1)^n=$$
$$=C_1\left(\tfrac{1}{2}+\tfrac{\sqrt{5}}{2}\right)^n+C_2\left(1-\left(\tfrac{1}{2}+\tfrac{\sqrt{5}}{2}\right)\right)^n=C_1\left(\tfrac{1+\sqrt{5}}{2}\right)^n+C_2\left(1-\left(\tfrac{1+\sqrt{5}}{2}\right)\right)^n=$$
$$=C_1\lambda_1^n+C_2\left(-\tfrac{1}{\lambda_1}\right)^n=C_1\left(\tfrac{1}{2}+\tfrac{\sqrt{5}}{2}\right)^n+C_2\left(-\frac{1}{\frac{1}{2}+\frac{\sqrt{5}}{2}}\right)^n=C_1\left(\tfrac{1+\sqrt{5}}{2}\right)^n+C_2\left(-\tfrac{2}{1+\sqrt{5}}\right)^n$$

These linear combinations form a two-dimensional linear vector space; the original Fibonacci sequence can be found in this space.

Initial conditions are $F(0)=0$ and $F(1)=1$, than $C_1=\frac{1}{\sqrt{5}}$ and $C_2=-\frac{1}{\sqrt{5}}$.

Establishing the base cases of the recursion, proving that

$$F(n)=\frac{\lambda_1^n-(1-\lambda_1)^n}{\sqrt{5}}=\frac{\left(\frac{1}{2}+\frac{\sqrt{5}}{2}\right)^n-\left(\frac{1}{2}-\frac{\sqrt{5}}{2}\right)^n}{\sqrt{5}}=\frac{\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n}{\sqrt{5}}=\frac{\left(1+\sqrt{5}\right)^n-\left(1-\sqrt{5}\right)^n}{2^n\cdot\sqrt{5}}$$

for all natural n.

Johannes Kepler observed that the ratio of consecutive Fibonacci numbers converges. He wrote that "as 5 is to 8 so is 8 to 13, practically, and as 8 is to 13, so is 13 to 21 almost", and concluded that the limit approaches the golden ratio.

More formally, it follows from the explicit formula that

$$\lim_{n\to\infty}\frac{F(n)}{F(n+1)}=\lim_{n\to\infty}\frac{\frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right)}{\frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right)}=$$
$$=\lim_{n\to\infty}\frac{\frac{1}{2^n}}{\frac{1}{2^{n+1}}}\cdot\frac{\left(1+\sqrt{5}\right)^n-\left(1-\sqrt{5}\right)^n}{\left(1+\sqrt{5}\right)^{n+1}-\left(1-\sqrt{5}\right)^{n+1}}=$$
$$=\lim_{n\to\infty}2\cdot\frac{\left(1+\sqrt{5}\right)^n\left(1-\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n\right)}{\left(1+\sqrt{5}\right)^n\left(1+\sqrt{5}-\left(1-\sqrt{5}\right)\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n\right)}=$$
$$=\lim_{n\to\infty}2\cdot\frac{1-\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n}{1+\sqrt{5}-\left(1-\sqrt{5}\right)\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n}=2\cdot\frac{1-0}{1+\sqrt{5}-\left(1-\sqrt{5}\right)\cdot 0}=$$
$$=\frac{2}{1+\sqrt{5}}\approx 0.618034\text{, because }\left|\frac{1-\sqrt{5}}{1+\sqrt{5}}\right|<1\text{ and thus }\lim_{n\to\infty}\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n=0;$$

or $\lim_{n\to\infty}\frac{F(n+1)}{F(n)}=\frac{1+\sqrt{5}}{2}\approx 1.61803398$.

This convergence does not depend on the starting values chosen, excluding 0, 0. For example, the initial values 19 and 31 generate the sequence 19, 31, 50, 81, 131, 212, 343, 555 etc. The ratio of consecutive terms in this sequence shows the same convergence towards the golden ratio.

Proof.

If $G(n) = C_1\left(\frac{1+\sqrt{5}}{2}\right)^n + C_2\left(\frac{1-\sqrt{5}}{2}\right)^n$, $C_1 \neq 0$ and $C_2 \neq 0$, than

$$\lim_{n\to\infty}\frac{G(n)}{G(n+1)} = \lim_{n\to\infty}\frac{C_1\left(\frac{1+\sqrt{5}}{2}\right)^n - C_2\left(\frac{1-\sqrt{5}}{2}\right)^n}{C_1\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - C_2\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}} =$$

$$= \lim_{n\to\infty}\frac{\frac{1}{2^n}}{\frac{1}{2^{n+1}}}\cdot\frac{C_1\left(1+\sqrt{5}\right)^n - C_2\left(1-\sqrt{5}\right)^n}{C_1\left(1+\sqrt{5}\right)^{n+1} - C_2\left(1-\sqrt{5}\right)^{n+1}} =$$

$$= \lim_{n\to\infty} 2\cdot\frac{\left(1+\sqrt{5}\right)^n\left(C_1 - C_2\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n\right)}{\left(1+\sqrt{5}\right)^n\left(C_1\left(1+\sqrt{5}\right) - C_2\left(1-\sqrt{5}\right)\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n\right)} =$$

$$= \lim_{n\to\infty} 2\cdot\frac{C_1 - C_2\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n}{C_1\left(1+\sqrt{5}\right) - C_2\left(1-\sqrt{5}\right)\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n} = 2\cdot\frac{C_1 - C_2\cdot 0}{C_1\cdot\left(1+\sqrt{5}\right) - C_2\cdot\left(1-\sqrt{5}\right)\cdot 0} =$$

$$= \frac{2\cdot C_1}{C_1\cdot\left(1+\sqrt{5}\right)} = \frac{2}{1+\sqrt{5}} \approx 0.618034$$, because $\left|\frac{1-\sqrt{5}}{1+\sqrt{5}}\right| < 1$ and thus

$\lim_{n\to\infty}\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n = 0$; or $\lim_{n\to\infty}\frac{G(n+1)}{G(n)} = \frac{1+\sqrt{5}}{2} \approx 1.61803398$.

The Fibonacci Roulette System

The next number in the series is simply the sum of the previous two numbers. The starting number is 1. The second number calculated from $0+1$ (no number in front of the first 1) and is 1 again. The next number is $1+1$ or 2, then $1+2$ for 3, then $2+3=5$ and $3+5=8$, etc. The system works similarly to the Labouchere or cancellation system, only the player starts out with an empty line. If the first bet is won, then the sequence is over and the player has won. No numbers need to be written down. If the first bet is lost, then a line is started and a "1" is written down. The next number in the sequence represents the following wager size. If this bet is lost, then it is added to the end of the line. As each bet is lost, it is added to the end of the series. If a bet is won, the last number in the series is crossed out. An example here will help clarify things:

1.) Bet 1 unit and lose: 1 -1 unit
2.) Bet 1 unit and lose: 1-1 -2 units

3.) Bet 2 units and lose: 1-1-2 -4 units
4.) Bet 3 units and win: 1-x-x -1 unit
5.) Bet 1 unit and lose: 1-1 -2 units
6.) Bet 2 units and lose: 1-1-2 -4 units
7.) Bet 3 units and lose: 1-1-2-3 -7 units
8.) Bet 5 units and win: 1-1-x-x -2 units
9.) Bet 2 units and lose: 1-1-2 -4 units
10.) Bet 3 units and win: 1-x-x -1 unit
11.) Bet 1 unit and lose: 1-1 -2 units
12.) Bet 2 units and win: x-x +0 units
13.) Bet 1 unit and win: stop +1 unit
-Series has been won-

Our player starts with a one unit loss, so a "1" is recorded to start the line. Another "1" is added after the second wager of one unit loses. The third stake requires a two-unit wager and loses, so a "2" is added. The fourth bet of three units finally wins and the "1-2" can be cancelled from the line. Because each wager adds up to the previous two bets, the last two numbers on the line can be crossed out when a bet wins. The next three bets lose, escalating our eighth stake up to five units. Our player experiences a win at this level, allowing him to cancel out the "2-3" at the end of the line. The ninth bet of two units loses, so the line grows to "1-1-2." A win, loss and win on the tenth, eleventh and twelve wagers finally wipe out the betting line. The player needs and gets a win at this point to go up a net profit of one unit and win the sequence.

With only five wins and eight losses, this particular sequence of wins and losses is tough, but our player is able to pull it out. On the eighth wager, the stake reaches a high of five units. If that bet had lost, our player would be twelve units in the hole. At a $5 unit size, that equates to a $60 deficit. The next wager from here would be eight units and another loss would put him back 20 units total. If you elect to use the Fibonacci, I would highly recommend that you limit your top bet to five units. If you lose your wager at this level, then abandon the series. Things get ugly too quickly from here. Stop and regroup. Let's take the Fibonacci up to twelve straight losses to see how quickly the wagers can mount:

1.) Bet 1 unit and lose: 1 -1 unit
2.) Bet 1 unit and lose: 1-1 -2 units
3.) Bet 2 units and lose: 1-1-2 -4 units
4.) Bet 3 units and lose: 1-1-2-3 -7 units
5.) Bet 5 units and lose: 1-1-2-3-5 -12 units
6.) Bet 8 units and lose: 1-1-2-3-5-8 -20 units
7.) Bet 13 units and lose: 1-1-2-3-5-8-13 -33 units
8.) Bet 21 units and lose: 1-1-2-3-5-8-13-21 -54 units
9.) Bet 34 units and lose: 1-1-2-3-5-8-13-21-34 -88 units
10.) Bet 55 units and lose: 1-1-2-3-5-8-13-21-34-55 -143 units
11.) Bet 89 units and lose: 1-1-2-3-5-8-13-21-34-55-89 -232 units
12.) Bet 144 units and lose: 1-1-2-3-5-8-13-21-34-55-89-144 -376 units

This last example demonstrates how the bets can mount in a losing string of twelve losses. The chances of losing twelve straight on a double zero roulette wheel are

$\left(\frac{20}{38}\right)^{12} = 0.0004518$, or about 1 shot in 2213. The purpose here was to show a range of cumulative losses and let the system player decide where to draw the line. Some authors show the Fibonacci sequence and omit the first "1" in the series. That's fine, but the shortened version is a little more aggressive than the "full" Fibonacci. You will lose a bit more money on average with this abbreviated variation. Overall, the Fibonacci sequence does not fare to badly. This system can be fun and not too damaging if you limit your top bet to five units.

Fibonacci numbers in popular culture

This sequence has appeared many times in popular culture. Fibonacci numbers have been mentioned in novels, films, episodes of television shows, and songs. The numbers have also been used in the creation of music and visual art.

Cinema

* In the film Dopo Mezzanotte (After Midnight), the sequence appears as neon numbers on the dome of the Mole Antonelliana in Turin, Italy, and is also used to select numbers in a lottery, ultimately winning it.
* Along with the golden rectangle and golden spiral, the Fibonacci sequence is mentioned in Darren Aronofsky's independent film π. They are used to find the name of God.(1998)
* In "The Da Vinci Code" the numbers are used to unlock a safe.
* In Mr. Magorium's Wonder Emporium (2007), Magorium hires accountant Henry Weston (Jason Bateman) after an interview in which he demonstrates knowledge of Fibonacci numbers.
* In Death Note: L Change the World (2007), a child sets up marshmallows in the shape of a grid showing a Fibonacci Spiral.
* In 21 (2008), the first eight numbers of the sequence are written in icing on a birthday cake which has twenty-one candles.
* Mentioned in The Oxford Murders (2008).
* Mentioned in Taken (2002)

Literature

* A youthful Fibonacci is one of the main characters in the novel Crusade in Jeans (1973). He was left out of the 2006 movie version, however.
* The Fibonacci sequence plays an important role in the plot of the children's book The Wright 3 by Blue Balliett.
* The Fibonacci sequence plays a small part in the bestselling novel and film The Da Vinci Code.
* The Fibonacci sequence plays a part in unravelling the Atlantis Code in Stel Pavlou's bestselling novel Decipher.
* A part of the Fibonacci sequence is used as a code in Matthew Reilly's novel Ice Station.
* In Philip K. Dick's novel VALIS, the Fibonacci sequence (as well as the Fibonacci constant) is used as identification signs by an organization called the "Friends of God".

* In the collection of poetry alphabet by the Danish poet Inger Christensen, the Fibonacci sequence is used to define the number of lines in each poem.
* The Fibonacci sequence is one of many mathematical topics in Scarlett Thomas's novel PopCo whose main character has an affinity for mathematics.
* The Fibonacci sequence is one of the main sources of math-based magic for the main character, Reason Cansino, in Justine Larbalestier's trilogy, Magic or Madness
* The Fibonacci sequence is mentioned in the children's book Math Curse by Jon Scieszka.
* It was briefly included (and recognized by Charles Wallace) in the television film adaptation of A Wrinkle in Time.
* Fibonacci and the Fibonacci numbers are mentioned as a code to unlock a vessel in Dan Brown's best selling novel, The Da Vinci Code, and its movie adaptation.

Music

* Tool's song "Lateralus" from the album of the same name features the Fibonacci sequence symbolically in the verses of the song. The syllables in the first verse count 1, 1, 2, 3, 5, 8, 5, 3, 13, 8, 5, 3. The missing section (2, 1, 1, 2, 3, 5, 8) is later filled in during the second verse.
* Ernő Lendvai analyzes Béla Bartók's works as being based on two opposing systems, that of the golden ratio and the acoustic scale. In the third movement of Bartok's Music for Strings, Percussion and Celesta, the opening xylophone passage uses Fibonacci rhythm as such: 1:1:2:3:5:8:5:3:2:1:1.
* The Fibonacci numbers are also apparent in the organisation of the sections in the music of Debussy's Image, Reflections in Water, in which the sequence of keys is marked out by the intervals 34, 21, 13 and 8.
* Polish composer Krzysztof Meyer structured the values in his Trio for clarinet, cello and piano according to the Fibonacci sequence.
* American musician BT also recorded a song titled Fibonacci Sequence. The narrator in the song goes through all the numbers of the sequence from 1 to 21 (0 is not mentioned). The song is featured on the second disc of the Global Underground 013: Ibiza compilation mixed by Sasha.

Architecture

The sequence has been used in the design of a building, the Core, at the Eden Project, near St Austell, Cornwall, England.

Chimney of Turku Energia, Turku, Finland featuring Fibonacci sequence in 2m high neon lights. By Italian artist Mario Merz for an environmental art project (1994)

References

1. K. Vogel, Biography in *Dictionary of Scientific Biography* (New York 1970-1990).

2. Biography in *Encyclopaedia Britannica.* [Available on the Web]

Books:

3. J. Gies and F. Gies, *Leonard of Pisa and the New Mathematics of the Middle Ages* (1969).
4. H. Lüneburg, *Leonardi Pisani Liber Abbaci oder Lesevergnügen eines Mathematikers* (Mannheim, 1993)

Articles:

5. A. Agostini, Leonardo Fibonacci (Italian), *Archimede* **5** (1953), 205-206.
6. A. Agostini, L'uso delle lettere nel 'Liber abaci' di Leonardo Fibonacci, *Boll. Un. Mat. Ital.* (3) **4** (1949), 282-287.
7. I. G. Basmakova, The 'Liber quadratorum' of Leonardo of Pisa (Russian), in *History and methodology of the natural sciences* **XX** (Moscow, 1978), 27-37.
8. P. K. Chong, The life and work of Leonardo of Pisa, *Menemui Mat.* **4** (2) (1982), 60-66.
9. M. Dunton and R. E. Grimm, Fibonacci on Egyptian fractions, *Fibonacci Quart* **4** (1966), 339-354.
10. R. Franci and L. Toti Rigatelli, Towards a history of algebra from Leonardo of Pisa to Luca Pacioli, *Janus* **72** (1-3) (1985), 17-82.
11. P. Freguglia, The determination of π in Fibonacci's 'Practica geometriae' in a fifteenth-century manuscript (Italian), in *Contributions to the history of mathematics* (Modena, 1992), 75-84.
12. S. Glushkov, On approximation methods of Leonardo Fibonacci, *Historia Math.* **3** (1976), 291-296.
13. A. F. Horadam, Fibonacci's mathematical letter to Master Theodorus, *Fibonacci Quart.* **29** (2) (1991), 103-107.
14. A. F. Horadam, Eight hundred years young, *The Australian Mathematics Teacher* **31** (1975) 123-134.
15. G. Loria, Leonardo Fibonacci, *Storia delle mathematiche* **I** (Turin, 1929), 379-410.
16. H. Lüneburg, Fibonaccis aufsteigende Kettenbrüche, ein elegantes Werkzeug mittelalterlicher Rechenkunst, in *Séminaire Lotharingien de Combinatoire* (Strasbourg, 1991), 135-149.
17. H. Lüneburg, Fibonaccis aufsteigende Kettenbrüche, ein elegantes Werkzeug mittelalterlicher Rechenkunst, *Sudhoffs Arch.* **75** (2) (1991), 129-139.
18. E. A. Marchisotto, Connections in mathematics: an introduction to Fibonacci via Pythagoras, *Fibonacci Quart.* **31** (1) (1993), 21-27.
19. E. Picutti, Leonardo of Pisa's congruous-congruent numbers (Italian), *Physis - Riv. Internaz. Storia Sci.* **23** (2) (1981), 141-170.
20. E. Picutti, The 'Book of squares' of Leonardo of Pisa and the problems of indeterminate analysis in the Palatine Codex 557 of the National Library in Florence : Introduction and comments (Italian), *Physis - Riv. Internaz. Storia Sci.* **21** (1-4) (1979), 195-339.
21. S. Shalhub, The calculations and algebra of abu Kamil Shuja- ibn Aslam and his effects on the work of al-Karaji and on the work of Leonardo Fibonacci (Arabic), in *Deuxième Colloque Maghrebin sur l'Histoire des Mathématiques Arabes* (Tunis, 1990), A23-A39.
22. J. Weszely, Fibonacci, Leonardo Pisano (c. 1170-c. 1240) (Romanian), *Gaz. Mat. Mat. Inform.* **1** (3) (1980), 124-126.

Methods used for Price Testing

Petra Coufalová

Department of Statistics and Probability
University of Economics, Prague
nám. W. Churchilla 4, Praha 3
Czech Republic
E-mail: CoufalovaPetra@seznam.cz

Abstract: The article focuses on methods for price testing used in market research (such as Gabor-Granger, Price Sensitivity Meter, Brand-Price Trade-Off). It concentrates on description of methods, condition of use, advantages and disadvantages of these methods.

Keywords: Purchase Intention, Gabor-Granger price test, Price Sensitivity Meter, Brand-Price Trade-Off

1 INTRODUCTION

The price is part of a marketing mix that includes:

- Product – e.g. taste tests
- Placement – distribution, retail audit
- Promotion – advertising tests
- Package – packaging tests
- People – lifestyle test, U&A studies
- Price – Price Sensitivity Meter, Brand-Price Trade-Off, Conjoint analysis

The price is the only element of the marketing mix that is the source of income, and the others produce costs.
The price influences the company's profitability, sales, market share size, shopping behaviour of customers.
Properly managing the price of a product or service should lead to optimum setting of the price. If the method used to set the price is not optimum, there is a risk that the price will be too high and the company will not gain or will lose customers, or the price will be too low and the company will not utilize the yield potential of the market.

1.1 WHAT PRICE TESTS MEASURE

The basic information price tests must provide is a quantitative expression of customers' reaction to the price or to a price changes - price elasticity (flexibility of demand, price sensitivity). Besides the flexibility of demand, we are also interested in

the existence and position of breaking points of price sensitivity curve, otherwise referred to as price thresholds. In the price threshold point is occurring a major change of flexibility curve directive (it is the point of the change in price sensitivity). Most of the assignments working with the price need answers to the questions "what will happen when... (the price of our product changes from … to … / the price of a competitor's product changes from … to …)" how will the shares of products on the market change when...?. Must be available price tests that are able to simulating so-called "what if" scenarios.

The current development of price tests focuses mainly on making more realistic outputs, which means transforming estimates of shares of preference to estimates of actual market shares or sold volumes under price conditions simulated by the given scenario. The actuality and reliability of results dramatically influences the ability of clients to perform price tests.

1.2 WHAT PROMPTED THE NEED FOR PRICE TESTS?

The mid 1960´s are considered the start of the need for price tests, when in Great Britain the law on a single / minimum price was abolished (resale price maintenance that means agreement between manufacturer and seller that determines for what price a seller can sell goods for). The single / minimum price rule was created in 1880 and had served as a guarantee of profit for all sellers. Sellers began to have the opportunity to compete by offering lower prices, and this led to the need for valid predictions about how consumers would react to various price strategies. Effectiveness of retail increases, the size of shops grows (counter sales → small self-service shops → supermarkets → hypermarkets) and costs for having shop employees decline. The development of retail is furthered by improving technical equipping of consumers:

– the concentration of automobiles dramatically increases (people begin travelling greater distances to shop instead of walking to the shop around the corner). This enables large shops to spring up in shopping zones

– refrigerators and freezers gradually become standard household appliances (food can be stored longer).

In the 1960´s people begin dramatically more spending money than in previous years, but they also become more price sensitive. These conditions have led to prices becoming a "weapon in battles" to gain consumers, and in the 1970´s prices became a part of the marketing mix. The need arose for reliable price tests that provide valid information about how consumers react to price changes.

1.3 DEVELOPMENT OF PRICE TESTS

In the 1950´s, Great Britain: first documented studies focusing on examining the influence of prices in consumer behaviour (exclusively FMCG).

In the 1960´s, Great Britain: (1961-1966) Andre Gabor and Clive Granger developed an indirect method of testing prices with the help of the purchase intention probability scale: **Gabor-Granger price test.**

Early 197´0s, Great Britain: a different direct method was also developed at the university in Nottingham that instead of measuring the probability of purchases measures respondents actual reactions (respondents select from a set of products at certain prices): **Brand-Price Trade-Off (BPTO) .**

Mid 1970´s (1976), Holland: van Westendorp reveals a new method of price tests: **Price Sensitivity Meter**

Late 1970´s and early 1980´s gradual introduction of all three methods (Gabor-Granger price test, BPTO and Price Sensitivity Meter) in all of Western Europe. ESOMAR 1982 congress recognizes these price tests as "implemented" in practice.

In the 1980´s, world: price tests gradually expand to the whole world, for example in 1990 Asia reports a large expansion of Gabor-Granger price test (but still does not know BPTO).

From 1990´s till present, world: with the development of computer technology, **conjoint methodology** begins being applied as part of price tests (its development is still unfinished).

2. METHODS FOR PRICE TESTING

2.1 GABOR-GRANGER PRICE TEST

Gabor-Granger price test is an indirect method (we determine the probability that a product or service will be purchased at a particular price on a 5-point scale, and then we determine actual shopping behaviour based on the respondent's answers). It is a monadic test and addresses the product or service or the price of a product or service without defining the relationship with the prices of other products or services or the competition (respondent himself/herself intuitively puts together a set of products and their prices that could compete with the tested product). Tests the product as a whole and it cannot be tested as a combination of individual elements or characteristics. Isolated price points are tested. The resulting flexibility of demand is formed by an interpolation between the tested prices, it cannot be extrapolated below the lowest or above the highest tested price. This method for estimating price is a simple for implementation and analysis. For described product respondents are asked:

"If this product or service were available for the price CZK …, how likely would you be to buy it?"

Answers on 5-point purchase intention scale (PI scale):

- definitely would buy
- probably would buy
- not sure whether would buy or not
- probably would not buy

- definitely would not buy.

Then a new price is proposed, and willingness to purchase is asked again. The process continues across pre-defined price levels (usually 3-5 price levels for one product or service). From the results we are able to construct expected demand curve at each price level. We need to transform the answers on the PI scale to the demand curve. For it we must determine which portion of respondents in individual points on the scale will actually buy the particular product (it always depends on the category and position that we are addressing - see examples in the Table 1.).

Table 1. Transformation of answers on PI scale to the demand curve

Category: PI scale	FMCG	Telco	finance	highly involved target group (services)
definitely would buy	75%	70-80%	60-70%	80%
probably would buy	35%	20-30%	20%	50%
not sure whether would buy or not	15%	0-5%	0%	15%
probably would not buy	5%	0%	0%	0%
definitely would not buy	0%	0%	0%	0%

We display the calculated projections as expected demand curve. Using this estimate also the expected revenue will be calculated and optimum price level for product or service (that maximizes revenue).

Then data are plotted into a **Gabor-Granger price volume curve** (Fig. 1.).

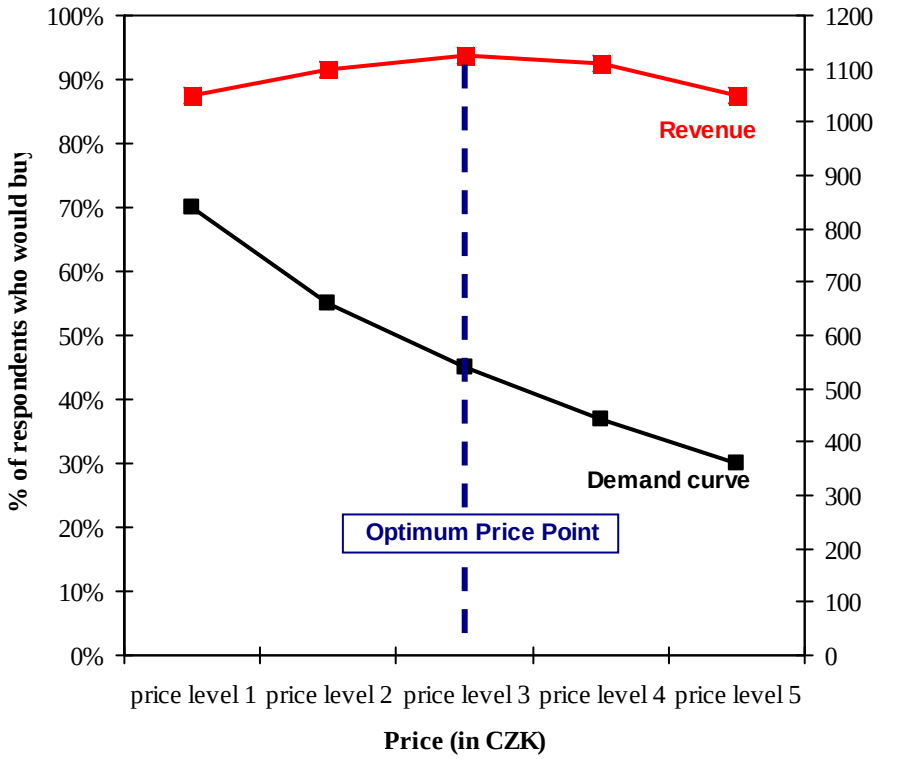


Fig. 1. Gabor-Granger price volume curve

2.2 PRICE SENSITIVITY METER

More sophisticated variation of the Gabor-Granger technique was presented by Peter H. van Westendorp and called Price Sensitivity Meter (sometimes Price Sensitivity Measurement). Price Sensitivity Meter is similar to Gabor-Granger in respect of indirect method, monadic test and test the product as a whole (does not separately consider its individual elements). Price Sensitivity Meter is a simple method to assess consumers' price perception, it is based on the premise that there is a range of prices bounded by a maximum that a consumer is prepared to spend and a minimum below which credibility is in doubt. Respondents are asked for the following four price-related questions about a product or service:

Q1. At what price would you consider this product or service so expensive that you would not consider buying it? (referred to below as **"Too expensive"**)

Q2. At what price would you consider this product or service to be priced so low that you would begin to question its quality? (referred to below as **"Too cheap"**)

Q3. At what price would you consider this product or service to start getting expensive, but still a possible purchase? (referred to below as **"Expensive"**)

Q4. At what price would you consider this product or service to be a bargain - a great buy for the money? (referred to below as **"Cheap"**)

From responses to these four questions, cumulative frequency distributions are derived and plotted. The "Not cheap" and "Not expensive" curves are counted as 100% - cumulative % of responses to the question "Cheap" or "Expensive". Also various intersections on the curves provide inputs for pricing decisions (Fig. 2.).

The **Indifference Price Point (IPP)** is a point at which the number of respondents who regard the price as cheap is equal to the number of respondents who regard the price as expensive. According to Van Westendorp, this generally represents either the median price actually paid by consumers or the price of the product or service of an important market leader. IPP is based on customers' experiences with price levels in the market and will change with market conditions.

The **Optimal Price Point (OPP)** is a point at which the number of customers who see the product or service as too cheap is equal to the number who see the product or service as too expensive. This is typically the recommended price for product or service. Also most clients´question is that OPP is definitive optimal price for a product or service. The questions asked in PRICE SENSITIVITY METER force respondents to choose a range of prices (as opposed to just one) that they consider to be acceptable.

The **Point of Marginal Cheapness (PMC)** is the point at which number of respondents who consider the product or service as not cheap is equal to the number of respondents who regard the product or service as too cheap.

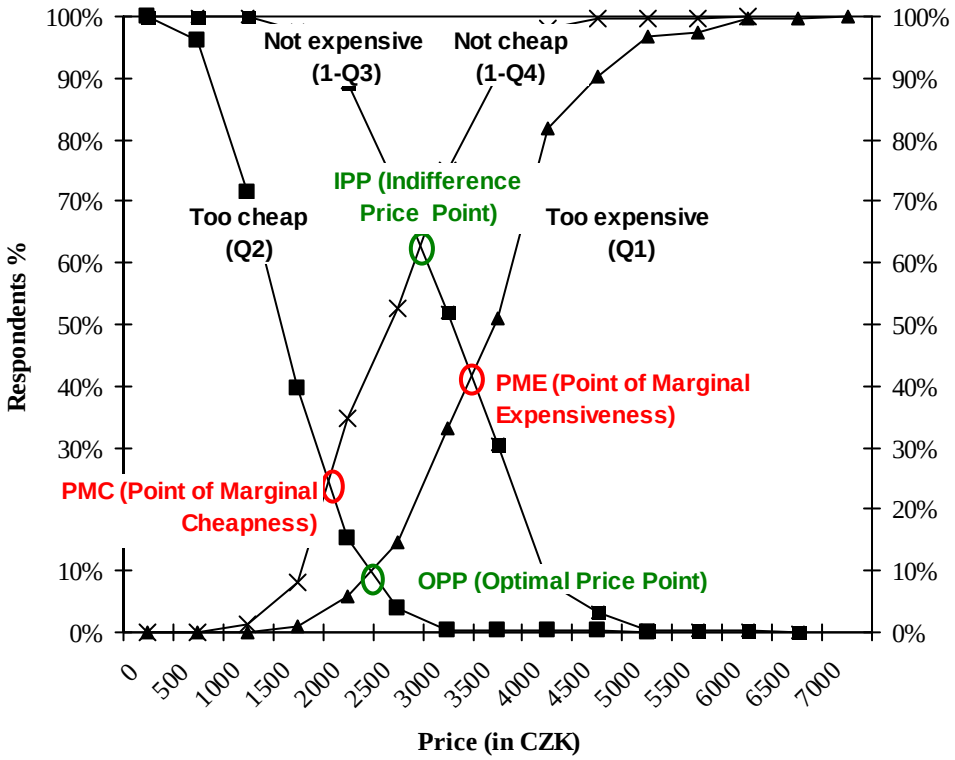


Fig. 2. Results from Price Sensitivity Meter questions

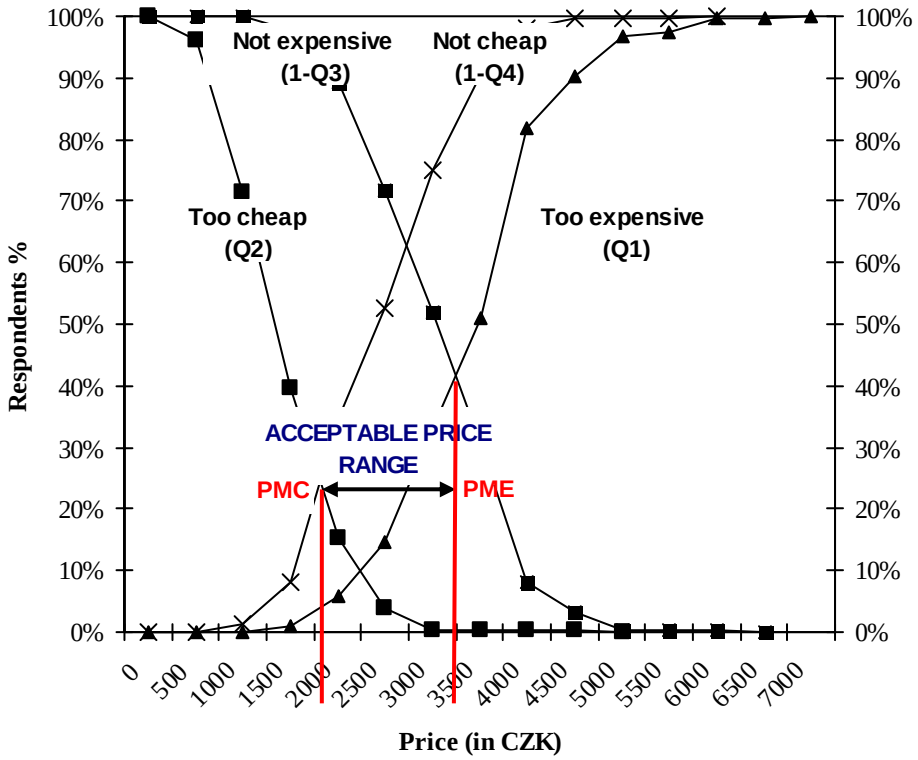


Fig. 3. Price Sensitivity Meter - Acceptable Price Range

The **Point of Marginal Expensiveness (PME)** is the point at which number of respondents who consider the product or service as not expensive is equal to the number of respondents who regard the product or service as too expensive.

The range of prices between the **Point of Marginal Cheapness (PMC)** and **Point of Marginal Expensiveness (PME)** is considered the **Acceptable Price Range** (Figure 3). According to Van Westendorp, in well established markets, few competitive products are priced outside this range.

If the clients´ goal is to maximize market share or penetration, price must be set somewhere between PMC and OPP. If the clients´goal is to maximize revenue, price must be set somewhere between OPP and PME.

2.3 BRAND-PRICE TRADE-OFF (BPTO)

BPTO measures the price sensitivity of brands in a competitive environment. It is a direct method (respondents in an interview actually select the product that they would buy for the defined prices). This method can be used only in categories where we can predict that purchase decisions will depend on two factors: price and brand (brand here includes all other marketing variables besides price).

For BPTO testing we must define competitive environment. We need main competitive brands in tested category (total market share of tested brands could cover at least 80% of market), usually 10-20, for each brand must be defined actual market price, price levels which are used interviewing (usually 5-8 price levels, price levels must equidistant, size of price step - usually 2%-10%). Market price when reducing prices could be set on price level P5 or P6 and the market price when increasing prices could be set on price level P2 or P3 price level (example in Table 2.).

For questioning all tested products are in front of the respondent. The products are labelled with price tags (all at the 1st price level at the beginning of the test).

Basic BPTO question:

Q1. If these brands of products or services were available in the shop for these prices, which product or service would you buy under those circumstances? You can select from any of the product or service brands shown.

Table 2. Price levels offered for interviewing

Brands	**P1 CZK**	**P2 CZK**	**P3 CZK**	**P4 CZK**	**P5 CZK**	**P6 CZK**	**P7 CZK**
Brand 1	13	15	17	19	21	23	25
Brand 2	25	27	29	31	33	35	37
Brand 3	25	27	29	31	33	35	37

Q2. And if these brands of products or services were available in the shop for these prices, which product or service would you buy now? You can select from any of the product or service brands shown.

Etc.

The interviewing continues till the respondent for some brand achieve maximal price level or say that do not choose any of offered brands.

Resultant matrix of purchase decisions for one respondent is shown in Table 3.

Table 3. Example of purchase decisions for one respondent

Brands	**P1 CZK**	**P2 CZK**	**P3 CZK**	**P4 CZK**	**P5 CZK**	**P6 CZK**	**P7 CZK**
Brand 1	1st	2nd	3rd	4th	9th	10th	11th
Brand 2	5th	6th	7th				
Brand 3	8th						

Main BPTO results:

- Preference Share – how does the consumer decide between different products or services in certain price situations
- Brand Switching (Table 4. and Table 5.)
- Chosen at first – the first brand that respondents chose at the very lowest price level (Table 6.)
- Chosen at all – those who chose each product or service at least once in the test (Table 6.)
- Total Occurrence – the average number of accepted price levels for the brand (Table 6.)
- Before 1st switching – the average number of price levels accepted before initial switch (Table 6.)
- Price elasticity (Fig. 5., Table 7.)

BPTO results will be presented in the following BPTO analysis (for analysis we use brands and price levels from Table 2., actual market price is on price level P2).

Table 4. Brand switching if price level for brand 1 increases from price level P1 to P2 (brand 2 and brand 3 stay on their actual market prices)

	Switch to (chosen second)				
Switch from (chosen first)	Brand 1	Brand 2	Brand 3	Refused	**Total**
Brand 1	**149**	**1**			**150**
Brand 2		**42**			**42**
Brand 3			**8**		**8**
Refused					**0**
Total	**149**	**43**	**8**	**0**	**200**

Table 5. Brand switching if price level for brand 1 increases from price level P2 to P3 (brand 2 and brand 3 stay on their actual market prices)

	Switch to				
Switch from	Brand 1	Brand 2	Brand 3	Refused	**Total**
Brand 1	**132**	**6**	**1**	**10**	**149**
Brand 2		**43**			**43**
Brand 3			**8**		**8**
Refused					**0**
Total	**132**	**49**	**9**	**10**	**200**

In Fig. 4. we could see how preference shares change if we increase a price of brand 1 (brand 2 and brand 3 stay on their actual market prices).

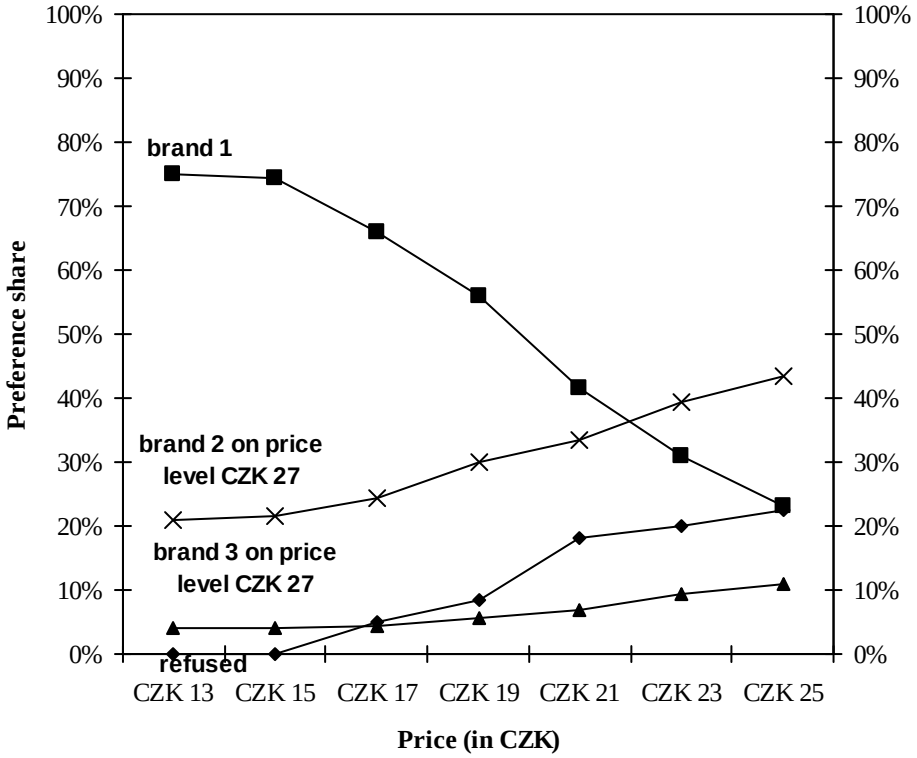


Fig. 4. Change of preference shares if increases price of brand 1.

Recommendation resulting from Fig. 4. and 5. and from Tables 4. and 5. is to keep the price of brand 1 on actual market price (CZK 15), because price increasing would cause switching from brand 1 to brand 2.

Table 6. Chosen at first, Chosen at all, Total Occurrence, Before 1st switching

	Chosen at first	**Chosen at all**	**Total Occurrence**	**Before 1st switching**
Number of respondents	**200**	**200**	**200**	**200**
	%	*%*	*mean*	*mean*
Brand 1	**67,5**	**90**	**5,0050**	**4,29**
Brand 2	**25,5**	**65,5**	**1,9550**	**3,27**
Brand 3	**7**	**39**	**0,9150**	**2,57**

Table 7. Price elasticity table for brand 1

Price level	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Price (CZK)	**13**	**15**	**17**	**19**	**21**	**23**	**25**
Price change (CZK)	-2	x	2	2	2	2	2
Volume (respondents)	**150**	**149**	**132**	**112**	**83**	**62**	**46**
Volume change (respondents)	0,7 %	x	-11,4 %	-15,2 %	-25,9 %	-25,3 %	-25,8 %

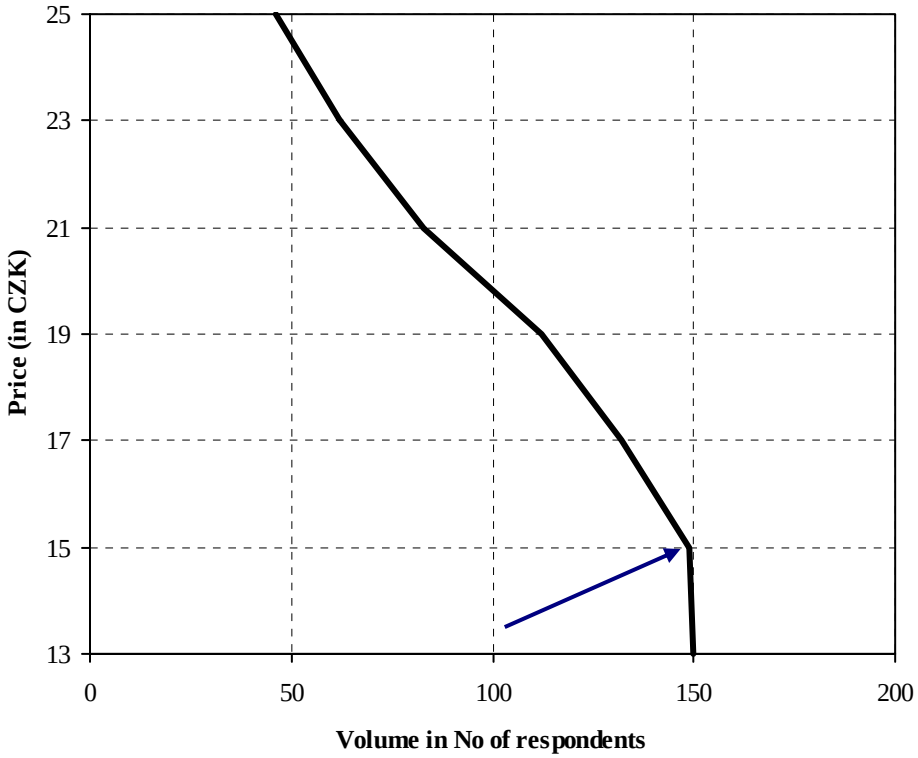


Fig. 5. Price Elasticity

3 CONCLUSIONS

Each of three tests presented above have some advantages as well as limitations (disadvantages).

Gabor-Granger price test can be used for new products or product variations. Objections have been existed arguing that the answers to the second and subsequent price are influenced by answers to the first tested price and that therefore one

respondent should only test one price. But independent measuring of “many” vs. “one” price per respondent has shown that measuring several prices for one respondent gives better and better interpreted results.

Expectations for Price Sensitivity Meter are small perception about price and rational shopping behaviour. Main advantages of Gabor-Granger and Price Sensitivity Meter methods are quick, easy to execute (part of another test, consists only of one or four questions) and easy to understand. We need not know the specified price beforehand. Price Sensitivity Meter determine how customers perceive the price of a product in a particular category and what expectations are based on price.

Price Sensitivity Meter is especially advantageous for testing prices of new or innovated products or in categories where the competition does not exist (typically the pharmaceutical industry – new types of pharmaceuticals, new technologies – new types of mobile phones, digital or interactive television, video-on-demand etc.)

It can also be used successfully where direct competition exists, but where during the shopping process comparisons of product prices usually are neither expressed nor can be compared directly (e.g. petrol prices at a service station or extra fees for credit / debit card insurance).

The system of four Price Sensitivity Meter questions perform well where the concept “too cheap prices” makes sense, which raises doubts about the quality of a product and could lead to a consumer’s refusal to buy it (e.g. products in the FMCG category). Also exist areas where this concept does not perform well, typically the financial sector, where some services can be provided for free (customers can even directly expect that the particular service will be free of charge) and the question regarding too cheap price can then be misleading. In this situation we do not ask the question about too cheap (the interval of acceptable prices begins at CZK 0).

An enormous advantage of Price Sensitivity Meter is that it measures the price sensitivity for the continual range of prices, other price tests always work with several specific prices, among which we must interpolate (and beyond which it would be difficult to extrapolate).

Since we are not comparing the product with any competition, the price test can be performed as early as during the initial phases of product development (a description of the concept or the rough design of the product is sufficient for the Price Sensitivity Meter).

Limitations of Price Sensitivity Meter: Price Sensitivity Meter does not replicate the actual shopping process. The results depend on respondents' experience with price levels in the market. If respondent does not have a good reference price, this method often causes the underestimation of a product's ability to command a premium price. Price Sensitivity Meter does not simulate market situations and does not consider competition.

Expectations for BPTO are 100% distribution and awareness of tested products. Respondents know the prices of products (buy them). Must be well determined prices that we test.

Advantages BPTO are that BPTO includes competition. Considers together brand and price (various combinations). Enable us to simulate market situations. Possibility of products mutually influencing each other (brand switching price thresholds). BPTO is close to real buying situations.

Limitations of BPTO method: Gradual price increases make respondents more sensitive to higher prices. Coverage of the market of the tested category (market share of tested products). Limitation in case of new products, premium and common (usual) brands.

References

1. Baker, Michael J.: The Marketing Book. 4th Edition. Butterworth-Heinemann (1999)
2. Burrow, James L: Marketing. 3rd Edition. South Western Educational Publishing (2008)
3. Farace, W.: What Should We Charge? Setting Price. Satisfaction Management Systems Inc. (2008)
4. Frain, John: Introduction to marketing. London: International Thomson Business Press (1999)
5. Xu, Jie: Market Research Handbook. iUniverse (2005)

Time of Crystal Balls
(GDP predictions in crisis)

Jan Čadil

Unicorn College, V Kapslovně 2, Praha 3
jan.cadil@unicorncollege.cz

Abstract. The paper deals with the inability of contemporary methods used by financial institutions to efficiently estimate the GDP development in crisis and to give the prior warning of upcoming crisis as well. The paper is not of scientific nature its only purpose is to bring attention to necessity of change of economic models and methods which could bring better estimates and what is most important could save a lot of money in the future especially in cases when the economy sustains some bubble bust or other shock.

Keywords: GDP, growth estimates

1 Introduction

The purpose of this paper is not to present any serious economic research but only to point at one contemporary analytical problem that is connected to economic crisis – the loss of faith in analysts and analyses. The paper consists of two parts – in the first part the prediction ability of financial and economic research institutions is discussed on the case of GDP growth estimates. In the second part a very simple model that can be easily used for similar predictions is introduced to stress the need of change in analytical approach to get better predictions and to enable investors but also economic policy makers to react fast and in appropriate way. I believe that good analytical work and proper predictions may save lots of money (and values) that would be otherwise loss and possibly prevent any economic crisis even to occur[1].

[1] As in the US before the housing crises started. Some economist warned that the crisis is on stake. See Case, K.E., Schiller, R.J. (2003) and the paper discussion - Mayer,Ch.,Quigley,J.M. (2003) or Noord, P. (2006). Levy Economic Institute is also drawing attention to this danger in beginning of 2006. On the contrary some institutions were denying the risk of housing bubble – see Harvard University, Joint Center for Housing Studies 2006 for example. Today we already know who was right alas too late.

2 Financial institutions ability to predict – case of Czech Republic

Contemporary economic situation and its development has risen many questions and challenges for economic analysts who were usually not well prepared for the bubbles and crisis development predictions. Because of that they were sometimes (falsely) blamed by the public for the crisis along with financial institutions and their credibility sustained serious hit. Figure 1 is giving a brief illustration of Czech GDP growth predictions made by institutions in 2007, 2008 and 2009 as reported to Bloomberg database[2]. We can clearly see that when economy was behaving "normally" predictions were quite accurate but under the crisis shock in 4Q 2008 and 1Q 2009 they have become inaccurate and also the variation among predictions increased[3]. On the other hand it should be the very core work of the analyst to reveal and predict economic behavior in these times because in "normal" times it is quite easy to make predictions almost with using common sense or a simple model as will be shown later[4].

2 These predictions were made after quarterly release of industrial data which could have (and should have) influenced predictions. Data are not available for all institutions for all periods. Therefore especially Figure 2 must be interpreted with care.

3 There is an exception in the first quarter of 2008 when institutions also failed to predict accurately. It was probably caused by large CSU data revisions of historical time series on which the models usually depend.

4 The most popular phrase among Czech analysts says *"I do not have a crystal ball"*.

Table 1: Finanacial Institutions GDP Growth Estimates (Czech Republic)

	7.9.07 (2Q 07)	7.12.07 (3Q 07)	15.2.08 (4Q 07)	15.5.2008 (1Q 08)	14.8.08 (2Q 08)	14.11.08 (3Q 08)	13.2.09 (4Q 08)	15.5.09 (1Q 09)	14.8.09 (2Q 09)
Institution	Estimation								
4Cast				5,3	5,7				-1,4	
Atlantik	5,1	5	5,3	4,4	6	3,9	0	-2,9	-4,2
CSOB	5,8		5,6		5,9	4	1,8		
Citi		6,2	5,3	5,6	4,4				-4,2
DZ Research	6,5	5,6	5,9				1,1	-1,7	
Generali PPF		5,7	5,7	5,2	4,8	3,8	-0,5	-2,4	-3,7
IDEA global		5,7	5,9	4	5,2	4,2		-3	
ING	5,9	5,6	5,4	5,3	4,9	3,6	2,3		-5,5
JP Morgan	5,5		5,5		5	3,2		-2	-4,5
KB	5,5	5,5	5,7	5,5	5,2		0	-1,4	-4,2
Morgan Stanley	5,5	6	6,5	5,6			0,9	-1,6	
Nordea Markets	5,7		5,5						
Raiffeisenbank	6	5,8	6,1	6,2	5,4	3,5	-0,7		-4,8
Unicredit		5,8	5,2	4,3	4,8	3,8	-1,5	-4	-4,1
Ceska Sporitelna	5,2	5,8			5	4,1	-0,8		
Danske Bank	5,6	6,2					0,2	-1,8	
Merryl Lynch	5,8	5,8		5,4			0,5	-2,5	-4,2
Patria Finance	6,5	5,8		5,5	4,5	4,1	1,8	-1,7	-5,5
Bloomberg Median	5,8	5,8	5,6	5,5	5	3,8	0,5	-1,8	-4,2
CSU Actual	5,6	6	6,1	3,8	4,6	3,1	-1	-3,4	-4,9
Variance	0,18	0,09	0,13	0,43	0,24	0,1	1,31	0,62	0,36

Source: Bloomberg, own calculations

The main problem of accurate predictions in these times is probably in methods that are being used and also lack of relevant information or that the information is not reliable. For a good prediction it is also vital to be flexible and take a greater extent (like world demand development which is affecting exports in future) into account. If we make a list of most and least successful predictors according to Figure 1 estimates we get "winners" and "losers" in analytical competition as Figure 2 displays[5]. However not all institutions were observed in the same time (and some predictions were much tougher than others) and provided the database with all predictions – for example Raffeisenbank was measured eight times but Danske Bank only four times. That is the reason why we should not make any serious conclusions based on Table 2.

Table 2: Best and worst Estimations

5 The higher the deviation (calculated simply as an average of differences) the worse predictions are made by the institution at average.

Istitution	Deviation (Average)
4Cast	1,57
Patria Finance	1,13
DZ Research	1,06
Danske Bank	1
Morgan Stanley	1
CSOB	0,95
ING	0,95
Citi	0,93
KB	0,88
Merryl Lynch	0,85
Bloomberg Median	0,83
Atlantik	0,81
Generali PPF	0,71
Raiffeisenbank	0,58
Unicredit	0,55
JP Morgan	0,5
Ceska Sporitelna	0,44
IDEA global	0,4

Of course one can say that financial institutions use mainly simple models to predict GDP because it is not as vital for them to make accurate predictions of this factor and that a use of more sophisticated model (with lots of equations in practice) would be more efficient[6]. Naturally there are institutions that use such models in the Czech Republic like Ministry of Finance or the Czech National Bank. Nevertheless it does not seem that using complicated models yields much better results. For example for the critical period - 4th quarter of 2008 and 1st quarter of 2009 the Ministry of Finance predicted in considerable advance to others a GDP growth of 2,8% (in January 2009) and -1% (April 2009) exhibiting much worse results then the financial institution estimates. Czech National Bank predicted a growth of 2,0% for the 4th quarter of 2008 (in February 2009) and -2,5% for the 1st quarter of 2009[7] again lagging behind some financial institutions.

6 Nice overview of macroeconometric models that were used in practice offers Hušek, R., Pelikán, J.: Aplikovaná ekonometrie, teorie a praxe, pp. 175-213.

7 These values are referred as „centred". Czech National Bank is using spread which is quite large and to me quite useless. For example in 1st qurter of 2009 the 90% spread was between -4,1% and -0,9%, for the preceding period it was between 0,4% and 3,7% (but the real growth was very different).

On the other side it should be stressed here that wrong prediction is not always a failure of analysts. Czech Statistical Office is revisiting the time series quite often and in the critical period the revision was substantial (quarterly revisions for 2008 were around 1,4% points)[8].

3 GDP growth estimate – a very simple "analyst like" model

There are lots of methods that can be used for GDP estimations like VAR models, C-D function estimates (not often used), ARIMA models etc9. For purpose of this paper I decided for a very simple model with respect to information available by analysts in time of making their predictions and also to speed when making their predictions (which is usually very high)10. I would like to demonstrate that one is able to make a model by himself that exhibits same accuracy but also the same problems. The institution or analyst should probably offer something more especially in times of uncertainty when your model does not work. Quarterly data of seasonally adjusted GDP, industrial production (not including water and electricity), households consumption and exports were used for estimates. Czech Statistical Office was the data source; estimations were made in EViews 6.1. The model is specified in first differences as follows

$$\Delta \ln(gdp_t) = \alpha \Delta \ln(ex)_{t-1} + \beta \Delta \ln(cons)_{t-1} + \gamma \Delta \ln(ind)_t + c \qquad (1)$$

where *gdp* is GDP, *ex* stands for exports, *cons* for consumption and *ind* for industrial production. Industry is taken as a flash estimate published by the CZSO as a first variable long before the GDP (usually several weeks before the GDP release). I should stress here one more time that the analysts' estimates were made always after these releases. All series are stationary according to ADF test on 1% level of statistical significance. Equation (2) and Figure 2 shows the results when using ordinary least squares method on (1).

$$\Delta \ln(gdp_t) = 0{,}211\Delta(ex)_{t-1} + 0{,}498\Delta \ln(cons)_{t-1} + 0{,}168\Delta \ln(ind)_t - 0{,}014 \qquad (2)$$

8 See Macroeconomic Prediction of the Czech Republic, Ministry of Finance, April 2009

9 See Heij, Ch., Boer, P., Franses, P. H., Kloek, T., Dijk, H. K. (2004), *Ecometric Methods with Application in Business and Economics* for example.

10 To develop and test the model itself took 32 minutes.

Table 3: Estimation Representation

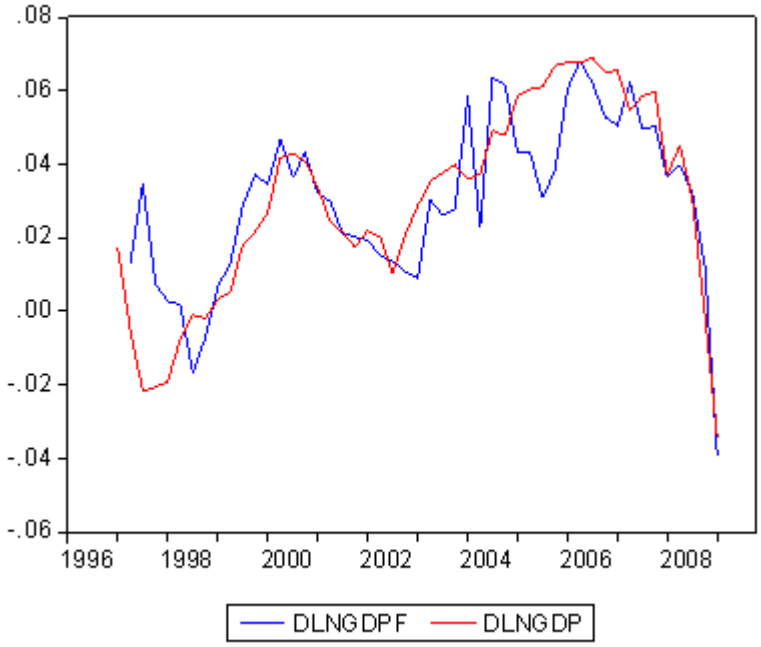
Dependent Variable: DLNGDP
Method: Least Squares
Date: 08/20/09 Time: 13:31
Sample (adjusted): 1997Q2 2009Q1
Included observations: 48 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
DLNEX(-1)	0,210	0,039	5,400	0,000
DLNCONS(-1)	0,498	0,132	3,755	0,001
DLNIND	0,168	0,037	4,582	0,000
C	-0,014	0,006	-2,286	0,027

R-squared	0,686	Mean dependent var	0,030
Adjusted R-squared	0,665	S,D, dependent var	0,027
S.E. of regression	0,016	Akaike info criterion	-5,386
Sum squared resid	0,011	Schwarz criterion	-5,230
Log likelihood	133,274	Hannan-Quinn criter.	-5,327
F-statistic	32,105	Durbin-Watson stat	0,996
Prob(F-statistic)	0,000		

Figure 1 then displays fore casted values for period 1997-2009 (1nd quarter). Static one step ahead forecast was used here.

Figure 1: GDP growth forecast



When comparing this model estimations in previously selected period 2007-2009 to other institutions it seems like it is in the better half with average deviation equal to 0,8 pp (better than Atlantik but worse than Generali). Nevertheless the model was not as many others able to predict the downturn of Czech economy towards recession in last quarter of 2008 and will have probably tendency to keep on the trend lines because of lags in the model. The model is also not very suitable for longer predictions on the other hand I think it is generally extremely difficult to make serious long time predictions under unstable conditions with standard models that are now being used.

Conclusions

As I have tried to show in this paper it seems that not even long time but also short time predictions made by financial, state or independent research institutions do not work well when the economy suffers shock like the housing crisis which has started the worldwide economy crisis. Relying just on standard procedures, models and

variables seems inefficient in these times at least until the situation returns to normal conditions. The institution which is able to give reliable predictions and – that is mostly important – is able to predict the future changes in trend (form negative to positive and vice versa) will gain competitive advantage. The question which is now striking analysts over the world is "When is the growth positive again?" and "How fast will the economy grow?". Until now nobody still knows exactly to what extent the crisis has struck the US (and other foreign) banks and when and where another bank or financial institution will fall although the threat decreases in time. Also the total effect of the crisis on national economies remains unclear and predictions of the end of crisis are very heterogeneous, although now it seems like the majority believes that we are in post-crisis times already. It is the most important now to estimate the crisis impact on economy as a whole (GDP, unemployment, budget deficit etc.), on economic sectors, social groups and on regions because the crisis will struck the regions as well as social groups asymmetrically. But the crucial experience we should take from the crisis is to not let it happen again. Good analyses and in time bubbles and crises identification should be the future of analytical work in my opinion enabling to save more than any (usually controversial) regulations.

There are three ways out of this "analytical crisis" situation. First we may wait until situation gets stable again and use standard models and predictions as always and prey the crisis will not come again. Second we will try to develop some better analytical approaches able to identify incoming bubbles and other shocks and their possible economic impact in advance. Third we may buy a crystal ball.

References

1. Harvard University, Joint Center for Housing Studies: The State of Nation´s Housing, (2006).
2. Heij, Ch., Boer, P., Franses, P. H., Kloek, T., Dijk, H. K. (2004), *Econometric Methods with Application in Business and Economics*. Oxford University Press, (2004)
3. Hušek, R., Pelikán, J. (2003). *Aplikovaná ekonometrie.* Professional Publishing (2003)
4. Case, K. E., Schiller, R. J. (2003), "Is There a Bubble in the Housing Market?" *Brookings Papers on Economic Activity* 2003, no. 2, pp. 299-342.
5. Noord, P., "Are House Prices Nearing a Peak? A Probit Analysis for 17 OECD Countries." OECD Economic Department Working Paper No. 488, OECD, (2006).
6. Papadimitriou, D. B., Chilcote, E., Zezza, G. "Are Housing Prices, Household Debt and Growth Sustainable?" The Levy Economic Institute, (2006).
7. Czech Statistical Office www.czso.cz
8. Czech National Bank www.cnb.cz
9. Czech Ministry of Finance www.mfcr.cz
10. Bloomberg www.Bloomberg.com

On the Hartwick´s Rule in More Dimensional Case

Anton Dekrét

Cikkerova 11, 97401 Banská Bystrica, Slovakia
anton.dekret@umb.sk

Abstract. In literature there are two main approches to the formulation of the Hartwick´s rule in economy with sevaral renewable and nonrenewable capital goods: one without and the other with price relations. In this paper these approaches are analysed and compared by the dynamical differential system formalism using the cotangent prolongation of the basic dynamical system.

Keywords: control dynamical system, cotangent prolongation, Hartwick´s rule.

JEL Classification: D90, O11, O41

1 Introduction

Hartwick (1977) [3] formulated his rule by the production function in economy with one renewable and one non-renewable capital goods. This approach was developed by several authors (see for examle [1], [4], [5], [7]). Dixit, Hammond and Hoel in [2] formulated the Hartwick´s rule by price relations. They recognized that the Hartwick´s rule can be also expressed by an assertion that the valuation of net investment (including the dis-investment in the exhaustible resources) is zero at each date. They already supposed economics with several renewable and non-renewable capital goods. They generalized the demand of zero valuation of net investment by constant one. They did not use dynamical system formalism which was involved in this investigation later first of all in the connection with optimal utility.

The main aim of this paper is to show natural mathematical relations between the both formulations of the Hartwick´s rule. The first one expressed by the production function, and the second one expressed by price relations. We strictly use control

dynamic differential system formalism (see [8]). In our investigation the esential tool we use is the so-called cotangent prolongation of the basic control dynamical system (see [6], [9]). This prolongation is also present in the Pontryagin´s maximum principle though it is not explicitly mentioned in literature dealing with this subject.
The structure of this paper is as follows. In the second chapter the basic dynamic of the studied economy with its cotangent prolongation is introduced. In the third chapter both the 1Hart-path in the spirit of the Hartwick´s approach and also the 2Hart-path corresponding to the Dixit-Hammond-Hoel definition of the Hartwick´s rule are introduced. The both these paths are connected by the cotangent prolongation and can be therefore compared. As for the 1Hart-path it is connected only with basic variables whereas the 2Hart-path is defined by the relation with basic and price coordinates. In this chapter Jacobian of the basic system is also used for characterization of such properties as competition, equity, stationarity. Between the both approaches are signs of duality.

2 A control model of economy with renewable and non-renewable capital goods

We use control theory formalism to model economy. Roughly speaking it consists of stating the basic dynamic system for basic variables and control or optimization conditions. We will omit optimization. First of all we want to find out what information is possible to obtain about the Hartwick´s rule from the cotangent prolongation of the basic system.
We treat the following dynamical system of economy with n renewable and m non-renewable capital goods which is generally accepted in literature:

$$(1) \qquad dk_i/dt = f_i(k,r) - \varepsilon_i k_i - c_i \equiv F_i(k,r,c)\ ,\ i = 1,...,n$$

$$ds_j/dt = -r_j\ ,\ j = 1,...,m\ ,$$

where $k = (k_1,...,k_n) \in K_n$ is the n-couple of the renewable capitals in n capital sectors, $s = (s_1,...,s_m) \in S_m$ is the m-couple of the non-renewable capitals in m source sectors, $\varepsilon = (\varepsilon_1,...,\varepsilon_n)$ is the n-couple of amortization rates, $c = (c_1,...,c_n)$ is the n-couple of consumptions in renewable capital sectors, $r = (r_1,...,r_m)$ is the m-

couple of extraction rates in m non-renewable source sectors, f_i (k,r) is the production function in the i-th renewable capital sector.

The differential system (1) is a basic dynamical system for the basic variables k and s with the control parameters $c = (c_1,..., c_n) \in U_n$, $r = (r_1 ,..., r_m) \in U_m$, where U_n , U_m are open admissible regions of parameters.

Let K°_n denote the space of the change rates in capital (net investments) $k^\circ = (k^\circ_1,.. ,k^\circ_n)$, $k^\circ_i = dk_i/dt$ and S°_m the space of extraction rates $s^\circ = (s^\circ_1,..., s^\circ_m)$, $s^\circ_j = ds_j/dt$. Let $(K^\circ_n)^*$, $(S^\circ_m)^*$ be the dual spaces to K°_n , S°_m , respectively. This means that they are the vector spaces of all linear maps $\xi: K^\circ_n \to R$, $\xi(k^\circ) = \sum_{i=1}^{n} (\xi_i \ k^\circ_i)$, $\psi: S^\circ_m \to R$, $\psi (s^\circ) = \sum_{j=1}^{m} (\psi_j \ s^\circ_j)$. Then the variables ξ and ψ are co-variables (dual) which are refered to as the prices variables. We state the Hamiltonian of the basic system (1) as the function

$$H(k, s, \xi,\psi ,r,c) = \xi(k^\circ) + \psi (s^\circ) = \sum_{i=1}^{n} \xi_i F_i (k,r,c) - \sum_{j=1}^{m} \psi_j r_j.$$

It means that the Hamiltonian of the system (1) is an evaluation of the change rates (net investments and extraction rates).

We introduce the so - called T^* - prolongation (cotangent prolongation) of the system (1) (we refer our readers to [6] or [9] for details). It is of the form

$$(2) \qquad \begin{aligned} k^\circ_i &= F_i (k,r,c) \ , \quad \xi^\circ_i = -\partial H/\partial k_i \ , \ i = 1,..., n \\ s^\circ_j &= -r_j \quad , \quad \psi^\circ_j = -\partial H/\partial s_j = 0, \text{ i.e. } \psi_j(t) = \text{const}, \\ j &= 1,..., m, \end{aligned}$$

where const includes extraction costs in the j-th exhaustible source.

Both systems (1) and (2) are autonomous. A curve u(t) = (c(t), r(t)) in $U_n \times U_m$ is called to be the control path. Systems (1) and (2) become non-autonomous when we put c = c(t) , r = r(t). Solutions (k(t), s(t)) or (k(t) , r(t), ξ(t), ψ(t)) of such systems are brifly said „b-output“ or „T^*-outputs“ for the path u(t) or outputs corresponding to u(t).

The functions F_i (k, r, c) determine the map $F: K_n \times U_n \times U_m \to K^\circ_n$. Its Jacobian matrix is a block matrix of the form

$$J = ((\partial F_i/\partial k_v) = J_1 , (\partial F_i/\partial r_j) = J_2 , (\partial F_i/\partial c_v)).$$

The matrices J_1 , J_2 state the linear maps $J_1 : K^\circ_n \to K^\circ_n$, $J_2 : S^\circ_m \to K^\circ_n$ and its dual maps $J^*_1 : (K^\circ_n)^* \to (K^\circ_n)^*$, $J^*_2 : (K^\circ_n)^* \to (S^\circ_m)^*$.

Recall that the dual map J^*_2 is characterized by the relation

(3) $$\xi(J_2(r)) = J^*_2(\xi)(r)$$

A control path u(t) = (c(t),r(t)) with its corresponding b-output (k(t), s(t)) determine the matrice $J_2(t)$ and then its derivation with respect to t states the map $T J_2(t): S^\circ_m \rightarrow K^\circ_n$.

3 On the Hartwick´s Rule

We now turn to the Hartwick´s rule. Roughly speaking by this rule all extracts from the non-renewable capitals are the only sources for the net investments to the renewable capitals. In literature there are different mathematical formulations of this rule. We will deal with two of them.

Definition 1. The control path u(t) = (c(t),r(t)) which together with its corresponding b-output (k(t),s(t)) satisfy the condition

(4) $$k^\circ(t) = J_2(r(t))$$

will be called the 1Hart-path.

This condition corresponds with the original Hartwick´s form, see [1] , [3] .

Remark 1. As the system (1) is a subsystem of the system (2), i.e. every solution of (2) projects into the solution (k(t), s(t)) of (1), then the terms „b-output“ can be replaced by the terms „T^*-output“ in the definition 1. This will be prefered in what follows.
Let u(t) = (c(t), r(t)) be a 1Hart-path. Then $J_2(r(t) = F(k(t), r(t), c(t))$. Derivating it with respect to t we get $T J_2(r(t)) = J_1(k^\circ(t)) - c^\circ(t)$. A control path u(t) is called totaly equitable if c(t) = const. If 1Hart-path u(t) is totaly equitable then it holds

(5) $$TJ_2 – J_1 J_2)(r(t)) = 0 .$$

The control path u(t) which together with its corresponding b-output or T^*-output satisfy (5) is said to be competitive.
Evidently it holds:

Proposition 1. *A 1Hart-path is totaly equitable iff it is competitive.*

Definition 2. A control path u(t) = (c(t), r(t)) which together with its corresponding T^*-output $\gamma(t) = (k(t), r(t), \xi(t), \psi(t))$ of the system (2) satisfy the condition $H(\gamma(t), u(t)) = \text{const}$ will be called the 2Hart-path. If const = 0, i.e. if $\xi(t)(\dot{k}(t)) = \psi(t)(r(t))$ then we call the notion „2Hart-0-path".

Remark 2. The notions 2Hart-path , 2Hart-0-path correspondwith the formula of the Hartwick´s rule in [2] . We see that the condition $H(\gamma(t), u(t)) = \text{const}$ is dependent on basic variables k, s and also on price variables ξ, ψ on the contrary to the relation (4) involving only the basic variables.

Using the system (2) we get in the case of a 2Hart-path u(t)

$$d\,H(\gamma(t), u(t))/dt = d_cH + d_rH = 0\ ,\ d_cH = \sum(\partial H/\partial c_i)\,\dot{c}_i = -\,\xi(\dot{c}),$$
$$d_rH == \sum(\partial H/\partial r_j)\,\dot{r}_j = \xi(J_2(\dot{r}(t)) - \psi(\dot{r}(t)),$$

d_cH and d_rH are the changes of Hamiltonian H due to the changes of consumptions and extraction rates. Then a path u(t) = (c(t), r(t)) is a 2Hart-path iff together with its corresponding T^*- output $\gamma(t) = (k(t), r(t), \xi(t), \psi(t))$ satisfy the relation

$$\xi(\dot{c}) = \xi(J_2(\dot{r}(t)) - \psi(\dot{r}(t)). \tag{6}$$

Evidently every constant path is a 2Hart-path. We will say that u(t) = (c(t), r(t)) is
equitable if $d_cH = 0$, i.e. if $\xi(\dot{c}) = 0$,
r-stationary if $d_rH = 0$, i.e. if $\xi(J_2(\dot{r}(t)) - \psi(\dot{r}(t)) = 0$,
quite r-stationary if $\partial H/\partial r_j = 0$, j = 1,...,m, i.e. if $\psi = J^*_2(\xi)$,
r-margin-stationary if $J^*_2(\xi) = \text{const}$
along the path u(t) and along its coresponding T^*- output $\gamma(t) = (k(t), r(t), \xi(t), \psi(t))$.
Apparently a quite r-stationary path is r-stationary and r-margin –stationary.

Remark 3. The relation $\psi = J^*_2(\xi)$ is dual to that in (4). We can say that the quite r-stationary path is dual to the 1Hart-path.

The relation (6) immediately gives

Proposition 2. *If a control path u(t) is equitable and r-stationary then it is a 2Hart-path. If u(t) is r-stationary then it is a 2Hart-path iff it is r-stationary. A 2Hart-path is equitable iff is r-stationary.*

From the relation (4) it follows

(7) $$\xi(\overset{\circ}{k}) = \xi(J_2(r)) .$$

The control path u(t) which together with its corresponding T^*- output $\gamma(t)$ satisfy the relation (7) will be called to be 1Hart-ev-path. Evidently every 1Hart-path is 1Hart-ev-path but the converse is not true.

Proposition 3. *A quite r-stationary 1Hart-ev-path is a 2Hart-0-path*.

Proof. If u(t) is a quite r-stationary 1-Hart-ev-path then the relations (7) and (3) give $\xi(\overset{\circ}{k}) = \xi(J_2(r)) = J^*_2(\xi)(r) = \psi(r)$. Proof is finished.

We deal now with the r-margin-stationary path. Derivating the relation $J^*_2(\xi) = \text{const}$ with respect to t and realizing that $\xi(t)$, $\psi(t)$ satisfy the system (2) we get

(8) $$(TJ_2 - J_1 J_2)^*\xi = 0 ,$$

where the map $(TJ_2 - J_1 J_2)^*: (K^\circ_n)^* \to (S^\circ_m)^*$ is dual to $TJ_2 - J_1 J_2 : S^\circ_m \to K^\circ_n$. By (8) we can say that this relation is dual to the relation (5). Then the notion „r-margin-stationary“ is dual to the notion „competitive“.

Corollary of Proposition 3. If u(t) is a quite r-stationary 1Hart-path then it is an equitable 2Hart-0-path.

Proof. By Proposition 3 u(t) is a 2Hart-0-path. Then the relation (6) is true. As u(t) is r-stationary then $\xi(\overset{\circ}{c}) = \xi(J_2(\overset{\circ}{r}(t)) - \psi(\overset{\circ}{r}(t)) = 0$. Proof is finished.

Acknowledgment: The work was supported by the Slovak grant agency, grant No.1/4633/07.

References

1. Buchholz,W., Dasgupta,J. and Mitra,T.: Intertemporal equity and Hartwick´s rule inExhaustible resource model, Scandinavial Journal of Economics, Vol. 107, No.3 (2005), 547-561.
2. Dixit, A., Hammond, P.and Hoel, M.“ On Hartwick´s rule for regular maximin path of capital accumulations and resource depletion, Review of Economic Studies, Vol. 47 (1980), 551-556.
3. Hartwick, J. M.: Intergenerational equity and the investing of rents from exhaustible resources, American Economic Review,Vol. 67, No.5 (1977), 972-974.
4. Hamilton, K.: Sustainable development, the Hartwick´s rule and optimal growth, Enviromental and Resource Economics, No. 5 (1995), 393-411.
5. Jurča, P.: On sustainability constraint in models with non-renewable resources, Proceedings of the conference „Mathematical methods in economics and industry 2007“, Herlany, Slovak Republic.
6. Kolar, I., Michor, P.W., Slovak, J.: Natural operations in differential geometry, Springer - Verlag, 1993.
7. Martinet,V.: The Hartwick´s rule and the characterization of constant consumption paths in the presence of an exhaustable resource, Working paper, THEMA, 2004.
8. Nijmeijer, H., Schaft,A.: Nonlinear dynamical system, Springer - Verlag, New York, 1990.
9. Yano, K., Ishihara, S.: Tangent and cotangent bundles, M.Dekker Inc New York, 1973.

Changes in the Age Structure of the Population in the Czech Republic and its Economics Consequences

Tomáš Fiala[1], Jitka Langhamrová[1] and Jana Langhamrová[2]

[1]Department of Demography, [2]Student,
Faculty of Informatics and Statistics, University of Economics, Prague
nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
fiala@vse.cz, langhamj@vse.cz, janalanghamrova@seznam.cz

This article came into being within the framework of the long-term research project 2D06026, "Reproduction of Human Capital", financed by the Ministry of Education, Youth and Sport within the framework of National Research Program II.

Abstract. Population development may influence the economy and the economy, on the other hand, may influence population development. On the basis of the data on the sex-and-age structure of the population it is possible to anticipate relatively well the long-term development and foresee the future requirements, for instance, in the fields of education, the health service, social services, etc. Changes in the sex-and-age structure of the population are dependent on the levels of the birth and death rates and on migration. When the average life expectancy of people is lengthening and at the same time fertility is declining, the population will gradually change into a population of older people. In consequence this will have a marked influence on the economy, social and health insurance, etc.

Keywords: ageing of population, population projections, pension security

1 Introduction

In all economically developed countries the ageing of the population has been taking place for many years. Forecasts expect that in the future the life expectancy will rise still further, whereas fertility will more probably stagnate below the replacement level. The ageing of the population will therefore continue in this century.

The ageing of the population has a number of consequences in many areas of the life of society. One of the most frequently mentioned is the impact of the population ageing on the field of pension security systems. Also in the sphere of the financing of health care the population ageing must be taken into account. Low fertility follows in lower numbers of schoolchildren and applicants for studies etc.

The paper provides a simple analysis of the changes in the age structure and population ageing of the Czech Republic in the past century and offers estimates of the ageing in the future based on the latest version of the population projection.

2 Source Data and Methodology

All analyses are based on the time series of the sex-and-age structure of the population of the Czech Republic. Until the end of 2008 the real demographic structure of the population has been used (data of the Czech Statistical Office, see [7]). Since 2009 the data of the population projection are used.

The projection has been computed in the Department of Demography of the Faculty of Informatics and Statistics of the University of Economics, Prague. The classical component method (see e.g. [6]) with simplified migration (no emigration supposed, immigration equal to net migration) has been used for computations.

Initial demographic structure for the projection has been that of 1st January 2009. Three variants (low, medium and high) of the development of fertility, life expectancy and migration have been taken into account.

Preliminary demographic data of the first quarter of 2009 show that after several years of fertility growth the total fertility rate this year will probably be the same or even a little bit lower than in the previous year. The low variant of fertility development is therefore based on stabilization of the fertility at the present level with total fertility rate equal to 1.5. According to medium or high variant the total fertility rate will slowly continue to grow to the value 1.7, or 1.9 respectively. The fertility structure is supposed to converge to the fertility of the Netherlands, where the transition of the fertility has been finished and the fertility seems to be relatively stable.

The previous increase of the life expectancy is supposed to be continuing all the time. The variants differ only in the rate of growth. The life expectancy of males in 2050 is supposed to be for low variant 83 years, for medium variant 85.5 years and for high variant 88 years. The corresponding values for females are equal to 87, 89.5 and 92 years, respectively. The difference between the life expectancies of females and males is therefore supposed to diminish.

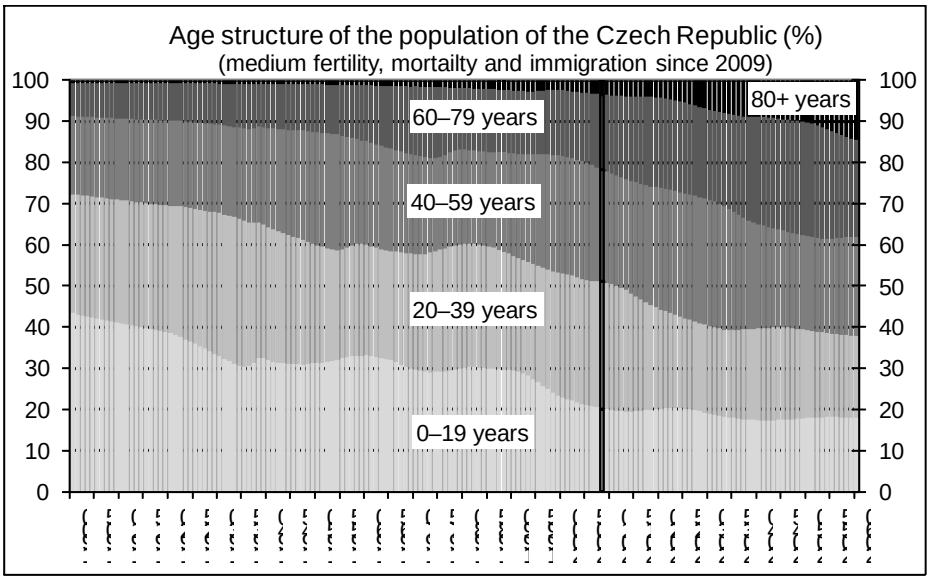
It is very difficult to predict net migration at the present time. Preliminary data indicate that this year the net migration may be at the level of only one third of the value of the previous year. One of the main reasons for this may be the continuing economic crisis. Because of this fact the projected annual net migration is supposed to be lower than in previous years in all variants, i.e. 10 000, 25 000 and 40 000 persons per annum, respectively. The demographic structure of immigrants is supposed to converge to the structure of the immigrants into the EU. See [1] and [4].

It is evident that minimal ageing of the population will occur in the case of high fertility and high net migration and simultaneously low life expectancy. And on the other hand low fertility and low net migration in combination with high life expectancy will lead to the highest ageing of the population.

3 Changing of the age structure of the population of the Czech Republic

The population of the Czech Republic has been ageing recently. In 1900 there were still more than 40 % of all citizens of the Czech Republic under the age of 20 and only about 10 % older than 60 years.

The population ageing in the future will depend on the development of fertility, mortality and migration. According to medium variant of the projection it is probable that in the future the proportion of persons under the age of 20 will stay permanently below the level of 20 %. The proportion of people of main productive age, i.e. the age of 20–59 years, will drop from the present 60 % to less than 45 % in the year 2060. There will therefore be a considerable decline in workforces. And the proportion of seniors will continue to grow. The proportion of people aged 60 and over will increase from the present 20 % to almost 40 %. In fifty years time every seventh citizen of the Czech Republic will be over 80 years of age. (See the Fig. 1.)



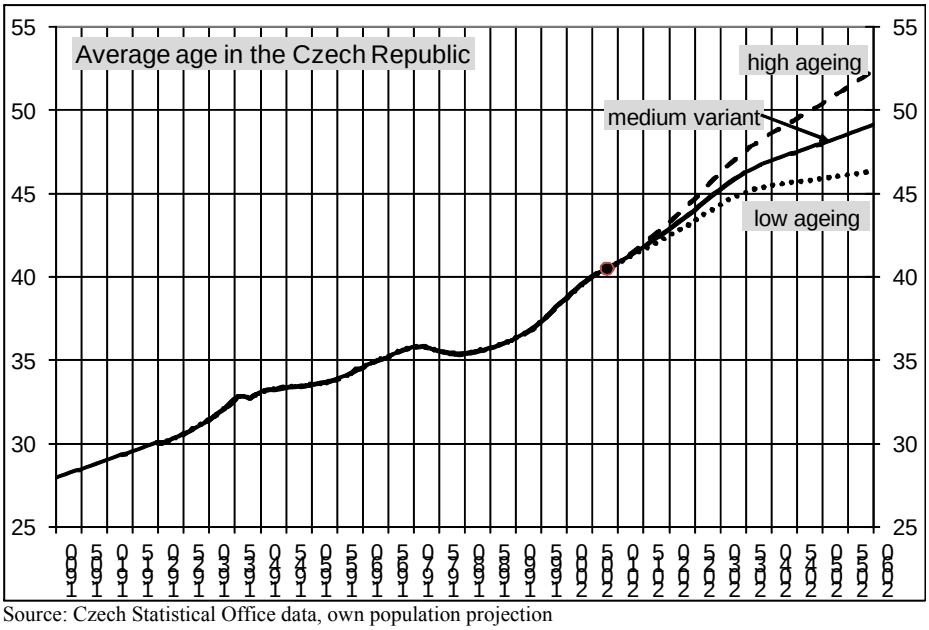
Source: Czech Statistical Office data, own population projection

Fig. 1. Age structure

The average age of the inhabitants of the Czech Republic at the beginning of the previous century was about 28 years. Several years ago its value has reached 40 years. And in the next decades the average age will continue to grow and be coming close to 50 years. (See the Fig. 2.)

The changes in the sex-and-age structure in the previous century are apparent also from the differences in the age pyramids (see the Fig. 3). At the beginning of the 20th century the demographic structure of the Czech population was very regular and it was a typical structure of the population of progressive type of population with prevailing share of young people.

The present age structure of the population of the Czech Republic is very irregular. It is characterized in particular e.g. in the higher proportion of those aged from 50–64 years. This has been caused by the growth in the number of those born during the German occupation and mainly after the 2nd World War. On the other hand the Czech Republic has a lower proportion of those aged 35–49. This is the result of the decline in natality in the late fifties and in the sixties, when the weak population cohorts of those born in the thirties (at the time of the economic depression) were at the age of highest fertility. There is also a markedly higher proportion here of those aged 25–34 years. This marked the climax of the fertility of the strong population cohorts of those born after the 2nd World War, also influenced by the pro-population measures in the seventies. (From the viewpoint of a demographer these measures were quite unsuitably timed and resulted in a further increase in the already great irregularity of the age structure.) In the Czech Republic there is also a lower proportion of children at the age about 10, caused by the decline in fertility as a result of the political, economic and social changes after 1989. Fertility has been dropping far below the so-called replacement level and young people have been postponing the birth of a child until a later age as in the majority of the countries of Western Europe.

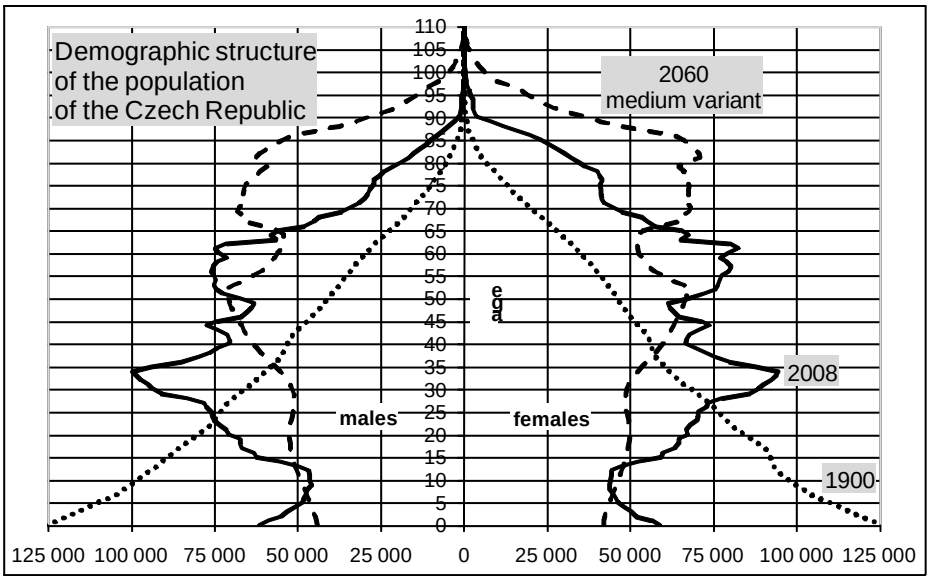


Source: Czech Statistical Office data, own population projection

Fig. 2. Average age

This irregular age structure of the population brings with it irregular course of the ageing of the Czech population in the nearest future. In the half of this century a very high proportion of seniors is expected to live in the Czech republic because the strong cohorts born in the seventies will reach the retirement age. And on the other hand the proportion of people at the age about 30 would be much lower than today because another baby-boom wave like in the seventies is not expected. (Preliminary data indi-

cate that probably since the year 2009 the number of births will decline for several decades.) The Czech population will be among the oldest countries in the world.



Source: Czech Statistical Office data, own population projection

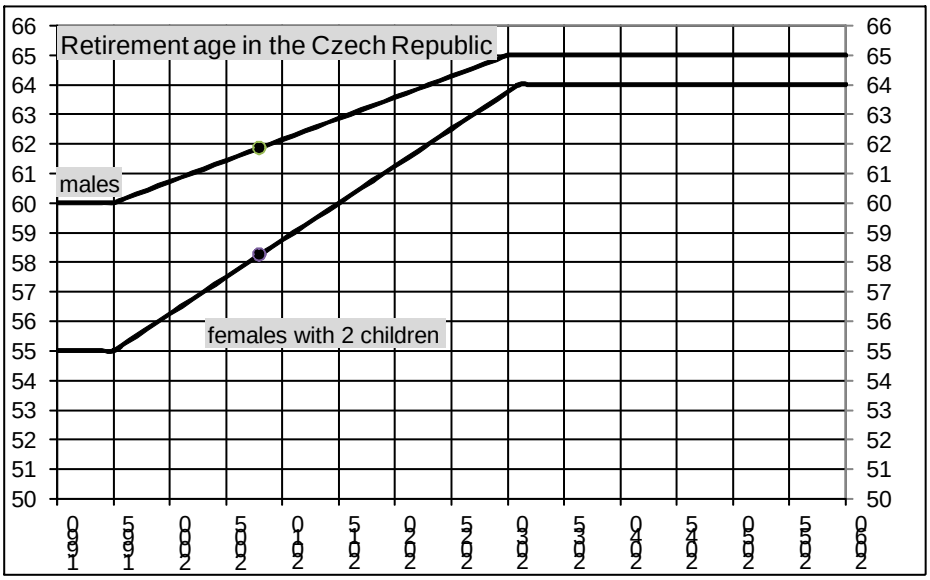
Fig. 3. Demographic structure

4 Consequences in the sphere of pension security

The ageing of the population will have a number of consequences in many areas of the life of society. One of the most frequently mentioned consequences is the impact on the field of pension security. At present pension security guaranteed by the state in the Czech Republic is still based exclusively on the so-called PAYG (pay-as-you-go) system. Economically active people pay contributions to pensions into the system of pension security and this is immediately re-allocated to payments to present pensioners. Saving for one's own pension in pension funds is so far merely voluntary. A suitable degree of burden on the current system is therefore the so-called old-age-dependency ratio (the ratio of the number of persons of retirement age to the number of persons of productive age).

Until the end of 1995 the retirement age in the Czech Republic was 60 years for males and 55 years for females with two children. For simplicity we have supposed that the retirement age of all women is the same as for women having two children. Since 1996 the retirement age is increasing according to the year of birth: for each subsequent year of births the pension age is 2 months higher (for males), or 4 months higher (for females) than for the previous year of births. According to present legal regulations this increase will continue until the retirement age will reach 65 years for males and 64 years for females with two children, respectively.

It is evident that the average annual increase of the retirement age in time is 1/7 years for males and 1/4 years for females. The dependence of the retirement age on time is shown in the Fig. 4 (for simplicity the linear instead of stepwise function was used). See [2].



Source: The Collection of Law, own computations

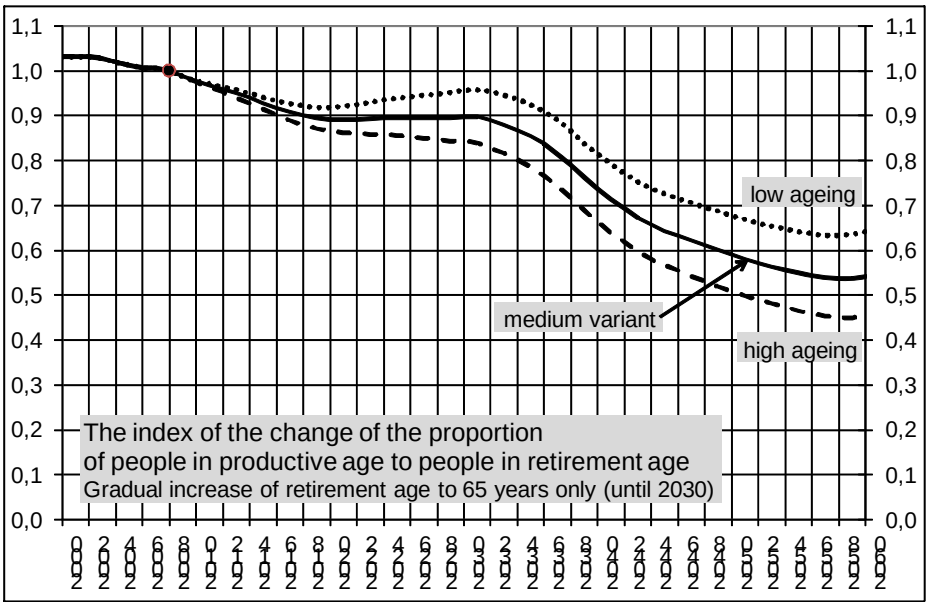
Fig. 4. Retirement age

The values of the old-age-dependency ratio have been computed taking into account the increase of retirement age in time. At present for 100 people of productive age there are about 37 persons of pension age. The gradual raising of the retirement age will only slow down, not prevent the increase of the old-age-dependency ratio. Around 2030 there may already be about more than 40 pensioners for every 100 persons of productive age. And in the following decades when the retirement age is supposed not to grow more, the old-age-dependency ratio would rise very rapidly. In 2060 it might already be as many as 60–80 pensioners depending on the variant of demographic development. (See the Fig. 5.)

The inverse value of the proportion mentioned above – the proportion of people in productive age to people in retirement age – and especially its changes can serve as a simple indicator of a threat of decrease of the value of pensions. The Fig. 6 indicates that in approximately in 10 years despite the increase of retirement age the proportion of people of productive age to one pensioner will drop to about 90 % of the present value. And after 2030 the decline will continue so that in 2060 the number of people "earning for one pensioner" will be only about 45–65 % in comparison with the present situation. Without changing the system there is a real threat of rapid decrease of pensions in the future.

Source: Czech Statistical Office data, own population projection

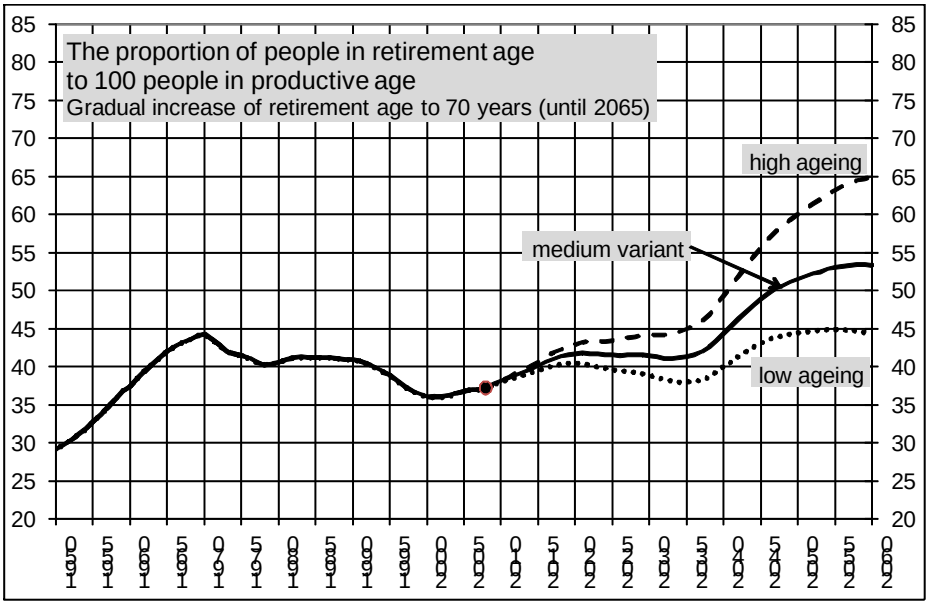
Fig. 5. The proportion of people in retirement age to people in productive age (increase of retirement age only to 65 years supposed)



Source: Czech Statistical Office data, own population projection

Fig. 6. The index of the change of the proportion of people in productive age to people in retirement age (increase of retirement age only to 65 years supposed)

It is of course possible and probable that the legal regulations will change and the increase of retirement age will continue permanently in the same way as today for males (i.e. for each subsequent year of births the pension age will be 2 months higher than for the previous year of births). In 2065 it would reach the level 70 years. But the Fig. 7 shows that even this permanent increase will not be satisfactory to eliminate the influence of population ageing. It will only cause that the increase of the old-age-dependency ratio will not be as rapid as in the previous case. In 2060 the value of the old-age-dependency ratio would be about 45–65 persons of retirement age to 100 persons of productive age. The main reason is that after 2040 the strong population cohorts born in the seventies will reach the retirement age. And in 2060 the number of people in productive age per one person in retirement age will be (despite the permanent increase of retirement age) 20–40 % lower than today. (See the Fig. 8.)



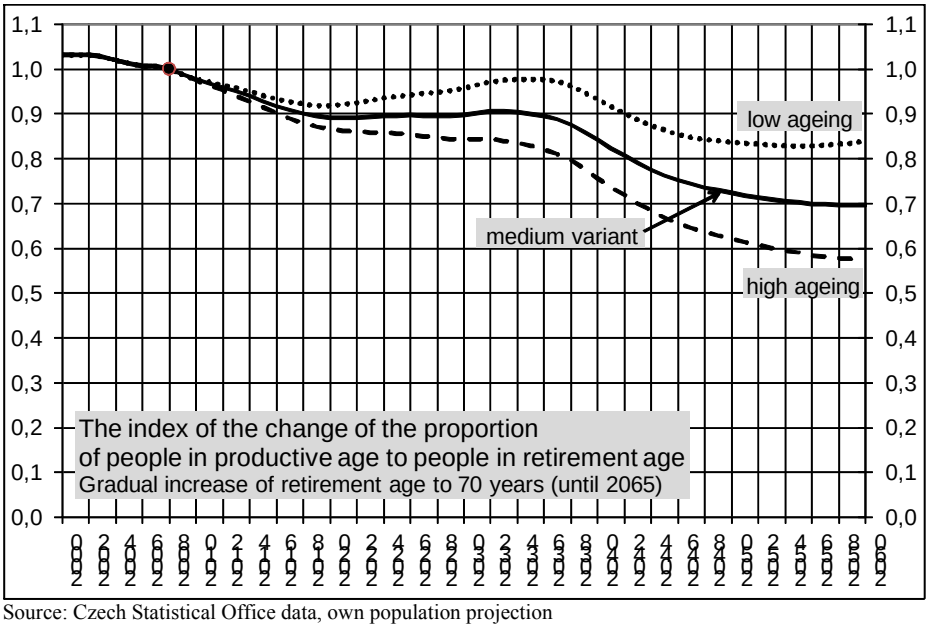
Source: Czech Statistical Office data, own population projection

Fig. 7. The proportion of people in retirement age to people in productive age (increase of retirement age to 70 years supposed)

If the increase of retirement age should quite eliminate the impact of the population ageing, it should be more rapid than at present. (See the Fig. 9.) The retirement age would reach 65 years of age about the year 2020, in 2045 the retirement age should be about 70 years and in 2060 its value should be about 75 years. In this case the value of the old-age-dependency ratio would stay at the present level (about 37 %) all the time.

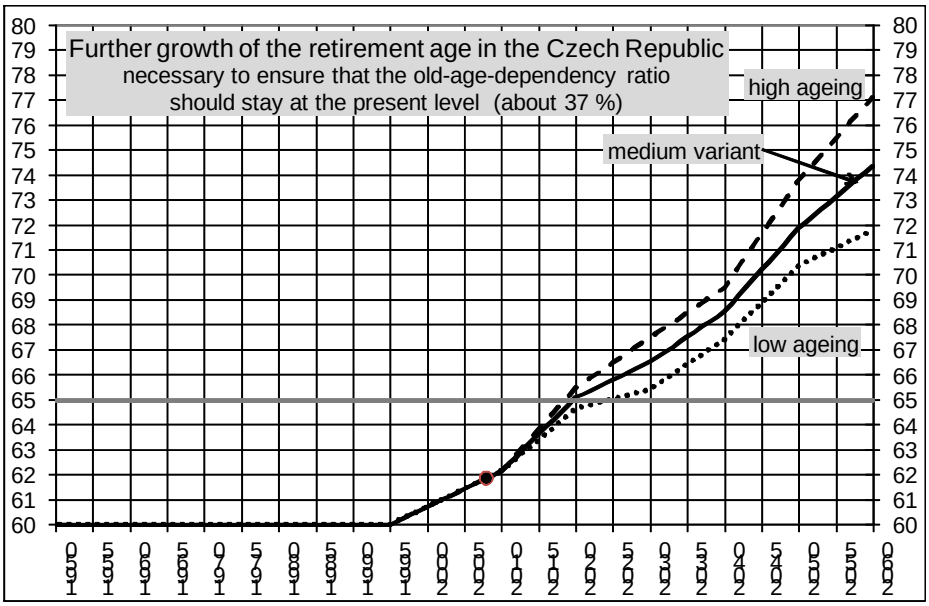
The question is, of course, whether the seniors would be able to work in such a high age and whether there would be sufficient suitable jobs for them.

A number of further solutions suggest themselves: the introduction of obligatory deductions to pension funds, the raising of deductions for pension security, greater support for additional pension insurance, etc.



Source: Czech Statistical Office data, own population projection

Fig. 8. The index of the change of the proportion of people in productive age to people in retirement age (increase of retirement age to 70 years supposed)



Source: Czech Statistical Office data, own population projection

Fig. 9. Retirement age necessary for maintaining the stability of the pension system

An interesting idea consists in the following combination of PAYG system and obligatory contributions to pension funds. While the rate of contributions to PAYG system would be the same for all, the rate of *obligatory contributions to pension fund* would be indirectly dependent on the number of children. Childless people would receive pension only from the fund, other people would have pension both from the fund (dependent on contributions, i.e. indirectly dependent on the number of children) and from the PAYG system (dependent on the number of children) so that the total pension would practically not depend on the number of children. People having 4 and more children would be regarded as people with 3 children only. (This pension reform has been proposed by James Hyzl, Martin Kulhavý and Jiří Rusnok, [3].) This idea could not only insure the stability of the PAYG pension system but also maybe increase the fertility.

It may also be assumed that labour productivity will increase with the rise in the education of the population. There is, then, some measure of hope that in the future it will be possible to maintain the present material standard of society even with a lower proportion of economically active persons and higher proportion of seniors than at present.

5 Conclusions

It is without doubt that the ageing of the population will continue. It is impossible to prevent the ageing but it is possible to slow it down by increase in fertility and immigration.

We should find the ways and create some solutions to harmonize professional career and parents' role of people. Not only in the sense that the child must not be an obstacle for the parents in their professional career but mainly that the professional career must not be an obstacle for the parents in the care and education of their children. It would be good take advantage of the possibility of work at home, flexible working time etc. as much as possible.

We should find the ways and create some solutions how to diminish the fear and negative distance between immigrants and the society. Immigrants should not be mainly the source of low-cost labour force. It is necessary to improve their working and living condition, eliminate illegal work, corruption in obtaining visa and other documents etc.

In any case the proportion of seniors in the society will increase. We should find the ways and create some solutions that the society will not only admire economic effect, success, youth, beauty, but also will be able to recognize the specific qualities of seniors and accept them. So that the seniors could feel their own personal dignity again as it is in some so called "primitive" societies (e. g. traditional American Indians).

Let's hope that the thinking of the society will change to be more concentrated and to pay more attention not only to economic growth, but also to social relations, environmental problems, culture, spirituality etc. Because if this change will come true, we are able to solve successfully not only the problem of population ageing but all the problems of the present time.

References

1. Arltová, M., Langhamrová, Jitka, Langhamrová, Jana: Population Ageing in European Union Countries and Migration. In: EURISBIS'09, pp. 220--221. Tilapia, IMSI (2009). ISBN 978-88-89744-13-0.
2. Fiala, T., Langhamrová, J., Marek, L.: Consequences of population ageing in an area of burdening of the pension insurance system. In: EURISBIS'09, pp. 195--196. Tilapia, IMSI (2009). ISBN 978-88-89744-13-0.
3. Hyzl, J., Kulhavý, M., Rusnok. J.: Penzijní reforma pro Českou republiku (inovativní přístup). Škoda Auto Vysoká škola, Mladá Boleslav (2005)
4. Kačerová, E.: International migration and mobility of the EU citizens in the Visegrad group countries: Comparison and bilateral flows. In: European Population Conference, p. 142. EPC, Barcelona (2008)
5. Koschin, F.: The Ageing of the Population from the Viewpoint of Potential Demography. In: ICABR 2008 [CD-ROM], p. 77. Mendel University in Brno, Brno (2008). ISBN 978-80-7375-154-8.
6. Readings in Population Research Methodology. Vol. 5. Population Models, Projections and Estimates. Bogue, D.J., Arriaga, E.E., Anderton, D.L. (eds.). United Nations Population Fund, Social Development Center, Chicago, Illinois (1993)
7. Czech Statistical Office, http://www.czso.cz

Capital Services in Supply and Use Framework[1]

Jaroslav Sixta, Jakub Fischer

Dept. of Economic Statistics, W. Churchill Sq. 4 ,
13067 Prague, Czech Republic
sixta@vse.cz, fischerj@vse.cz

Abstract. The issues of capital services were deeply discussed when the revision of SNA 1993 and ESA 1995 was prepared. The inclusion of capital services into national accounts has a lot of advantages relating to the valuation of capital stocks and productivity measurement. If capital services for other non-market producers were estimated and the cost method was changed by substituting of consumption of fixed capital by capital services, the value of output and the value of government consumption expenditures would be significantly influenced. There are important impacts on the total level of aggregates and industrial composition of output and value added as well. Therefore the figures in supply and use tables and symmetric input-output tables will be influenced by capital services for non-market producers because the current concept, when net operating surplus of other non-market producers equals zero would be changed.

Keywords:

1 Introduction

The system of national accounts provides lot of interconnected information that are widely requested by users. The system has its own history and development and currently a new standard is being prepared – System of National Accounts 2008. Supply and use tables and symmetric input-output tables represent an integral part of national accounts and moreover are used for other purposes like balancing commodities and deflation. Changes in national accounting standards will therefore influence supply and use tables and input-output tables. The new approach to R&D, military expenditures etc. will redraw the time series of GDP and other indicators. Capital services also have the effects on indicators for other non-market producers. In the end, capital services will not be compiled as a integral part of accounts but there we strongly effort to implement them. We deal with capital services and we tried to estimate their impact on GDP and other indicators and how they will be reflected in supply and use tables. This paper focuses on the issue of capital service only.

[1] This paper is prepared under the support of the project „Capital Services in National Accounts and its Impact on GDP of the Czech Republic“ of the Czech Science Foundation, project No. 402/07/0387.

2 Concept of Capital Services

Capital services show the contribution of assets to the production process and they should be more accurate to be used in a total productivity production function. Capital services consist of two parts:

1. Consumption of fixed capital and
2. Return on capital.

Consumption of fixed capital represent the wear and tear of fixed capital and it is the amount that investor has to obtain back from the investment otherwise he would never invest in such asset. On the other hand, return to capital represents the amount that creates the profit from investment. Therefore the capital services are linked with operating surplus. Capital services should equal the gross operating surplus (SNA 2008 draft) and return to capital to net operating surplus. The problem consists in the value of an asset. According to national accounts rules' assets should be valued at market prices (to 1.1. or 31.12.). This is very hard to survey because companies mostly have in their accounts historic prices and therefore model approach in used. In the European Union, perpetual inventory method (PIM) is widely used. The method is based on the revaluation of past investment into constant prices and then by applying of so-called mortality curves. Then the value of assets that still serve (and are included in the capital stock) is estimated. From this point of view there is not clear problem. When we adopt another assumption that the value of the asset should correspond with its discounted future income, the problems occur in non-market sector. Suppose that the value of the asset (V) is given by following formula:

$$V_t = \sum_{i=1}^{T} \frac{f_{t+i-1}}{(1+r)^i}, \tag{1}$$

where Vt is the value of the asset at 1.1.t
f is a future income from the use of the asset,
r is a discount factor.

Future incomes are given by quantities of capital services multiplied by their prices, e.g. ton-kilometres x price per ton-kilometre. Total value of capital services corresponds with the value of asset. An approach to the value of assets, described by formula (1), leads to the alternative way of PIM. On the basis of expected service-life and discount rate it is easy to derive age-efficiency and age-price schemes used for estimation of stock of capital in efficiency unit and net capital stock. Due to the discount factor, the linear decrease of quantity of capital services is followed by non-linear decrease of price of the asset, see figure 1.

Position of figure 1

The figure assumes that efficiency is decreasing 10% per year, e.g. the car can not provide so many services (ton-kilometres) because of frequent maintaining, brake-downs etc. The concept of deriving price from future income is possible for market producers who have net operating surplus. That means that capital services equals to the gross operating surplus. This is not valid for other non-market producers, like government units. The value of capital is in many countries based on PIM and it does not correspond to the operating surplus. Current standard SNA assumes that gross operating surplus of non-market producer equals to his consumption of fixed capital. Net operating surplus is therefore zero.

3 Non-market producers and their output

Current approach to other non-market producers (covering mainly government units) assumes that their output is a sum of intermediate consumption, compensation of employees, consumption of fixed capital and other net taxes on production. It means that gross value added is a sum of compensation of employees, consumption of fixed capital and other net taxes on production. Gross operating surplus consists of consumption of fixed capital only; net operating surplus is zero. With respect to the formula (1) it means that the assets in sector of non-market producers have lower value than the same assets in the sector of market producers. The difference is in net operating surplus, non-market producers have no return to capital, only consumption of fixed capital and therefore their future discounted incomes are lower. This strong assumption may or may not be correct. The supporters of the concept of capital services argue that the difference in the value of very similar property can not be justified. An example could be a school, either public or private; it could produce the same quality of services. This means that when non-market producer sells assets to the market producer, the value of capital increase and of value added as well. Capital services for non-market producers is the possible solution, it means the change of the so-called cost method of estimating output. Return to capital should be added to cost method and the capital services will be completed. Then the net operating surplus will be no longer zero.

Generally, there are two methods of estimating return to capital:

a) to use rates of return derived from market producers (internal),
b) to use rates of return for assets with low risk, like government bonds (external).

The advantage of internal method is that the rates are connected with the economy; it means the property has the same return to capital disregard who owns it. On the other hand, some assets are not owned by market producers. In the Czech Republic, the infrastructure (roads and railways) is owned by the state. Then the only solution represents external method. Internal rate of return (r) is derived by formula 2:

$$r = \frac{GOS_t - CFC_t + NHG_t}{K_t}, \tag{2}$$

where GOS … gross operating surplus,
CFC … consumption of fixed capital,
NHG … nominal holding gain,
K … capital in efficiency unit.

4 Supply and Use Tables Connected With Capital Services

The imputation of return to capital to other-non-market producers will influence not only operating surplus but final consumption expenditures, as well. The cost method, mentioned above, has to be therefore changed, instead of consumption of fixed capital, capital services will be included. By definition, other non-market output is consumed by other non-market producer, in the system it is recorded as government consumption expenditures. From the point of input-output analysis, the change will appear in input-output coefficient; the value of output will be increased and the value of final consumption as well. When such change in the methodology is applied, past figures have to be changed. The change of the figures could be significant and it could redraw the history. Input-output tables have a lot of users that rely on them and who use them for their economic research. Capital services are able to change the results of such research because they can influence nearly all industries. When the output of public administration industry is change, it is quite less problematic because this commodity is consumed mainly by government. The problem is in mixed industries, like transport and education and health.

Effect of imputation of capital services on input-output coefficients will be different from commodity to commodity. The most import problem is connected with mixed industries. Transport industry (or transport commodity) will be highly affected by return to capital on roads and railways. If structures were owned by market producer, the net operating surplus will be above zero. Therefore imputation significantly increases the level of output. In our example, we used 3% rate of return that was applied to the capital valued at efficiency units. The 3% rate of return was set on the basis of rate of government bonds because there is no equivalent in the sector of market producers.

Similarly, the industry of health and education is generally mixed industry but the share of market producers in the Czech Republic is very low.

5 Capital Services for the Czech Republic

We are solving project aimed at imputation of capital services for other non-market producers, the following figures were calculated within the project. The main purpose is to show, how national accounts' figures can be influenced by capital services. We adopted following assumptions that we use for computation:

a) Capital services cover only fixed assets; inventories, valuables and non-produced assets are not included in the model,

b) When no rate of return from market producers is available, 3 % rate of return is used,
c) Linear decrease of age-efficiency was derived from the official data on average service-lives,
d) Mixed income is not split into work and profit part; it is a part of total operating surplus,
e) Estimation of capital stock in efficiency unit was done. Our estimates are based on alternative PIM and time series of gross fixed capital formation for past years were based on the splitting of gross capital stock as published by the Czech Statistical Office.

The following figure 2 shows average rate of return for total economy, including industries with external rate of return. The development of total rate of return is given mainly by the development of net operating surplus of market producers because the rate of return of non-market producers is quite steady. The reason is a high share of capital with external rate of return, more than 86% in 2006. Internal rate of return was derived by formula (2). The sharp decrease in 2000 – 2002 was caused by the decrease of net operating surplus. Generally, it means that the imputation of capital services has very low effects on nominal and real GDP growth rate. The effects are highly shown in the level of GDP. In our example, we estimated that the level of GDP will be increased by 5 % in average for 1995 to 2006.

Position of figure 2

The impact on GDP growth rate is very low, we estimated the change of real GDP growth rate for the period 1996 – 2006, see figure 3. The highest positive change of real GDP growth rate was estimated for 1997 (+ 0.3 %); similarly, the lowest for 1995 (- 0.3%).

Position of figure 3

If the change of output occurs, the input coefficients must change, as well. The level of value added is increased (operating surplus). The change of input coefficient for transport commodity is shown in table 1. It is based on published symmetric input-output table (product x product) for 2005 by the Czech Statistical Office and adjusted by capital services for other-non market producers.

Position of Table 1

6 Conclusion

The aim of this short paper was to describe the effects of capital services on the time-series comparability. Capital services for other non-market producers have a lot of advantages, namely the consistent approach to the property disregard who owns it and consistent approach to operating surplus. On the other hand the biggest disadvantage is given by the changes of data and problems for the users and comparability between

countries. Capital services are nowadays partly included in the system of national accounts, in consumption of fixed capital. But there is no return to capital for other non-market producers. An impact of implementation consist in the increase of output of selected products, mainly transport, public administration, health and education are influenced. Although the imputation of return to capital for other non-market producers have nearly no effect on GDP growth rate, the changed level of GDP influences input-output analyses. Input coefficients are changed (matrix A) for selected products (or industries) and because of the high values of imputed rent on capital the impact on Leontief inverse matrix ((I-A)-1 may be significant. The estimated increase of the level of output and GDP by imputation is about 5%. Input-output tables is not avaible now but we expect quite significant changes in the Leontief inverse matrix. Such methodological revisions are dangerous at least from two points of view, at first the users who are using input-output tables for econometric modelling may be unpleasantly surprised that their models providing different results, especially when the sensitiveness of data is high for small changes. The second danger is given by the possible incomparability of countries' GDP and national accounts. The importance of modelling is increased and without any standardisation the figures can be influenced by totally different assumptions among the world. The last what should not be forgotten is that such models require a lot of detailed data that are very hard to acquire. The standard PIM that gives clear and simple view on capital is applied in many countries and the change to more model based approach – alternative PIM requires also a lot of experiences from economic modelling and from economic theory. Therefore it seems not possible that such approach can be applied in the near future except for a few of very developed countries. We can remind the complication with adoption of SNA 1993 in the world because a lot of countries are still using SNA 1968. From Czech experience (country that had to change its macroeconomic statistical system from Material Product System (MPS) to SNA 1993 (respectively ESA 1995)) correct implementation of capital services seems too be much ambitious.

References

1. European system of accounts (ESA 1995). Eurostat, Luxembourg 1996.
2. FISCHER, J, SIXTA, J. Implementation of the New ESA Standard into the Czech National Accounts: Capital Services and Related Issues. International Statistical Institute. In: Bulletin of the International Statistical Institute 56th Session [CD-ROM]. Lisabon, 2007.
3. HARRISON, A. Measuring the contribution of non-financial assets to non-market production. OECD, 2004.
   URL:http://unstats.un.org/unsd/nationalaccount/AEG/papers/m2assets.pdf
4. Links between Business Accounting and National Accounting. United Nations 2000.
5. Measuring Capital. A Manual of the Measurement of Capital Stocks, Consumption of Fixed Capital and Capital Services. OECD, Paris 2001.

6. SIXTA, J. The Estimates of Consumption of fixed capital. 8th International Scientific Conference - Applications of Mathematics and Statistics in Economy (AMSE 2005) in Wroclaw.
7. SIXTA, J. The Influence of the Revision of Consumption of Fixed Capital on GDP. In: Statistics: Investment in the Future. http://www.czso.cz/sif/conference2004.nsf/i/national_accounts.
8. System of National Accounts 1993. United Nations, IMF, OECD, Eurostat, World Bank, New York 1993.
9. System of National Accounts 2008, volume 2, Pre-edit white-cover version of the 2008 SNA, downloaded at http://unstats.un.org/unsd/sna1993/draftingphase/volume1and2.asp.
10. www.czso.cz
11. www.euklems.net

Appendix

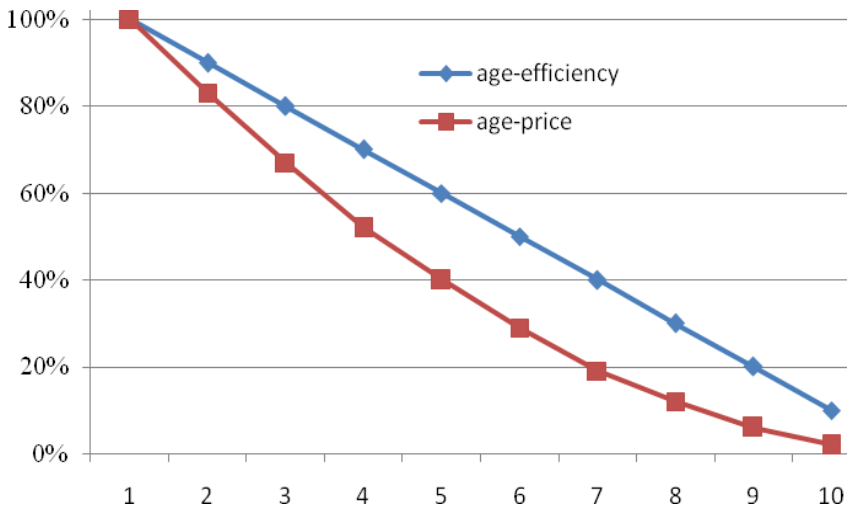


Fig. 1 Age-efficiency and age-price

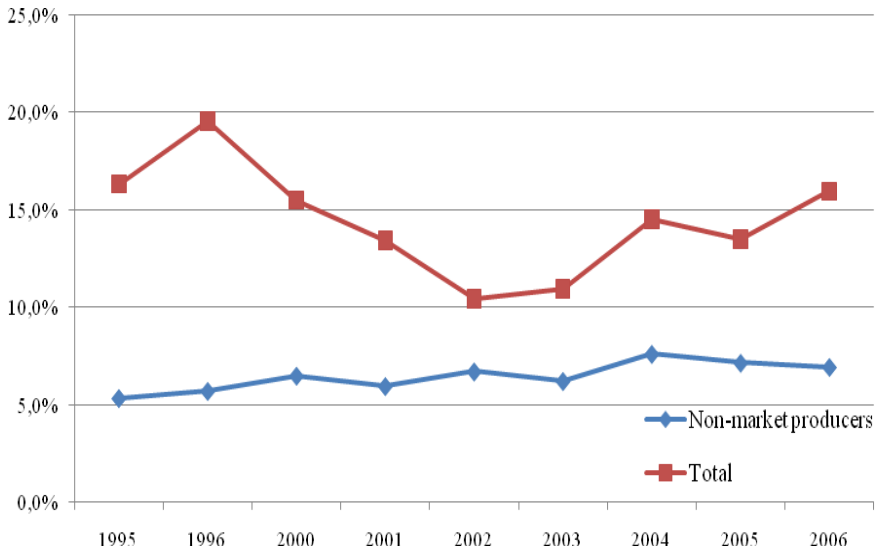


Fig. 2 Rates of return for non-market producers and total economy, %

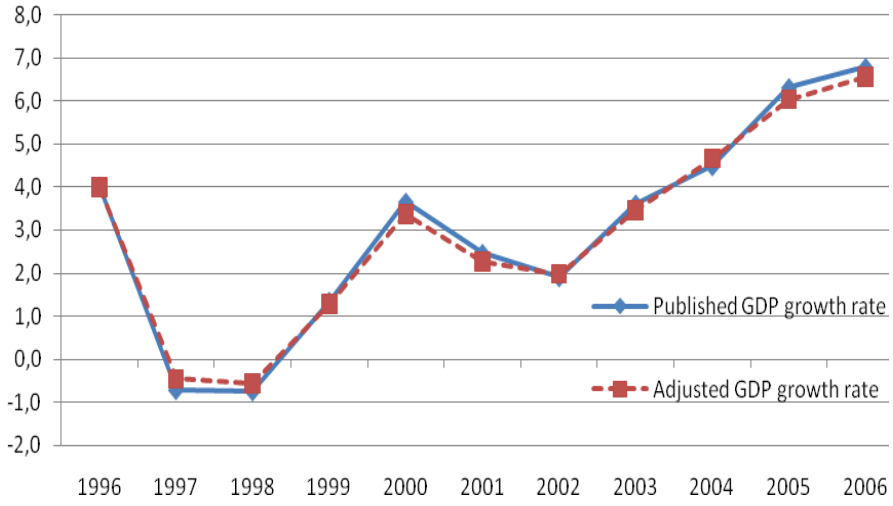


Fig. 3 Real GDP growth rate, published and adjusted, %

Table 1. Input coefficient for transport commodity, 2005, %

	Inputs	
	Published	Revised
All products	58,4%	48,5%
Net taxes on products	2,6%	2,1%
Gross value added	39,0%	49,3%
Output	100,0%	100,0%

Generalized binomial distribution – dependent version

Stanisław Heilpern

University of Economics, ul. Komandorska 118/120, 53-345 Wrocław, Poland
stanislaw.heilpern@ue.wroc.pl

Abstract. The generalization of the binomial distribution is investigated in this paper. The classical assumption of independence of Bernoulli random variables is omitted and it is weakened by exchangeability. The combination between dependence and strict dependence is studied. The case where the dependence structure is described by copula is investigated. The Markov binomial distribution is examined.

Keywords: Generalized binomial distribution, dependent structure, copula, Markov binomial distribution

1 Introduction

In this paper we investigate the generalization of binomial distribution. We mainly omit the assumption of independence of the Bernoulli random variables. This assumption is very "nice" from mathematical point of view, we can prove many facts in the easy way in this case, but it is often nonrealistic. The variables and processes occurred in economic models are generally dependent. For instance, the common external factor, e.g. natural calamity, affects different variables making them dependent. The generalized binomial model presented in this paper describe reality better then the classical model.

The presented generalized binomial distribution may be applied in many actuarial and financial models, e.g. the excess-of-loss reinsurance and the credit risk management models (see [6], [7] and [5]). In the excess-of-loss reinsurance model, we investigate the portfolio consisted of the n claims $Y_1, \ldots, Y_n$ with retentions $d_1, \ldots, d_n$. We want to analyze the random variable

$$X = \sum_{j=1}^{n} I_j ,$$

where I_j are the following Bernoulli random variables

$$I_j = \begin{cases} 0 & Y_j \le d_j \\ 1 & Y_j > d_j \end{cases}.$$

The variable X can be interpreted as the number of claims the reinsurer has to pay for. The random variables Y_j can be often dependent in practice.

In the credit risk management models we investigate the dependent Bernoulli random variables $I_1, \ldots, I_n$. The variable I_j reflects the default indicator for obligor j at time T.

In Section 2 we present the general model, where the Bernoulli random variables can be dependent and they can have different probability of "success". In Section 3 we assume that the dependent structure is described by copula and in Section 4 we assume that the Bernoulli random variables are exchangeable. The combination of independence and strict dependence are investigated in the Section 5. The dependent structure induced by Archimedean copulas are studied in the Section 6. Finally, we examine Markov binomial model. The impact of degree of dependence on the form of the generalized binomial distribution is investigated in the presented examples.

2 General model

Let $I_1, \ldots, I_n$ be finite sequence of Bernoulli random variables. They represent the results of trials. We will denote probability of "success" in the j-th trial I_j as symbol p_j and probability of "defeat" as $q_j = 1 - p_j$, i.e.

$$p_j = P(I_j = 1) \text{ and } q_j = P(I_j = 0).$$

In the classical theory it is assumed that the random variables I_j are independent and they have the same distribution. But in this section we omit these assumptions.

The joint distribution of Bernoulli variables I_j is described by the probability mass function (p.m.f.)

$$f(i_1, \ldots, i_n) = P(I_1 = i_1, \ldots, I_n = i_n),$$

where $i_j \in \{0, 1\}$ or by the cumulative distribution function (c.d.f.)

$$F(x_1, \ldots, x_n) = P(I_1 \le x_1, \ldots, I_n \le x_n).$$

The probability generating function (p.g.f.) of the random vector $(I_1, \ldots, I_n)$ is of the form

$$G(t_1,\ldots,t_n) = \sum_{i_1,\ldots,i_n \in \{0,1\}} f(i_1,\ldots,i_n) t_1^{i_1} \ldots t_n^{i_n} .$$

The joint c.d.f. F of random variables $I_1, \ldots, I_n$ is univocally determined by the values in the points of jump. We can interpret every point of jump $(i_1, \ldots, i_n)$ of c.d.f. F as the subset $A \subset \{1, \ldots, n\}$ such, that $i_j \in A$ iff $i_j = 1$. We will use the notation $\mathbf{1}_A = (i_1, \ldots, i_n)$. The number of elements of subset A, denoted by $|A|$, is equal the number of "1" in the point of jump $(i_1, \ldots, i_n)$. Moreover, we can calculate the value of joint p.m.f. f using the values of joint c.d.f. F in the points of jump ([1], [5]):

$$f(\mathbf{1}_A) = \sum_{j=0}^{k} (-1)^j \sum_{D \subset A, |D| = k-j} F(\mathbf{1}_D) . \tag{1}$$

We will investigate the following random variable in our paper:

$$X = I_1 + \ldots + I_n.$$

This is the number of "successes" in the n trials. We will call the distribution of such random variable as a **generalized binomial distribution**. The (1) and the values of the joint c.d.f. F lets us to derive the p.m.f. f_X of this random variable ([1], [5]):

$$f_X(k) = P(X = k) = \sum_{|A|=k} f(\mathbf{1}_A) = \sum_{j=0}^{k} (-1)^j \binom{n-k+j}{j} \sum_{|A|=k-j} F(\mathbf{1}_A) .$$

The g.p.f. of random variable X is described by formula:

$$G(t) = \sum_{i_1, \ldots, i_n \in \{0,1\}} f(i_1, \ldots, i_n) t^{i_1 + \ldots + i_n} = \sum_{k=0}^{n} \sum_{|A|=k} f(\mathbf{1}_A) t^k ,$$

the expected value and the variance of X are given by

$$E(X) = \sum_{j=1}^{n} p_j \qquad \text{and} \qquad V(X) = \sum_{j=1}^{n} V(I_j) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Cov(I_i, I_j) .$$

3 Copulas

We can describe the dependent structure of the random variables $I_1, \ldots, I_n$ by copula functions. The copula C_n is the n-dimension c.d.f. on $[0, 1]^n$ with the uniform marginal distribution ([8], [4]). This function is the link between the join c.d.f. F and the marginal c.d.f. F_j and it satisfies the following relation:

$$F(x_1, \ldots, x_n) = C_n(F_1(x_1), \ldots, F_n(x_n)).$$

If the k arguments of n-dimension copula C_n are equal 1, then we obtain the $(n - k)$-dimension marginal copula.

We investigate discrete random variables I_j, so the copula is univocally determined in the points of jump only ([8]). The copula determines the c.d.f. F only, so if we want to obtain the p.m.f. f, we must use the formula (1). The values of the marginal c.d.f. F_j the point of jump are equal

$$F_j(0) = P(I_j = 0) = q_j, \qquad F_j(1) = 1.$$

For instance, we have $F(1, 0, 0, 1, 0) = C_5(1, q_2, q_3, 1, q_5)$.

The covariance $\mathrm{Cov}_{i,j}$ between random variables I_i and I_j is determined by two-dimensional marginal copula $C_{i,j}$ induced by copula C_n. The k-th argument of copula C_n, for k different than i or j, is equal 1 in this case. For instance, for $n = 5$ we have $C_{2,3}(u, v) = C_5(1, u, v, 1, 1)$. Using (1) for $n = 2$, we obtain the covariance

$$\begin{aligned}\mathrm{Cov}_{i,j} &= E(I_iI_j) - E(I_i)E(I_j) = P(I_i = 1, I_j = 1) - p_ip_j = \\ &= 1 - C_{i,j}(q_i, 1) - C_{i,j}(1, q_j) + C_{i,j}(q_i, q_j) - p_ip_j = C_{i,j}(q_i, q_j) - q_iq_j,\end{aligned}$$

and the coefficient of correlation $\rho_{i,j}$ between the pair of random variables I_j and I_j

$$\rho_{i,j} = \frac{C_{i,j}(q_i, q_j) - q_iq_j}{\sqrt{p_iq_ip_jq_j}}.$$

If the random variables $I_1, \ldots, I_n$ are independent then the dependent structure is described by copula

$$\Pi(u_1, \ldots, u_n) = u_1 \cdot \ldots \cdot u_n.$$

The second extreme case – strict dependence, called comonotonicity, induces the following copula:

$$M(u_1, \ldots, u_n) = \min(u_1, \ldots, u_n).$$

Example 1. Let the dependent structure induced by Bernoulli random variables I_1, I_2, I_3 is described by Farlie-Gumbela-Morgenstern copula ([6], [4]):

$$\begin{aligned}C(u_1, u_2, u_3) = u_1u_2u_3(1 &+ \alpha_{12}(1 - u_1)(1 - u_2) + \alpha_{13}(1 - u_1)(1 - u_3) + \\ &+ \alpha_{23}(1 - u_2)(1 - u_3) + \alpha_{123}(1 - u_1)(1 - u_2)(1 - u_3)),\end{aligned}$$

where $-1 \leq \alpha_{12}, \alpha_{12}, \alpha_{12}, \alpha_{123} \leq 1$. The parameters $\alpha_{12}, \alpha_{12}, \alpha_{12}$ reflect the degree of dependence between these random variables, because we have

$$\rho_{i,j} = \sqrt{p_i q_i p_j q_j}\,\alpha_{ij},$$

where $i,j \in \{1,2,3\}$. When these parameters are equal zero, then we obtain independence.

If we want to describe the distribution of random variable

$$X = I_1 + I_2 + I_3,$$

then we must first compute the values of c.d.f. F in the points of jump, i.e. the values $F_0 = F(0, 0, 0)$, $F_i = F(\mathbf{1}_{\{i\}})$, $F_{ij} = F(\mathbf{1}_{\{i,j\}})$ and using (1) compute the values of p.m.f. f: f_0, f_i, f_{ij}. So, we obtain the p.m.f. of random variable X taking the sum of values of f for the subsets with the same number of elements. For instance, we have

$$F_2 = F(0, 1, 0) = C(q_1, 1, q_3) = u_1 u_3(1 + \alpha_{13}(1 - u_1)(1 - u_3)),$$
$$f_2 = F_2 - F_0 = F_2 - C(q_1, q_2, q_3),$$
$$F_{13} = F(1, 0, 1) = C(1, q_2, 1) = q_2, \quad f_{13} = F_{13} - F_1 - F_3 + F_0$$

and $f_X(2) = f_{12} + f_{13} + f_{23}$.

Let $q_1 = 0.8$, $q_2 = 0.6$, $q_1 = 0.7$, $\alpha_{12} = 0.3$, $\alpha_{13} = 0.8$, $\alpha_{23} = 0.6$ and $\alpha_{123} = -0.1$, then we obtain the following p.m.f. of random variable X:

$$f_X(0) = 0.3835, \quad f_X(1) = 0.3779, \quad f_X(2) = 0.1935, \quad f_X(3) = 0.0451.$$

The expected value and variance of X are given by

$$E(X) = 0.9, \qquad V(X) = 0.7473$$

and the coefficients of correlation between the pairs of random variables I_j are equal

$$\rho_{1,2} = 0.0588, \quad \rho_{1,3} = 0.1466, \quad \rho_{2,3} = 0.1347.$$

4 Exchangable Bernoulli random variables

In the classical model it is assumed that the random variables I_j have the same distribution with the probability of "success" p and they are independent. But we weaken this assumptions and we assume that these random variables are exchangeable only. The random variables $I_1, \dots, I_n$ are exchangeable if their joint distribution do not depend on the order of them, i.e.

$$P(I_1 \le x_1, \dots, I_n \le x_n) = P(I_{\pi(1)} \le x_1, \dots, I_{\pi(n)} \le x_n),$$

for every permutation π of set $\{1, \dots, n\}$. If $|A| = |B| = k$, then we obtain very comfortable in many computations relation

$$F(\mathbf{1}_A) = F(\mathbf{1}_B) = P(I_{k+1} = 0, \dots, . I_n = 0) = F_{k,n}.$$

Using (1) we can derive the values of p.m.f. f. They take the following form in this case:

$$f_{k,n} = P(I_1 = 1, \dots, I_k = 1, I_{k+1} = 0, \dots, I_n = 0) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} F_{k-j,n}\,.$$

The exchangeable random variables $I_1, \dots, I_n$ are equicorrelated, i.e.

$$\mathrm{Cov}(I_i, I_j) = \mathrm{Cov}(I_h, I_k) = \mathrm{cov},$$

for $i \neq h$ and $j \neq k$. Moreover, Pearson coefficient of correlation $\rho = \rho_{i,j}$ satisfies the following inequalities ([6]):

$$-\frac{1}{n-1} \le \rho \le 1.$$

The p.g.f. of the general binomial distributed random variable X takes the form

$$G_X(t) = \sum_{k=0}^{n} f_{k,n} t^k\ .$$

Using the values of functions $F_{i,n}$ we can derive the values of p.m.f. f_X of random variable X in the following way:

$$f_X(k) = \sum_{j=0}^{k} (-1)^j \frac{n!}{(n-k)!\, j!\, (k-j)!} F_{k-j,n}\ .$$

The expected value and variance of random variable X are given by

$$E(X) = np, \qquad V(X) = n(pq + (n-1)\mathrm{cov}).$$

For exchangeable random variables $I_1, \ldots, I_n$ the copula C_n, which describes their dependent structure satisfies the condition:

$$C_n(u_1, \ldots, u_n) = C_n(u_{\pi(1)}, \ldots, u_{\pi(n)}).$$

Then we have

$$F_{k,n} = C_n\left(\underbrace{1,\ldots,1}_{k}, \underbrace{q,\ldots,q}_{n-k}\right) = C_{n-k}(q,\ldots,q)$$

and

$$\mathrm{Cov} = C_2(q, q) - q^2,$$

where C_k is the marginal, k-dimension copula of C_n. Moreover, the coefficient of correlation between the pair random variables I_j is equal

$$\rho = \frac{C_2(q,q) - q^2}{pq}.$$

5 Combination between independence and comonotonicity

We assume now, that the random variables $I_1, \ldots, I_n$ have the same distribution and the dependent structure is described by the convex combination of copulas Π and M, induced by the independence and comonotonicity, i.e.

$$C = (1-\rho)\Pi + \rho M, \tag{2}$$

where $0 \le \rho \le 1$. This case was studied first by Tallis in [9] and next by Kolev and Paive in [6].

If the random variables $I_1, \ldots, I_n$ are independent, i.e. $\rho = 0$, then we obtain the classical binomial distribution with known p.m.f.

$$f_X(k) = \binom{n}{k} p^k q^{n-k},$$

probability generating function

$$G_X(t) = (q + pt)^n$$

and variance $V(X) = npq$. Then

$$F_{k,n} = q^{n-k}, \qquad f_{k,n} = p^k q^{n-k},$$

and $\rho_{i,j} = 0$.

We know from the classical course of probability theory, that if $np = \lambda > 0$ and n tends to infinity, then the binomial distribution tends to the Poisson distribution with the parameter λ ([3]).

The second component of combination – comonotonicity is described by copula

$$M(u_1, \ldots, u_n) = \min(u_1, \ldots, u_n).$$

Then

$$F_{k,n} = \begin{cases} q & \text{for} \quad k < n \\ 1 & \text{for} \quad k = n \end{cases}, \qquad f_{k,n} = \begin{cases} q & \text{for} \quad k = 0 \\ 0 & \text{for} \quad 0 < k < n \\ p & \text{for} \quad k = n \end{cases}$$

and

$$f_X(k) = \begin{cases} q & \text{for} \quad k = 0 \\ 0 & \text{for} \quad 0 < k < n \\ p & \text{for} \quad k = n \end{cases}.$$

The p.g.f. of the random variable X is equal

$$G_X(t) = q + pt^n$$

in this case and variance is given by $V(X) = n^2pq$. There is a strict dependence and the coefficient of correlation between the pairs of variables I_j takes the most values, i.e. $\rho_{i,j} = 1$.

The p.g.f. of the general binomial distribution X with copula defined by (2) is the convex combination of component p.g.f.:

$$G(t) = (1 - \rho)(q + pt)^n + \rho(q + pt^n).$$

The coefficient of the convex combination ρ is equal the coefficient of correlation between the pairs of random variables I_j. Because using (2) we obtain

$$\rho_{i,j} = \frac{(1-\rho)q^2 + pq - q^2}{pq} = \rho .$$

The variance of X is given by formula

$$V(X) = npq(1 + \rho(n - 1)).$$

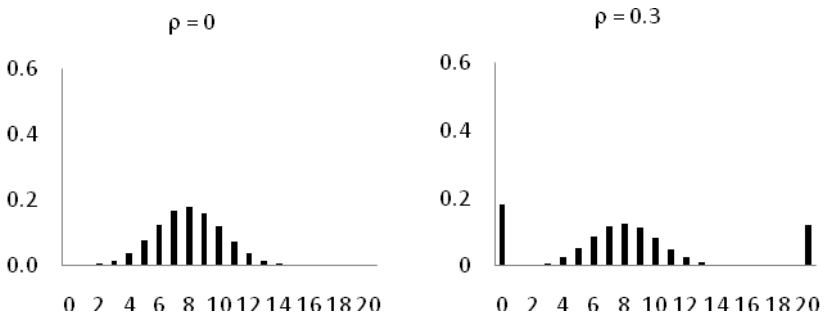
In the asymptotic case, if $np = \lambda > 0$ and n tends to infinity, then the general binomial distributed random variable X tends to the random variable

$$(1 - \rho)X_\lambda + \rho\delta_{\{0\}},$$

where X_λ is Poisson random variable with parameter λ and $\delta_{\{0\}}$ is the random variable focused in zero.

Example 2. We analyze of the influence of the degree of dependence, i.e. the value of coefficient of correlation ρ, on the form of the generalized binomial distributed random variable X. Let $n = 20$, ρ is equal 0 (independence), 0.3, 0.7 and 1 (strict dependence) for probability of "success" $p = 0.4$. The particular p.m.f. of random variable X are presented on fig. 1

We see that degree of dependence represented by the coefficient of correlation ρ radically influences on the form of p.m.f. of such generalized binomial distribution. The graphs vary from the classical, modal form to the graph focused in the two extreme point.



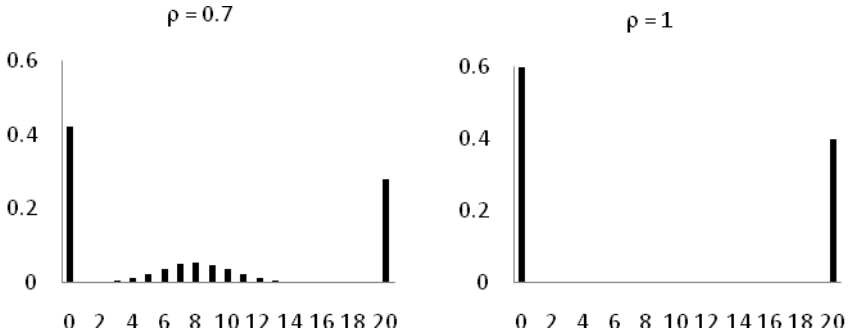


Fig. 1. Generalized binomial distribution for different values of the coefficient of correlation ρ

6 Archimedean copulas

The Archimedean copulas are characterized by simple form, so they are often used in the various applications. These functions are induced by the generator φ, which is the decreasing, convex function satisfying conditions: $\varphi(0) = \infty$, $\varphi(1) = 0$. The Archimedean copulas take the quasi-additive form ([8], [4]):

$$C_n(u_1, \dots, u_n) = \varphi^{-1}(\varphi(u_1) + \dots + \varphi(u_n)).$$

The c.d.f. $F_{k,n}$ takes the following values in this case:

$$F_{k,n} = \varphi^{-1}((n-k)\varphi(q)).$$

In practice we often use the families of Archimedean copulas characterized by some parameters. These parameters reflect degree of dependence and there are relations between the values of the parameters and the Kendal or Spearman coefficients of correlation ([8], [4]). Every Archimedean copula C_n for $n > 2$ satisfies the following inequality:

$$\Pi_n(u_1, \dots, u_n) \leq C_n(u_1, \dots, u_n).$$

Now, we present the main, often used in practice families of copulas.

a) Clayton family:

$$C_n(u_1,..,u_n) = (u_1^{-\alpha} + \dots + u_n^{-\alpha} - n + 1)^{-1/\alpha},$$

for $\alpha > 0$, with generator $\varphi(u) = (u^{-\alpha} - 1)/\alpha$. Limit value of parameter $\alpha = 0$ corresponds with independence, $\alpha = \infty$ implies comonotonicity and

$$F_{k,n} = ((n-k)q^{-a} - n + k + 1)^{-1/\alpha}.$$

b) Gumbel family:

$$C_n(u_1,..,u_n) = \exp(-((-\ln u_1)^{\alpha} + ... + (-\ln u_n)^{\alpha})^{1/\alpha}),$$

for $\alpha \geq 1$, with generator $\varphi(u) = (-\ln u)^{\alpha}$. For $\alpha = 1$ we obtain independence and for $\alpha = \infty$ strict dependence. Moreover

$$F_{k,n} = q^{(n-k)^{1/\alpha}}.$$

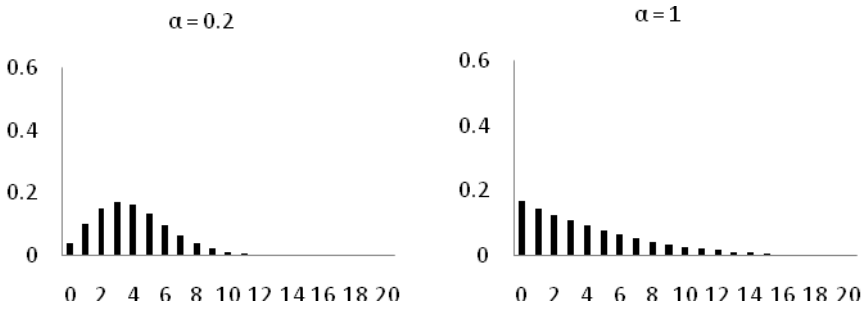
c) Frank family:

$$C_n(u_1,..,u_n) = -\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha u_1} - 1)...(e^{-\alpha u_n} - 1)}{(e^{-\alpha} - 1)^{n-1}}\right),$$

for $0 \leq \alpha$, with generator $\varphi(u) = -\ln((e^{-\alpha u} - 1)/(e^{-\alpha} - 1))$. Limit value $\alpha = 0$ corresponds with independence and $\alpha = \infty$ implies strict dependence. Then

$$F_{k,n} = -\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha q} - 1)^{n-k}}{(e^{-\alpha} - 1)^{n-k-1}}\right).$$

Example 3. Let the dependent structure is described by Clayton copulas with parameters α equals 0.2, 1, 2, 10 and $p = 0.2$. The values of parameter correspond to the values of the Kendal coefficient of correlation equal 0.17, 0.33, 0.5 and 0.83. The particular p.m.f. of random variable X are presented on fig. 2.



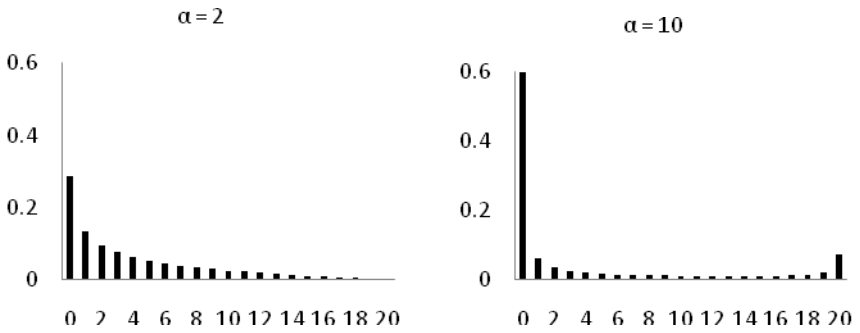


Fig. 2. Generalized binomial distribution induced by Clayton copula for different values of the parameter α

We se that degree of dependence represented by the parameter α radically influences on the form of p.m.f. of generalized binomial distribution induced by Clayton copula.

7 Markov binomial distribution

Let I_0, I_1, I_2, ... be stationary Markov chain with binary state space {0, 1}. Now, we determine the transition probabilities $p_{ij} = P(I_{k+1} = j | I_k = i)$, where $i, j \in \{0, 1\}$, knowing the coefficient of correlation ρ between random variables I_k i I_{k+1} and the probability of "success" $p = P(I_k = 1)$. This process is stationary, then these probabilities do not depend on $k = 0, 1, \ldots$. The coefficient of correlation is given by

$$\rho = \frac{E(I_{k+1}I_k) - E(I_{k+1})E(I_k)}{pq} = \frac{P(I_k = 1,\ I_{k=1} = 1) - p^2}{pq}.$$

Then

$$p_{11} = \frac{P(I_k = 1,\ I_{k+1} = 1)}{P(I_k = 1)} = p + \rho q\ ,$$

and $p_{10} = 1 - p_{11} = q - \rho q$. So, we obtain

$$P(I_k = 0, I_{k+1} = 1) = p - P(I_k = 1, I_{k+1} = 1) = pq - \rho pq,$$

i.e. $p_{01} = p - \rho p$ and $p_{00} = q + \rho p$. Then the transition probability matrix takes the form

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} q + \rho p & p - \rho p \\ q - \rho q & p + \rho q \end{pmatrix}.$$

We can treat the coefficient of correlation as the parameter which reflects the degree of dependence of the random variables inducing Markov chain. The coefficient of correlation between any pairs of random variables I_j are equal $\rho_{k,k+h} = \rho^h$ [2]. For $\rho = 0$ we obtain the classical case of independence with

$$\mathbf{P} = \begin{pmatrix} 1 - p & p \\ 1 - p & p \end{pmatrix}.$$

Moreover for $\rho = 1$ we have strict dependence with the identity matrix **P**.

Now, we define random variable

$$X_n = I_1 + ... + I_n.$$

The distribution of such random variable is called a **Markov binomial distribution**. We can compute the values of the p.m.f. of random variable X_n using the recurrence method [2]. First, we derive the conditional probabilities

$$f_n(j|\, i) = P(X_n = j|\, I_0 = i),$$

where $j = 0, 1, ... , n$ and $i = 0, 1$, using the formulas

$$f_n(0 \mid 0) = p_{00}^n, \qquad f_n(0 \mid 1) = p_{10} p_{00}^{n-1},$$
$$f_n(n \mid 0) = p_{01} p_{11}^{n-1}, \qquad f_n(n \mid 1) = p_{11}^n,$$
$$f_n(j|\, 0) = p_{00} f_{n-1}(j|\, 0) + p_{01} f_{n-1}(j - 1|\, 1),$$
$$f_n(j|\, 1) = p_{10} f_{n-1}(j|\, 0) + p_{11} f_{n-1}(j - 1|\, 1),$$

for $j = 0, 1, ... , n$. The variance of the random variable X_n is equal

$$V(X_n) = npq + 2pq \frac{\rho}{1 - \rho} \left(n - 1 - \frac{\rho(1 - \rho^{n-1})}{1 - \rho} \right),$$

and p.g.f. is given by formula [10]:

$$G_n(t) = (q, pt) \begin{pmatrix} q + \rho p & (1 - \rho) pt \\ q - \rho q & (p + \rho q) t \end{pmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

If $np = \lambda > 0$ and n tends to infinity, then the Markov binomial distributed random variable X_n tends to the compound Poisson-Geometric distributed random variable X ([10], [2]). This random variable takes the form

$$X = \begin{cases} \sum_{j=1}^{N} Z_j & \text{for} \quad N > 0 \\ 0 & \text{for} \quad N = 0 \end{cases},$$

where $N \sim \text{Po}((1-\rho)\lambda)$ and $P(Z_j = k) = \rho^{k-1}(1-\rho)$ for $k = 1, 2, \ldots$.

Example 4. Let the dependent structure is described by Markov chain with $p = 0.4$ and ρ equal 0, 0.5, 0.7 and 0.8 for $n = 20$. The particular graphs of p.m.f. of random variable X are presented on fig. 3. We see that for bigger values of the coefficient ρ we do not obtain the unimodal graphs.

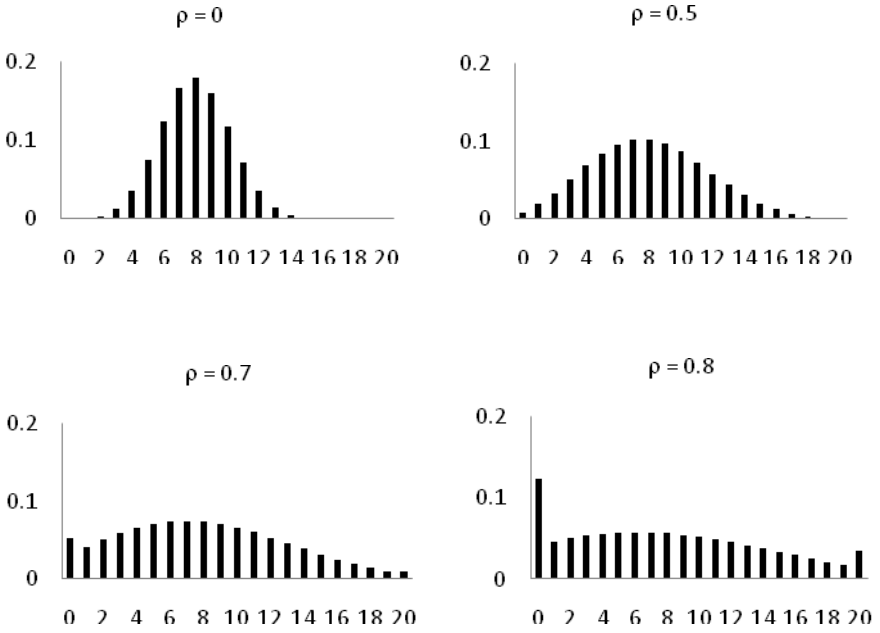


Fig. 3. Markov binomial distribution for different values of the coefficient of correlation ρ

Example 5. Let $n = 100$, $p = 0.1$ and $\rho = 0.6$. The graphs of p.m.f. of Markov binomial distribution and the limit compound Poisson-Geometric distribution are presented on the fig. 4. We can observe the good consistence of these graphs.

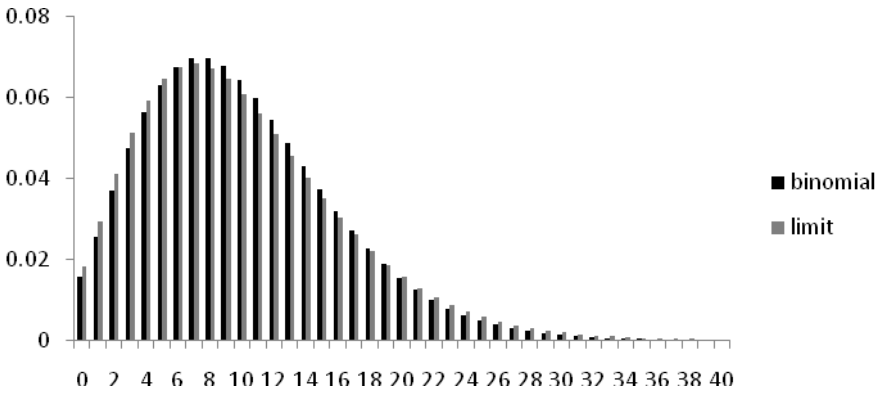


Fig. 4. Markov binomial distribution and the limit compound Poisson-Geometric distribution

8 Conclusion

The aim of this paper is presentation and investigation of the generalized binomial distribution when the assumption of independence is omitted. We study different dependence structures and the impact of the degree of dependence on the form of p.m.f. of the generalized binomial distribution. This is the starting point to further investigations, e.g. the ruin problem based on the dependent risk process.

References

1. Cossete H., Gaillardetz P., Marceau E., Rioux J.: On Two Dependent Individual Risk Models. Insurance: Mathematics and Economics 30, 153-166 (2002). ISSN 0167-6687
2. Cossete H., Landriault D., Marceau E.: Ruin Probabilities in the Compound Markov Binomial Model. Scand. Actuarial J. 4, 301-323 (2003). ISSN 0346-1238
3. Feller W.: An Introduction to Probability Theory and Its Applications, vol. 1, 3rd ed. Willey, New York (1968). ISBN 978-0471257080
4. Heilpern S.: Funkcje laczace. AE Wroclaw Press, Wroclaw (2007). ISBN 978-8370118778
5. Heilpern S.: Zalezny rozklad dwumianowy. Badania Operacyjne i Decyzje 1, 45-61 (2007). ISSN 1230-1868
6. Kolev N., Paiva D.: Multinomial Model for Random Sums. Insurance: Mathematics and Economics 37, 494-504 (2005). ISSN 0167-6687
7. McNeil A.J., Frey R., Embrechts P.: Quantitative Risk Management: Concepts, Techniques and Tools. Princeton Un. Press, Princeton (2005). ISBN 0-691-12255-5
8. Nelsen R.B.: An Introduction to Copulas. Springer, New York (1999). ISBN 0-387-986235
9. Tallis G.M.: The Use of Generalized multinomial Distribution in the Estimation of Correlation in Discrete Data. J. R. Stat. Soc., Ser. B 24, 530-534 (1962). ISSN 0035-9246

Wang Y.H.: On the Limit of the Markov Binomial Distribution. J. of Applied Probability 18, 937-942 (1981). ISSN 0021-90

Analysis of longitudinal data

Eva Jarošová

University of Economics, Prague, W. Churchill Sq. 4,
130 67 Prague 3, Czech Republic

Abstract. The paper deals with the evaluation of experimental data that are obtained through repeated measurement of the same specimens over time. The specimens are subjected to different experimental conditions and a linear model with mixed effects is used. Several methods of constructing confidence and prediction intervals based on the mixed effects model are compared together with the intervals obtained by means of the general linear model regardless of the true covariance structure of the data. For illustration, data representing the growth of yeast colonies under optimal and stress cultivation conditions are used.

Key words: experimental design, repeated measures, confidence and prediction intervals, Satterthwaite's method, Kenward-Roger's method

1 Introduction

Longitudinal data are a special case of repeated measures when observations are performed repeatedly on the same unit. Values of the corresponding response variable are not independent and the covariance structure of these data is more complex. Observations do not form a simple time series because there is more than one experimental unit in an experiment. Time series are structured by some type of the experimental design and usually effects of experimental factors on the response variable are examined. Commonly the mean response profile as a function of time is modeled, mean profiles under different experimental conditions are compared and future values based on some model are predicted. The mixed effects model is recommended for analysis. In this model both fixed parameters and parameters determining the data covariance structure are estimated. Beside conventional estimators which do not take account of uncertainty in estimating variability, various methods such as the Satterthwhite's method and the Kenward-Roger's method have been derived to improve statistical properties of estimators.

The aim of the paper is to examine properties of different estimators based both on a large and a small sample. The linear mixed effects model is applied to experimental microbiological data coming from the University of Chemical Technology [3]. Growth curves obtained after four days of cultivation of giant colonies of yeast are approximately linear and simplified approach would consist in comparison of the mean profiles under various experimental conditions using fitted coefficients of straight lines or applying a general linear model. Such analysis is not quite

appropriate due to the complex covariance structure. Differences between results based on the linear mixed effects model and those based on the general linear model not taking the true covariance structure into account are examined whereas various techniques of estimation are used in the mixed effects model. Although prediction of future values of the response variable would not be of interest in such microbiological example, prediction intervals are computed to show the major advantage of the mixed effects model. With respect to the intended examination of estimators' behavior both a large and a small data sample are used. The large sample includes all experimental data, while the small sample is formed by part of them.

2 Linear mixed effects model

The linear mixed effects model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ (Nx1) is a vector of responses, $\mathbf{X}$ (Nxp) is a design matrix linking $\boldsymbol{\beta}$ to $\mathbf{y}$, $\boldsymbol{\beta}$ (px1) is a vector of unknown parameters (fixed effects), $\mathbf{Z}$ (Nxq) is a design matrix linking $\mathbf{b}$ to $\mathbf{y}$, $\mathbf{b}$ (qx1) is a vector of unknown random effects and $\mathbf{e}$ (Nx1) is a vector of random errors. Assuming $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, and $\mathbf{b}$ and $\mathbf{e}$ independent, the mean profile is given by $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and the covariance structure is defined by means of matrices $\mathbf{G}$ and $\mathbf{R}$, namely $\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$. The elements of $\mathbf{V}$ are assumed to be functions of parameters $\boldsymbol{\theta}$ (hx1) that are estimated either by the maximum likelihood method or by the restricted maximum likelihood method. Logarithms of the corresponding likelihood functions are

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r} - \frac{N}{2}\log(2\pi), \tag{2}$$

$$l_R(\boldsymbol{\theta}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{r}^T\mathbf{V}^{-1}\mathbf{r} - \frac{N-p}{2}\log(2\pi) \tag{3}$$

where $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ and p is the rank of $\mathbf{X}$.

Solving the mixed model equations

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{4}$$

first the best linear unbiased estimator and predictor are obtained assuming that variance parameters are known and then the variance parameters are replaced with

their estimators. The resulting estimator (predictor) is called the empirical best linear unbiased estimator (predictor) and is expressed as (see e.g. [5], [6])

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y} \tag{5}$$

$$\tilde{\mathbf{b}} = \hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{6}$$

Typically, a linear combination $\boldsymbol{\lambda}^T\boldsymbol{\beta} + \boldsymbol{\delta}^T\mathbf{b}$ of the model's fixed and random effects is of interest.

3 Statistical properties

The covariance matrix of the fixed and random effects ($\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, $\tilde{\mathbf{b}} - \mathbf{b}$) is

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-}. \tag{7}$$

It is a common practice to obtain $\hat{\mathbf{C}}$ by putting $\hat{\mathbf{G}} = \mathbf{G}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{R}} = \mathbf{R}(\hat{\boldsymbol{\theta}})$. An estimator of $\mathbf{C}$ can be written in form (see e.g. [6])

$$\hat{\mathbf{C}} = \begin{bmatrix} \hat{\mathbf{C}}_{11} & \hat{\mathbf{C}}_{21}^T \\ \hat{\mathbf{C}}_{21} & \hat{\mathbf{C}}_{22} \end{bmatrix} \tag{8}$$

where

$$\hat{\mathbf{C}}_{11} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}$$
$$\hat{\mathbf{C}}_{21} = -\hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}$$
$$\hat{\mathbf{C}}_{22} = (\mathbf{Z}^T\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{G}}^{-1})^{-1} - \hat{\mathbf{C}}_{21}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{Z}\hat{\mathbf{G}}.$$

Because no account is made for uncertainty in estimating **G** and **R**, the true variability tends to be underestimated.

Conventionally, confidence limits for $\mathbf{k}^T\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{b}\end{bmatrix}$ are given by

$$\mathbf{k}^T\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \tilde{\mathbf{b}}\end{bmatrix} \pm t_{1-\alpha/2}(\nu)\sqrt{\mathbf{k}^T\hat{\mathbf{C}}\mathbf{k}} \tag{9}$$

where $\mathbf{k}$ ((p+q)x1) is an arbitrary vector, $t_{1-\alpha/2,\nu}$ is the $1-\alpha/2$ quantile of the t-distribution with ν degrees of freedom, $\nu = N - rank\,(\mathbf{X}\,\mathbf{Z})$, $\hat{\mathbf{C}} = \mathbf{C}(\hat{\boldsymbol{\theta}})$. Due to more than one random variables involved in model (1) $\mathbf{k}^T\hat{\mathbf{C}}\mathbf{k}$ includes estimators of at least two variance components and $t = \mathbf{k}^T\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \tilde{\mathbf{b}}\end{bmatrix}/\sqrt{\mathbf{k}^T\hat{\mathbf{C}}\mathbf{k}}$ does not have t-distribution any more.

One of the solutions is the Satterthwaite's approximation of the distribution of $\mathbf{k}^T\hat{\mathbf{C}}\mathbf{k} = M$. The variable $\frac{\nu M}{E(M)}$ with $\nu = 2\frac{E(M)^2}{V(M)}$ has approximately $\chi^2(\nu)$ distribution. Using a Taylor series expansion about $\boldsymbol{\theta}$ with only first-degree terms

$$M = \mathbf{k}^T\hat{\mathbf{C}}\mathbf{k} \cong \mathbf{k}^T\mathbf{C}\mathbf{k} + \sum_{j=1}^{h}(\hat{\theta}_j - \theta_j)\frac{\partial M}{\partial \theta_j}, \tag{10}$$

we can write

$$E(M) = \mathbf{k}^T\mathbf{C}\mathbf{k} \qquad V(M) = \mathbf{g}^T\,\mathrm{var}(\hat{\boldsymbol{\theta}})\mathbf{g} \tag{11}$$

where $\mathbf{g}$ is the gradient of $M = \mathbf{k}^T\hat{\mathbf{C}}\mathbf{k}$ with respect to $\boldsymbol{\theta}$. After the unknown parameters are replaced with their estimates $\hat{\boldsymbol{\theta}}$, the degrees of freedom of the t-quantile are adjusted according to

$$\nu = \frac{2(\mathbf{k}^T\hat{\mathbf{C}}\mathbf{k})^2}{\mathbf{g}^T\,\mathrm{var}(\hat{\boldsymbol{\theta}})\mathbf{g}}. \tag{12}$$

Another procedure consisting in applying an adjusted estimator of $\mathbf{C}$ was proposed by Kenward and Roger [2]. Based on the Taylor series expansion with second-order terms

$$\hat{\boldsymbol{\Phi}} \cong \boldsymbol{\Phi} + \sum_{j=1}^{h}(\hat{\theta}_j - \theta_j)\frac{\partial \boldsymbol{\Phi}}{\partial \theta_j} + \frac{1}{2}\sum_{i=1}^{h}\sum_{j=1}^{h}(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)\frac{\partial^2 \boldsymbol{\Phi}}{\partial \theta_i \partial \theta_j} \tag{13}$$

it can be seen that the estimator of $\boldsymbol{\Phi} = \mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ is biased because

$$E(\hat{\boldsymbol{\Phi}}) \cong \boldsymbol{\Phi} + \frac{1}{2}\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}\frac{\partial^2 \boldsymbol{\Phi}}{\partial \theta_i \partial \theta_j} \tag{14}$$

where $\mathbf{W} = \mathrm{var}(\hat{\boldsymbol{\theta}})$. Therefore the adjusted estimator of $\boldsymbol{\Phi}$ is

$$\hat{\mathbf{\Phi}}_A = \hat{\mathbf{\Phi}} - \frac{1}{2}\sum_{i=1}^{h}\sum_{j=1}^{h} W_{ij}\frac{\partial^2 \mathbf{\Phi}}{\partial\theta_i \partial\theta_j}. \tag{15}$$

4 Model building

Building a mixed model we must determine not only the form of its "fixed" part, i.e. the form of **X** but also its "random" part, i.e. the form of **Z**, **G** and **R**. Usually t-tests of F-test are used to test hypotheses about fixed effects under given covariance structure. The likelihood ratio test can be applied, too, but in this case maximum likelihood method (and not restricted maximum likelihood method) must be used. The latter is also suitable for testing hypotheses about null parameters of the covariance matrices on condition that fixed parts of the compared models are the same. Furthermore, models can be compared using information criteria.

The statistic

$$t = \frac{\mathbf{k}^T \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{pmatrix}}{\sqrt{\mathbf{k}^T \hat{\mathbf{C}} \mathbf{k}}} \tag{16}$$

has approximately t-distribution with $\nu = N - rank(\mathbf{X}\ \mathbf{Z})$ degrees of freedom. When underestimation of variability is to be taken into account, degrees of freedom are adjusted by means of the Satterthwaite's method.

The statistic

$$F = \frac{\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{pmatrix}^T \mathbf{L}^T (\mathbf{L}\hat{\mathbf{C}}\mathbf{L}^T)^{-1} \mathbf{L} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{pmatrix}}{J} \tag{17}$$

where **L** (Jx(p+q)) is arbitrary with rank J, has approximately F-distribution with J and ν degrees of freedom. To reach an appropriate F approximation a scaled form of F may be used and degrees of freedom are modified. Details are described in [2].

Suitable covariance models are chosen based on the information criteria AIC and BIC (the lower the values of AIC and BIC the better).

$$AIC = -2l + 2(p + h) \tag{18}$$

$$BIC = -2l + (p + h)\log N \tag{19}$$

where l is the (restricted) log-likelihood at convergence, p is the number of fixed parameters and h is the number of parameters in **G** and **R**.

5 Case study

Giant colonies of yeast were cultivated at 10 or 20 °C for 14 days. Different NaCl concentrations (0, 1, 2 %) were added to the media as a stress factor. Six specimens were observed under the same treatment conditions. The area of the colony was measured by the method of image analysis and the equivalent diameter was derived (the diameter of a circle having the same area as the colony). The camera and illuminating system used in the experiment did not enable to monitor the initial stages of colony growth due to a low contrast between the colony and the background. The first results were obtained after five days ($t_0 = 5$). From then on, most growth curves exhibited linear dependence of equivalent diameter on time (Fig. 1).

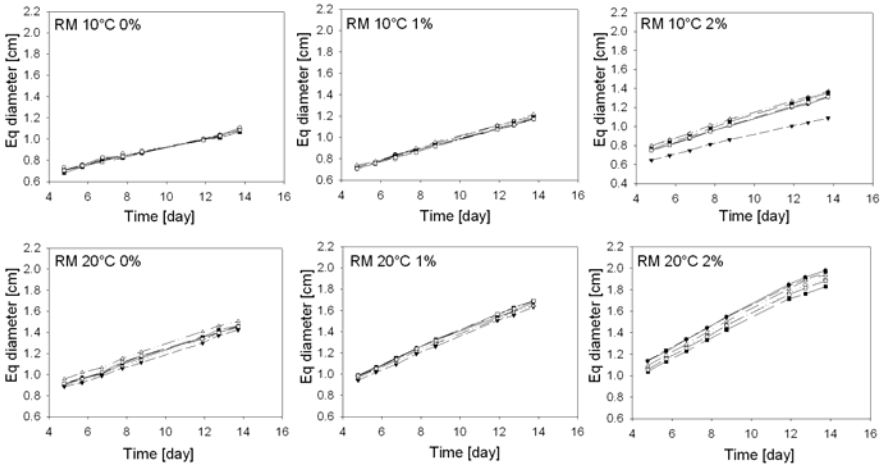


Fig. 1. Growth curves of giant colonies at 10 and 20 °C and 0, 1, 2 % NaCl concentration

The equivalent colony diameter was a response variable. With time taken as continuous and on the assumption that our measurements record the period of linear growth, the diameter of the i-th colony at the j-th level of temperature, the k-th level of NaCl concentration and time t_l can be expressed in the form

$$y_{ijkl} = \beta_{0,jk} + b_{0,i(j,k)} + (\beta_{1,jk} + b_{1,i(j,k)})(t_l - t_0) + e_{ijkl} \qquad \textbf{(20)}$$

where $\beta_{0,jk}$ corresponds to the mean diameter at $t = t_0$, $\beta_{1,jk}$ denotes the mean growth rate (in cm per day) in the "linear" period, random effects $b_{0,i(jk)}$ represent variation of line intercepts around $\beta_{0,jk}$ and random effects $b_{1,i(jk)}$ correspond to variation of line slopes around $\beta_{1,jk}$. Random errors e_{ijkl} denote departures of observations from the model. For $t \geq t_0$ the mean profile at T_j and $NaCl_k$ has the form

$$E(Y(t) \mid T_j, NaCl_k) = \beta_{0,jk} + \beta_{1,jk}(t - t_0) \quad (21)$$

The aim of this case study is to compare statistical properties of estimators and predictors based on linear mixed effects model (LME) using various methods of accounting for the uncertainty in variance estimation and to present differences between results obtained by means of LME and those based on a general linear model (GLM) where the true covariance structure is not taken into account. Confidence and prediction intervals are constructed using various methods. Two alternatives are considered:

a) Model based on observations from all six groups determined by different experimental conditions (large sample, 48 specimens, 288 observations).
b) Model based on observations corresponding only to one of the six groups (small sample, 6 specimens, 48 observations). In this model time *t* is a single explanatory variable. The observations obtained at 20°C and 2 % NaCl have been chosen.

As for LME, confidence intervals for the mean diameter in the chosen group (20°C and 2 % NaCl) at various times, i.e. for $\mathbf{x}^T\boldsymbol{\beta}$, and prediction intervals for the diameter of a chosen specimen belonging to this group, i.e. for $\mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\mathbf{b}$ are constructed. GLM considered is expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (22)$$

where $\mathbf{X}$ is the same matrix as in the mixed-effect alternative and $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. It should be noted that unlike with LME, prediction intervals based on GLM apply for an arbitrary specimen under given experimental condition.

Two statistical packages were used. Fixed and random effects were estimated both in SPlus and in SAS (with identical results). The output of SPlus procedure *lme* was chosen to illustrate the model fitting because of its better transparency. Estimates of the mean profiles' parameters and their confidence intervals were computed based on the estimates of the fixed effects and their covariance matrix obtained in SPlus. SAS was used to compute all confidence and prediction limits because SPlus does not offer required procedures.

6 Results

a) Model based on all observations

Both parameters $\beta_{0,jk}$ and $\beta_{1,jk}$ were supposed to be affected by a treatment (note that $\beta_{0,jk}$ does not represent the intercept at t = 0, but at t = 5); the indices j and k stand for various treatment conditions. Columns of $\mathbf{X}$ in Eq. (1) correspond to the fixed part of the model (in notation used in S Plus)

$$T_j + NaCl_k + T_j * NaCl_k + time + (T_j + NaCl_k + T_j * NaCl_k) * time \qquad (23)$$

where T is a two-level factor (j = 1, 2) and $NaCl$ is a three-level factor (k = 1, 2, 3). *Time* is a continuous variable.

Variances $\sigma^2_{0,jk}$ and $\sigma^2_{1,jk}$ of random effects $b_{0,i(jk)}$ and $b_{1,i(jk)}$, respectively, lying on the main diagonal of **G** may or may not differ for various treatments. In dependence on whether random effects $b_{0,i(jk)}$ and $b_{1,i(jk)}$ are or are not dependent, matrix **G** is or is not diagonal. Various forms of **G** were considered. Differences among $\sigma_{0,jk}$ (standard deviations of random effects $b_{0,i(jk)}$) under different treatment conditions were apparent. One of the curves at 10°C and 2 % NaCl which differed substantially from the others was omitted to make an analysis more transparent. Then the systematic effect of T on $\sigma_{0,jk}$ was clearly distinguishable. Non-parallel curves corresponding to non-zero $\sigma_{1,jk}$ were evident at 20°C and 2 % NaCl but after fitting the model, the value of $\hat{\sigma}_{1,jk}$ were practically zero. Therefore only two parameters corresponding to two different temperatures were included. The variance matrix of random effects had the form $\mathbf{G} = diag\left\{\sigma^2_{0,1}, \sigma^2_{0,2}\right\}$ with $\sigma_{0,1} = 0.0135$ and $\sigma_{0,2} = 0.0326$.

The expected autocorrelation of repeated measurements corresponding to the same specimen was represented by AR(1) scheme of the matrix **R** with autocorrelation coefficient $\phi = 0.6375$ and residual standard deviation $\sigma = 0{,}0171$.

The estimates of the fixed effects of the model are displayed in Table 1. According to the chosen coding of categorical factors the fixed effect denoted by NaCl2 (NaCl3) represents an increase of the mean response corresponding to the change of NaCl from 0% to 1% (from 0% to 2%) etc. Using these estimates and their approximate covariance matrix (not displayed) estimates of $\beta_{0,jk}$ and $\beta_{1,jk}$ including conventional confidence limits were computed (Table 2).

Table 1. S Plus, procedure lme, table of estimated fixed effects (treatment contrasts). large sample

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.7092815	0.00823268	239	86.15442	<.0001
NaCl2	0.0072617	0.01164276	29	0.62371	0.5377
NaCl3	0.0657890	0.01221103	29	5.38767	<.0001
temp2	0.2001370	0.01681874	29	11.89964	<.0001
time	0.0413214	0.00099064	239	41.71166	<.0001
NaCl2temp2	0.0630001	0.02378529	29	2.64870	0.0129
NaCl3temp2	0.1221470	0.02406856	29	5.07496	<.0001
time:temp2	0.0199723	0.00140098	239	14.25593	<.0001
NaCl2time	0.0110260	0.00140098	239	7.87022	<.0001
NaCl3time	0.0218462	0.00146936	239	14.86783	<.0001
NaCl2temp2time	0.0045004	0.00198129	239	2.27145	0.0240
NaCl3temp2time	0.0100141	0.00203022	239	4.93253	<.0001

Considering the same initial mean diameter regardless of the treatment as a reasonable assumption, differences among the estimates $\hat{\beta}_0$ (Table 2) reflected different mean rates under various treatment conditions in the previous growth period. In addition to the effect of temperature a positive effect of higher NaCl concentration was observed both at lower and higher temperature. The effect was more distinguishable at the higher temperature. Effects of temperature and NaCl concentration on $\hat{\beta}_1$ were similar (Table 2).

Table 2. Estimated parameters of the mean structure model with conventional 95% confidence limits

T [°C]	NaCl[%]	$\hat{\beta}_0$	lcl	ucl	$\hat{\beta}_1$	lcl	ucl
10	0	0.7093	0.6931	0.7255	0.0413	0.0394	0.0433
10	1	0.7165	0.7003	0.7328	0.0523	0.0504	0.0543
10	2	0.7751	0.7573	0.7928	0.0632	0.0610	0.0653
20	0	0.9094	0.8805	0.9383	0.0613	0.0593	0.0632
20	1	0.9797	0.9508	1.0086	0.0768	0.0749	0.0788
20	2	1.0974	1.0685	1.1262	0.0932	0.0912	0.0951

The estimated mean profile in the group with 20°C and 2 % NaCl is expressed as

$$\hat{E}(Y(t) \mid T_2, NaCl_3) = 1.0974 + 0.0932(t-5) \tag{24}$$

and its values for t = 5, ..., 14 (denoted pred) can be seen in Table 3. The random effects corresponding to one of the specimens in this group are $b_{0,1} = 0$, $b_{0,2} = 0.0492$, and the predicted growth curve of this specimen is given by

$$\tilde{Y}(t) \mid T_2, NaCl_3 = 1.1465 + 0.0932(t-5). \tag{25}$$

Its values for t = 5, ..., 14 (denoted BLUP) can be seen in Table 4. Fitted values of GLM (denoted y_hat) are shown in Table 3. Corresponding confidence limits for the mean (lclm,uclm) are displayed in Table 3, prediction limits (lclp, uclp) are in Table 4. Conventional, Satterthwaite-type and Kenward-Roger-type confidence limits are practically the same (Table 3). Confidence limits based on GLM are nearer than those based on LME so that the true confidence level must be lower than stated 95 %.

Table 3. Large sample, 95% confidence limits for means based on LME and GLM

time	pred	lcl_c	ucl_c	lcl_s	ucl_s	lcl_kr	ucl_kr	y_hat	lclm	uclm
5	1.097	1.068	1.126	1.066	1.128	1.067	1.128	1.101	1.087	1.115
6	1.190	1.162	1.219	1.160	1.221	1.160	1.221	1.195	1.183	1.207
7	1.284	1.256	1.312	1.253	1.314	1.254	1.314	1.289	1.279	1.299
8	1.377	1.349	1.405	1.347	1.407	1.347	1.407	1.383	1.374	1.391
9	1.470	1.442	1.498	1.440	1.500	1.440	1.500	1.477	1.468	1.485
12	1.749	1.721	1.778	1.719	1.780	1.719	1.780	1.758	1.747	1.769
13	1.843	1.814	1.871	1.812	1.873	1.812	1.873	1.852	1.840	1.865
14	1.936	1.907	1.965	1.905	1.967	1.905	1.967	1.946	1.931	1.961

Table 4. Large sample, 95% prediction limits based on LME and GLM

time	y	BLUP	lcl_c	ucl_c	lcl_s	ucl_s	lcl_kr	ucl_kr	lclp	uclp
5	1.135	1.147	1.125	1.168	1.123	1.170	1.123	1.170	1.043	1.159
6	1.223	1.240	1.219	1.261	1.217	1.262	1.217	1.262	1.137	1.253
7	1.337	1.333	1.312	1.354	1.311	1.355	1.311	1.355	1.231	1.346
8	1.444	1.426	1.406	1.446	1.404	1.448	1.404	1.448	1.325	1.440
9	1.549	1.519	1.499	1.539	1.498	1.541	1.498	1.541	1.419	1.534
12	1.851	1.799	1.778	1.820	1.777	1.821	1.776	1.821	1.701	1.816
13	1.916	1.892	1.870	1.913	1.869	1.915	1.869	1.915	1.794	1.910
14	1.978	1.985	1.963	2.007	1.962	2.008	1.961	2.009	1.888	2.005

Fig. 2 shows the clear superiority of BLUP as compared with Fig. 3. All three types of confidence limits based on the mixed effect model are almost identical (Table 4). Only conventional confidence limits are displayed in Fig. 2. Not only is the width of the prediction interval from GLM more than twice as long as the width of the interval from LME but the true confidence level must be lower judging by the preceding case.

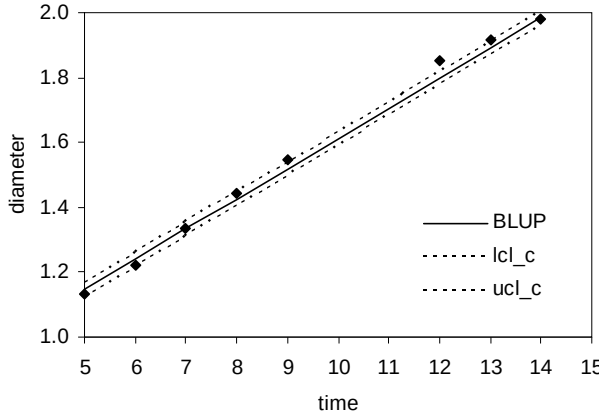


Fig. 2. Large sample, BLUP and prediction limits based on LME

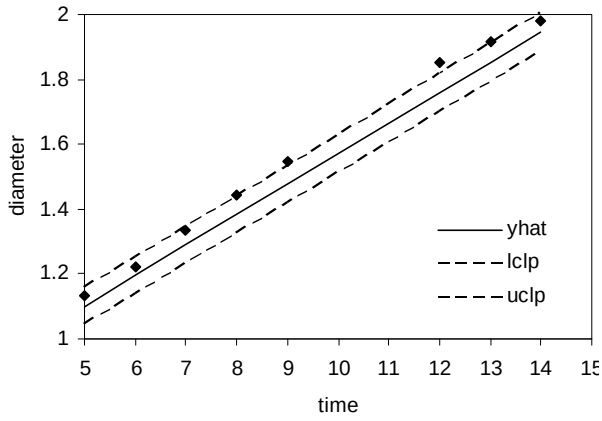


Fig. 3. Large sample, fitted line and prediction limits based on GLM

b) Model based on one group

The variance matrix of random effects was reduced to a scalar with $\sigma_0 = 0.0374$. The matrix **R** corresponded to AR(1) scheme with autocorrelation coefficient $\phi = 0.8598$ and residual standard deviation $\sigma = 0.0370$. The estimates of the fixed effects of the model are displayed in Table 5.

Table 5. S Plus, procedure lme, table of estimated fixed effects (treatment contrasts), small sample

```
Fixed effects: d ~ time
               Value  Std.Error DF  t-value p-value
(Intercept) 1.089511 0.02102933 41 51.80911  <.0001
       time 0.093409 0.00175594 41 53.19611  <.0001
```

The mean profile of this group is only slightly different from the previous one based on all the data

$$\hat{E}(Y(t) \mid T_2, NaCl_3) = 1.0895 + 0.0934(t-5) \tag{26}$$

its values for t = 5, ..., 14 (denoted pred) can be seen in Table 6. Random effect corresponding to the same specimen as before is $b_0 = 0.0334$, and the predicted growth curve of this specimen is expressed as

$$\tilde{Y}(t) \mid T_2, NaCl_3 = 1.1229 + 0.0934(t-5)\ . \tag{27}$$

Its values for t = 5, ..., 14 (denoted BLUP) are in Table 7.

Table 6. Small sample, 95% confidence limits for the means based on LME and GLM

time	y	pred	lcl_c	ucl_c	lcl_s	ucl_s	lcl_kr	ucl_kr	lclm	uclm
5	1.135	1.089	1.047	1.132	1.039	1.140	1.040	1.139	1.077	1.125
6	1.223	1.183	1.141	1.224	1.133	1.233	1.136	1.230	1.174	1.215
7	1.337	1.276	1.236	1.317	1.226	1.326	1.231	1.322	1.271	1.306
8	1.444	1.370	1.330	1.410	1.319	1.420	1.325	1.414	1.367	1.398
9	1.549	1.463	1.424	1.503	1.413	1.513	1.419	1.507	1.462	1.491
12	1.851	1.743	1.703	1.784	1.693	1.794	1.697	1.789	1.740	1.777
13	1.916	1.837	1.795	1.878	1.787	1.887	1.789	1.885	1.830	1.874
14	1.978	1.930	1.887	1.973	1.880	1.981	1.880	1.980	1.921	1.972

Again, the three types of confidence intervals based on LME are very similar (Table 6). As in case of the large sample, the width of the confidence intervals for the mean based on GLM is shorter than that of the confidence intervals based on LME indicating their lower confidence level.

Table 7. Small sample, 95% prediction limits based on LME and GLM

time	y	BLUP	lcl_c	ucl_c	lcl_s	ucl_s	lcl_kr	ucl_kr	lclp	uclp
5	1.135	1.123	1.071	1.175	1.039	1.207	0.971	1.275	1.000	1.202
6	1.223	1.216	1.165	1.267	1.132	1.301	1.062	1.370	1.095	1.295
7	1.337	1.310	1.259	1.360	1.225	1.395	1.154	1.465	1.189	1.388
8	1.444	1.403	1.353	1.453	1.318	1.489	1.247	1.560	1.283	1.482
9	1.549	1.497	1.447	1.546	1.411	1.582	1.340	1.653	1.377	1.576
12	1.851	1.777	1.726	1.827	1.692	1.861	1.622	1.931	1.658	1.858
13	1.916	1.870	1.819	1.922	1.786	1.954	1.718	2.023	1.752	1.953
14	1.978	1.964	1.911	2.016	1.880	2.047	1.813	2.114	1.845	2.048

Different types of prediction limits based on LME are evidently distinguishable (Fig. 4). The width of Satterthwhite-type intervals is more than 1.5 times longer than the width of the conventional intervals and the Kenward-Roger-type intervals are almost 3 times longer. Comparison with GLM is not so unambiguous as in preceding cases. The conventional prediction intervals are shorter than the prediction intervals based on GLM, the opposite is true for the Kenward-Roger-type intervals. Satterthwhite-type intervals and GLM prediction intervals are very similar here (Table 7).

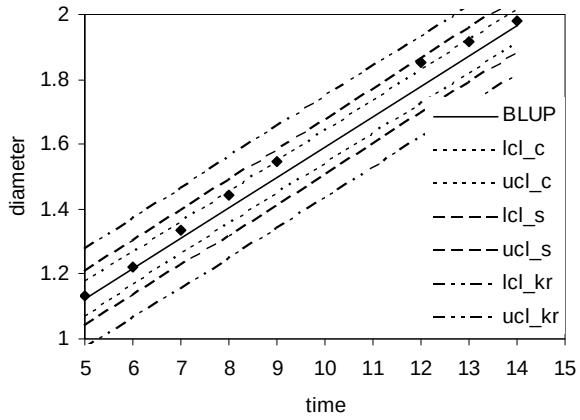


Fig. 4. Comparison of three types of prediction limits based on LME

7 Discussion

Although the presented case study comes from microbiology, mixed effects model may be appropriate in many other areas where data consist of a collection of

independent sets, for example in predicting the wear of components in quality control or for improving efficiency of estimators in small administrative areas in economical surveys. By using information from other small data sets estimators obtained by means of the mixed effects model are more efficient compared to those using only an individual data set.

Procedures for the mixed effects model are implemented in major statistical packages such as SPlus or SAS. In S Plus only approximate confidence intervals and tests, i.e. methods using the conventional estimate of the covariance matrix, are used. SAS offers all methods applied in the paper. Results of the case study confirmed that asymptotic behaviour of all these methods is similar and adjustments are not needed in samples containing a large number of items (or sets). Performing of these techniques in smaller samples is questionable. For example the simulation study in [1] shows that in dependence on relative size of variance components the bias of the Kenward-Roger-type estimator may be quite large. This might explain relatively wide prediction intervals in the second part of this case study.

Another question is in what situations in practice it is worth to use the linear mixed effects model as opposed to a simpler procedure in GLM. LME is recommended for files with unbalanced data. Although our results based on balanced data are more likely in favour of the mixed effects model, the situation may differ when the correlation structure is simpler.

References

1. Harville, D.A., Jeske, D.R.: Mean Squared Error of Estimation or Prediction Under a General Linear Model. JASA 87, 724--731 (1992)
2. Kenward, M.G., Roger, J.H.: Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. Biometrics 53, 983--997 (1997)
3. Krulikovská, T.: Srovnání metod chemické analýzy s metodami analýzy obrazu při studiu populací eukaryotních buněk (Comparison of chemical analysis methods with image analysis during the research of eukaryotic cells populations). Dissertation Thesis, VŠCHT Praha (2009)
4. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in S and S-PLUS. Springer, New York (2000). ISBN 0-387-98957-9.
5. Prasad, N.G.N, Rao, J.N.K.: The Estimation of the Mean Squared Error of Small-Area Estimators. JASA 85, 163--171 (1990)
6. SAS Institute Inc., SAS/STAT: User's Guide, Version 9.1. SAS Institute Inc., Cary, NC (2003)

Unemployment in the Czech Republic and its predictions based on the Box-Jenkins methodology

Věra Jeřábková

Department of Economic statistics, University of Economics Prague
W. Churchill sq. 4, Prague 3
vera.jerabkova@vse.cz

Abstract. Unemployment is one of the main economic indicators which show us the economic situation in a country. Unemployment rate has intensively increased since 1996 - mainly because of economic, demographic and social aspects. The peak was reached at the end of 2003. Since 2004 the Czech Republic has become a full member of the European Union, which means a revival of the labour market in conjunction with the decrease in unemployment. But in these days there is a financial crisis that is increasing joblessness. Thanks to time series analysis, predicting values, especially based on the Box-Jenkins methodology, and making sense of reasons of existing unemployment, we can take measures to eliminate its negative aspects in the economy.

Key words. Unemployment rate, Box-Jenkins methodology

1 Introduction

Due to transformation from a centrally planned system to a market economy during the 90s of 20th century, the Czech Republic started to face up the unemployment again after more than 40 years. The socialist system provided the citizens of certainty and stability of the job through strong regulation of labour-law provisions by which the system ensured the full employment. However, that full employment was characterized by a high overemployment which was associated with a low productivity of workers and inefficient management. The work was created artificially (e.g. finished goods in factories were remelted) and the central regulation of wages kept them at a very low level. That caused a technical and technological gap.

Early 90‘s of the 20th century, the transformation of political, economic and social reforms began to ensure an effective functioning of the market economy and opening of the labour market. This process, however, meant the “revival” of the unemployment. Structural changes (decline), especially in areas of heavy industry (mining, heavy engineering), significantly affected districts in which the economic interconnection to the sector was particularly strong. The structure of employment, sector classification of the region and demographic factors together with the low mobility of workers, generous social benefits and high tax burden on labour caused the unemployment rate exceeded 10 % in 2003. The accession of the Czech Republic

into the European Union in 2004 meant not only the revitalization of the economy, the inflow of new investment and the creation of new jobs, but also the adoption of national action plans of the employment. The new Act No. 435/2004 Coll., About Employment, made more strict the rules for the unemployment benefit and the length of time maintenance, and focused more on the active employment policy, the main aim of which it is to ensure the balance of the labour market and facilitate the access of the unemployed into the labour market. Due to these factors, the unemployment rate recorded a decline. However, in the second half of 2008 a financial crisis erupted in the U.S.A. which subsequently spread to Europe. The economic downturn, which affected particularly the export business and engineering (metallurgy, metal production, machinery production), has resulted in significant staff redundancies and reduced the cost of staff (reducing overtime, the rewards and benefits, the collection of forced leave, change of working time), which caused a re-growth in the unemployment rate.

Monitoring of the development in unemployment rate is a highly current social phenomenon because of it being considered a sign of economic health. By analyzing time series we can identify some regularity in its behaviour that will enable us to predict future developments. Due to this we can take measures which will lead to the reduction of negative consequences of the unemployment.

2 Objective and methodology

By analyzing the development of the unemployment in terms of its time and spatial distribution, structure and sector, which contribute at the most to the labour market, together with the detection of a period and estimated future projections we can take such measures in the economic and social fields that will lead to a reduction or at least a stabilisation of unemployment.

The modeling of seasonal time series is based on Box-Jenkins methodology and was conducted for the period between January 2000 and March 2008 (the period before the financial crisis) and for the period between January 2000 and March 2009, due to the impact of financial crisis on the unemployment. Software Give Win 2.30, esp. PCGive and X12ARIMA modules, was used for total analysis of 15 time series of the unemployment rate (14 regions and the Czech Republic). The quality of estimated models have been verified on the basis of diagnostic tests and under its short verification, i.e. comparing estimates with actual values for the period from April to June 2008 and from April to June 2009. Due to this we will explore if SARIMA models based on the Box-Jenkins methodology are suitable in term of sudden changes in economy.

3 Results

Labour market is affected not only by the economic conditions and social policy of the state, but also by the demographic structure, education, skills and health of the

population. In the analysis of the unemployment in the Czech Republic we should consider specific characteristics of each region and its economic potential. Higher regional differentiation in the unemployment rate was showed particularly in the second half of the 90s, when in the context of the restructuring a significant downturn of traditional industries (heavy industry, mining, steel...) took place. Some areas faced dramatic increase in the unemployment rate, while others reached a low level (e.g., October 1997 Most 10.6 %, Prague 0.82 %, Czech Republic 4.8 %).

3.1 Regional differences

Long-term lowest rate is in Hlavní město Praha region where the value does not exceed 5%. The regions, where the unemployment rate is below the state average, in addition to Hlavní město Praha are Jihočeský, Středočeský, Plzeňský and Královéhradecký. About the average range the Liberecký, Vysočina, Karlovarský, Pardubický, Zlínský and Jihomoravský regions. The highest rate of unemployment is in Ústecký region together with Moravskoslezský and Olomoucký. High unemployment rates are caused by restructuring (decline) in the field of heavy industry and metallurgy in Ústecký and Moravskoslezský regions.

From the labour market's viewpoint the most involved is the tertiary sector in almost all regions (except Vysočina, Liberecký and Zlínský regions in which secondary sector is prevailing). In terms of future developments it is assumed that a dramatic decrease of workers in the primary sector will no longer occur. On the other hand, the secondary sector will intensify pressure to reduce costs, optimize production processes, increase productivity and reduce jobs. This may cause the rise of unemployment in Vysočina, Liberecký and Zlínský regions.

In terms of sector of employed in national economy the most jobs are involved in the manufacturing sector, which contributes nearly 30% to the total employment in the Czech Republic. It has also the highest share in all regions, except Hlavní město Praha. Agricultural sector records the continuous decline in employment (especially Vysočina) as well as extraction of mineral resources. The financial crisis is affecting particularly the export business and engineering.

Regional characteristics of education and age structure of the unemployed in 2008 are in Table 1. It is obvious that the highest proportion of unemployed by education is almost the same in all regions. People with secondary education without "A"levels contribute to the total unemployment at intervals 41-51 %. On the other hand the age structure is different in regions. The unemployed in the age interval of 55-59 years contribute by 16-18 % to the total employment only in Hlavní město Praha and Královéhradecký regions. The most part of the unemployed in the youngest age interval of 20-24 years are in Vysočina and Zlínský regions (17 % of the total unemployment). They can be low qualification people and graduates who do not have the necessary professional knowledge and practical skills, work habits, or are not willing to perform certain jobs, contribute to the unemployed at the most. Continuous decline of the unemployed is noted in the age interval of 15-19 years for reason of the extending possibilities of the education.

Table 1. The regional characteristics - the highest proportion of the unemployed by education and age in 2008

Region	Education		Age	
	Portion of the total unemployment (%)	Type	Portion of the total unemployment (%)	Age interval
Hlavní město Praha	35	secondary	16	55-59
Středočeský	43	secondary without "A"levels	16	30-34
Jihočeský	47	secondary without "A"levels	16	35-39
Plzeňský	41	secondary without "A"levels	18	30-34
Karlovarský	46	primary	16	30-34
Ústecký	47	primary	16	45-49
Liberecký	41	secondary without "A"levels	20	30-34
Královéhradecký	43	secondary without "A"levels	18	55-59
Pardubický	38	secondary without "A"levels	18	50-54
Vysočina	47	secondary without "A"levels	17	20-24
Jihomoravský	43	secondary without "A"levels	17	30-34
Olomoucký	41	secondary without "A"levels	15	35-39
Zlínský	51	secondary without "A"levels	17	20-24
Moravskoslezský	42	secondary without "A"levels	14	30-34

3.2 SARIMA models

For the prediction of the development in the unemployment rate in the Czech Republic and its regions estimates of the SARIMA models have been used - for the period of January 2000 to March 2008 (the period before the financial crisis). Subsequently the SARIMA models were estimated for the period of January 2000 to March 2009. Due to this we can compare the predictions and find out if estimated models based on Box-Jenkins methodology are suitable for the unemployment forecasting in case of sudden changes in economy, especially at the time of financial crisis, or not.

The estimated models are represented by both seasonal and non-seasonal moving average processes and autoregressive processes. From the viewpoint of difference for most regions, except Hlavní město Praha, non-seasonal differences are included in the models. The resulting estimated models are shown in Table 2. While SARIMA models on the bases of the first period are estimated with no problems, the sudden changes in the development of the unemployment make more difficult to find a suitable SARIMA model for the second period. All mentioned SARIMA models in Table 2 meet the criteria of diagnostic tests, i.e. homoscedasticity and non-autocorrelation of the nonsystematic component, but no normality.

Table 2. Estimated SARIMA models for regions and the Czech Republic

Region	SARIMA model – period January 2000 - March 2008	SARIMA model – period January 2000 - March 2009
Prague	SARIMA(1,0,0)(0,1,1)12	SARIMA(2,0,0)(1,1,0)12
Středočeský	SARIMA(1,1,0)(1,0,1)12	SARIMA(0,1,1)(1,0,0)12
Jihočeský	SARIMA(1,1,0)(1,0,1)12	SARIMA(0,2,1)(1,0,0)12
Plzeňský	SARIMA(0,1,1)(2,0,0)12	SARIMA(1,1,1)(2,0,0)12
Karlovarský	SARIMA(0,1,2)(2,0,0)12	SARIMA(0,2,1)(1,0,1)12
Ústecký	SARIMA(1,1,0)(2,0,0)12	SARIMA(2,1,0)(1,0,1)12
Liberecký	SARIMA(1,1,1)(1,0,1)12	SARIMA(1,1,1)(1,0,0)12
Královéhradecký	SARIMA(1,1,0)(2,0,0)12	SARIMA(1,1,0)(1,0,0)12
Pardubický	SARIMA(0,1,1)(2,0,0)12	SARIMA(0,2,1)(1,0,0)12
Vysočina	SARIMA(1,1,0)(2,0,0)12	SARIMA(0,2,0)(1,0,0)12
Jihomoravský	SARIMA(1,1,1)(1,0,1)12	SARIMA(1,1,0)(1,0,0)12
Olomoucký	SARIMA(0,1,2)(1,0,0)12	SARIMA(0,2,1)(1,0,0)12
Zlínský	SARIMA(1,1,1)(1,0,0)12	SARIMA(1,1,0)(1,0,0)12
Moravskoslezský	SARIMA(1,1,0)(1,0,0)12	SARIMA(1,2,0)(2,0,1)12
Czech republic	SARIMA(1,1,0)(1,0,0)12	SARIMA(1,2,0)(1,0,1)12

For verification of the estimated SARIMA models, the predictions of the unemployment rate in regions were compared with their actual values for the period from April to June 2008, and thereafter for the period from April to June 2009. The resulting values of the actual unemployment rate and forecasts (including interval forecasts at 5% level of significance) are given in Annex 1. It can be noted that the SARIMA models relating to the period before the financial crisis appear to be good. The predictions are completely consistent with the actual values or ranges of 95% confidence interval.

However, we obtain other results for the SARIMA models relatively to the period between January 2000 and March 2009. In the case of Jihočeský, Pardubický, Vysočina and Olomoucký regions, the estimated SARIMA models are not suitable at all. They seem to be useful only in the Středočeský, Ústecký, Liberecký, Královéhradecký and Jihomoravský regions. But the follow-up comparing of forecasts and actual values in July 2009 shows us that the above mentioned SARIMA models are suitable only in Středočeský, Ústecký and Královéhradecký regions. So we find out that SARIMA models based on Box-Jenkins methodology are not suitable for unemployment forecasting in case of sudden changes in economy, especially at the time of financial crisis.

For illustration the table 3 shows the expected values of the unemployment rate in December 2009 based on the Box-Jenkins methodology (both periods). While the forecast of the unemployment rate is 3,6 % for the Czech Republic in December 2009 based on the first period, the "newer" forecast is almost triple (9,7 %). In all regions the estimated rates relating to the second period are almost double (except Plzeňský a Královéhradecký regions, where forecasts are triple). But as mentioned above the second estimated SARIMA models are not useful. So that the "newer" forecasts are not correct.

Table 3. Forecasts of unemployment rates (%) for December 2009 on the basis of SARIMA models

Region	XII.09 (period I.2000-III.2008)	XII.09 (period I.2000-III.2009)
Hlavní město Praha	1,7	3,1
Středočeský	3,3	6,1
Jihočeský	3,7	-
Plzeňský	3	7,8
Karlovarský	5,2	11,5
Ústecký	6,7	13,4
Liberecký	5,8	11,5
Královéhradecký	3	8,1
Pardubický	3,9	-
Vysočina	4	-
Jihomoravský	5,5	9,5
Olomoucký	4,4	-
Zlínský	4,3	9,7
Moravskoslezský	5,8	13,1
Czech Republic	3,6	9,7

4 Conclusion

Although unemployment is a natural phenomenon in a market economy, its high degree causes economic and social problems. People are not only excluded from the main stream of society, but also suffer from psychological problems and are at risk of poverty. But in particular, they are lagging behind their qualifications. From an economic perspective, state costs are rising due to the preservation, social peace (income support, social benefits…) and expenditures on active employment policy, especially in time of financial crisis, when many companies reduce the production and dismiss their employees.

By analyzing the development of the unemployment rate in terms of its time and spatial distribution, structure of the joblessness and sectors, which have the majority on the labour market together with an estimate of future predictions by using SARIMA models for the period before the financial crisis, we saw that rising unemployment threatened Liberecký, Zlínský and Vysočina regions.

Almost in all regions (except Hlavní město Praha, Karlovarský and Ústecký regions), persons with secondary education without „A“ levels participated at the most in the level of unemployment in 2008. But the age structure of the unemployed was different among regions.

While the estimated SARIMA models for the period of January 2000 to March 2008 seemed to be suitable for all regions, the estimated SARIMA models for the

period of January 2000 to March 2009 were not so good. In terms of that second estimates, the unemployment rate will reach the 10% limit in the December 2009 in Karlovarský, Ústecký, Liberecký and Moravskoslezský regions. The forecast for the unemployment rate in the Czech Republic at the end of the year 2009 will be also about 10%. Nevertheless we found out that SARIMA models based on the Box-Jenkins methodology are not suitable for unemployment forecasting in case of sudden changes in economy, especially at the time of financial crisis. Due to this we can not take measures which will lead to the reduction of negative consequences of the unemployment only on the basis of SARIMA forecasts. If the financial crisis did not appear, the unemployment rate would be probably estimated to be 3.6 % in the Czech Republic. The way how to make suitable SARIMA models which would incorporate the financial crisis and work for forecasts as well should be a theme of other paper.

References

1. Arlt, J., Arltová, M.: Ekonomické časové řady. Grada Publishing, Praha (2007).
2. Arlt, J., Arltová, M., Rublíková, E.: Analýza ekonomických časových řad s příklady. Vysoká škola ekonomická, Praha (2002).
3. Buchtová, B. a kol.: Nezaměstnanost – psychologický, ekonomický a sociální problém. Grada Publishing, Praha (2002).
4. Kotýnková, M.: Trh práce na přelomu tisíciletí. Oeconomica, Praha (2006).
5. Mareš, P.: Nezaměstnanost jako sociální problém. Perfekt, Praha (1997).
6. Czech Statistical Office, http://www.czso.cz

Appendix 1: Actual values and forecasts of the unemployment rate in the period IV.-VI.2008 and IV.-VI.2009

Region	Period	Actual value	Forecast	Lower forecast	Upper forecast	Region	Period	Actual value	Forecast	Lower forecast	Upper forecast
Hlavní město Praha	IV.08	2,0	2,0	1,9	2,2	Pardubický	IV.08	4,6	4,5	4,2	4,8
	V.08	2,0	2,0	1,7	2,2		V.08	4,5	4,1	3,6	4,6
	VI.08	2,0	1,9	1,6	2,2		VI.08	4,5	4,0	3,3	4,7
	IV.09	2,7	2,6	2,4	2,8		IV.09	7,8	8,0	7,7	8,4
	V.09	2,8	2,5	2,2	2,9		V.09	7,7	8,5	7,9	9,1
	VI.09	3,0	2,5	2,1	2,9		VI.09	7,8	9,1	8,2	10,0
Středočeský	IV.08	3,7	3,6	3,4	3,9	Vysočina	IV.08	4,8	4,6	4,3	4,8
	V.08	3,6	3,4	3,1	3,8		V.08	4,6	4,2	3,7	4,7
	VI.08	3,6	3,4	2,9	3,9		VI.08	4,6	4,1	3,4	4,7
	IV.09	5,6	5,6	5,3	5,9		IV.09	8,5	8,9	8,6	9,3
	V.09	5,6	5,5	5,0	6,0		V.09	8,5	9,5	8,7	10,3
	VI.09	5,7	5,5	4,9	6,2		VI.09	8,6	10,4	9,0	11,7
Jihočeský	IV.08	3,8	3,7	3,4	3,9	Jihomoravský	IV.08	6,0	5,8	5,5	6,1
	V.08	3,6	3,3	2,9	3,7		V.08	5,7	5,4	5,0	5,9
	VI.08	3,5	3,2	2,6	3,8		VI.08	5,7	5,3	4,7	5,9
	IV.09	6,5	6,8	6,5	7,1		IV.09	8,6	8,6	8,2	9,1
	V.09	6,4	7,1	6,6	7,7		V.09	8,6	8,6	7,9	9,3
	VI.09	6,4	7,7	6,8	8,5		VI.09	8,7	8,6	7,6	9,6
Plzeňský	IV.08	3,9	3,8	3,6	4,1	Olomoucký	IV.08	6,0	5,8	5,5	6,1
	V.08	3,7	3,6	3,2	4,0		V.08	5,6	5,3	4,8	5,9
	VI.08	3,7	3,5	3,1	4,1		VI.08	5,6	5,1	4,4	5,8
	IV.09	6,7	6,8	6,5	7,1		IV.09	10,0	10,4	10,0	10,7
	V.09	6,8	7,1	6,6	7,5		V.09	10,1	11,1	10,3	11,8
	VI.09	6,9	7,5	6,8	8,2		VI.09	10,2	12,0	10,9	13,1
Karlovarský	IV.08	6,6	6,5	6,1	6,8	Zlínský	IV.08	5,3	5,1	4,8	5,5
	V.08	6,4	6,1	5,5	6,6		V.08	5,1	4,8	4,2	5,3
	VI.08	6,3	5,9	5,2	6,7		VI.08	5,0	4,6	3,9	5,6
	IV.09	10,1	9,7	9,4	10,1		IV.09	8,8	8,6	8,2	9,0
	V.09	10,1	10,0	9,4	10,5		V.09	8,9	8,7	8,0	9,4
	VI.09	10,2	10,4	9,6	11,1		VI.09	9,2	8,8	7,8	9,8
Ústecký	IV.08	9,8	9,6	9,5	10,4	Moravskoslezský	IV.08	8,3	8,3	7,9	8,8
	V.08	9,5	9,5	8,8	10,2		V.08	8,1	8,0	7,3	8,7
	VI.08	9,3	9,4	8,5	10,3		VI.08	8,0	7,8	6,9	8,7
	IV.09	12,4	12,1	11,8	12,5		IV.09	11,1	11,4	11,0	11,7
	V.09	12,4	12,2	11,6	12,8		V.09	11,3	11,8	11,1	12,5
	VI.09	12,4	12,4	11,6	13,3		VI.09	11,5	12,7	11,5	13,8
Liberecký	IV.08	5,8	5,5	5,3	5,8	Czech Republic	IV.08	5,2	5,2	4,9	5,6
	V.08	5,6	5,3	4,9	5,7		V.08	5,0	5,0	4,5	5,5
	VI.08	5,6	5,3	4,7	5,8		VI.08	5,0	4,9	4,2	5,5
	IV.09	10,1	10,1	9,7	10,4		IV.09	7,9	8,0	7,7	8,3
	V.09	10,2	10,3	9,6	11,0		V.09	7,9	8,4	7,8	8,9
	VI.09	10,4	10,5	9,5	11,6		VI.09	8,0	9,0	8,1	9,8
Královéhradecký	IV.08	4,0	3,9	3,7	4,2						
	V.08	3,8	3,6	3,2	4,1						
	VI.08	3,8	3,5	2,9	4,1						
	IV.09	6,8	6,8	6,5	7,1						
	V.09	6,7	6,9	6,3	7,5						
	VI.09	6,8	7,1	6,2	7,9						

Migration in Central Europe

Eva Kačerová

University of Economics nam. W. Churchilla 4, 130 67 Prague 3, Czech Rep.
kacerova@vse.cz

Abstract. In May 2004 the four Visegrad Group countries, the Czech Republic, Slovakia, Poland and Hungary, entered (among others) the European Union and they have become a part of the single internal market with four freedoms such as free movement of goods, services, capital and people. Difficulty in monitoring of migration processes is one of the most important problems connected with research on this phenomenon in various countries. The definition of migrant is not the same in selected countries. This article came into being within the framework of the long-term research project 2D06026, "Reproduction of Human Capital", financed by the Ministry of Education, Youth and Sport within the framework of National Research Program II.

1 Introduction

Based on the United Nations estimation, migrants make up 3 % of the world population (about 175 million people). The number of immigrants has raised significantly in the First World countries since 1970. More then 14 million people will move there during next years, according to the World Bank forecast. This will lead to increase of 1.8 % in national income in these countries and simultaneously a 0.4 % rise in countries which are main source of immigrants [3].

In May 2004 the four Visegrad Group countries, the Czech Republic, Slovakia, Poland and Hungary, entered (among others) the European Union and they have become a part of the single internal market with four freedoms such as free movement of goods, services, capital and people. The movement of people between the new and old EU Member States has been a very important topic of many research studies as well as it has become a hot political issue and remained with partial restrictions of a free movement of workers until today.

However, there are also other international migration topics which, according to our opinion, deserve our interest. The aim of this paper is to evaluate the international migration and mobility of the EU citizens - from the old Member States as well as from the new Member States in four selected countries. These flows have not been restricted since the enlargement and we can evaluate whether this moment has had any effect on the immigration flows.

Difficulty in monitoring of migration processes is one of the most important problems connected with research on this phenomenon in various countries. The definition of migrant is not the same in selected countries [2], [5], [6], [7], [8].

Table 1. Natural increase of population per 1 000 population

Country	1980	1990	1995	2000	2005	2006	2007	2008
Czech Rep.	1.8	0.1	-2.1	-1.8	-0.6	0.1	1.0	1.4
Hungary	0.3	-1.9	-3.2	-3.7	-3.8	-3.1	-3.5	-3.1
Poland	9.7	4.1	1.2	0.3	-0.1	0.1	0.3	0.9
Slovakia	8.9	4.8	1.6	0.5	0.2	0.1	0.1	0.8

Source: www.eurostat.eu

Table 2. Net migration per 1 000 population

Country	1980	1990	1995	2000	2004	2005	2006	2007	2008
Czech Rep.	-4.0	-5.7	1.0	0.6	1.8	3.5	3.4	8.1	6.7
Hungary	0.0	1.8	1.7	1.6	1.8	1.7	1.9	2.0	1.7
Poland	-0.7	-0.3	-0.5	-10.7	-0.2	-0.3	-0.9	-0.5	-0.4
Slovakia	-2.3	-0.4	0.5	-4.1	0.5	0.6	1.7	1.3	1.3

Source: www.eurostat.eu

Table 3. Stocks of foreign population in selected countries (thousands)

	2001	2002	2003	2004	2005	2006	2007
Czech Rep.	210.8	231.6	240.4	254.3	278.3	321.0	392.0
% of total population	2.0	2.3	2.4	2.5	2.7	3.1	3.8
Hungary	116.4	115.9	130.1	142.2	..	..	..
% of total population	1.1	1.1	1.3	1.4	..	..	..
Poland		49.2	..		..	..	..
% of total population	..	0.1	..	..	..	..	..
Slovak Rep.	29.4	29.5	29.2	22.3	25.0	..	..
% of total population	0.5	0.5	0.5	0.4	0.5	..	..

Source: www.eurostat.eu

Table 4. Foreigners – employees (%)

Country of destination→ Country of origin ↓	Czech. Rep. (2005)	Slovakia (2005)	Hungary (2004)	Poland (2001)
Czech Republic	x	42.5	< 0.7	< 1.2
Slovakia	55.7	X	6.6	< 1.2
Hungary	< 0.3	< 2	x	< 1.2
Poland	6.7	2.1	0.7	x
Total (selected countries)	< 62.7	< 46.6	< 8.0	< 3.6

Source: www.eurostat.eu

2 Empirical observations

Most of the international migration in the Visegrad Group countries is related to their historical and geographical ties (table 1, 2). Thus the role of migrants from the EU 15 in the total immigration flows is relatively small. Anyway, the number of EU 15 citizens has been gradually rising with the deeper economic relations of Visegrad Group countries with the European Union during the 90s. This migration shows predominantly economic motivation. The citizens of old EU Member States usually work in highly skilled positions as managers, professionals or entrepreneurs. These migration flows are related to the trade and investment flows from the source countries. For example, the regional distribution of EU 15 citizens is highly correlated with the foreign direct investment location within the regions in the Czech Republic [1].

The second part is devoted to the migration and mobility from the new Member States to the Visegrad Group countries and is mainly focused on the bilateral migration flows of the selected four countries. We have found out that despite their geographical proximity and former economic integration within CEFTA, the four Visegrad countries are not significantly interconnected with international migration flows except the relation between the Czech and Slovak Republic. Although there are insufficient data and difference in migration definitions we have found out that the migration from the Slovak to the Czech Republic is the strongest bilateral migration flow (approximately 97 thousand of workers in 2007), followed by number of Poles in the Czech Republic (almost 21 thousand workers in 2007) and the Czechs in the Slovakia (2 thousand workers in 2006). The rest of the bilateral flows are rather small.

For the comparison of the form and depth of regional integration we used the relative share of number of foreigners from the rest three Visegrad countries in the total number of foreigners (measured as foreign workers) for every single Visegrad country and we also counted the share of imports and exports with the three Visegrad countries in total imports and exports for every single Visegrad country. We have found out that the Czech and Slovak Republics are also significantly interconnected with labour migration. There is also relation between Slovakia and Hungary with regard to labour force. The migration relations between the Czech and Slovak republics are stronger than the trade flows although both countries are relatively more integrated in the regional trade than Hungary and Poland. For the latter countries it is typical that if they are integrated in regional economy they are more likely trade than migration flows. Poland is an important labour exporter but these workers are mainly active in the old EU member states. The strong Czech and Slovak regional participation can be explained mainly by their strong bilateral economic ties.

In this part we look at available data as seen by the end user. We investigate immigration and emigration data by organising them in a way that allows us to compare data reported by sending and receiving countries and to evaluate international comparability of data provided by individual countries. We analyse two types of information: the double entry matrix containing the flows between selected country and time series of flows between selected pairs of countries.

In order to illustrate the problems with data on international migration flows we have constructed a double entry matrix for the year 2003 and for the year 2005 (tables 5 and 6). Unfortunately latest data are not available. The idea of double entry migration

matrix is to present the data on immigration, reported by the receiving countries, and those on emigration, reported by the sending countries, in one table. The cells in tables 5 and 6 representing migration from country A to country B contain two entries: the upper one includes immigration (I) form country A reported by country B and the lower one includes emigration (E) to country B reported by country A. For a better understanding the data in a double entry matrix we have calculate I/E ratio and I – E differences, where I and E are the flows reported by the receiving and by the sending country. The figures reported by the receiving country are often several times higher than those reported by the sending country. Large I/E ratio (tables 7 and 8] have been observed for flows from Slovakia to Czech Republic (I/E = 54 in 2003, 14 in 2005) and from Poland to Czech Republic (I/E = 36 in 2003, 25 in 2005).

The general believe is that immigration data are better than those concerning emigration. The flow from Slovakia to the Czech Republic in 2003 was according to Czech Republic 24 385 people; the value reported by Slovakia was only 448. The flow from Czech Republic to Slovakia was 18 262 according the Czech data source and 650 according to Slovakia data source. So, both countries had a positive net migration. The flow from Slovakia to the Czech Republic in 2005 was according to Slovakia 734 people; the value reported by Czech Republic was 10 133. The flow from Slovakia to the Czech Republic was 1 144 according the Czech data source and 1 950 according to Slovakia data source. So, both countries had a positive net migration.

Identifying and counting expatriates is not without difficulties and different methods may produce different estimates. There are three main types of estimates, each of them with it advantages and shortcomings: emigration survey in origin countries and compilation of statistics from receiving countries and population census.

Other interesting observations can be made by looking at the figures presenting the evolution of the flows between pairs of countries over time reported by each of both countries. Such graphs are very helpful when trying to understand international migration trends and prepare a forecast (Figure 1a-f). Dates for Hungary are not available.

Table 2. Migration flows between selected countries according to receiving (I) and sending (E) countries in 2003.

Sending country		Receiving country			
		Czech Rep.	Hungary	Poland	Slovak Rep.
Czech Rep.	I	–	…	46	650
	E	–	35	1 040	18 262
Hungary	I	58	–	20	25
	E	…	–	…	…
Poland	I	1 653	…	–	36
	E	46	6	–	10
Slovak Rep.	I	24 385	…	19	–
	E	448	18	10	–

Source: prepared on data from Eurostat
… data not available

Table 3. Migration flows between selected countries according to receiving (I) and sending (E) countries in 2005.

Sending country		Receiving country			
		Czech Rep.	Hungary	Poland	Slovak Rep.
Czech Rep.	I	–	…	60	1 144
	E	–	4	138	1 935
Hungary	I	28	–	21	248
	E	…	–	…	…
Poland	I	1 246	…	–	311
	E	49	13	–	5
Slovak Rep.	I	10 133	…	31	–
	E	734	28	6	–

Source: prepared on data from Eurostat
… data not available

Table 4. Ratios of flows reported by the receiving and sending countries (I/E) in 2003

Sending country	Receiving country			
	Czech Rep.	Hungary	Poland	Slovak Rep.
Czech Rep.	–	…	0.04	0.04
Hungary	…	–	…	…
Poland	35.93	…	–	3.60
Slovak Rep.	54.43	…	1.90	–

… data not available

Table 5. Ratios of flows reported by the receiving and sending countries (I/E) in 2005

Sending country	Receiving country			
	Czech Rep.	Hungary	Poland	Slovak Rep.
Czech Rep.	–	…	0.43	0.59
Hungary	…	–	…	…
Poland	25.43	…	–	62.50
Slovak Rep.	13.80	…	5.17	–

… data not available

Table 6. Person born in selected countries and residing in another country

Country of origin	Country of residence			
	Czech Rep.	Hungary	Poland	Slovak Rep.
Czech Rep.	–	2 494	6 200	75 585
Hungary	6 200	–	1 344	17 293
Poland	24 707	2 685	–	3 473
Slovak Rep.	285 372	37 439	1 514	–

Source: The latest population census around 2000

The direct comparison of flows between Poland and Slovakia reported by the sending and receiving countries reveals important feature of the statistics based on the concept of permanent place of residence, namely the underestimation of emigration flows. The data reported by the receiving country are higher both for flows from Poland to Slovakia and from Slovakia to Poland.
A very low level of both immigration and emigration is reported by Slovakia and Poland during the whole period for which the data are available and it does not allow the identification of the changes in the flow magnitude observed by the partner country.
The flows among Czech Republic and Slovak Republic are the same until 1993, when former Czechoslovakia was split to two separated countries. After the dissolution of Czechoslovakia on 1 January 1993 the previously internal movements between the territories of the Czech Republic and Slovakia became international migration flows. In 2002 and 2003 flows from Slovakia and Poland are average 41 times higher than those reported by the sending countries. As concerns the sudden jumps observed in the Czech data, they might be explained by the changes in the definitions. In the Czech Republic until 2000 the statistics covered permanent migration only, as registered in the population register, similarly to Poland and Slovakia. Since 2001, data from the aliens register were used as well: immigration statistics covered persons who stayed over one year (the exact criteria varied over time) and emigration statistics included data on permits that expired, in addition to self-reported departures for permanent stay abroad.

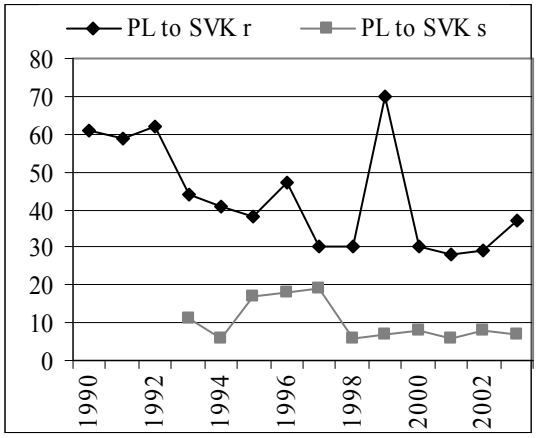


Fig. 1a. Migration between Poland and Slovak Republic

r – data according to the receiving countries
s – data according to the sending countries

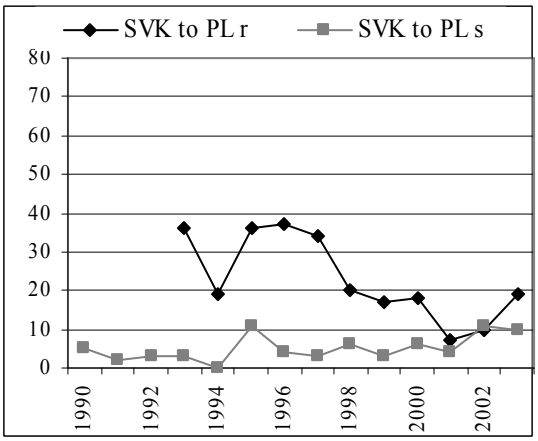


Fig. 2b. Migration between Slovak Republic and Poland

r – data according to the receiving countries
s – data according to the sending countries

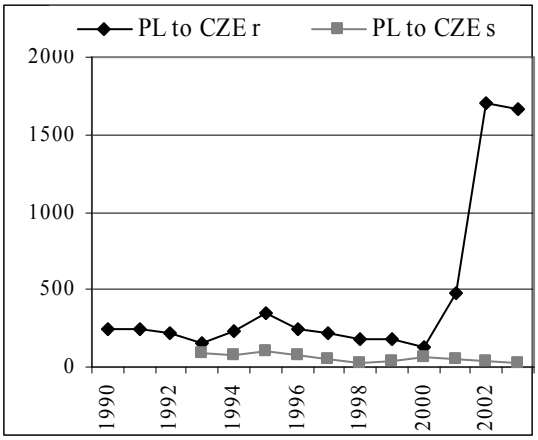


Fig. 3c. Migration between Poland and Czech Republic

r – data according to the receiving countries
s – data according to the sending countries

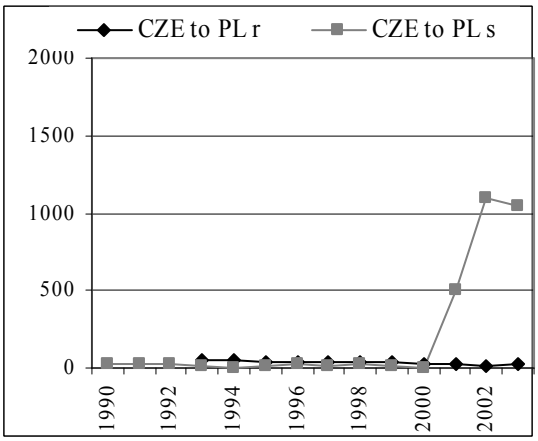


Fig. 4d. Migration between Czech Republic and Poland

r – data according to the receiving countries
s – data according to the sending countries

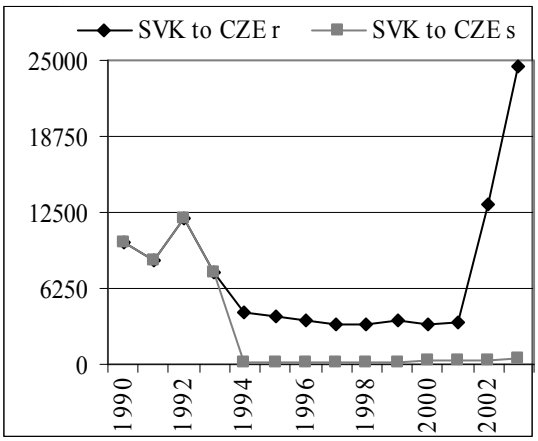


Fig. 5e. Migration between Slovak Republic and Czech Republic

r – data according to the receiving countries
s – data according to the sending countries

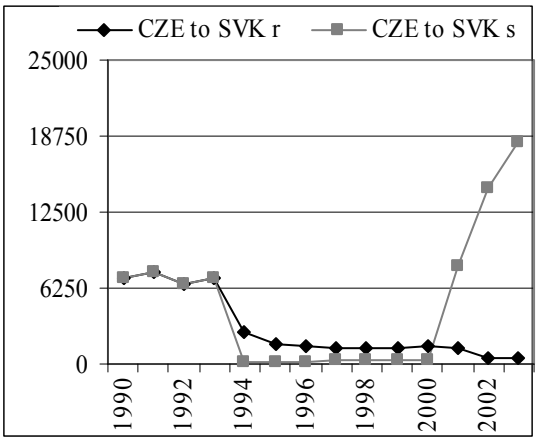


Fig. 6f. Migration between Czech Republic and Slovak Republic

r – data according to the receiving countries
s – data according to the sending countries

Conclusion

Migration in the region will be seen as a consequence of "interplay" of three kinds of imbalances in particular countries or between countries: demographic, economic and political.
It can be expected that after the EU enlargement and the relaxation of migration rules the number of other Visegrad countries' citizens have grown especially in border regions with differences in economic level and unemployment. This effect has probably been stronger in the Czech and Slovak Republic which are more regionally integrated than in Hungary and Poland where most of the migration flows come from neighbouring countries which are not yet the EU members.
Achieving comparability of international migration statistics is a difficult task. The legislation and administrative procedures concerning registration, which is the main source of information on migration flows in the selected countries, will continue to differ. It should be noted that the lack of comparability of statistics on international migration flows is strictly linked with the lack of comparability of statistics on population stocks, so both problems should be solved simultaneously.
And at the end some recommendations for end users of the data of international migration:
- try to find out what is the real content of the data
- do not rely on one source
- do not draw conclusion without taking the definition into account.

References

1. Drbohlav, D.: Migration Trends in Selected EU Applicant Countries: Czech Republic. International Organization for Migration, Wien (2004)
2. Holá, B. (2007): The Comparability of International Migration Statistics, Czech Demography 2007/1, Praha
3. Jennissen, R.P.W.: Macro-economic determinants of international migration in Europe. Rijksuniversiteit Groningen, Dutch University Press, Amsterdam, Kupiszewska, D., Nowok, B. (2006): Official international migration statistics in the EU – data availability and comparability, EPC 2006, Liverpool (2004)
4. Trends in International Migration (2004): SOPEMI 2004. OECD, Paris 2005. Ujhazy, K.: The Connection Between Economic and Income (Wage) Levels Convergence in the CR, Germany, Austria, the United Kingdom and Ireland and Specialists Employment Migration Abroad (Characteristics Summary). Prague, RILSA (2005)
5. Czech Statistical Office, www.czso.cz
6. Central Statistical Office of Poland, www.stat.gov.pl/english
7. Statistical Office of Slovak Republic, www.portal.statistics.sk
8. Statistical Office of Hungary, www.portal.ksh.hu
9. Eurostat, http://epp.eurostat.ec.europa.eu

Limiting Probability Distribution of Random Sample Maximum

Kahounová Jana, Vojtěch Jan

Vysoká škola ekonomická v Praze
Katedra statistiky a pravděpodobnosti
nám. W. Churchilla 4
130 67 Praha 3

Abstract. The extreme value theory is the most appropriate approach to an inherently difficult problem – predicting the possibility that an extreme event will occur. Broadly speaking, there are two kinds of models for extreme values. The first group of models are ones for a distribution of normalized maximum (minimum) of the sequence of independent identically distributed random variables. The second, more modern group of models are the peak-over-threshold (POT) models. These are models for all large observations which exceed a threshold. This paper is concentrated on the first type of model. Here, the maximum river level is treated.

Keywords: extreme value distribution, Gumbel distribution, method of point estimation, moment estimator, maximum likelihood estimator, quantile function.

1 Introduction

The extreme value theory (EVT) has been employed to a answer question relating to the distribution of extremes (i.e. rare events which occur once in a long time period but which may have catastrophic consequences, in other words, rare events are observations with small probability of occurrence). This subject has a rich mathematical theory and a long tradition of applications in a variety of areas involving especially natural phenomena such as windthrow disasters, extreme earthquakes, floods, rainfalls, air pollution etc. For example, EVT answers the question – what height of a river will be exceeded with probability 0.01 in a given year? – this quantity is often called the 100-year flood. 14th August 2002 is kept in our minds as a day of disaster. The worst flood of last centuries spoiled vast area of the Czech Republic including the capital, Prague.

During the last 30 years, many new techniques have been developed concerned with exceedances over high thresholds, the dependence among extreme events in various types of stochastic processes and multivariate extremes. Nowadays, methods of EVT are also gained ground in economic sphere. The statistical analysis of extremes becomes one of the essential tools for integrated risk management dealing with problems related to finance and insurance. On the other hand, EVT is not a panacea for risk managers and actuaries either. There are many theoretical issues that are unresolved till now.

As extreme value distributions are generally considered the three following families, which are described by distribution function $F(x)$:

Type 1 – *Gumbel distribution*

$$F(x) = \exp\left(-\mathrm{e}^{-\frac{x-\alpha}{\beta}}\right), \quad -\infty < x < \infty. \tag{1}$$

Type 2 – *Fréchet distribution*

$$\begin{aligned} F(x) &= 0, && x < \alpha, \\ &= \exp\left[-\left(\frac{x-\alpha}{\beta}\right)^{-k}\right], && x \geq \alpha. \end{aligned} \tag{2}$$

Type 3 – *Weibull distribution*

$$\begin{aligned} F(x) &= \exp\left[-\left(\frac{\alpha-x}{\beta}\right)^{k}\right], && x \leq \alpha, \\ &= 1, && x > \alpha. \end{aligned} \tag{3}$$

where $\alpha, \beta > 0$ and $k > 0$ are parameters.

The three types may be combined into a single Generalized Extreme Value distribution (GEV)

$$F(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\alpha}{\beta}\right)\right]^{-1/\xi}\right\} \tag{4}$$

for $1+\xi(x-\alpha)/\beta > 0$, where $\alpha \in R$ is the location parameter, $\beta > 0$ the scale parameter and $\xi \in R$ the shape parameter. The shape parameter ξ governs the tail behavior of the distribution. The sub-families defined by $\xi = 0$, $\xi > 0$ and $\xi < 0$ correspond respectively, to the *Gumbel*, *Fréchet* and *Weibull* families. In some fields of application the GEV distribution is known as the *Fisher-Tippett distribution*. (For Fisher-Tippett theorem see, e.g. Embrechts [1]). In this paper we will concentrate on the Gumbel distribution of (normalized) sample maximum (1).

Suppose we have a sequence of independent identically distributed random variables $X_1, X_2, \ldots, X_n, \ldots$ whose common distribution function is $F(x)$. Let $X = X_{(n)} = \max(X_1, X_2, \ldots, X_n), \quad n \geq 1$ denotes the sample maximum. Extreme value distributions were obtained as limiting distributions of the greatest (or the least) values in random samples of increasing size $n \to \infty$. For that purpose, it is necessary to perform a linear transformation with coefficients which depend on the sample size. This process is analogous to standardization as in central limit theorems.

The GEV distribution (4) is an extreme value distribution for sample maximum. A GEV distribution for sample minimum, i.e. for order statistic $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$ can be obtained by substituting $(-X)$ for X in the

distribution function (4) and this yields a separate family of distributions.

2 Gumbel distribution of sample maximum

In this section we will consider type 1 distributions. Corresponding to (1) the probability density function is

$$f(x)=\frac{1}{\beta}\mathrm{e}^{-\frac{x-\alpha}{\beta}}\exp\left(-\mathrm{e}^{-\frac{x-\alpha}{\beta}}\right),\quad -\infty<x<\infty, \tag{5}$$

$-\infty<\alpha<\infty$ and $0<\beta<\infty$. If $\alpha=0,\beta=1$, or equivalently, the distribution of random variable $Y=(X-\alpha)/\beta$, we have the distribution (1) in the standard form with distribution function

$$F(y)=\exp\left(-\mathrm{e}^{-y}\right), \tag{6}$$

respectively with probability density function

$$f(y)=\exp(-y-\mathrm{e}^{-y}),\quad -\infty<y<\infty. \tag{7}$$

Since random variable

$$Z=\exp\left(-\frac{X-\alpha}{\beta}\right)=\mathrm{e}^{-Y} \tag{8}$$

has obviously the exponential distribution with probability density function

$$f(z)=\mathrm{e}^{-z},\quad z\geq 0 \tag{9}$$

it follows that the moment-generating function of random variable Y, denoted by $m_Y(t)$, is

$$\begin{aligned} m_Y(t)&=E\left[\mathrm{e}^{\frac{t(X-\alpha)}{\beta}}\right]=E\left[\left(\mathrm{e}^{-\frac{X-\alpha}{\beta}}\right)^{-t}\right]=E(Z^{-t})= \\ &=\int_0^\infty z^{-t}\mathrm{e}^{-z}\,\mathrm{d}z=\Gamma(1-t),\quad t<1. \end{aligned} \tag{10}$$

Clearly, considering random variable $X=\beta Y+\alpha$, we have

$$m_X(t)=E\left[\mathrm{e}^{t(\beta Y+\alpha)}\right]=\mathrm{e}^{t\alpha}m_Y(t\beta)=\mathrm{e}^{t\alpha}\Gamma(1-t\beta),\quad |t|\beta<1 \tag{11}$$

and the cumulant-generating function is

$$k_X(t)=\ln m_X(t)=t\alpha+\ln\Gamma(1-t\beta). \tag{12}$$

Then the first cumulant, i.e. mathematical expectation $E(X)$

$$\kappa_1(X) = E(X) = k'_X(0) = \alpha + \frac{\mathrm{d}}{\mathrm{d}t}\ln\Gamma(1-t\beta)\,|_{t=0} = \alpha - \beta\frac{\Gamma'(1)}{\Gamma(1)} = \alpha - \beta\psi(1).$$

The function $\psi(x)$ is so called the *digamma* function. It is defined as the derivative of the logarithm of the gamma function, i.e.

$$\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Putting $x = 1$, we have

$$\psi(1) = \Gamma'(1) = -\gamma,$$

where γ is the *Euler-Mascheroni* constant, $\gamma \doteq 0.57722$. Thus

$$\kappa_1(x) = \alpha + \beta\gamma \doteq \alpha + 0.57722\beta. \tag{13}$$

In a similar manner we find the cumulant of the second order

$$\kappa_2(x) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\ln\Gamma(1-t\beta)\big|_{t=0} = (-\beta)^2\psi_1(1),$$

where $\psi_1(x)$ is the *trigamma* function. It is the second of the polygamma functions and is defined by

$$\psi_1(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}\ln\Gamma(x).$$

If $x = 1$, we have

$$\psi_1(1) = \frac{\pi^2}{6}.$$

Hence

$$\kappa_2(X) = \frac{\beta^2\pi^2}{6}. \tag{14}$$

The variance of X will be denoted by $D(X)$ and we define $D(X)$, if it exists, by

$$D(X) = E\{[X - E(X)]^2\}. \tag{15}$$

We can compute $D(X)$ by using the well-known relations

$$D(X) = E(X^2) - E^2(X), \tag{16}$$

$$E(X^2) = \kappa_2(X) + [\kappa_1(X)]^2. \tag{17}$$

Thus

$$E(X^2) = \frac{\beta^2\pi^2}{6} + (\alpha + \beta\gamma)^2 \tag{18}$$

and in accordance with (16)

$$D(X) = \frac{\beta^2\pi^2}{6} + (\alpha + \beta\gamma)^2 - (\alpha + \beta\gamma)^2 = \frac{\beta^2\pi^2}{6} = \kappa_2(X) = 1.64493\beta^2. \tag{19}$$

In general the cumulant of r th order

$$\kappa_r(X) = (-\beta)^r \psi_{r-1}(1), \quad r \geq 2, \tag{20}$$

where $\psi_{r-1}(X)$ is called the polygamma function of order $r-1$ (is defined as the r th derivative of the logarithm of the gamma function).

The moments, respectively moment characteristics of the Gumbel distribution can be computed directly by „classic“ way. The n th raw moment is

$$E(X^n) = \frac{1}{\beta}\int_{-\infty}^{\infty} x^n \mathrm{e}^{\frac{\alpha - x}{\beta}} \exp\left(-\mathrm{e}^{\frac{\alpha - x}{\beta}}\right)\mathrm{d}x. \tag{21}$$

Considering the substitution in (21)

$$u = \exp\left(\frac{\alpha - x}{\beta}\right),$$

we have

$$\begin{aligned} E(X^n) &= -\int_{\infty}^{0} (\alpha - \beta\ln u)^n \mathrm{e}^{-u}\mathrm{d}u = \int_0^{\infty} (\alpha - \beta\ln u)^n \mathrm{e}^{-u}\mathrm{d}u = \\ &= \sum_{k=0}^{n} \binom{n}{k} (-1)^k \alpha^{n-k}\beta^k \int_0^{\infty} (\ln u)^k \mathrm{e}^{-u}\mathrm{d}u = \\ &= \sum_{k=0}^{n} \binom{n}{k} (-1)^k \alpha^{n-k}\beta^k I_k, \end{aligned} \tag{22}$$

where

$$I_k = (-1)^k \int_0^{\infty} (\ln u)^k \mathrm{e}^{-u}\mathrm{d}u = (-1)^k \Gamma^{(k)}(1) \tag{23}$$

($\Gamma^{(k)}(u)$ is the k th derivative of the gamma function) are *Euler-Mascheroni* integrals. If we set $k = 0,1,2$, then

$$I_0 = \int_0^\infty \mathrm{e}^{-u}\mathrm{d}u = 1,$$
$$I_1 = -\int_0^\infty \ln u\ \mathrm{e}^{-u}\mathrm{d}u = \gamma,$$
$$I_2 = \int_0^\infty (\ln u)^2 \mathrm{e}^{-u}\mathrm{d}u = \gamma^2 + \frac{\pi^2}{6}$$

thus according to (22)

$$E(X) = \alpha\int_0^\infty \mathrm{e}^{-u}\mathrm{d}u - \beta\int_0^\infty \ln u\ \mathrm{e}^{-u}\mathrm{d}u = \alpha + \beta\gamma,$$
$$E(X^2) = \alpha^2\int_0^\infty \mathrm{e}^{-u}\mathrm{d}u - 2\alpha\beta\int_0^\infty \ln u\ \mathrm{e}^{-u}\mathrm{d}u + \beta^2\int_0^\infty (\ln u)^2\mathrm{e}^{-u}\mathrm{d}u =$$
$$= \alpha^2 + 2\alpha\beta\gamma + \beta^2\left(\gamma^2 + \frac{\pi^2}{6}\right) = (\alpha + \beta\gamma)^2 + \frac{\beta^2\pi^2}{6}$$

and $D(X)$ is (19).

Note that α and β are purely location and scale parameters, respectively. All distributions (5) have the same shape.

The quantile function $Q(P)$ of continuous random variable X is the inverse of the distribution function, i.e.

$$Q(P) = F^{-1}(P), \qquad 0 < P < 1. \tag{24}$$

If we consider the distribution function (1) then obviously

$$Q(P) = \alpha - \beta\ln(-\ln P). \tag{25}$$

Values of this function for given $0 < P < 1$ are P th quantiles x_p.

The probability that the random variable X takes a value greater than x is called survival function $S(x)$,

$$S(x) = P(X > x) = 1 - F(x). \tag{26}$$

In our case

$$S(x) = 1 - \exp\left(-\mathrm{e}^{-\frac{x-\alpha}{\beta}}\right). \tag{27}$$

Survival functions are most often used in reliability and related fields.

3 Point estimation of the quantile of the asymptotic distribution of the maximum river level

We have yearly maximum values of level of Labe (maximum water state in cm) for

last 69 years (period 1940-2008) at our disposal. The source of the data is Povodí Labe, state enterprise. From sample data we computed the sample mean $\overline{x} = 596.5507$ and the sample standard deviation $s = 152.5514$. To estimate the quantile of the distribution (5) first of all is necessary to estimate parameters α, β of this distribution.

The moment estimators of parameters will be denoted by the symbols α^+, β^+. We solved simultaneously the two equations

$$\overline{x} = \alpha^+ + 0.57722\beta^+, \tag{28}$$

$$s^2 = \frac{1}{6}\pi^2(\beta^+)^2. \tag{29}$$

The solutions are found to be

$$\beta^+ = \frac{\sqrt{6}}{\pi}s, \tag{30}$$

$$\alpha^+ = \overline{x} - 0.57722\frac{\sqrt{6}}{\pi}s. \tag{31}$$

In our case

$$\beta^+ = 118.9438,$$
$$\alpha^+ = 527.8939.$$

According to [2] the Rao-Cramér lower bound of variances of unbiased statistics for parameters α and β are given by $1.1087\beta^2 n^{-1}$ and $0.6079\beta^2 n^{-1}$ respectively. It has been shown in [4] that $D(\alpha^+) \approx 1.1678\beta^2 n^{-1}$, $D(\beta^+) \approx 1.1\beta^2 n^{-1}$. Then the moment estimator α^+ has about 95% efficiency while the moment estimator β^+ has only 55% efficiency. The estimators α^+ and β^+ are both consistent.

The maximum likelihood method provides estimators with better properties. Likelihood function of random sample is

$$L(\alpha, \beta) = \beta^{-n} \mathrm{e}^{-\frac{\sum_{i=1}^{n}(x_i - \alpha)}{\beta}} \exp\left(-\sum_{i=1}^{n} \mathrm{e}^{-\frac{x_i - \alpha}{\beta}}\right). \tag{32}$$

Maximum likelihood estimators will be denoted by symbols $\hat{\alpha}, \hat{\beta}$. They satisfy the equations

$$n - \sum_{i=1}^{n} \mathrm{e}^{-\frac{x_i - \hat{\alpha}}{\hat{\beta}}} = 0, \tag{33}$$

$$-n\hat{\beta}+\sum_{i=1}^{n}(x_i-\hat{\alpha})(1-\mathrm{e}^{-\frac{x_i-\hat{\alpha}}{\hat{\beta}}})=0. \tag{34}$$

Equation (33) can be rewritten as

$$\hat{\alpha}=-\hat{\beta}\ln\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{-\frac{x_i}{\hat{\beta}}}\right). \tag{35}$$

The relation (35), when used in equation (34), yields the following equation

$$\hat{\beta}=\overline{x}-\frac{\sum_{i=1}^{n}x_i\mathrm{e}^{-\frac{x_i}{\hat{\beta}}}}{\sum_{i=1}^{n}\mathrm{e}^{-\frac{x_i}{\hat{\beta}}}}. \tag{36}$$

The equation (36) has to be solved by using a numerical iterative procedure. We used the *Newton-Raphson method*. The number of iterations required depends strongly on the initial values chosen. Our initial value is the moment estimator β^{+} and we have chosen stop rule $|\beta_{k+1}-\beta_k|<10^{-12}$. The Newton-Raphson method converges quickly and after 5 iterations we have

$$\hat{\beta}=129.9253.$$

By substituting value $\hat{\beta}$ in (35) we obtain

$$\hat{\alpha}=525.6531.$$

The maximum likelihood estimator of the P th quantile is given by

$$\hat{x}_P=\hat{\alpha}-\hat{\beta}\ln(-\ln P),\quad 0<P<1. \tag{37}$$

If we now wish to estimate value which maximum level of Labe does not exceed with high probability, for example $P=0.99$, then according to (37)

$$\hat{x}_{0.99}=525.6531-129.9253\ln(-\ln 0.99)\doteq 1123.329$$

According to (27) the estimate of probability that the maximum level of Labe exceeds the third degree of flood activity i.e. value $x=600$, will be

$$S(600)=1-\exp\left(-\mathrm{e}^{-\frac{600-\hat{\alpha}}{\beta}}\right)\doteq 0.4312.$$

Conclusion

The extreme value theory is important for assessing risk for highly unusual events, such as 100-year floods. EVT has two significant results. First, the limiting distributions for the maximum or the minimum of a very large collection of independent identically distributed observations. It is shown that asymptotic distribution of the series of maxima (minima) under certain conditions converges to the Gumbel, Fréchet, or Weibull distributions.

The second significant result is the peak-over-threshold (POT) models. This more modern approach has been developed largely in the insurance engineering, where the insurance company is interested in modeling of the excess loss behavior once a high threshold is reached. The POT models are generally considered to be the most useful for practical applications.

This paper is concentrated on the Gumbel distribution, on its properties and parameters estimators. Then we used this distribution as the model for the distribution of the random variable – maximum level of Labe river.

References

1. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events. Wien, Springer Verlag 1997.
2. Downton, F.: Linear estimates of parameters in the extreme value distribution. Technometrics 1966, 8, p. 3-17.
3. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, Second Edition. New York, John Wiley & Sons 1995.
4. Tiago de Oliveira, J.: Decision results for the parameters of the extreme value (Gumbel) distribution based on the mean and the standard deviation. Trabajos de Estadíctica, 1963, 14, p. 61-81.

The Tendency of Slovak University Students in Their Future Economic Activities

Alena Kaščáková[1], Gabriela Nedelová[1]

[1] Department of Quantitative Methods and Information Systems, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica, Slovakia
alena.kascakova@umb.sk, gabriela.nedelova@umb.sk

Abstract. The purpose of this paper is to analyze the influential factors on enterprising of the Slovak University students. Reliable data was obtained by means of a questionnaire. Questions were divided into three sections – the desire to run a business, the education and support available for establishing a business, and the background profile of the respondents. The preferences of the respondents were tested and the findings were interpreted. The logistic regression was used to identify the most significant factors. The self-confidence, personal experience, and business achievements of close relatives and friends were found to be the greatest factors influencing the probability of the respondent becoming an entrepreneur.

Keywords: Slovak University students, economic activity, questionnaire, enterprising, hypothesis testing, logistic regression.

1 Introduction

At the end of the 20th century Europe was significantly confronted with new trends, possibilities and challenges of a globalizing world. An effort to follow the new world trends with all their obstacles and challenges while keeping harmony with previous development of EU economy in the worldwide context, compelled the highest representatives of EU members to admit and agree to new strategic goals for the Union. The main goal of the so called Lisbon Strategy of 2000 was supposed to be a plan to change the European Union into the most competitive and dynamic knowledge oriented economy of the world, while by the year 2010 one of its points of realization has been the support of the business environment [1].

During the evaluation of the results of this strategy, conducted in the middle of the first decade of this century, some deficits and reserves were found, mainly in "the loss of business spirit" of Europeans. This economically and socially negative fact is becoming more and more widespread amongst the youth. This is why business education has become an object of interest and reformation.

1.1 Opinions of the Business Environment in Slovakia

Enterprising and good business abilities are irreplaceable in running of any economy, especially a market oriented one. Motivation plays an important role in the process of stimulating and activating a business environment.

Business has always been the basic element of a national economy. That is why economic theory has always paid great attention to its functionality. Knowledge that came from the results of the executed business analysis influenced preparing and using the tools of economical politics to stimulate the directions in economy.

The business environment in Slovakia, as in the other post communist countries, has passed through dynamic development and various modifications, mainly in its legislative and juridical settings during its transforming period. Nowadays the problem of building the optimal business structure and the need for building small and medium business is still current. From this point of view the establishment of business by young people is as important as their ideas and opinions of enterprising and related topics.

The problem of integrating young people, mainly University students, into the business environment was the object of analysis of the research project VEGA 1/0795/08 "Potential of young people in the business environment in Slovakia and wider European region and possibilities of its use". The aim of the project was social-economic analysis of the economical life, business support by the state, and comparison of the influence of these factors with the other European countries, mainly of the Vysegrad Four. The research organized under the terms of the international scientific research project was followed by the project „Sensibilité a la création d´enterprise". A database of the questionnaire data is one of its outputs. These were distributed in year 2007 and collected in Slovak Universities. The sample contained the respondents' answers – students of management and marketing in one group and students of the technical and natural sciences in the other.

1.2 The Questionnaire Study

The questions in the survey questionnaire were thematically divided into three parts: desire to run a business, the education and support for establishing a business, and the background profile of respondents. The questions in the first section were oriented to the importance of certain aspects in the professional life (for example job security, a stable income, amount of work, etc.). This section also focused on the respondent's opinion on how entrepreneurship can benefit them, their ability to solve problems while starting a business, their attitude and opinion regarding business decisions, the intent of a respondent to be an entrepreneur or to become employed after completing their education, their confidence in their abilities, and the attraction to the idea of running a business. The respondent was asked to answer every question based on a seven-level evaluation scale.

The second section of questions was given to verify the respondent's opinion of the support available in the form of a business course when planning to start a business after completing their education.

The third section of questions focused on the respondent's profile, his parents and the close acquaintances and their business activities.

The total number of questions was 299, with 73 questions in the first section, 13 questions in the second section, and 15 profile questions, that means that the database contains over 20 000 results.

2 The Results of the Research

The statistical program SPSS was used for data analysis. The proportion of categories was considered using frequency tables, their level was compared using the mean and median and the Friedman test was used [1][2] for verification of the equality of ordinal variables. The significance of differences between the pairs of categories was evaluated using the Wilcoxon pair test [1][2]. The results were interpreted on a 10 % significance level.

Description of the Sample. The respondents consisted of Slovak University students, 150 were studying economy while the remainder were studying applied and natural sciences, that means they had a non-economic specialization of study. 35 % of the respondents came from rural and 65 % from urban areas, while 53.2 % of the respondents were female and 46.8 % male. Approximately a quarter of students financed their education by themselves and 86.3 % have worked for a business. Monitoring the business activity among close relatives and friends of the respondents found that in a proportion of 35 % of students at least one of the respondent's parents has already started a business. Up to 81 % of the students know someone who has started a business while 86.2 % of the respondents considered the business of their parents or relatives to be successful.

Desire to Run a Business. The questionnaire began with a series of questions labeled as 'the desire to run a business.' In the first question, the respondent was asked to rank the importance of a series of statements to her/his professional life. The question consisted of 23 statements whose importance the respondents evaluated according to their opinion. The Friedman test rejected the equality in evaluation of individual statements. The Wilcox test concluded that the respondents placed the highest importance in their professional life on having an enjoyable job. In second place, with an almost identical value of importance, was having a job with: the opportunity for career development; enough time for rest, family and friends; the opportunity to realize one's dreams; and good job security. On the other hand, the statements ranked as least important were not having many responsibilities, having an easy and uncomplicated job, and not having a lot of work.

The second question was focused on finding the respondents opinion on the benefits of entrepreneurship. The question consisted of 23 statements that the students were asked to evaluate as to the likelihood of their occurrence. The same steps as in the first question were used to confirm the assumption of inequality in evaluation of individual statements. The students selected multiple statements, which they

considered as the most likely benefits of entrepreneurship, these are: the opportunity to utilize their creativity, having responsibility, being their own boss, being a part of the entire process, turning their dreams into a reality, being independent at work, and having an enjoyable job.

In the third question the respondents were asked to evaluate their ability to complete tasks necessary to start a business (they were offered 14 tasks). Again, the equality of evaluations was rejected. The respondents indicated the most confidence in their ability to sacrifice body and soul to the project, and secondly in their ability to find qualified coworkers and to lead a team. In association to starting a business that has potential, the lowest level of confidence was given to the following tasks: to attract the business associates, to obtain financing from a bank, and to accumulate funds from close acquaintance.

The next two questions focused on the business attitude in the respondents' environments and the importance of its influence on the respondents. According to the results of the survey, people who are important to the respondents and their friends would provide the most support for a potential business. Family support was valued as secondary, teachers placed completely differently from the previous assessments. The opinions of important people in respondents' surroundings would be most significant for future entrepreneurs, followed by the opinions of family and friends.

Mean and median point evaluations of answers in the next questions demonstrated that the intention to find employment is somewhat stronger than the intention to start a business. The attractiveness of the idea of having a business is stronger than the confidence in one`s own skills to start it.

The key question of the survey was to compare the intention of respondents to be an employer or an employee after completing their education. The proposed scale contained 3 leveled choices regarding options for employment, one neutral choice, and 3 choices regarding owning a business. 123 respondents chose one of the options from the employment range, while the same number of respondents preferred one of the business options. 53 of the respondents were undecided. The distribution of answers is shown on the following graph.

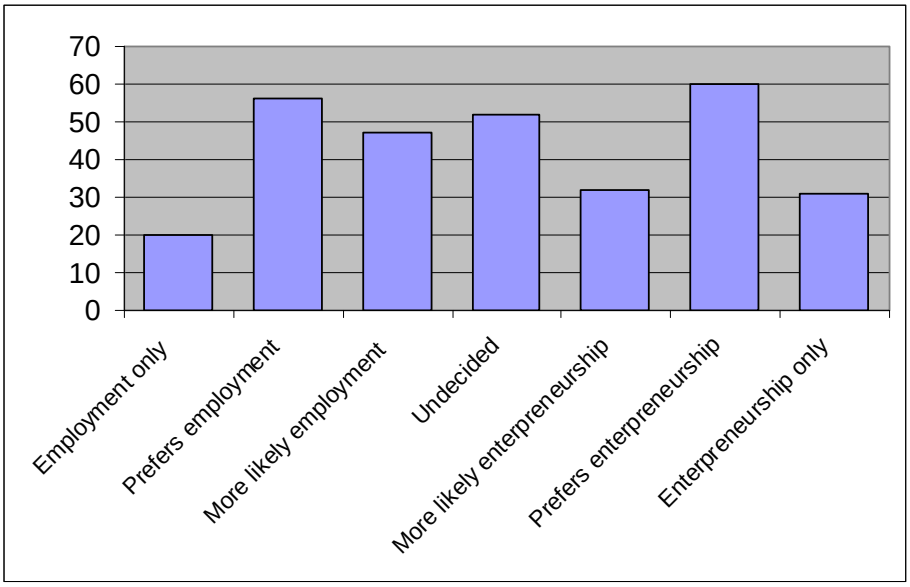


Fig. 1. The distribution of the answers of the respondents according to their intention to run a business after graduation.

Education and assistance in establishing a business. This part of the survey questionnaire includes questions dedicated to the respondents' opinion on the importance of assistance to potential entrepreneurs during the university education. The proportion of those who find the business lectures during their university education important was 82 %. Almost 90 % among them would opt for a lecture based on experiences of founders of existing business enterprises; 87 % would prefer lectures in that they would work on projects to establish a virtual business; 77 % would like to listen to lectures about founding a business, or attend other forms of education like workshops or internships in business enterprises. 61 % of all respondents have not taken a part in any education related to founding a business. Opinion of 64% of the respondents is that these lectures should not be mandatory.

3 Modeling of Probability to Run a Business

The crucial questions regarding the tendency to run a business by the University students in Slovakia were focused on determination of strength of the intension, attraction of the idea, and confidence in the personal ability. They lead to a natural conclusive question: what is the probability that a student will start her/his own business after the university graduation. Multiplicative factors influence the individual decision to establish a business. Various factors impact the final decision with a different intensity level. The main groups of factors that were identified are: economical factors (size of market, tax and levy state policy, support to businesses, etc.); educational factors (professional readiness, experiences and skills obtained

during education); psychological factors (confidence in the individual ability, ability to make decisions and to be in charge of a business, etc.); other factors as influence by business experiences and skills by respondent's environment.

Our intention was to create a model estimating the probability of running a business by students that will graduate on Slovak universities that determines the contribution of multiple significant factors. The statistical technique of logistic regression was used for modeling the outcomes of the survey questionnaire [3]. The implementation of the method modeling the probability requires a binary dependent variable. It was necessary to transfer the scale of answers in 7 levels (from least to most probable) to pair of values expressing the probability that a student will run a business after her/his graduation. The independent variables were chosen mostly from the second and the third parts of the questionnaire, so the variables regarding students background profile, confidence to their own abilities for running a business and students' opinions on help from university in preparation for business activities. We selected the method Enter. The classification ability of the model was 70.6%. The result of the used method is seen in the Table 1. Only the most significant variables are interpreted.

Table 1. Basic statistical characteristics on independent variables used in the model (outcome SPSS)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CAPACITE	,800	,184	18,932	1	,000	2,226
	SEXE	,306	,423	,525	1	,469	1,358
	DOMAINE(1)	-,225	,430	,273	1	,601	,799
	MILIEU(1)	,639	,414	2,384	1	,123	1,895
	CREA1	,381	,388	,963	1	,326	1,464
	CREA2	1,083	,715	2,292	1	,130	2,953
	CREA3	1,153	,542	4,518	1	,034	3,167
	FORMNEC	,275	,528	,271	1	,603	1,316
	FORMDEJk	,104	,385	,072	1	,788	1,109
	FAMILLE2k	-,094	,154	,374	1	,541	,910
	AMIS2k	-,198	,171	1,336	1	,248	,820
	PROFS2k	,115	,130	,772	1	,380	1,121
	GENS2k	,047	,134	,122	1	,726	1,048
	JOB1k	,859	,580	2,193	1	,139	2,361
	JOB2k	-,625	,443	1,995	1	,158	,535
	ASSO1k	-,794	,420	3,563	1	,059	,452
	Constant	-5,188	2,162	5,758	1	,016	,006

a. Variable(s) entered on step 1: CAPACITE, SEXE, DOMAINE, MILIEU, CREA1, CREA2, CREA3 FORMNEC, FORMDEJk, FAMILLE2k, AMIS2k, PROFS2k, GENS2k, JOB1k, JOB2k, ASSO1k.

As seen from the outcomes, there are three major variables (the level of significance is 10 %) influencing the probability that a student will run a business after university graduation: confidence in personal abilities (the variable CAPACITE), business achievements in close surroundings of the respondent (CREA3), and a matter of fact that the respondent has already been a member of a corporation or a business association (ASSO1k), thus her/his own experience. While

interpreting the obtained results we can conclude that one-level increase in confidence on 7-level scale means increasing the chance to run a business by 2.23 times. Business experiences in close surroundings also increase the chance to run a business by 3.17 times. Significant is also the personal experience of the student as a member of a corporation or a business association. In regards to coding of the answers, it is necessary to understand the result as decrease of the probability to run a business in about a half, if the student has not had any experiences as a member of an association. The following facts did not have any influence on the tendency to run a business: taking a part in preparation to start a business; an opinion on importance of business lectures; sex; urban or rural living areas; specialization in education; opinions of relatives and environment; existing experience in working for a business; self-financing during education. It is interesting that even the fact that parents, relatives, or acquaintance have run a business in past did not influence the decision of students to run a business in the future, while the fact that the past business was successful, influenced the tendency significantly.

4 Conclusion

Slovak Republic implemented aims of the Lisbon strategy to Employment National action plan for years 2004 - 2006 and National strategic reference frame for years 2007 – 2013 which should primarily direct to development of weak parts of economy the Slovak Republic and to reduction of regional disparities. The weak parts of the Slovak economy encompass inadequate competitiveness of the Slovak small and medium enterprises and low share of innovating enterprises in industry, inflexible labor market, insufficient connections between educational system and actual requirements of the labor market and not least troubled situation regarding the interest and participation of the Slovak enterprises on research and technological development as well insufficient infrastructure of science and research and implementation of the results in practice [4]. Achievement of such aims constitutes one of the presumptions of building functioning business environment in the Slovak republic.

This article aimed to examine whether the young generation of university students have ambitions to run business enterprise, to participate on creation and gradual development of the business environment as well as to describe the factors capable to support such tendencies. The survey questionnaire proved the young people want to be employed and after finishing their studies expect interesting work wish possibility of professional growth together with employment stability. Approximately 40% of students is considering possibility of own business enterprise, whereas this group has lowest self confidence in acquiring the necessary financing from bank, acquaintances, family, etc., and to impress the shareholders. The students appreciate organizing of the business lectures. Among the important factors for making a decision on running own business belong also positive experiences of the closest surroundings and self confidence. According to the findings is still very important question of quality preparation of students for business activities and by that also the appeal to the Slovak university education to organize such lectures as part of university studies. Distrust of the students in the possibilities to obtain financial support is a call for economic

policy to support small and medium enterprises by the state and for creation sufficient guaranties for financing the starting entrepreneurs by the commercial banks. In the final effect, the future business activities are motivated also by positive personal experiences with running successful business in the close surroundings.

There is also emerging the need of resolving the macroeconomic problem of enduring collectivistic mentality, enhancing state interventions and reduced individual responsibility which are the causes of lost of the business spirit.

It is necessary the National Strategic Reference Frame contains specific and effective measures to improve the situation primary in the fields which could enhance the development of the business environment and by that improve the employment situation in the Slovak republic.

References

1. Pecáková, I.: Statistika v terénních průzkumech. Professional Publishing, Praha (2008)
2. Řezanková H.: Analýza dat z dotazníkových šetření. Professional Publishing, Praha (2007)
3. Stankovičová, I., Vojtková, M.: Viacrozmerné štatistické metódy s aplikáciami. Iura Edition, Bratislava (2007)
4. Návrh inštiticionálneho zabezpečenia koordinácie lisabonskej agendy v Slovenskej republike. http://www.finance.gov.sk/
5. Lisabonská stratégia. http://www.europskaunia.sk/lisabonska_strategia

This work was supported by grant VEGA 1/0795/08 and grant KEGA 3/5214/07.

Usage of static and dynamic control charts in company financial proceeding

Martin Kovářík, Petr Klímek

Tomas Bata University in Zlín, Faculty of Management and Economics
Department of Statistics and Quantitative Methods
Mostní 5139, 760 01 Zlín
{m1kovarik, klimek}@fame.utb.cz

Abstract. We will deal with company financial proceeding using statistical process control in this paper. Especially, we will use Shewhart´s control charts operating with the constant mean and also control charts with non-constant mean and finally process capability indices. We will need to define the center line, UCL and LCL for the control chart construction. The regulated process is not allowed to cross the UCL and LCL boundaries. We will use an Altman´s model for this construction as the most popular index of company financial stability (so called Z-score). We will demonstrate described situation on two case studies. The first one will be focused on financial flow regulation for one company. The second one will describe situation for six companies. Two special types of control charts (CUSUM and EWMA) will be introduced towards the end of this paper. These charts are sensitive on the mean shift. The practical applications will be displayed on two case studies. We would like to point out that the control charts can be used not only in manufactoring but also in company financial proceeding.

Keywords: Altman's Z- Score, Statistical Process Control, Shewhart's control charts, process capability indices, control chart EWMA, control chart CUSUM.

1 Introduction

Statistical financial flow proceeding means the cash flow management in company. We can avoid possible loss by the cash flow monitoring. This loss can be caused by nondelivery goods, bad financial investment, etc. We have chosen the Altman´s model for the financial analysis of the company. Financial analysis should be done once a year. For our example, we will introduce monthly values for unknown company (see Case Study No 1). In other example, we will describe situation in six unknown companies using also monthly values (see Case Study No 2). The end of this paper will be dedicated to the dynamic control charts together with the practical examples (see Case Study No 3 and No 4).

Models for Prediction of Possible Financial Problems of Company

These models are the possibility how to evaluate health of a company via one simple number (index). These numbers try to include all the components of the financial analysis (profitability, liquidity, indebtedness and capital structure). Each component has its own weight. This weight is a picture of importance in the health of a company. These weights for the index components are based on empirical research. A lot of simple and multiple models use it for the prediction of finacial problems (for example Beaver´s test, Edmister´s analysis, Altman´s test, Tamar´s risk index, coefficient ZCR, Lis´ index, Taffler´s index, Springate-Gordon´s index, Fulmer´s index, Index IN 95, Index IN) in a form of numerical intervals. We describe a few of the most used models in the following paragraphs of this paper [1], [2].

Altman´s Model

Altman´s Z-score was found in 1968 as a prognostic index of a company solvency. The base is discriminant analysis of 60 companies being listed at the same time on NYSE. The aim of the paper is to gain a tool for the bancrupt prediction (or more precisely future problems with liquidity) [2].

2 Z-Score and indices

2.1 General Z-Score and Indices

The Altman´s Z-Score is based on the discriminant analysis principle. General notation of discriminant function is (1) [6]:

$$Z = a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_4 + a_5 X_5 + a_6 X_6, \quad (1)$$

where
a_i are discriminant coefficients, i = 1, 2, …, 6;
X_i are discriminant variables, i = 1, 2, …, 6.

Discriminant variables are the same for all listed Z-Score variations [6].

$$x_1 = \text{working capital / total assets;} \quad (2)$$

$$x_2 = \text{revenue after taxation+indivisible revenue (last years) / total assets;} \quad (3)$$

$$x_3 = \text{EBIT / total assets;} \quad (4)$$

$$x_4 = \text{market value of shares / total debts;} \quad (5)$$

$$x_5 = \text{returns / total assets;} \quad (6)$$

$$x_6 = \text{Undertakings after deadline / returns.} \quad (7)$$

2.2 Z-Score Model for Joint-stock Companies

Z-Score model for joint-stock companies is original Altman´s Z-Score model which was built in 1968. Model was tested and built for US companies. But the basic parameters for Czech companies differ significantly from american ones. Therefore informative value of the model could be very low for Czech companies. We will use term Z_1-Score for Z-model for joint-stock companies [1].

The Z_1-Score model is displayed by the formula (8) [6]:

$$Z_1 = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5 + 0.0X_6. \quad (8)$$

Index X_6 is equal to zero, therefore we can replace the formula (8) by (9):

$$Z_1 = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5. \quad (9)$$

Classification Intervals:

According to the Z-Score model result, company can be classified into the one of the following intervals:

$Z_1 > 2.99$ → *Safe Zone* – company is financially strong
$Z_1 \in \langle 1.81; 2.98 \rangle$ → *Grey Zone* – company has some small financial problems
$Z_1 < 1.80$ → *Distress Zone* – company has serious financial problems

Healthy companies in good condition belong to the safe zone. Companies from the distress zone face serious problems and possible bankruptcy. Companies in the grey zone have some problems, but it is hard to tell if the situation gets better or not [3].

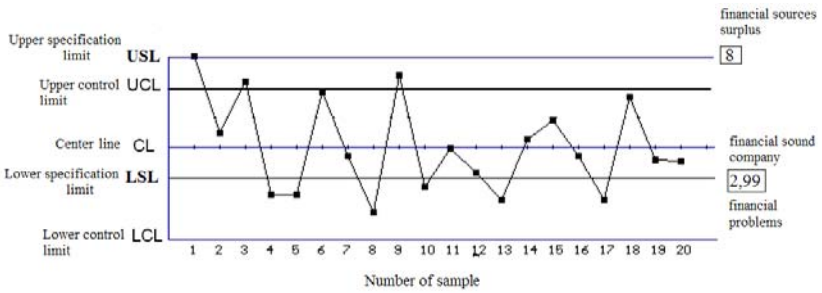


Fig. 1. Altman´s Index for Joint-stock Companies in the Shewhart´s Concept [own processing]

2.3 Z-Score for the Czech Economy

The insolvency is very important in the Czech company economy. Therefore X_6 variable was added into Z-Score model. Disadvantage of this model is a small number of bancrupted companies. Therefore does not exist satisfactory big sample on that this model can be tested. Z-Score model modified for the Czech economy will be marked as Z_1_CZ [6].

The Z-Score is in the following form (10) [6]:

$$Z_{1\_cz} = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5 + 1.0X_6. \quad (10)$$

The classification intervals of model Z_{1_CZ} are the same as in the previous Z-Score (Z_1) in Chapter 2.2.

2.4 Z-Score Model for the Others "Non-joint-stock" Companies

The questions about the use of Z_1-Score model for non-joint-stock companies appeared after its publishing in 1968. Modification of Z_1-Score model was based on some changes in X_4 index. All the coefficients changed and also classification criteria had to be changed. This new model was published in 1983. We can call it Z_2-Score [6].

The Z_2-Score is described by the formula (11) [6]:

$$Z_2 = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X_4 + 0.998X_5 + 0.000X_6. \quad (11)$$

The discriminant coefficient of index X_6 is also equal to zero (see above). Model (11) can be written as the formula (12):

$$Z_2 = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X_4 + 0.998X_5. \quad (12)$$

Classification Intervals:

Classification intervals for this model were modified:

$Z_2 > 2.90$ → Safe zone
$Z_2 \in \langle 1.23; 2.90 \rangle$ → Grey zone
$Z_2 < 1.23$ → Distress zone

The grey zone in the Z_2 is wider than in the Z_1 model (see Chapter 2.2).

2.5 Z-Score Model for Nonproduction Companies and for Growing Markets

This model (published in 1995) does not include index X_5. This index was omitted so that the influence of the industries was minimized. This influence appears by the sensitive variables as index X_5. All the coefficients of the variables from X_1 to X_4 were changed. This model is also useful for the industrial company comparison with the different kind of assets financing. This model is defined as the Z_3-Score [6]:

$$Z_3 = 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4. \quad (13)$$

Classification Intervals:

$Z_3 > 2.60$ → Safe zone
$Z_3 \in \langle 1.10; 2.60 \rangle$ → Grey zone
$Z_3 < 1.10$ → Distress zone

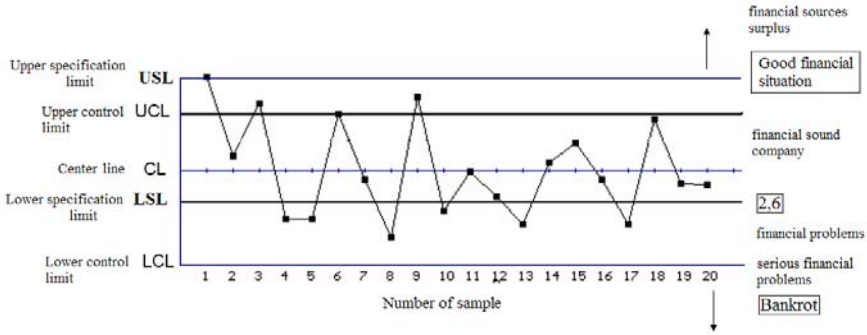


Fig. 2. Updated Model for Nonproduction, Trading and Beginning Companies in the Shewhart´s Concept [own processing]

According to the Altman´s model experience, it is good enough for prediction. Model successfully predicts bankruptcy two years before its appearance. More distant future is not so statistically significant. More modified models are derived from this Altman´s model. For example Springate-Gordon´s Model and Fulmer´s model (see following Chapters 2.6 and 2.7).

2.6 Springate-Gordon´s Model

Model is based on principles of the integral Altman´s model. It was tested on data from 40 companies. Originally, 19 proportional indices were tested. Discriminant analysis chose only four of them. Springate-Gordon´s model is then defined by the formula (14)[1] [6]:

$$S = 1.03X_1 + 3.07X_2 + 0.66X_3 + 0.4X_4, \quad (\mathbf{14})$$

where

$$X_1 = \text{net working capital / property;} \quad (\mathbf{15})$$

$$X_2 = \text{EBIT / property;} \quad (\mathbf{16})$$

$$X_3 = \text{EBT / short-term undertakings;} \quad (\mathbf{17})$$

$$X_4 = \text{returns / property.} \quad (\mathbf{18})$$

If the values of S index are smaller than 0.862 it is possible to expect some problems. Company is classified then as "failed".

[1] SANDS, E.G. - SPRINGATE, G.L.V. – TURGUT, V.: Predicting Business Failures, In. CGA Magazine, May 1983, pages 24–27.

2.7 Fulmer´s Model

This model is suitable for small companies. Originally, 40 indices were analysed on data gained from the selected 40 companies. The half of them were successful, the other half failed. Fulmer´s model is then defined by the formula (19)[2] [6]:

$$F = 5.528X_1 + 0.212X_2 + 0.073X_3 + 1.270X_4 - 0.120X5 + 2.335X_6 + +0.575X_7 + 1.083X_8 + 0.894X_9 - 6.075, \quad (19)$$

where

$$X_1 = \text{indivisible revenues / property;} \quad (20)$$

$$X_2 = \text{returns / property;} \quad (21)$$

$$X_3 = \text{EBT / capital;} \quad (22)$$

$$X_4 = \text{cash flow/ total debt}^3; \quad (23)$$

$$X_5 = \text{total debt / property;} \quad (24)$$

$$X_6 = \text{shot-term undertakings / property;} \quad (25)$$

$$X_7 = \text{property;} \quad (26)$$

$$X_8 = \text{net working capital / total undertakings;} \quad (27)$$

$$X_9 = \text{EBIT / cost interests.} \quad (28)$$

If the value of F in (19) is negative, company can expect some serious problems in future.

3 Statistical process control – Shewhart´s control charts

Statistical process control is one possibility of effective use of statistical methods for company financial flow management. We have to respect variability of financial flow. If we use the same type of calculation we never get the same results – values from the Altman´s model. Control charts consist of the **center line** (CL) placed in a reference value and also the **upper control limit** (UCL) and the **lower control limit** (LCL). These control limits are also called the action limits. UCL and LCL are boundaries of random cause of process variability. They are a decision rule for the process regulation. These lines CL, UCL and LCL are drawed in software tools for the SPC (e.g. QC Expert). If the process is "in control", between the UCL and LCL lies 99.7 % values of the sample.

---

[2] FULMER, J.G.Jr.- MOON,J.E. – GAVIN, T.A. – ERWIN, M.J.: A Bankruptcy Classification Model For Small Firms. In: Journal of Commercial Bank Lending, July 1984, pages 25– 37

[3] Total undertaking is considered as a total debt

Engineering limits – they are usually given by the **upper specification limit** (USL) and by the **lower specification limit** (LSL) according to the Altman´s model. We can calculate in above mentioned examples: USL = 8 and LSL = 2.99. If the value is lower than 2.99 the company has serious problems and if the value is higher than 8 company has financial surplus [3].

W. A. Shewhart is an author of the basic concepts of control charts. Control charts are an graphical aid to separate identifiable causes from random causes of process variability. Control chart construction has a mathematical and statistical basis. Classic control charts belong to a group of control charts without memory because in an actual value are not included the previous ones. Therefore this type of control charts is suitable for identification of sporadic mistakes/deviations in the process (deviations greater than 2_σ from needed level).

How to Built a Shewhart´s Control Chart:

1. We choose a part of process we would like to analyse. We prepare data from observed process.
2. According to the data from the Step 1, we calculate statistical model (represented by the sample mean and the sample standard deviation). We test statistical conditions for the Shewhart´s control chart use.

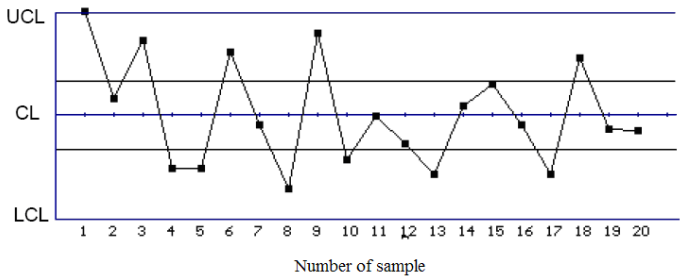


Fig. 3. Shewhart´s Control Chart [QC Expert 2.5Cz]

3. Control chart is constructed on the basis of parameters from the Step 2 (sample mean and sample standard deviation). Center line, upper, and lower control limits are displayed on this chart.
4. Data from selected process are plotted in constructed chart. We focused on "strange cases" which signalize unexpected change in process behaviour. A basic "strange case" is the UCL or the LCL lines crossing.
5. "Strange cases" are registered and their cause is searched. If we find this cause we look for precautions [5].

3.1 Control Chart for Individual Values

We use Shewhart´s control chart for individual values in cases when subgroups are not defined. It is called x-individual. We calculate directly with the x-values instead of the x-means. R-chart is used as a graph of process variability. The range between

consequent values is used instead of the range in subgroup. This value is called moving range (MR) and is calculated as $MR_i = |x_i - x_{i-1}|$. Logically, the first value is not defined. The next formulae (29), (30), and (31) are defined for the CL, UCL, and LCL [4], [5]:

$$UCL = \bar{\bar{x}} + 3\frac{\overline{MR}}{d_2}, \tag{29}$$

$$CL = \bar{\bar{x}}, \tag{30}$$

$$LCL = \bar{\bar{x}} - 3\frac{\overline{MR}}{d_2}. \tag{31}$$

Statistical features of moving range are the same as for the subgroup range where n = 2. Coefficient d_2 is equal to value 1.128. The following formula (32) is valid for the standard deviation calculation:

$$\sigma = \frac{\overline{R}}{d_2} \tag{32}$$

3.2 Control Charts $(\bar{x}, R)$

This chart is suitable for small samples where sample volume is from 2 to 10. This is given by the fact that the R (range) is not a good process variability estimate for larger samples (if n > 10) [4], [5].

Control Chart for Sample Means $(\bar{x})$

Values of sample mean $\bar{x}_j$ from sample with the constant number of observations n are drawn into the control charts $(\bar{x})$. This mean is calculated as shown in formula (33):

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \tag{33}$$

where x_{jj} is i^{th} measured value of regulated quantity in a j^{th} sample.

If we set risk to α = 0,0027 and if we do not know the target values μ_0 and σ_0 we can determine the CL from (34):

$$CL = \hat{\mu}_0 = \bar{\bar{x}} = \frac{1}{k}\sum_{j=1}^{k} \bar{x}_j. \tag{34}$$

We can use the following equations to compute the UCL and the LCL:

$$UCL = \bar{\bar{x}} + \frac{3}{\sqrt{n}}.\frac{\overline{R}}{d_2} = UCL = \bar{\bar{x}} + A_2.\overline{R} \tag{35}$$

and

Values A_2 and d_2 for n from 2 to 25 can be found in the regulation ČSN ISO 8258 (Czech standard ISO) [4], [5].

Control Chart for Sample Range (R)

Values of sample range R_j are used to construct a control chart R. If $\alpha = 0.0027$ and if we do not know the target values μ_0 and σ_0 we can determine the CL from the formula (36):

$$\mathrm{CL} = \overline{\mathrm{R}} = \frac{\sum_{j=1}^{k} \mathrm{R_j}}{\mathrm{k}}, \tag{36}$$

where k is number of samples used for $\overline{R}$ calculation (at least 20), R_j is a sample range in a j^{th} sample. We need to determine standard deviation of sample range for the UCL and the LCL. This standard deviation $\hat{\sigma}_R$ is defined by (37):

$$\hat{\sigma}_R = d_3 . \frac{\overline{R}}{d_2}, \tag{37}$$

where d_3 is a constant for $\hat{\sigma}_R$ estimate. Value of d_3 depends on the number of observations n and it was derived for the regulated quantity from the normal distribution. The formulae (38) and (39) are valid for the UCL and LCL:

$$UCL = CL + u_{0,99865} . \hat{\sigma}_R = \overline{R} + 3 . d_3 . \frac{\overline{R}}{d_2} = (1 + \frac{3 . d_3}{d_2}) . \overline{R}, \tag{38}$$

$$LCL = CL - u_{0,99865} . \hat{\sigma}_R = \overline{R} - 3 . d_3 . \frac{\overline{R}}{d_2} = (1 - \frac{3 . d_3}{d_2}) . \overline{R}. \tag{39}$$

If we replace expressions in brakets in (38) by D_4 and in (39) by D_3 we get more simple formulae (40) and (41) than previous ones (38) and (39):

$$UCL = D_4 . \overline{R}, \tag{40}$$

$$LCL = D_3 . \overline{R}. \tag{41}$$

where coefficients D_4 and D_3 can be found in the regulation ČSN ISO 8258 for n from 2 to 25 [4], [5].

3.3 Process Capability

Process capability indices can be divided into two groups: the indices that measure the processes potential capabilities, and the ones that measure their actual capabilities. The potential capability indices determine how capable a process is if certain conditions are met – essentially, if the mean of the process´natural variability is centered to the target of engineered specifications. The actual capability indices do not require the process to be centered to be accured [4], [5].

Process capability analysis assumes two basic conditions:

- process has to be statistically under control (we can see it from a control chart),
- data must be normally distributed (a use of histogram or/and tests of normality: chi-square test, Kolmogoroff-Smirnoff test, Shapiro-Wilk´s test, etc.).

The most often used process capability indices are **Cp** and **Cpk**. They measure the processes potential and actual capabilities permanently produce nondefective products (in control limits). Nowadays, if the value of Cp ≥ 1.33 or Cpk ≥ 1.33 the process is considered as potentially/actually capable [4], [5].

3.4 Process Capability Indices

a) Capability Index c_p

Capability index c_p is a measure of potential capability. A process is said to be capable if the spread of the natural variations fits in the spread of specified limits. It can be calculated only in cases when both limits are specified. The value of index c_p is a ratio of the maximally permissible and the real variability of quality values regardless their placement in tolerance belt. Index c_p is defined in the formula (42):

$$C_p = \frac{USL - LSL}{6\sigma}, \tag{42}$$

where σ is a standard deviation, LSL is the lower specified limit and USL is the upper specified limit. Real variability of quality characteristic is value equal to 6σ (6sigma) which is the area where lies 99.73 % of all values in a case of normal distribution of the variable. For example, value $c_p = 1$ indicates that the reached probability of defective units is equal to 0.27 % (i.e. 2 700 parts per million). This minimal value of cp will be reached only when the mean of observed quality feature is in the centre of the tolerance belt [4], [5].

b) Capability Index c_{pk}

If the process mean is not centered to the specified target, c_p would not be very informative because it would only tell which of the two ranges (process control limits and engineered specified limits) is wider, but it would not be able to inform on whether the process is generating defects or not. In that case, another capability index is used to determine a process´ ability to respond to customers´requirements. The c_{pk} measures how much of the production process really conforms to the engineered specifications. The k in c_{pk} is called k-factor; it measures the level of deviation of the process mean from the specified target. We use the formulae (43) or (44) to calculate this index [4], [5].

$$C_{pk} = \frac{USL - \mu}{3\sigma}, \tag{43}$$

$$Cpk = \frac{\min\left(USL - \bar{x}, \bar{x} - LSL\right)}{6s}, \tag{44}$$

where μ is a population mean of observed quality. Value of index c_{pk} can be also negative. It can happen when μ crosses one of the control limits. Practically, it means that the process gives more than 50 % defective products [4], [5].

4 Practical part

The aim of this part is to discover if it is possible to use control charts for financial flow. We can found control chart limits by the aid of the Altman´s index to decide if the company is in good financial condition. When the process stability is corrupted it is necessary to look for cause why the index is low or high. Applications of statistical process control methods can indicate changes in financial flow in time before they become serious.

4.1 Case Study No 1 – Calculation of the SPC for One Company

We find out 20 values from two years from the balance sheet of certain company. These values were put in Altman´s model formula. Calculated values of the CL, UCL, and LCL are in the Table 2. Then, we construct a control chart for one company. We use a special control chart for the individual values x_i. The most common value was in the range from 3 to 4. It means that company has no serious financial problems and that it is quite healthy.

Table 1. Values for One Company

Rank	***Value***	***Rank***	***Value***
1	3.578	11	4.28
2	3.953	12	3.577
3	4.288	13	3.855
4	4.191	14	3.605
5	3.129	15	3.7
6	3.039	16	3.415
7	3.525	17	3.535
8	4.595	18	3.455
9	3.915	19	3.21
10	3.757	20	3.355

Given tolerances are: LSL = 8 and USL = 2.99. Moving range is calculated from two neighbouring values, so we find n = 2 in special tables. For n = 2 are coefficients $d_2 = 1.128$, $D_4 = 3.267$, $D_3 = 0$ and formula (32) is valid for the standard deviation calculation – it is equal to 0.4177. Then we use the formulae (29), (30), and (31) for the CL, UCL and finally LCL calculation of x-individual control chart. Similarly, we use the formulae (36), (40), and (41) for CL, UCL, and finally LCL calculation of control chart R (see Figure 5).

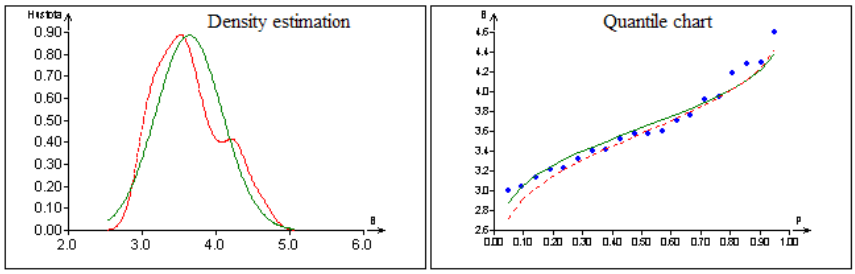


Fig. 4. Test of Normality – Density Estimation (histogram) and Quantile Chart [QC Expert 2.5Cz]

We have to test data normality before we construct control charts. We use histogram and quantile chart on Figure 4. We can compare ideal green curves with the red ones calculated for our data. Red and green curves are very close if data are normally distributed. We can see that in our example these curves are very close. This exploratory data analysis proved normality of example data. Therefore we are allowed to construct control charts.

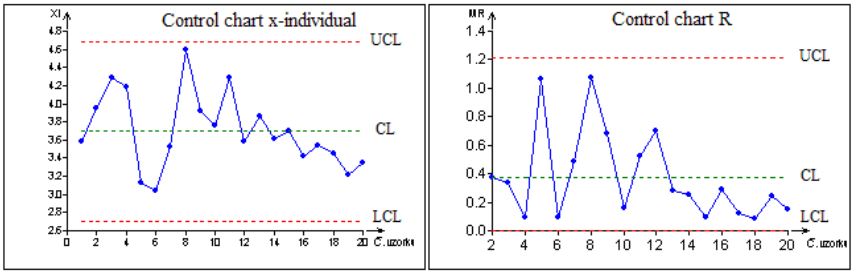


Fig. 5. Control Charts x-individual and R

Table 2. Control Limits Calculations

x-individual	***R***
CL = 3.6978	CL = 0.3724
UCL = 4.6884	UCL = 1.2168
LCL = 2.7022	LCL = 0

Table 3. Capability Indices Calculations

capability index cp	Cp = 1.999
stability of financial flow c_{pk}	Cpk = 0.5730

Calculated value of the cp index is 1.999. It means that our tested company was doing well and that it is financially healthy. But the value of the c_{pk} is 0.5730. It means that the financial flow is not under control. The x-individual control chart shows declining tendency to the boundary of -2σ. It could a warning before the control limits crossing.

4.2 Case Study No 2 – Calculation of the SPC for Six Companies

We gained example data from the balance sheets of six companies. Data were observed monthly. We put data into the Altman´s model. Results can be found in the Table 5. Given tollerances are: LSL = 8 and USL = 2.99. Formula (32) is valid for the standard deviation. It is equal to 0.4177.

Table 4. Values for Six Companies

Rank	X_1	X_2	X_3	X_4	X_5	X_6
1	3.255	3.215	3.426	3.129	3.217	3.511
2	3.436	3.888	3.366	3.273	4.036	3.327
3	3.012	4.164	4.326	3.792	3.43	3.6
4	3.292	3.576	3.686	3.235	3.601	3673
5	3.155	4.347	3.081	4.221	3.262	3.769
6	3.705	4.214	3.427	3.466	3.424	3.188
7	3.33	3.407	3.334	3.126	3.541	3.527
8	3.235	4.742	3.766	3.759	3.433	3.383
9	3.201	3.993	3.43	3.91	3.633	3.396
10	3.98	3.836	3.58	3.394	2.453	3.204
11	4.118	3.519	3.495	3.945	3.243	3.191
12	3.486	4.482	3.336	3.644	3.573	3.741

We use the formulae (36), (40), and (41) for the center line (CL), UCL, and LCL calculation of the control chart x-mean. Formulae (36), (40), and (41) are for the center line (CL), UCL, and LCL calculation of the control chart R. We must test data normality before the control charts construction (we use exploratory data analysis again – see Figure 6).

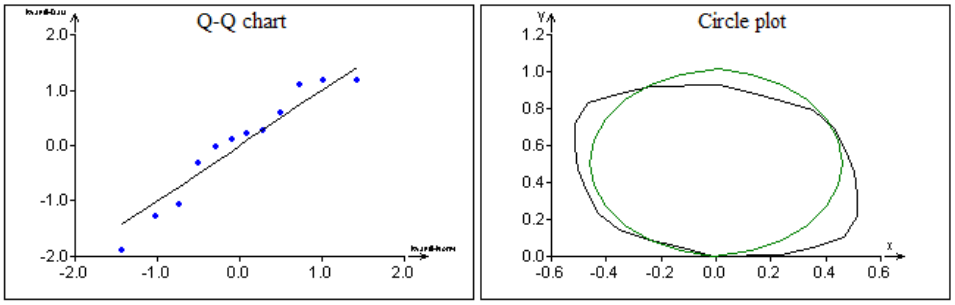


Fig. 6. Test of Normality – Q-Q Chart and Circle plot [QC Expert 2.5Cz]

Q-Q graph for normally distributed data without outliers is line shaped, for normally distributed data with outliers is line shaped with ending points outside the line. Circle plot serves for a visual evaluation of data normality based on skewness and kurtosis. A green circle (ellipse) is an optimal shape for the normal distribution,

black circle represents analysed data. Both circles are the same in a case of normal distribution. This exploratory data analysis proved normally distributed data. We can construct the control charts x-mean and R (see Figure 7).

Table 5. Control Limits Calculations

x-mean	***R***
CL = 3.5568	CL = 0.9596
UCL = 4.0277	UCL = 1.9231
LCL = 3.0859	LCL = 0

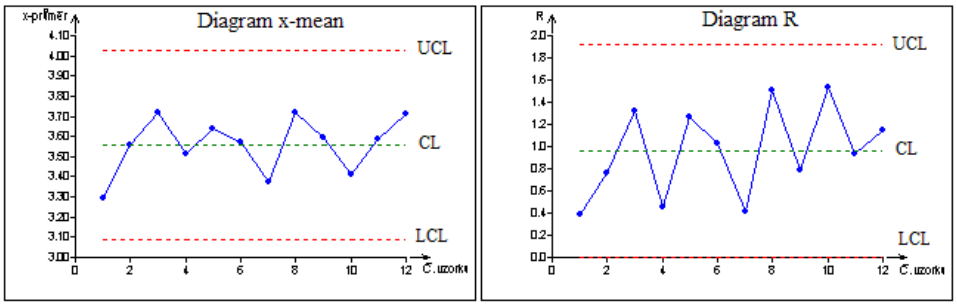


Fig. 7. Control Chart x-mean and R [QC Expert 2.5Cz]

Table 6. Capability Indices Calculations

capability index c_p	$C_p = 6.007$
stability of financial flow c_{pk}	$C_{pk} = 1.3592$

The value of the c_p index is equal to 6.007. It means that companies were doing well in observed months and that they are financially healthy. The value of c_{pk} is equal to 1.3592. It indicates that the financial flow in company is under control and that there are no financial problems in the company. Variability of results among six companies is between the control limits UCL and LCL (i.e. process variability does not cross the control lines – see Figure 7).

5 Other types of control charts

5.1 Control Charts CUSUM

The CUSUM control charts are based on the cumulative sums. They were introduced by Page in 1954. Their main advantage is a very quick detection of relatively small shift in the process mean. This detection is significantly quicker than by the Shewhart´s control charts. The sequential sums of deviations from μ_0 are used

for the CUSUM control chart construction. If μ_0 is a target value for the population mean and if X_j is a sample mean then the CUSUM control chart is constructed by

$$S_i = \sum_{j=1}^{i} (X_j - \mu_0) \tag{45}$$

plotting of variables of the type. This process is called a random walk [5].

5.2 CUSUM – Chart for Individual Values and for Samples Means from Normally Distributed Data

Values of x_i are independent with the same normal distribution $N(\mu,\sigma^2)$ with the known population mean and with the known population standard deviation σ. We assume logical subgroups with the same volume *n*.

Cumulative sum – **CUSUM C_n** is defined for individual values (n = 1) as:

A) on a base of original scale:

$$C_n = \sum_{j=1}^{n} (x_j - \mu). \tag{46}$$

B) on a base of normal distribution where the mean = 0 and the standard deviation = 1:

$$U_j = \frac{(x_j - \mu)}{\sigma}, \tag{47}$$

$$S_n = \sum_{j=1}^{n} U_i. \tag{48}$$

The CUSUM C_n is almost the same as CUSUM S_n measured in the units of standard deviation σ. Equation for C_n can be written recurrently:

$$C_0 = 0, \tag{49}$$

$$C_n = C_{n-1} + (x_n - \mu); \tag{50}$$

and with the same principle for S_n

$$S_0 = 0, \tag{51}$$

$$S_n = S_{n-1} + U_n . \tag{52}$$

Suppose that the original distribution of observed variable $N(\mu,\sigma^2)$ changes into $N(\mu + \delta,\sigma^2)$ distribution for integer *t* (at certain moment). It means that the population mean μ will face a certain shift of δ. It also means that the shift starts at point (m, C_m) and it grows linearly with the slope δ. But the population mean shift can be more complicated. The CUSUM control chart can reflect all these changes [5].

5.3 Case Study No 3 – A Use of the CUSUM Control Chart

- $\mu_0 = 10$, n = 1, $\sigma = 1{,}0$
- We would like to detect the shift $1.0\sigma = 1.0(1.0) = 1.0$ (d = 1,0)
- A process mean which is out of control: $\mu_1 = 10 + 1 = 11$
- K = d/2 = 1/2 a H = $5\sigma = 5$ (recommended)
- Equations for the statistics C_i^+ and C_i^- are then:

$$C_i^+ = \max\left[0,\ x_i - 10.5 + C_{i-1}^+\right]$$
$$C_i^- = \max\left[0,\ 10.5 - x_i + C_{i-1}^-\right] \qquad \textbf{(53)}$$

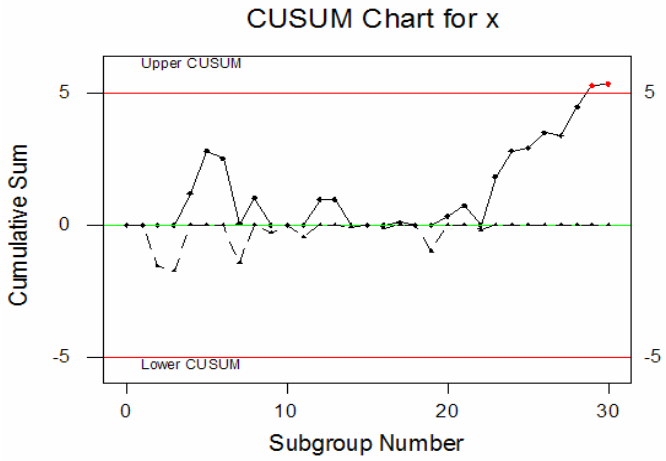


Fig. 8. The Control Chart CUSUM for Case Study No 3 [QC Expert 2.5Cz]

The CUSUM chart shows that the process is out of control (see Figure 8). Following steps are to find the main cause(-s), accept precausion(-s) and to start a new CUSUM again. If the process was adjusted it could be useful to estimate mean of the process caused by the shift.

5.3 CUSUM for Sample Means $\bar{x}_j$

We have considered mainly the individual values until now. Now, we will consider subgroups with *m* observations and we calculate the sample means from this subgroups. We have to work with the sample mean standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}}$. A shift of mean Δ will not be measured in the units of σ but in the units of $\sigma_{\bar{x}}$ in this case. We will substitute the individual values of x_i with the sample means $\bar{x}_j$ and the

process standard deviation σ with the sample mean standard deviation $\sigma_{\bar{x}}$ in above mentioned formulae [4], [5].

New Process Mean Estimate

If there is a shift we can estimate a new process mean from the next formula:

$$\hat{\mu} = \begin{cases} \mu_0 + K + \frac{C_i^+}{N^+}, & \text{pro} \quad C_i^+ > H \\ \mu_0 - K + \frac{C_i^-}{N^-}, & \text{pro} \quad C_i^- < -H \end{cases}, \tag{54}$$

where N^+ and N^- is a number of selected points from a moment when $C_n^+ = 0$, let us say, when $C_n^- = 0$ [4], [5].

Comparsion of CUSUM and Shewhart´s Control Charts

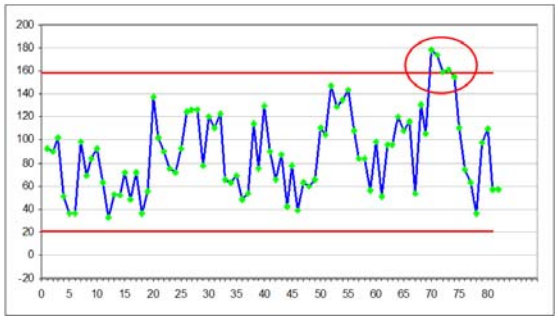


Fig. 9. Shewhart´s Control Chart [QC Expert 2.5Cz]

Fig. 10. Control Chart CUSUM [QC Expert 2.5Cz]

This example shows practically sensitivity of the CUSUM control chart in comparison with the Shewhart´s control chart for the sample means. The CUSUM control chart detects process mean deviation towards the lower values (around the subgroup 20 – see Figure 10) while the Shewhart´s control chart does not detect this

deviation (see Figure 9). It does not detect a shift towards the upper values (around the subgroup 56). It only detects a big shift around the subgroup 70 (see both Figures 9 and 10).

5.4 Dynamic Control Chart EWMA

Dynamic control charts EWMA (**E**xponentially **W**eighted **M**oving **A**verage) are used when the following conditions are fulfilled:

- are not independent, with possitive autocorrelation,
- mean is not constant, its changes are slow.

A sudden change in mean will only cause a control limits crossing. These dynamic charts provide not only the information about "in control" process but also about the process dynamic development. As we mentioned, we consider only data which are not independent with positive autocorrelation. We will explain it now. If the measured observations are influenced by the previous ones we can say that they are dependent. A special case of this dependence is so called autocorrelation of 1st degree when this dependence is linear. If there is a positive autocorrelation in data then the smaller value follows after the smaller value and the higher value follows after the higher value. Data have tendency to preserve their original values. Process is not stable in a case of negative autocorrelation. If there is a negative autocorrelation in data then the higher value follows after the smaller value and the smaller value follows after the higher value.

Suppose that we measure values $x_1, x_2, x_3, \ldots$ for the variable X in the process. We use so called one-step predictions $\hat{x}_k$ to construct the CL, UCL, and UCL for the control chart. The predictions are determined from the following equation $\hat{x}_k = \hat{x}_{k-1} + \lambda e_k$ for k = 1, 2, 3, ..., where the starting prediction value $\hat{x}_0$ is equal to the target value of μ_0. Parameter λ (level of "forgetting") is calculated by trying where the function $\sum_{k=1}^{n} e_k^2$ is minimal. Number n is equal to number of measured values of regulated variable. It is recommended that n is greater than 50. If the error values of one-step prediction of e_k for the optimal value of parameter λ are not correlated and if they have a normal distribution then the center line CL_k, control limits UCL_k and LCL_k for the dynamic control chart EWMA are calculated from the following formulae [4], [5]:

$$CL_k = \hat{x}_{k-1}, \tag{55}$$

$$LCL_k = \hat{x}_{k-1} - \hat{\sigma}_p u_{1-\frac{\alpha}{2}}, \tag{56}$$

$$LCL_k = \hat{x}_{k-1} + \hat{\sigma}_p u_{1-\frac{\alpha}{2}}, \tag{57}$$

$$\hat{\sigma}_p^2 = \frac{1}{n-1}\sum_{k-1}^{n} e_k^2, \tag{58}$$

where $\hat{\sigma}_p^2$ is a standard deviation of e_k estimate, while e_k values are determined for the optimal value of parameter λ [4], [5].

5.5 Case Study No 4 – a Use of Control Chart WMA

We have following financial data of a certain company (in millions of CZK). The starting value of *x* is 5.00. Now, we construct a control chart for data from the Table 7.

Table 7. Entering Data for Control Chart EWMA

k	x_k	*k*	x_k	*k*	x_k	*k*	x_k	*k*	x_k
1	5.01	11	4.85	21	4.80	31	4.77	41	4.65
2	5.11	12	4.99	22	4.84	32	4.60	42	4.67
3	5.04	13	5.05	23	4.78	33	4.51	43	4.50
4	5.12	14	5.38	24	4.82	34	4.58	44	4.50
5	4.94	15	4.97	25	4.88	35	4.67	45	4.44
6	5.01	16	4.84	26	4.78	36	4.51	46	4.44
7	5.11	17	4.77	27	4.80	37	4.57	47	4.53
8	5.18	18	4.78	28	4.81	38	4.56	48	4.53
9	5.04	19	4.82	29	4.88	39	4.57	49	4.57
10	5.18	20	4.78	30	4.75	40	4.60	50	4.34

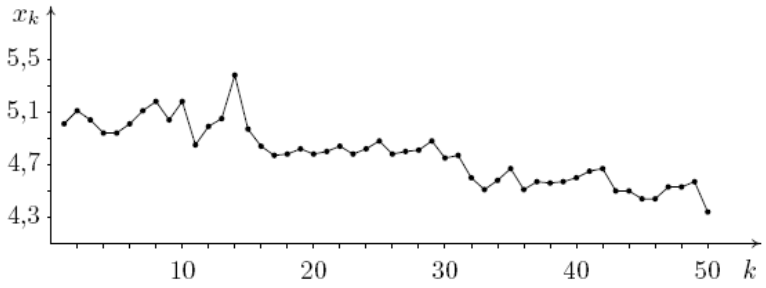


Fig. 11. Lineplot [QC Expert 2.5Cz]

It is evident from the Figure 11 that the values of the process are declining – therefore there is not a constant process mean. Further, we find out if there is an autocorrelation of the 1st degree in the process. We construct a correlation chart between the x_k and x_{k+1} values for k = 1, 2, 3, . . . , 49 (see Figure 12).

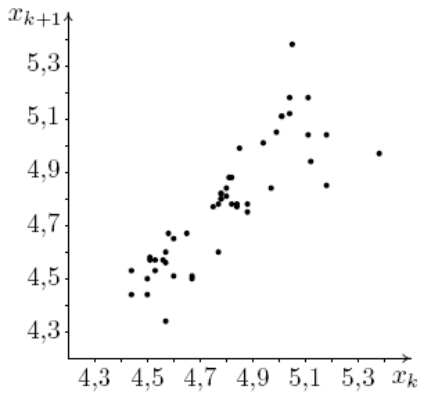


Fig. 12. Correlation chart between the x_k and x_{k+1} values (scatterplot) [QC Expert 2.5Cz]

We can see from the Figure 12 now that the scatterplot is ellipse-shaped and that the ellipse main axis forms an acute angle with the x-axis. We can conclude from this exploratory data analysis that there is a significant first degree positive autocorrelation between the variables x_k and x_{k+1}. We calculate exact coefficient of autocorrelation between the variables x_k and x_{k+1}. It is equal to 0.850. Thus, the strong (statistically significant) positive autocorrelation exists between the variables x_k and x_{k+1}. Therefore we can construct the EWMA dynamic control chart. Further, we will compute predicted values $\hat{x}_k$ for the empirically chosen values of the parameter λ from the interval <0;1>. A starting value was set to the number μ_0 = 5.00. We determine the value of function $S(\lambda) = \sum_{k=1}^{n} e_k^2$ (the sum of error squares) for chosen parameter λ at the same time. The calculation of the parameter λ = 0.48 is shown in Table 8. This value was found as optimal. Function $S(\lambda)$ is evidently parabolic with one extreme (minimum).

Table 8. Values of Function $S(\lambda)$

λ	$S(\lambda)$	λ	$S(\lambda)$
0.4	0.663138	0.47	0.658893
0.5	0.659108	0.48	0.658852
0.6	0.665795	0.49	0.658926

The Table 8 shows computation of the optimal λ parameter value. The values of λ can be found in the first and the third columns and the values of $S(\lambda)$ function in the second and the fourth columns. The values of $S(\lambda)$ function for 0.4, 0.5, and 0.6

are given in the first two columns of the Table 8. We can see that the function minimum is between values 0.4 and 0.5. Therefore we try values of λ equal to 0.47, 0.48 and 0.49. According to the previous computions, the function minimum $S(\lambda)$ is equal to 0.48. The final EWMA control chart is displayed on the Figure 13 where lower and upper control limits are drawed with bold full black lines, center line is drawed with line of dashes. Input values of quality feature are represented by the bold dots.

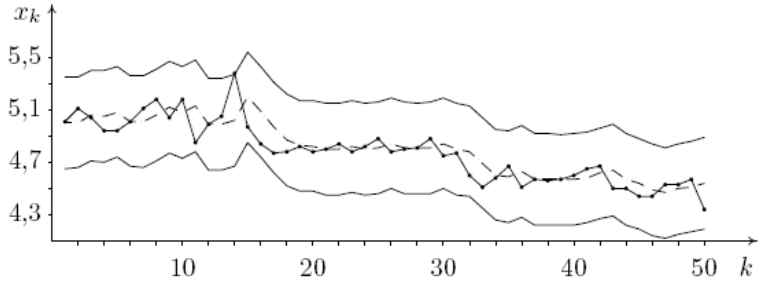


Fig. 13. Control Chart EWMA [QC Expert 2.5Cz]

We can conclude from the analysis of the Figure 13 that the process is under control except of the 14th value. Other values are between the control limits. The 14th value crossing can be caused by the process nonstability or by the measurement inaccuracy.

Summary

This paper dealt with company financial proceeding using statistical process control. Statistical financial flow proceeding means the cash flow management in company. One can avoid possible loss by the cash flow monitoring. We have chosen the Altman´s model for the financial analysis of the company. Financial analysis should be done once a year. We introduced monthly values for unknown company in the Case Study No 1. Further Case Study No 2 described situation in six unknown companies using also monthly values. The end of this paper was dedicated to dynamic control charts together with the practical examples in the Case Studies No 3 and No 4. These charts are sensitive on the mean shift.

References

1. ALTMAN, E. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance 23, September 1968, 589–609.
2. ALTMAN, E. Distress Prediction Models and Some Applications. [online]. [2009-2-15]. URL: <http://www.blackwellpublishing.com/content/BPL_Images/Content_store/Sample_chapter/0631225633/altman.pdf >

3. CÉZOVÁ, E. Statistické řízení finančních toků. ROBUST (23.1. – 27.1. 2006). Fakulta strojní ČVUT v Praze, Ústav technické matematiky. s. 43-50. ISBN 80-7015-073-4.
4. KUPKA, K. Statistické řízení jakosti. 1st ed. Pardubice: TriloByte, 2001. ISBN 80-238-1818-X.
5. TOŠENOVSKÝ, J., NOSKIEVIČOVÁ, D. Statistické metody pro zlepšování jakosti. 1st ed. Ostrava: Montanex, a.s., 2000. 362 pp. ISBN 80-7225-040-X.
6. WOHLMUTHOVÁ, H. Analýza vlastností Altmanova Z-Score. Bakalářská práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, katedra matematiky. Plzeň, 2007. 71

Predicting Financial Distress of Slovak Companies Using Fuzzy Set Theory

Pavol Kráľ[1], Vladimír Hiadlovský[2]

[1] Department of Quantitative Methods and Information Systems,
Faculty of Economics, Matej Bel University, Tajovského 10,
975 90 Banská Bystrica, Slovakia
pavol.kral@umb.sk

[2] Department of Corporate Economics and Management,
Faculty of Economics, Matej Bel University, Tajovského 10,
975 90 Banská Bystrica, Slovakia
vladimir.hiadlovsky@umb.sk

Abstract. The aim of this paper is to discuss substantial problems connected to the prediction of financial distress in the case of the Slovak companies. We focus our attention on the application of fuzzy set theory in the classification process. A fuzzy rule base emulating an expert decision-making process is used to classify companies. The proposed approach is compared with the classical one based on multivariate statistical methods, especially discriminant analysis and logistic regression.

Keywords: fuzzy set theory, fuzzy rule base, financial distress

1 Introduction

The detection of company operating and financial difficulties using financial ratios can be qualitative (managerial decision-making process) or quantitative (using for example an appropriate multivariate statistical method).

The qualitative approach is context dependent, i.e. it can be different with respect to each case (company) and the decision-maker (the analyst). It can use all possible information related to the specific company and accordingly selected parameters. Therefore the decision-making process often fits the selected company very well. But it is rather difficult to generalize it to specific industry.

On the other hand the quantitative approach seems to be easy applicable in general. We build our model once and we are able to make predictions for an arbitrary company. For example the bankruptcy prediction of companies using quantitative methods goes back to the end of sixties when E. I. Altman discussed in his papers the

ability of linear discriminant analysis to provide an appropriate classification model. Starting with 22 financial ratios he developed the following easy to understand so called Altman discriminant function for American publicly traded firms:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5, \quad (1)$$

where X_1 means working capital/total assets, X_2 means retained earnings/total assets, X_3 means earnings before interest and taxes/total assets, X_4 means market value of equity/book value of total liabilities, X_5 means sales/total assets. The general performance of this original model was more than 94 %. Using logistic regression or random forests it can be even better.

The problem is that the aforementioned model is applicable to American publicly traded entities only and it should be revised in order to use it for other companies or other countries. Moreover, the prediction ability of such model is not satisfactory for the period longer than three years (see [1, 2]). It is evident that such model should be revised almost every year. The model revision is completely impossible without data of good quality. The main problem with data in Slovakia is the very low number of companies identifiable as companies with financial difficulties. This fact affects negatively the prediction ability of such model.

In our opinion the following question is very interesting. Is it possible to build an easy applicable model which can be modified for a specific company without any data about other companies? If the answer is yes, such model will combine the positive properties of both approaches. In the presented paper we would like to give a partial answer to the above mentioned question.

It seems to us that a general framework for such model can be provided by fuzzy set theory. The model is constructed using the language and environment for statistical computing and graphics R (http://www.r-project.org/).

The paper is organized as follows:

in Section 2 we present some preliminaries of the fuzzy set theory, especially the basic terms, fuzzy inference and defuzzification, in Section 3 we briefly describe the R package “sets” with implemented terms from the fuzzy set theory and in Section 4 we build a simple predictive model based on financial ratios and fuzzy rule base.

2 Preliminaries

First of all we will recall some basic definitions used in the rest of our paper.

2.1 Basic terms of the fuzzy set theory

The fuzzy set theory has been developed since 1965 (see [6]). The main term of this theory is a fuzzy set. The fuzzy set is fully characterized by its membership function.

Definition 1 (e.g.[6])
Let U be a universal set. A membership function of a fuzzy set A on U is a mapping

$\mu_A : U \to [0,1]$. We denote the class of all fuzzy sets on *U* by *F(X)*.

It is also possible to extend the set operations (union, intersection, complement) to fuzzy sets. These set operations are defined pointwise using triangular norms (t-norms), triangular t-conorms and negators.

Definition 2 (e.g.[3])
A triangular norm *T* is a binary operation on the unit interval which is commutative, associative, increasing in both arguments and has 1 as a neutral element.

Definition 3 (e.g.[3])
A triangular conorm *S* is a binary operation on the unit interval which is commutative, associative increasing in both arguments and has 0 as a neutral element.

Definition 4 (e.g.[3])
A negator *N* is a unary operation on the unit interval satisfying the following conditions:
1. $N(1) = 0, N(0) = 1$,
2. for all $a,b \in R$, if $a \le b$ then $N(a) \ge N(b)$.

The most common negator is so called standard negator defined, for all $u \in [0,1]$, as follows: $N(u) = 1 - u$.

The set operations for fuzzy sets are then defined in the following way:

Definition 5 (e.g.[3])
Let T be a t-norm, S be a t-conorm and N be a negator. Then the generalized intersection $\cap_T$, union $\cup_S$ and complement co_N fuzzy sets are defined as follows: for all $A, B \in F(U)$ and for all $u \in U$,

$$\begin{aligned} \mu_{A \cap_T B}(u) &= T\left(\mu_A(u), \mu_B(u)\right), \\ \mu_{A \cap_S B}(u) &= S\left(\mu_A(u), \mu_B(u)\right), \\ \mu_{\mathrm{co}_N(A)}(u) &= N\left(\mu_A(u)\right) . \end{aligned} \tag{2}$$

We mainly use the special fuzzy sets, called fuzzy numbers, in practice.

Definition 6 (e.g.[3])
A fuzzy number is an arbitrary fuzzy set $A \in F(R)$ for which it holds

1. μ_A is normal, i.e. there exits $u \in R$ such that $\mu_A(u) = 1$,
2. $A_\alpha = \{u \mid u \in R \text{ and } \mu_A(u) \geq \alpha\}$ is closed interval for each $\alpha \in]0,1]$,
3. supp(A) = $\sup_\alpha \{A_\alpha\}$ is bounded subset of R.

There are two special classes of fuzzy numbers which are called triangular and trapezoidal fuzzy numbers. Both are piecewise linear functions. Each triangular (trapezoidal) number can be represented by a triple (quadruple) of real numbers.

2.2 Fuzzy inference

A fuzzy inference means the derivation of an output from the set of if-then rules, i.e. from the fuzzy rule base, and it is based on generalized Modus Ponens.
The starting point of the generalized Modus Ponens is the following scheme:

Rule: IF *X* is *A* THEN *Y* is *B*
Observation: *X is* A^*

Conclusion: *Y is* B^*

The above mentioned scheme is called generalized Modus Ponens if the following condition holds: IF $A = A^*$ THEN $B = B^*$ (see e.g. [3]).
Let us assume that *X* belongs to $F(U)$ and Y belongs to $F(V)$. The (fuzzy) rule in generalized Modus Ponens can be represented by a fuzzy relation *FR*, $FR \in F(U \times V)$, i.e. fuzzy relation *FR* is a fuzzy subset of $U \times V$. The output B^* related to an input A^* we compute using compositional rule of inference as the composition of the fuzzy relation *FR* and the input A^* (see e.g. [3]):

$$\mu_{B^*}(v) = \sup_{u \in U}\left(T\left(\mu_{FR}(u,v), \mu_{A^*}(u)\right)\right), \tag{3}$$

where T is a continuous t-norm.

In general we have more fuzzy rules and they form so called fuzzy rule base. The fuzzy rule base consisting of n if-then rules can be defined in the following way:

Definition 7 (e.g.[5])
Let Φ be a mapping from the input space *F*(*U*) to the output space *F*(*V*). A fuzzy rule base with *n* if – then rules, a single input and a single output has the following form:

Rule 1: IF *X* is A_1 THEN *Y* is C_1 and
...
Rule *n*: IF *X* is A_n THEN *Y* is C_n

where $X \in F(U)$ is a fuzzy input, $Y \in F(V)$ is the corresponding fuzzy output, $A_i \in F(U)$, $i = 1,2,\ldots,n$, are antecedents representing assumptions, domains of applicability or typical fuzzy inputs, $C_i \in F(V)$, $i = 1,2,\ldots,n$, are consequents.

A fuzzy rule base can be obtained for example by asking an expert. The output *Y* we obtain by a composition of the fuzzy relation *FR* representing our fuzzy rule base with the input *X*. We have many possibilities how to derive the fuzzy relation *FR* from our rule base. We restrict ourselves to the Mamdani-Asilian approach due to its simplicity and the fact that the obtained fuzzy relation is continuous.
In this case the fuzzy relation *FR* is defined as follows:

$$\mu_{FR}(u,v) = \max_i \left(\min\left(\mu_{A_i}(u), \mu_{C_i}(v) \right) \right), \tag{4}$$

where $i = 1,2,\ldots,n$, $u \in U, v \in V$.
We obtain the output Y using the following formula:

$$\mu_Y(v) = \max_i \left(\sup_{u \in U} \left(\min\left(\mu_{FR_i}(u,v), \mu_X(u) \right) \right) \right). \tag{5}$$

We use the more general fuzzy rule (with *k* inputs and a single output) in our model. It has the following form:

Rule 1: IF X_1 is A_{11} AND … AND X_k is A_{k1} THEN *Y* is C_1 and
...
Rule *n*: IF X_n is A_{1n} AND … AND X_k is A_{kn} THEN *Y* is C_n.

It is possible to show that the fuzzy rule base with multiple inputs and a single output can be transformed to the aforementioned fuzzy rule base with a single input and a single output (see [3, 5]).

2.3 Deffuzification

The output Y is not a fuzzy number in general. It means that it can not be directly interpreted. Often we need to represent the resulting fuzzy set by a single real number, i.e. defuzzify it, to find a useful interpretation. There exist many defuzzification methods but we restrict ourselves to the most common method – the center of gravity.

Definition 8 (e.g. [3])
Let A be a fuzzy set defined on an interval $[a,b]$, where $a \neq b$. Then the center of gravity of a fuzzy set A is the number *t* defined as follows:

$$t = \frac{\int_a^b \mu_A(u)u\,du}{\int_a^b \mu_A(u)\,du} \quad . \tag{6}$$

3 Fuzzy set theory in statistical environment R

Fuzzy set theory is covered by the R package "sets" (see [4]). This package offers a selection of an appropriate "fuzzy logic", i.e. the triple (t-norm, t-conorm, negator). Available logics are "drastic", "product", "Lukasiewicz", "Fodor", "Frank", "Hamacher" and "Zadeh logic". The default is "Zadeh logic" using the minimum as a t-norm, the maximum as a t-conorm and the standard negator. Zadeh logic is used in the above mentioned Mamdani – Asilian approach. The package provides us also with commands for definitions of fuzzy sets, fuzzy numbers, fuzzy variables, fuzzy rule bases. A fuzzy rule base and fuzzy variables can be combined into fuzzy systems and then a fuzzy inference can be performed. We can defuzzify the resulting output of the inference using several methods. The package "sets" supports the well known center of gravity method (called here "centroid") and several other methods – "meanofmax", "smallestofmax", "largestofmax".

4 Prediction of Financial Distress of Slovak Companies using a fuzzy rule base

We start the construction of our model with the selection of the most appropriate set of financial ratios. The selected financial ratios should be common in practice in Slovak economy and the number of them should not be too large. We decided to use the following five financial ratios: DCF – Debt Cash Flow, ROE – Return on Equity, DR – Debt Ratio, DRP – Debt Repayment Ratio, TIER – Times-Interest-Earned Ratio.

For each financial ratio we define the new domain as a subset of its original domain. The values outside the new domain are transformed to the lower or upper limit of the new domain. The following table lists the obtained limits:

Table 1. New domain of the selected financial ratios.

Financial ratio	Lower limit	Upper limit
DCF	0	20
ROE	-15	30
DR	30	95
DRP	30	100

TIER	0	10

For each financial ratio we assume a linguistic scale, for simplicity "low", "medium" and "high". To each value of the linguistic scale we assign a triangular fuzzy number such that the set of all assigned fuzzy numbers forms a fuzzy partition of the domain in the sense of Ruspini (i.e. $\sum_i \mu_A(u) = 1$, for all u from the domain of the selected financial ratio). For the financial ratio TIER it is illustrated in the following picture:

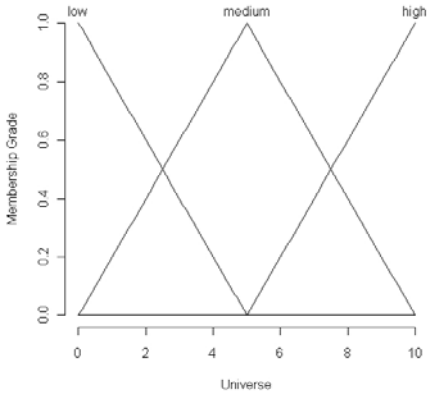


Fig. 1. The fuzzy partition of the domain for the financial ratio TIER

For the computational simplicity we transform the domain of each linguistic scale to the unit interval. In the next step we formulate a fuzzy rule base with multiple inputs – financial ratios and a single output – probability of financial distress (PFD). As in the case of financial ratios we assume a linguistic scale and a fuzzy partition for probability of financial distress. For simplicity we restricted ourselves to the following three rules simulating a possible expert decision:

IF DCF is low AND ROE is high AND DR is low AND DRP is low AND TIER is high THEN PFD is low

IF DCF is medium AND ROE is medium AND DR is arbitrary AND DRP is medium AND TIER is arbitrary THEN PFD is medium

IF DCF is high AND ROE is low AND DR is high AND DRP is high AND TIER is low THEN PFD is high,

where $\mu(x) = 1$, for each $x \in [0,1]$, represents the value arbitrary. Such representation assures that at least one rule is activated for the given set of inputs.

Using the equation (5) we can obtain the probability of financial distress as a fuzzy set. To obtain an easy interpretable value we use the equation (6).

We have applied the above mentioned algorithm to ten bankruptcy companies regardless of its size and industry using values of financial ratios one year prior to bankruptcy. We obtained the following probabilities of financial distress:

Table 2. The predicted probabilities of financial distress.

0.837	0.163	0.837	0.442	0.837
0.837	0.5	0.163	0.163	0.837

The last step of our analysis is the interpretation of the predicted probabilities of distress. We need to find an appropriate cut point for our classification. If we adopt the basic classification rule from the logistic regression and discriminant analysis assuming equal prior probabilities then the classification cut point is 0.5. If our predicted probability is greater than 0.5, the company is assumed to be bankruptcy. But as it was mentioned in the introduction such probability is very small in the case of Slovak companies, e.g. for available data it is 0.02 approximately. So in our opinion the cut point should be moved closer to 0. On the other hand our model should simulate the human analyst decision-making process. Let us assume that the predicted probability of financial distress is 0.2 for a company. Is this probability high or low? Is this company in danger or not? It seems to be quite low, but with respect to the prior probability it is very high. In our opinion we can solve this problem putting the cut point in the middle between 0.5 and the prior probability 0.02. We assume that the prior probability 0.02 can be underestimated so we prefer 0.3 as a new cut point instead of 0.26. Assuming the new cut point, eight companies are classified correct and three companies are misclassified. In our opinion such classification results are satisfactory because of the extremely simplified fuzzy rule base and they can be significantly improved with its further tuning.

Acknowledgments. This work was supported by grant VEGA 1/4634/07 and grant KEGA 3/5214/07. The data set was provided by SCB – Slovak Credit Bureau, s. r. o.

5 Conclusion

In this paper we have discussed the possibility to apply the fuzzy set theory to the prediction of financial distress of Slovak companies.

We have used a very simple fuzzy rule base to construct a simple classification algorithm. We presented here the initial foundation of the proposed algorithm mainly. Our approach is illustrated using the restricted number of financial ratios. For computational tasks the statistical environment R (the R package “sets”) has been used. The further tuning and careful verification of the proposed classification algorithm are goals of our future research. Moreover, in our future work we intend to study the limitations of its applicability and to compare its efficiency with other classification methods.

References

1. Altman, E. I.: Predicting financial distress of companies: Revisiting the Z-score and Z Model. Working paper (2000), www.stern.nyu.edu

2. Altman, E. I.: Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. Journal of Finance, 23(4), 589-609 (1968)

3. Kolesárová, A., Kováčová, M.: Fuzzy Sets and its Applications. Slovak University of Technology, Bratislava (2004)
4. Meyer, D., Hornik, K: Generalized and Customizable Sets in R. Journal of Statistical Software 31(2), 1–27 (2009),. http://www.jstatsoft.org/v31/i02/
5. Navara, M.: Fuzzy Control, http://cmp.felk.cvut.cz/~navara/fl/fc_Foligno04.pdf
6. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338—353 (1965)

Impact of Seasonal Variations on Tourism Businesses in Slovakia and Czech Republic

Peter Laco[1], Martina Bukovská[2]

[1]The Faculty of Economics, Matej Bel University, Banská Bystrica, Slovakia,
peter.laco@umb.sk
[2]T-Com, Košice, Slovakia, bukovska.martina@hotmail.com

Abstract. The paper compares the impact of seasonality on tourism in Slovakia and Czech Republic. Advantage is taken of the methods of econometric analysis, particularly the single-equation linear model which, apart from seasonal explanatory variables, also takes into account the factors reflecting the domestic and partner country`s economic development, as well as random errors. Factors reflecting countries` economic development include the domestic harmonized price index, partner country`s harmonized price index and gross domestic product and the movement of the two countries` currencies` mutual exchange rate. Seasonality factors include air temperature, sunlight duration, total precipitation and cloudiness.

Keywords: Tourism, Statistics, Seasonality, Econometrics

1 Introduction

The main aim of this article is to determine the existence of seasonal differences between the capacity usage of accommodation establishments in Slovakia and Czech Republic. We shall try to define and model the dependency of the number of accommodated tourists on factors based on natural seasonality, i.e. air temperature, sunlight duration, total precipitation and cloudiness. Econometric analysis has been used to elaborate the model. Applying econometric analysis to tourism is a rarely used method in Slovakia and there still exist only a small number of publications usable for putting it into practice. Various analytical methods are used to study and analyze time series and, in spite of being a tool for analyzing economic phenomena and time series development, there has been a significant absence in its utilization so far. Due to the largest portion of arriving tourists, and partially due to the length of the article, we have limited our scope of partner countries to certain European states.

2 Methods and material

We assume that the number of visitors staying at an accommodation establishment can be expressed by a linear relationship where the choice of explanatory variables is based on the Sorensen seasonal demand model for tourism with a correlation

coefficient. (Sorensen, 2004). This model defines the number of accommodated visitors as a function of three variables, namely the prices in tourism, the weather index and the unknown or unconsidered variables which can include social influences, for example. The three variables are explained below.

The prices in tourism can be considered as a variable consisting of the exchange rate (EXC) and the domestic and partner country`s price levels. Our model uses the harmonized index of consumer prices (HICP) to represent the domestic price level, for Slovakia in the first and for Czech Republic in the second case. The foreign price level is indicated by the partner country`s harmonized index of consumer prices. The domestic price level (HICPs/HICPc) expresses the costs of taking part in tourism whereas the foreign price level (HICPx) indicates the living costs of the visiting tourists.

Four indicators make the weather index, and these are the average quarterly air temperature in degrees Celsius, (TEMP), the quarterly sunlight length (SUN), the quarterly amount of precipitation in millimeters (RAIN) and the average quarterly cloudiness in tenths of sky coverage (CLOUDS).

For our seasonality analysis of the accommodation establishments in Slovakia, we have chosen the hydro-meteorological facility in Sliač to be our representative. We have generalized our assumption, considering the weather development recorded in this facility to reflect the overall average development in Slovakia.

Due to our inability to receive additional data from the contacted Czech hydro-meteorological facility, we have only worked with the temperature and sunlight variables for Czech Republic. Since we assume the presence of multicollinearity between the sunlight and temperature variables and the cloudiness and precipitation variables, we do not consider the absence of the latter two important. The necessary data has been acquired from the Czech hydro-meteorological facility`s website in the form of monthly precipitation amounts and monthly local air temperatures.

The data on the capacity usage of the accommodation establishments has been acquired from the European statistical office`s (Eurostat) databases, where we picked the indicator referring to the number of nights spent by residents and non-residents at the accommodation facilities. Our analysis shall only deal with the nights spent by non-residents.

When choosing the partner countries, we have focused on certain European states. For Slovakia, we picked Belgium, Czech Republic, Denmark, Germany, Greece, Spain, France, Italy, Hungary, the Netherlands, Austria, Poland, Finland and Great Britain. During the observation period, the proportion of the visitors from these countries equaled to 74-87% of all foreign visitors. For Czech Republic, we chose Belgium, Denmark, Germany, Spain, France, Italy, Hungary, the Netherlands, Austria, Poland, Slovakia, Finland, Sweden, Great Britain and Norway. During the observation period, the proportion of the visitors from these countries equaled to 65-80% of all foreign visitors. In both cases we were choosing from the countries of the European Union with the highest share of part taken in tourism.

The number of visitors is given as the average quarterly value of nights spent at hotel type collective accommodation establishments and other collective accommodation establishments. Considering the prescribed paper length, we have conducted the analysis for both the hotel and other collective accommodation establishments together. The number of visitors had originally been provided in

monthly numbers; these had to be converted into quarterly figures, but only after we refreshed them by calendar actualization.

After refreshing the data, we converted the sums accordingly to each quarter. The same procedure has then been repeated for all partner countries, as well as for Czech Republic with its partner states.

The originally conceived model has then been augmented by the GDP explanatory variable, expressing the level of Gross Domestic Product in stable prices. This value was obtained after using each partner country`s harmonized index of consumer prices to refresh the data on GDP in stable prices expressed in the domestic currency, available in Eurostat`s statistical database.

Based on the mentioned model, our regression scheme for Slovakia can thus be generally written as:

$$HN_t = \beta_0 + \beta_1 HICPs_t + \beta_2 HICPx_t + \beta_3 EXC_t + \beta_4 TEMP_t + \beta_5 SUN_t + \beta_6 RAIN_t + \beta_7 CLOUDS_t + \beta_8 GDP_t + \varepsilon_t \quad (1)$$

and similarly for Czech Republic as:

$$HN_t = \beta_0 + \beta_1 HICPc_t + \beta_2 HICPx_t + \beta_3 EXC_t + \beta_4 TEMP_t + \beta_5 RAIN_t + \beta_6 GDP_t + \varepsilon_t, \quad (2)$$

HN_t	the number of visitors from a partner country in time t
t	individual observation for the quarters of the 2003-2007 time series
β_j	j = 0, 1, 2,...,8 – regression coefficients
$HICPs_t$	domestic price level in time t (Slovakia)
$HICPc_t$	domestic price level in time t (Czech Republic)
$HICPx_t$	price level of a partner country in time t
EXC_t	domestic exchange rate in time t
$TEMP_t$	average quarterly air temperature in °C in time t
SUN_t	quarterly sunlight length in hours in time t
$RAIN_t$	quarterly amount of atmospheric precipitation in mm in time t
$CLOUDS_t$	average quarterly cloudiness in tenths of sky coverage in time t
GDP_t	gross domestic product in stable prices in time t
ε_t	random error in time t

We are observing the development during 2003-2007. Since we desired to avoid the single short-term random deviations not necessarily caused by factors influencing the movement of visiting tourists, we have preferred to use quarterly data instead of monthly. On the other hand, this has significantly decreased the number of observations which could bias our results due to a large number of explanatory variables put into our model. However, we assume the presence of multicollinearity between the TEMP, SUN and the RAIN, CLOUDS variables. It is also possible to expect a strong dependency between the domestic price level, the foreign price level and the exchange rate, casually also the level of GDP.

For modeling, we have used the freeware called R - environment for statistical computing and graphics, available at http://www.r-project.org and created in 1997 by Robert Gentleman and Ross Ihak from Auckland University`s department of statistics.

3 Application of the Chosen Methodology

For every country, virtually the same procedure has been applied, as will now be demonstrated on the example of Great Britain as Slovakia`s partner country. In this case we observe the seasonal movement of English tourists visiting Slovakia. We inserted the input data in the form of matrices containing the values of explanatory variables for individual observations into the R program.

Next, a correlation matrix was created to discover whether there exists a high correlation between variable couples. As mentioned before, we assume the presence of dependency between the variables indicating climate and those indicating the economic power of both countries.

Table 1 Correlation matrix of original model variables for Great Britain – 1st part

	HN	GBP	HICPs	HICPa	Temp
HN	1,0000000	-0,5622568	0,7075978	0,7066047	0,5890629
GBP	-0,5622570	1,0000000	-0,8652019	-0,9407626	-0,0149550
HICPs	0,7075978	-0,8652019	1,0000000	0,9445246	0,0458023
HICPa	0,7066047	-0,9407626	0,9445246	1,0000000	0,1211035
Temp	0,5890629	-0,0149551	0,0458023	0,1211034	1,0000000
Rain	0,3112578	0,0075632	0,1268704	0,0417216	0,4132420
Sun	0,5060081	0,0589820	-0,0770047	0,0014284	0,8764605
Clouds	-0,3313590	-0,2044767	0,3014529	0,1760638	-0,7023700
GDP	0,5352939	-0,9022993	0,9221770	0,9346081	-0,1134500

Table 2 Correlation matrix of original model variables for Great Britain – 2nd part

	Rain	Sun	Clouds	GDP
HN	0,3112578	0,5060081	-0,3313590	0,5352939
GBP	0,0075632	0,0589820	-0,2044767	-0,9022993
HICPs	0,1268704	-0,0770047	0,3014529	0,9221770
HICPa	0,0417216	0,0014284	0,1760638	0,9346081
Temp	0,4132420	0,8764605	-0,7023698	-0,1134502
Rain	1,0000000	0,2482628	0,0172090	-0,0207734
Sun	0,2482628	1,0000000	-0,9060857	-0,2986675
Clouds	0,0172090	-0,9060857	1,0000000	0,4394199
GDP	-0,0207730	-0,2986675	0,4394199	1,0000000

As anticipated, we may observe high correlations between the exchange rate development, the GDP level, the domestic price level, the English price level and also between the average temperature, sunlight and cloudiness. We will try to utilize differences to remove these dependencies. We calculate the differences of each two near-by numbers in a time series.

$$\Delta HN_t = \beta_1 \Delta HICPc_t + \beta_2 \Delta HICPx_t + \beta_3 \Delta EXC_t + \beta_4 \Delta TEMP_t + \beta_5 \Delta RAIN_t + \beta_6 \Delta GDP_t + \Delta \varepsilon_t \quad (3)$$

We shall therefore create a correlation matrix for the differences of explanatory variables.

Table 3 Correlation matrix of differentional model variables – 1st part

	HN	GBP	HICPs	HICPa	Temp
HN	1,0000000	0,4832591	-0,1334102	-0,1048773	0,8142635
GBP	0,4832591	1,0000000	0,1118323	0,0070810	0,3160411
HICPs	-0,1334102	0,1118323	1,0000000	-0,3761757	-0,2635210
HICPa	-0,1048773	0,0070810	-0,3761757	1,0000000	0,3195454
Temp	0,8142635	0,3160411	-0,2635205	0,3195454	1,0000000
Rain	0,3370397	0,0417155	-0,0521187	0,0751316	0,4239713
Sun	0,8323034	0,2696395	-0,0188774	0,1922799	0,8943220
Clouds	-0,8584919	-0,3592094	0,0818081	-0,0539982	-0,7548750
GDP	-0,6847449	-0,0939763	-0,1547729	-0,0834421	-0,7001280

Table 4 Correlation matrix of differentional model variables – 2nd part

	Rain	Sun	Clouds	GDP
HN	0,3370397	0,8323034	-0,8584919	-0,6847449
GBP	0,0417155	0,2696395	-0,3592094	-0,0939763
HICPs	-0,0521187	-0,0188774	0,0818081	-0,1547729
HICPa	0,0751316	0,1922799	-0,0539982	-0,0834421
Temp	0,4239713	0,8943220	-0,7548745	-0,7001279
Rain	1,0000000	0,2762470	-0,0560906	-0,3211382
Sun	0,2762470	1,0000000	-0,9180380	-0,8874821
Clouds	-0,0560906	-0,9180380	1,0000000	0,7803756
GDP	-0,3211382	-0,8874821	0,7803756	1,0000000

Although the differences helped us remove the high dependencies between the GBP, HICPs and HICPa, the climatic explanatory variables still remain highly correclated. We have therefore decided not to include cloudiness, sunlight and the GDP in our model anymore.

Table 5 Correlation matrix of differentional model variables after adjustment

	HN	GBP	HICPs	HICPa	Temp	Rain
HN	1,000000	0,483259	-0,133410	-0,104877	0,814264	0,337040
GBP	0,483259	1,000000	0,111832	0,007081	0,316041	0,041715
HICPs	-0,133410	0,111832	1,000000	-0,376176	-0,263521	-0,052119
HICPa	-0,104877	0,007081	-0,376176	1,000000	0,319545	0,075132
Temp	0,814264	0,316041	-0,263521	0,319545	1,000000	0,423971
Rain	0,337040	0,041715	-0,052119	0,075132	0,423971	1,000000

Table 5 represents the final matrix of variables that we will use for our testing. These are GBP, HICPs, HICPa, temperature and precipitation.

Since we are looking for linear relationship, we consider the high correlations between the individual explanatory variables and the number of visitors a good signal, without any need to change the model or to be highlighted by color in any of the correlation matrices.

After removing the undesired variables, we obtain the model for calculating the number of visitors in the accommodation establishments (HN_t) as a function:

$$\Delta HN_t = \beta_0 + \beta_1 \Delta HICPs_t + \beta_2 \Delta HICPa_t + \beta_3 \Delta GBP_t + \beta_4 \Delta TEMP_t + \beta_5 \Delta RAIN_t + \Delta\varepsilon_t \quad (4)$$

the testing of which now follows. Individual variables have been specified above.

We test them for statistical significance, removing two other variables after the test ($\Delta HICPs_t$ and $\Delta RAIN_t$) and obtaining the resulting model which looks:

$$\Delta HN_t = 7112,3 - 10015,7*\Delta HICPa_t + 1550,1*\Delta GBP + 846,1*\Delta TEMP_t. \quad (5)$$

The testing showed the price level in Great Britain, the GBP exchange rate and the temperature to be statistically significant. a rise in the English level of prices will be reflected by a decrease in the number of English tourists visiting Slovakia. This fact can be explained by rise in living costs, i.e. higher costs of participating in tourism. On the other hand, a positive relationship between the temperature and the participation of the English in Slovak tourism is demonstrated by an increased number of visitors arriving in the period of higher temperatures. We also see the impact of the exchange rate, whose rise brings more English. Such result had been expected, because we quoted the exchange rate directly, i.e. in the amount of domestic currency per one unit of foreign currency – with Slovakia as the domestic country. This means that the increase in the exchange rate means depreciation of Slovak crown, which brings about lower prices for foreigners and increased inflow of tourists.

This model, however, needs to be tested for autocorrelation, heteroscedasticity, normality of residues and correct specification of the model as a whole. Durbin-Watson test is used to determine autocorrelation, Breusch-Pagan test for heteroscedasticity, Jarque-Bera test is utilized to find out whether residues come from normal distribution and the RESET-test is used to test the correct specification of the model.

All the tests mentioned will be elaborated in the R-program interface and p-value will be used to decide whether to accept or reject the alternative hypothesis in all cases. (Gazda, 2009). In order to reject the alternative hypothesis, this value has to be greater than α=0,05. Therefore, if all tests end up with the p-value greater than 0,05, this model can be considered correctly specified, without heteroscedasticity, multicollinearity; we also assume that resisdues come from normal distribution.

In our case, the p-value for individual tests is evaluated as follows:

Table 6 Test results for heteroscedasticity, multicollinearity, normality of model residues and model specification

Test	P-value
DW-test	0,6346
BP-test	0,4155
JB-test	0,9930
RESET-test	0,0512

As apparent, the model has passed all demanded tests, but the resulting RESET-test value is relatively low, so some fine-tuning of the model might be convenient.

The last statistics reported in our results is the determination index (Hatrák, 2007), stating the percentage of values that can be explained by the conceived model. For Great Britain, the index value is 0,8510465, which means that the model can explain 85,1 % of the real values.

4 Results for Slovak republic

The procedure described in the previous text has been carried out separately for each partner country. Tables 7,8 and 9 show the results for Slovakia and tables 10, 11 and 12 in the next chapter show the results for Czech Republic. They include β – coefficients for individual explanatory variables with asterisks indicating whether the given parameter is statistically significant, and to what extent. *** stand for a parameter of 0,1% level of significance; ** for 1% level; * for 5% level; + for 10% level of significance.

Table 7 Regression model parameters and test results for Slovakia – 1st part

	Great Britain	Belgium	Czech Rep.	Denmark	Finland
Intercept	7112,3+	5350,13**	156677*	3772,037**	-1460,78+
EUR	1550,1**	2675,961*	-	17027,42*	-
HICPs	-	-1059,81	-62405,9	-981,784	1512,26*
HICPx	-10015,7**	-3883,49*	-	-2272,74	-
Temp	846,1***	360,044***	37406,09***	183,7564*	235,81***
Rain	-	-13,311	-721,37	14,6568+	-
Sun	-	-	-	-	-
Clouds	-	-	-	-	-
GDP	-	-18,845***	-91,48***	-1,1236**	-
DW-test	0,6346	0,695	0,952	0,9665	0,8345
BP-test	0,4155	0,4428	0,6817	0,3119	0,705
JB-test	0,993	0,611	0,117	0,945	0,478
RESET-test	0,0512	0,9035	0,6586	0,8407	0,2305
R2	0,851047	0,941337	0,886175	0,885078	0,576438

Table 8 Regression model parameters and test results for Slovakia – 2nd part

	Hungary	Austria	Spain	Italy
Intercept	25320,00*	-139,74	3495,19***	9721,4***
EUR	-	-	1447,64+	-8030593**
HICPs	-	-	-630,86	-2496,6*
HICPx	-16190,00*	-	-2435,15***	-6841,2***
Temp	-774,40	575,76***	230***	653,3***
Rain	-	-	-	-
Sun	-	-	-	-
Clouds	-	-	-	-
GDP	-1,11***	-	-	-
DW-test	0,7838	0,9653	0,2254	0,7857
BP-test	0,1753	0,7455	0,3162	0,3811
JB-test	0,284	0,337	0,711	0,339
RESET-test	0,1267	0,1594	0,8918	0,515
R2	0,8830619	0,769304	0,8174574	0,84634

The table makes it clear that the most statistically significant factor in all cases is the temperature, with positive relationship between it and the number of visitors present almost in all cases. Regarding Hungary, a negative relationship is manifested, which can be explained by a stronger rise in temperature in Hungary than in Slovakia, making the Hungarians prefer staying at home for their holiday to traveling to Slovakia. Unlike other countries, however, temperature is not a statistically significant factor for Hungary. The partner country`s level of GDP is another statistically significant factor, but not for all countries. The relationship between the rise in GP and the number of visitors from that particular country is negative, though. For Hungary, Spain, Italy, Great Britain and Belgium, the partner country`s price level (i.e. in one of the mentioned countries) is also a statistically significant variable with negative relationship observed – if the price level in the partner country rises, it causes a decrease in the number of people from that country participating in Slovakia`s tourism.

The Slovak price level proved to be a statistically insignificant factor, with a logical observation of a negative relationship between its level and the number of visitors in Slovak tourism – if the Slovak price level rises, it worsens the inflow of tourists. The only opposite situation happens in the case of Finland, where an increase in the Slovak price level attracts more Finnish tourists to Slovakia.

In the case of Great Britain, Belgium, Denmark, Spain and Italy, we also perceive the impact of the exchange rate. An increase in the exchange rate causes a rise in the number of tourist visitors coming from these countries, with Italy being the only country with a negative characteristic of this relationship.

The results for five countries (the Netherlands, France, Greece, Germany, Poland) speak of an incorrect model specification, in the case of Greece, the residues do not come from normal distribution. Table 9 shows these results.

Table 9 Wrong regression model parameters and test results for Slovakia

	Netherlands	France	Greece	Germany	Poland
Intercept	6590,842**	5150,6+	88,91	-5733	83285,86**
EUR	-	-	-	-	138740,8*
HICPs	-3572,38*	-	-	-	
HICPx	-	-9790,3*	-	-	-95065,1**
Temp	664,071***	843,7***	33,35*	5360***	-
Rain	-	-	-	-	-
Sun	-	-	-	-	-
Clouds	-	-	-	-	-
GDP	664,071***	-	-	-	-179,6***
DW-test	0,5565	0,716	0,936	1	0,3699
BP-test	0,2514	0,1016	0,9726	0,06792	0,1724
JB-test	0,736	0,51	2,20E-16	0,255	0,432
RESET-test	0,0253	0,01213	0,393	0,000201	0,001152

5 Results for Czech republic

The same analysis has been carried out for Czech Republic, with results shown in tables 10 and 11 and the results for incorrectly specified models contained in table 12.

Table 10 Regression model parameters and test results for Czech republic – 1st part

	Great Britain	Belgium	Denmark	Finland	France	Holland
Intercept	11425,4	3501,81*	7674,587	-1098,7	-13089,4	47081,58*
EUR	-	-	-	-4308,5	-	-60249,8*
HICPc	10889,1	-	16246,78+	-	-	-
HICPx	-52292,6*	-	-37508,3**	-	30898,9*	-63849,3+
Temp	5346,3***	1906,78***	6313,919***	1777***	7565,6***	6512,22**
Rain	-	-	-	-	-	-
GDP	389,5**	-63,95***	-3,678+	-	-	-676,1***
DW-test	0,1644	0,636	0,7916	0,7459	0,795	0,8782
BP-test	0,8392	0,2102	0,3571	0,7154	0,07973	0,05223
JB-test	0,621	0,332	0,841	0,219	0,696	0,529
RESET-test	0,1544	0,1813	0,06759	0,5724	0,7517	0,4977
R2	0,635198	0,963216	0,961808	0,858952	0,939534	0,962648

Table 11 Regression model parameters and test results for Czech republic – 2nd part

	Hungary	Norway	Poland	Spain	Italy
Intercept	902,30	-2720,6	19001,81	58868**	3907
EUR	-	-	129979,8**	-	-
HICPc	-	4256,8	39871,78*	-	45603*
HICPx	-	5641*	-62461,3**	-65201***	-45743**
Temp	1663,00***	2458,3***	2828,97**	12569***	6161***
Rain	-	-	-	-	-
GDP	0,06	-	-62,87**	-	-
DW-test	0,86	0,3778	0,4349	0,4102	0,9007
BP-test	0,20	0,8102	0,05059	0,2747	0,2665
JB-test	0,42	0,336	0,836	0,58	0,27
RESET-test	0,99	0,8062	0,9873	0,1543	0,4952
R2	0,635198	0,963216	0,961808	0,858952	0,939534

The seasonality analysis in Czech Republic`s accommodation establishments delivers us to results similar to those in Slovakia. There is a statistically significant relationship between the number of visitors and the temperature in Czech Republic. The second statistically significant factor seems to be the price level in partner countries, but the slope of the relationship varies. For Great Britain, Denmark, Netherlands, Poland, Spain and Italy, there is a negative influence of the change in the domestic price level on the number of visiting tourists coming to Czech Republic.

On the contrary, the relationship observed for France and Norway is positive. An increase in Czech prices positively influences the inflow of tourists.

Rising exchange rate is statistically significant for Poland, where an increase in this variably significantly positively influences the number of Polis tourists visiting Czech Republic. The influence for Finland and the Netherlands is negative, but these factors are little statistically significant for the given model.

Table 12 Wrong regression model parameters and test results for Czech republic

	Germany	Austria	Slovakia	Sweden
Intercept	-7147	-1849,69	1427	-4187,5
EUR	-	-	-	-125369*
HICPc	-	-	-	-
HICPx	-	-	-	-
Temp	23533,8***	1628,72***	3885***	3465,1***
Rain	1077,4+	-	-	-
GDP	-	54,43*	-	-
DW-test	1	0,9983	0,9964	0,8263
BP-test	0,7808	0,9295	0,2825	0,3834
JB-test	0,252	0,835	0,944	0,392
RESET-test	0.006997	0.02364	0.0001463	0.02796

6 Conclusion

The submitted analysis confirms our assumption about an undoubted influence of temperature on the movements of visitors in Slovakia`s and Czech Republic`s tourists. This factor has also proven to be statistically significant. Generally, an increase in temperature positively influences the inflow of visitors. From the climatic factors, we can also mention precipitation, but this has not presented itself as a statistically significant factor. Neither cloudiness and sunlight seem to be significant explanatory variables, mainly due to the presence of multicollinearity between the temperature and sunlight and the cloudiness and precipitation variables. However, if we used the presence of multicollinearity as an excuse for removing the precipitation and temperature variables and left the cloudiness and sunlight variables, we could expect similar results, that is a strong positive influence of sunlight and, on the contrary, a negligible influence of cloudiness.

The partner country`s harmonized index of consumer prices shows its negative influence on inflow of tourists into Slovakia or Czech Republic. This fact is explained by increased living costs and therefore a decrease in disposable resources for leisure.

For both countries, we got surprised by a dominant negative impact of the rise in GDP. It is not valid for all countries, neither it is statistically significant for all cases, but a rise in GDP is often accompanied by a fall in participation in tourism.

The development regarding the domestic price level is varied. While an increase in Czech price level increases the number of its visiting tourists, the same situation in Slovakia decreases the number of visiting tourists.

In most cases, the significant explanatory variables included the GDP, the harmonized index of domestic consumer prices, the partner country`s harmonized index of consumer prices and the exchange rate.

With respect to the analysis carried out, we have not observed significant differences in the seasonality of demand for accommodation establishments between Slovakia and Czech Republic, which is largely supported by the fact that the two countries have similar primary offer as far as tourism is concerned.

We realize that the time period is quite short and the panel data approach could bring interesting results. We plan to choose this type of analysis next time.

In the future, we plan to extend this analysis to one or more additional countries with climatic characteristics different from Central Europe, possibly maritime countries. Studying the development on the monthly basis would also be interesting.

References

1. Czech national bank : ARAD systém časových řad. [online database]. [cit. 2009-05-24]. Available at: <http://www.cnb.cz/docs/ARADY/HTML/index.htm>
2. Czech hydrometeorological institute : Information about climate. [online database]. Updated 22.05.2009 [cit. 2009-05-24]. Available at: <http://www.chmu.cz/meteo/ok/infklim.html>
3. Council Directive 95/57/EC of 23 November 1995 on the collection of statistical information in the field of tourism [online]. Available at: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0057:en:HTML> [cit. 2009-02-28].
4. Eurostat : Harmonized indices of consumer prices (HICP) [online database]. Updated 07.05.2009. [cit. 2009-05-24]. Available at: <http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database>
5. Eurostat : Nights spent by residents and non-residents [online database]. Updated 29.04.2009. [cit. 2009-05-24]. Available at: <http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database>
6. Eurostat : Quarterly national accounts [online database]. Updated 05.02.2009. [cit. 2009-05-24]. Available at: <http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database>
7. GAZDA, V. VÝROST, T. ŽELINSKÝ, T. 2009. Ekonometria. Košice : Technická univerzita v Košiciach, 2009. 24 s.
8. GÚČIK, M. 2004. Krátky slovník cestovného ruchu. Banská Bystrica : Slovensko-švajčiarskej združenie pre rozvoj cestovného ruchu. 175s. ISBN 80-88945-73.
9. HATRÁK, M. 2007. Ekonometria. Bratislava : Iura edition, 2007. 504 s. ISBN 978-80-8078-150-7.
10. National Bank of Slovakia: Exchange rates archive [online database]. [cit. 2009-05-24]. Available at: <http://www.nbs.sk/sk/statisticke-udaje/kurzovy-listok/denny-kurzovy-listok-ecb>
11. R [freeware]. Available at: <http://www.r-project.org>
12. SORENSEN, N. 2004. Model Selection, Forecasting and Monthly Seasonality of Hotel Nights in Denmark. In: ERSA conference papers ersa04p92, European Regional Science Association. [on-line]. 2004, [cit. 2009-05-22], s. 6, 8. Available at: <http://www-sre.wu-wien.ac.at/ersa/ersaconfs/ersa04/PDF/92.pdf>

On the necessary condition for the bifurcation of a torus in a macroeconomic model

Katarína Makovínyiová, Rudolf Zimka

Department of Mathematics and Descriptive Geometry, Technical University in Zvolen, T. G. Masaryka 2117/24, SK-960 53 Zvolen, Slovakia
kmakovin@vsld.tuzvo.sk
Department of Quantitative Methods and Informatics, Faculty of Economics, Matej Bel University, Tajovského 10, SK-975 90 Banská Bystrica, Slovakia
rudolf.zimka@umb.sk

Abstract. In the paper we deal with a macroeconomic model of a small open economy. The model describes the development of output, capital stock, interest rate and money stock. Sufficient conditions for the existence of two pairs of purely imaginary eigenvalues in the linear approximation matrix of the model, which are necessary for the bifurcation of a torus, are found. A numerical example is presented.

Keywords: macroeconomic model, equilibrium, matrix of linear approximation, eigenvalues, torus.

1 Introduction

In [2] Makovínyiová K. and Zimka R. introduced a nonlinear four dimensional dynamic macroeconomic model of a small open economy. It describes the development of output, capital stock, interest rate and money stock in a small open economy. The model has the classical IS-LM structure. It generalizes Schinasi's three dimensional model of a closed economy [4] and Asada's three dimensional model of a small open economy [1].

The model has the form

$$\begin{aligned} \dot{Y} &= \alpha \left[I(Y,K,R) + G - S(Y^D,R) - T(Y) + J(Y,\rho) \right], \\ \dot{K} &= I(Y,K,R), \\ \dot{R} &= \beta \left[L(Y,R) - M \right], \\ \dot{M} &= J(Y,\rho) + \gamma \left(R - R_f \right), \end{aligned} \tag{1}$$

where Y is output (income), K is capital stock, R is interest rate of domestic country, M is money supply, I are investments, G are constant government expenditures, S are savings, T are tax collections, J is net export, ρ is exchange rate,

L is money demand, R_f is constant interest rate of foreign country, α, β, γ are positive parameters, t is time and

$$Y^D = Y - T(Y), \dot{Y} = \frac{dY}{dt}, \dot{K} = \frac{dK}{dt}, \dot{R} = \frac{dR}{dt}, \dot{M} = \frac{dM}{dt}.$$

The economic properties of the functions in (1) are expressed by the partial derivatives

$$\frac{\partial I(Y,K,R)}{\partial Y} > 0, \frac{\partial I(Y,K,R)}{\partial K} < 0, \frac{\partial I(Y,K,R)}{\partial R} < 0, \frac{\partial S(Y^D,R)}{\partial Y^D} > 0,$$
$$\frac{\partial S(Y^D,R)}{\partial R} > 0, \frac{\partial T(Y)}{\partial Y} > 0, \frac{\partial J(Y,\rho)}{\partial Y} < 0, \frac{\partial J(Y,\rho)}{\partial \rho} > 0,$$
$$\frac{\partial L(Y,R)}{\partial Y} > 0, \frac{\partial L(Y,R)}{\partial R} < 0.$$

Remark 1. Reasons for the assumption $\frac{\partial J(Y,\rho)}{\partial Y} < 0$ are given e.g. in [1] on p. 242.

Remark 2. In this paper we concentrate on the analysis of a "fixed price economy". It means that the price level p is given exogenously and normalized to the value 1 (therefore it doesn't appear in the model). The interest rate of foreign country, R_f, is also given exogenously because of the small economy.

The question of the existence of an equilibrium in model (1) under fixed exchange rate and stability of this equilibrium was investigated [2]. In [3] the existence of business cycles is proved. In this paper we are concerned with the necessary condition for the existence of tori in model (1) under fixed exchange rate.
Tori can only appear if the linear approximation matrix of model (1) has two pairs of purely imaginary eigenvalues. In the paper sufficient conditions for the existence of two pairs of purely imaginary eigenvalues in the linear approximation matrix of the model are found. A numerical example of the gained result is presented.

2 Analysis of the model

Assume the following form of the functions in model (1):

$$\begin{aligned} I(Y,K,R) &= f_1(Y) - i_2K - i_3R, \\ S(Y^D,R) &= f_2(Y^D) + s_3R, \\ T(Y) &= t_1Y - t_0, \\ L(Y,R) &= f_3(Y) - l_3R, \\ J(Y,\rho) &= J(Y),\ \rho \text{ is constant}, \end{aligned} \tag{2}$$

where $f_1(Y), f_2(Y^D), f_3(Y), J(Y)$ are nonlinear functions with respect to Y of the type C^1, $\frac{df_1(Y)}{dY} > 0$, $\frac{df_2(Y^D)}{dY^D} > 0$, $\frac{df_3(Y)}{dY} > 0$, $i_2, i_3, s_3, t_0, t_1, l_3$ are positive constants, $0 < t_1 < 1$. After substituting (2) into model (1) we get

$$\begin{aligned}
\dot{Y} &= \alpha \left[f_1(Y) - i_2K - i_3R + G - f_2(Y^D) - s_3R - t_1Y + t_0 + J(Y)\right], \\
\dot{K} &= f_1(Y) - i_2K - i_3R, \\
\dot{R} &= \beta \left[f_3(Y) - l_3R - M\right], \\
\dot{M} &= J(Y) + \gamma \left(R - R_f\right).
\end{aligned} \tag{3}$$

Let model (3) has the unique positive equilibrium $E^*(\gamma) = (Y^*(\gamma), K^*(\gamma), R^*(\gamma), M^*(\gamma))$, $Y^*(\gamma) > 0$, $K^*(\gamma) > 0$, $R^*(\gamma) > 0$, $M^*(\gamma) > 0$.

Remark 3. Sufficient conditions for the existence of the equilibrium $E^*(\gamma) = (Y^*(\gamma), K^*(\gamma), R^*(\gamma), M^*(\gamma))$ were found in [2].

Let us transform the equilibrium $E^*(\gamma)$ into the origin $E_1^* = (Y_1^*, K_1^*, R_1^*, M_1^*) = (0, 0, 0, 0)$ by shifting

$$Y_1 = Y - Y^*, \ K_1 = K - K^*, \ R_1 = R - R^*, \ M_1 = M - M^*.$$

Then model (3) obtains the form

$$\begin{aligned}
\dot{Y}_1 &= \alpha \left[f_1(Y_1 + Y^*) - f_2(Y_1^D + (Y^*)^D - t_0) + J(Y_1 + Y^*)\right] \\
&\quad + \alpha \left[-t_1Y_1 - i_2K_1 - (i_3 + s_3)R_1\right] \\
&\quad + \alpha \left[-t_1Y^* - i_2K^* - (i_3 + s_3)R^* + t_0 + G\right], \\
\dot{K}_1 &= f_1(Y_1 + Y^*) - i_2K_1 - i_3R_1 - i_2K^* - i_3R^*, \\
\dot{R}_1 &= \beta \left[f_3(Y_1 + Y^*) - l_3R_1 - M_1 - l_3R^* - M^*\right], \\
\dot{M}_1 &= J(Y_1 + Y^*) + \gamma \left(R_1 + R^* - R_f\right).
\end{aligned} \tag{4}$$

The Jacobian matrix $\mathbf{A} = \mathbf{A}(\alpha, \beta, \gamma)$ of model (4) at the equilibrium E_1^* is

$$\mathbf{A}(\alpha, \beta, \gamma) = \begin{pmatrix} -\alpha\ K & -\alpha\ i_2 & -\alpha(i_3 + s_3) & 0 \\ f_{1Y} & -i_2 & -i_3 & 0 \\ \beta f_{3Y} & 0 & -\beta l_3 & -\beta \\ J_Y & 0 & \gamma & 0 \end{pmatrix}. \tag{5}$$

where

$K = -f_{1Y} + f_{2Y} - J_Y + t_1$, $f_{1Y} = \frac{df_1(E^*)}{dY}$, $f_{2Y} = \frac{df_2((E^*)^D)}{dY^D}(1 - t_1)$, $f_{3Y} = \frac{df_3(E^*)}{dY}$, $J_Y = \frac{dJ(E^*)}{dY}$.

Definition. A triple (α_0, β_0, γ_0) of parameters α, β, γ in (4) is called the critical triple of model (4) if the matrix $\mathbf{A} = \mathbf{A}(\alpha_0, \beta_0, \gamma_0)$ has two pairs of purely imaginary eigenvalues $\lambda_{1,2} = \pm i\omega_1$, $\lambda_{3,4} = \pm i\omega_2$, $i = \sqrt{-1}$.

The eigenvalues of (5) are the roots of the characteristic equation of $\mathbf{A}(\alpha, \beta, \gamma)$

$$\lambda^4 + a_1(\alpha, \beta, \gamma)\lambda^3 + a_2(\alpha, \beta, \gamma)\lambda^2 + a_3(\alpha, \beta, \gamma)\lambda + a_4(\alpha, \beta, \gamma) = 0, \tag{6}$$

where

$$\begin{aligned}
a_1 &= \alpha K + \beta l_3 + i_2, \\
a_2 &= \alpha\beta(l_3 K + f_{3Y}(i_3 + s_3)) + \alpha i_2(K + f_{1Y}) + \beta(\gamma + i_2 l_3), \\
a_3 &= \beta(\alpha\gamma K + \alpha(i_2 l_3(K + f_{1Y}) + i_2 s_3 f_{3Y} - J_Y(i_3 + s_3)) + \gamma i_2), \\
a_4 &= \alpha\beta\gamma i_2(K + f_{1Y}) - \alpha\beta i_2 s_3 J_Y.
\end{aligned}$$

Lemma. *Characteristic equation (6) has two pairs of purely imaginary roots if and only if*

$$\begin{aligned}
&(A.1) \quad a_1 = 0, \\
&(A.2) \quad a_2 > 0, \\
&(A.3) \quad a_3 = 0, \\
&(A.4) \quad a_4 > 0, \\
&(A.5) \quad a_2^2 - 4a_4 \geq 0.
\end{aligned}$$

Proof. Suppose the matrix $\mathbf{A} = \mathbf{A}(\alpha_0, \beta_0, \gamma_0)$ has two pairs of purely imaginary eigenvalues $\lambda_{1,2} = \pm \mathrm{i}\,\omega_1$, $\lambda_{3,4} = \pm \mathrm{i}\,\omega_2$, $\mathrm{i} = \sqrt{-1}$. Then characteristic equation (6) can be written in the form

$$(\lambda - \mathrm{i}\,\omega_1)(\lambda + \mathrm{i}\,\omega_1)(\lambda - \mathrm{i}\,\omega_2)(\lambda + \mathrm{i}\,\omega_2) = 0,$$

what gives

$$\lambda^4 + (\omega_1^2 + \omega_2^2)\,\lambda^2 + \omega_1^2\omega_2^2 = 0.$$

After comparing the last equation with (6) we have

$$\begin{aligned} a_1(\alpha_0,\beta_0,\gamma_0) &= 0, \\ a_2(\alpha_0,\beta_0,\gamma_0) &= \omega_1^2+\omega_2^2>0, \\ a_3(\alpha_0,\beta_0,\gamma_0) &= 0, \\ a_4(\alpha_0,\beta_0,\gamma_0) &= \omega_1^2\,\omega_2^2>0. \end{aligned}$$

By eliminating ω_2 we get from the second and the fourth equation

$$\omega_1^4-a_2\,\omega_1^2+a_4=0.$$

After substitution $z=\omega_1^2$ it goes to the quadratic equation

$$z^2-a_2\,z+a_4=0.$$

From here

$$z_{1,2}=\frac{a_2\pm\sqrt{a_2^2-4a_4}}{2}.$$

As $\omega_{1,2}$ are real numbers, $z_{1,2}$ have to be real and positive. This is fulfilled for $a_2^2-4a_4\geq 0$.

Conversely suppose the conditions $(A.1)-(A.5)$ hold. Then equation (6) has the form

$$\lambda^4+a_2\,\lambda^2+a_4=0.$$

Define $z=\lambda^2$. Then we have

$$z^2+a_2\,z+a_4=0,$$

$$z_{1,2}=\frac{-a_2\pm\sqrt{a_2^2-4a_4}}{2}.$$

As $a_4>0$ and $a_2^2-4a_4\geq 0$, we get $z_{1,2}$ are negative. Therefore

$$\lambda_{1,2}=\pm\sqrt{z_1},\ \ \lambda_{3,4}=\pm\sqrt{z_2}$$

are purely imaginary roots of characteristic equation (6). □

Let us denote

$$L=\gamma K+i_2l_3(K+f_{1Y})+i_2s_3f_{3Y}-J_Y(i_3+s_3).$$

Theorem. *Let* $i_2l_3-s_3\geq 0$ *and* $\gamma=\gamma_0$ *be such that*
(i) $l_3K+f_{3Y}(i_3+s_3)>0$*;*
(ii) $L<0$.
Then there exist the values α_0 *and* β_0 *such that the triple* $(\alpha_0,\beta_0,\gamma_0)$ *is the critical triple of model (4).*

Proof. Choose constants in model (4) and the value γ_0 such that they meet the conditions of Theorem. We will show that the conditions of Lemma are satisfied.

Condition (A.4) is always fulfilled.

Condition (A.3) requires

$$\beta(\alpha\gamma_0 K + \alpha(i_2 l_3(K + f_{1Y}) + i_2 s_3 f_{3Y} - J_Y(i_3 + s_3)) + \gamma_0 i_2) = 0.$$

As

$$\alpha(\gamma_0 K + i_2 l_3(K + f_{1Y}) + i_2 s_3 f_{3Y} - J_Y(i_3 + s_3)) = \alpha L = -\gamma_0 i_2,$$

and L is negative, relation (A.3) uniquely determines the positive value of parameter α,

$$\alpha_0 = -\frac{\gamma_0 i_2}{L}.$$

Condition (A.1) means

$$\alpha_0 K + \beta l_3 + i_2 = 0.$$

This equation uniquely determines the value of parameter β,

$$\beta_0 = -\frac{\alpha_0 K + i_2}{l_3}.$$

As

$$-\frac{\alpha_0 K + i_2}{l_3} = -\frac{-\frac{\gamma_0 i_2}{L} K + i_2}{l_3} = -\frac{i_2(L - \gamma_0 K)}{l_3 L}$$
$$= -\frac{i_2(i_2 l_3(K + f_{1Y}) + i_2 s_3 f_{3Y} - J_Y(i_3 + s_3))}{l_3 L} > 0,$$

value β_0 is positive.

Condition (A.2) is

$$\alpha_0\beta_0(l_3 K + f_{3Y}(i_3 + s_3)) + \alpha_0 i_2(K + f_{1Y}) + \beta_0(\gamma_0 + i_2 l_3) > 0.$$

This inequality is satisfied according to the condition (i) in Theorem.
Let us analyze condition (A.5):

$$(\alpha_0\beta_0(l_3 K + f_{3Y}(i_3 + s_3)) + \alpha_0 i_2(K + f_{1Y}) + \beta_0(\gamma_0 + i_2 l_3))^2$$
$$-4(\alpha_0\beta_0\gamma_0 i_2(K + f_{1Y}) - \alpha_0\beta_0 i_2 s_3 J_Y) \geq 0.$$

Utilizing the fact that

$$(a + b)^2 \geq 4ab$$

and condition (i) in Theorem, we can write

$$
\begin{aligned}
&(\alpha_0\beta_0(l_3K + f_{3Y}(i_3 + s_3)) + \alpha_0 i_2(K + f_{1Y}) + \beta_0(\gamma_0 + i_2 l_3))^2 \\
&-4(\alpha_0\beta_0\gamma_0 i_2(K + f_{1Y}) - \alpha_0\beta_0 i_2 s_3 J_Y) \\
&\geq 4(\alpha_0\beta_0(l_3K + f_{3Y}(i_3 + s_3)) + \alpha_0 i_2(K + f_{1Y}))\beta_0(\gamma_0 + i_2 l_3) \\
&-4(\alpha_0\beta_0\gamma_0 i_2(K + f_{1Y}) - \alpha_0\beta_0 i_2 s_3 J_Y) \\
&\geq 4\alpha_0\beta_0 i_2(K + f_{1Y})(\gamma_0 + i_2 l_3) \\
&-4(\alpha_0\beta_0\gamma_0 i_2(K + f_{1Y}) - \alpha_0\beta_0 i_2 s_3 J_Y) \\
&= 4\alpha_0\beta_0\gamma_0 i_2(K + f_{1Y}) + 4\alpha_0\beta_0 i_2^2 l_3(K + f_{1Y}) \\
&-4(\alpha_0\beta_0\gamma_0 i_2(K + f_{1Y}) - \alpha_0\beta_0 i_2 s_3 J_Y) \\
&= 4\alpha_0\beta_0 i_2(i_2 l_3(K + f_{1Y}) + s_3 J_Y).
\end{aligned}
$$

The last ordering is nonnegative under the assumption

$$i_2 l_3 - s_3 \geq 0.$$

The proof is completed. □

3 Numerical example

Consider the following functions:

$$
\begin{aligned}
I &= 0.1\sqrt{Y^3} - 10K - 200R + 150.79892, \\
S &= 0.8\sqrt{0.7Y + 4.2} + 100R - 1.35946, \\
T &= 0.19Y, \\
L &= 10\sqrt{Y} - 10R - 72.900054, \\
J &= -0.075\sqrt{Y} + 6,
\end{aligned}
$$

and assume the constants $G = 12, R_f = 0.001$.

Then, model (3) has the form

$$\begin{aligned}
\dot{Y} &= \alpha\left(\frac{1}{10}\sqrt{Y^3} - \frac{19}{100}Y - \frac{3}{40}\sqrt{Y} - \frac{4}{5}\sqrt{\frac{7}{10}Y + \frac{21}{5}} - 10K - 300R\right) \\
&\quad + \alpha\,\frac{8507919}{50000}, \\
\dot{K} &= \frac{1}{10}\sqrt{Y^3} - 10K - 200R + \frac{3769973}{25000}, \\
\dot{R} &= \beta\left(10\sqrt{Y} - 10R - M - \frac{36450027}{500000}\right), \\
\dot{M} &= -\frac{3}{40}\sqrt{Y} + \gamma(R - \frac{1}{100}) + 6.
\end{aligned} \tag{7}$$

The equilibrium E^* of (7) depends on parameter γ. For $\gamma_0 = 10^6$ the equilibrium is

$$(Y^* = 64, K^* = 20, R^* = 0.0099946, M^* = 7).$$

Let us verify the conditions of Theorem.

1) $i_2 l_3 - s_3 = 10.10 - 100 = 0$;

2) $l_3 K + f_{3Y}(i_3 + s_3) = 10\,(-6/5 + 1/25 - (-3/640) + 19/100) + 5/8(200 + 100) = 10\,(-3089/3200) + (5/8).300 = 56911/320 > 0$;

3) $L = 10^6\,(-3089/3200) + 10.10\,(-3089/3200 + 6/5) + 10.100.5/8 + (3/640)\,(200 + 100) = -7717301/8 < 0$;

Thus, all conditions are satisfied.

The critical values of parameters α, β are

$$\alpha_0 = -\frac{\gamma_0 i_2}{L} = -\frac{10^6.10}{-\frac{7717301}{8}} = \frac{80000000}{7717301},$$

$$\beta_0 = -\frac{\alpha_0 K + i_2}{l_3} = -\frac{\frac{80000000}{7717301}.(-\frac{3089}{3200}) + 10}{10} = \frac{5199}{7717301}.$$

Hence the critical triple is

$$(\alpha_0, \beta_0, \gamma_0) = \left(\frac{8}{7717301}.10^7\ ,\ \frac{5199}{7717301}\ ,\ 10^6\right).$$

Matrix of linear approximation at $(\alpha_0, \beta_0, \gamma_0)$ has the form

$$\mathbf{A} = \begin{pmatrix} \frac{77225000}{7717301} & -\frac{800000000}{7717301} & -\frac{24000000000}{7717301} & 0 \\ \frac{6}{5} & -10 & -200 & 0 \\ \frac{25995}{61738408} & 0 & -\frac{51990}{7717301} & -\frac{5199}{7717301} \\ -\frac{3}{640} & 0 & 1000000 & 0 \end{pmatrix}.$$

Matrix A has two pairs of purely imaginary eigenvalues

$$\begin{aligned} \lambda_1 &= 25.9815i, \\ \lambda_2 &= -25.9815i, \\ \lambda_3 &= 4.92743i, \\ \lambda_4 &= -4.92743i. \end{aligned}$$

This numerical result confirms the assertion of Theorem.

4 Conclusion

Model (1) is a dynamic four dimensional model of a small open economy under fixed exchange rate. Asada in [1] also introduced and analyzed a dynamic model of a small open economy under fixed exchange rate. His model is three dimensional one and describes the development of output, capital stock and money stock. The major difference between Asada's model and our model is in the view on the development of interest rate R. Asada assumes that the adjustment speed β in the dynamic adjustment equation for interest rate $\dot{R} = \beta(L(Y, R) - M)$ is infinite so that the interest rate is adjusted instantaneously to preserve the equilibrium $M = L(Y, R)$ on money market. This assumption is rather strong and not very real. In our model (1) the dynamic adjustment equation for the rate of interest is present. It enlarges the dimension of the model, what makes it more complex for analysis, but on the other side it enables to investigate the character of the development of variables in more detail. The four dimensional model (1) gives the possibility to study not only the question of the existence of business cycles, what was done in [3], but also to find out whether more complicated relations among output, capital stock, interest rate and money stock can happen. For example it is possible to study whether tori can arise around equilibrium. The necessary condition for arise of tori in model (1) is the existence of two pairs of purely imaginary eigenvalues in the linear approximation

matrix of the model. The theorem in the paper gives sufficient conditions for fulfilling this necessary condition. To solve completely the question of the existence of tori it would be necessary to derive the bifurcation equation of the model and to study its structure. This step is open and waits for realization.

Acknowledgment: The research was supported by the Slovak Grant Agency VEGA No. 1/4633/07.

References

1. Asada T.: Kaldorian Dynamics in an Open Economy. J. Econ., Vol. 62, No.3, pp. 239--269 (1995)
2. Makovínyiová K., Zimka R.: On stability in generalized Schinasi's macroeconomic model under fixed exchange rates. Tatra Mt. Math. Publ. 43, pp. 1--8 (2009)
3. Makovínyiová K., Zimka R.: On Kaldorian Dynamics in Generalized Schinasi's Macroeconomic Model under Fixed Exchange Rates. J. Differ. Equations, (submitted for publication)
4. Schinasi G. J.: Fluctuations in a Dynamic, Intermediate-Run IS-LM Model: Applications of the Poincaré-Bendixon Theorem. J. Econ. Theory 28, pp. 369--375 (1982)

Nonparametric Estimates of Survival Function from Interval Censored Observations

Ivana Malá

The University of Economics in Prague, W.Churchilla 4, Prague 3, Czech Republic
malai@vse.cz

Abstract. In the paper estimates of a survival function of a time to event random variable are discussed for data containing only interval censored values. In practice these data are usually transformed to datasets with exact and right censored data and well implemented methods in statistical software are used for estimation. In the paper four possible transformations are used and their properties are compared with results of proper methods for interval censored data. Kaplan–Meier and maximum likelihood estimates of survival function are computed. Only small samples are studied by simulation study for lognormal distribution.

Keywords: right censored data, interval censored data, Kaplan–Meier method, maximum likelihood estimate

1 Introduction

In the paper T is a time to event random variable. In survival analysis time of the occurrence of the event is known exactly or it is censored. For example for right censored time it is only known that the event occurs after a known time. In this paper no exact times are observed. The time of the occurrence of the event of interest is known only to occur in a time interval $(L, R\rangle$, $(L < T \leq R)$ – observation is interval censored or to occur after the end of the interval $(L < R < T < \infty)$ – observation is right censored at time R. This type of data frequently occurs in medical studies, where investigated variable could be time to the relapse of illness after surgery or time to the occurrence of complications. Suppose that patients are examined only at pre-scheduled visits and the occurrence of the event is observed only by these visits. Variable of this particular type is not frequently investigated time to deaths, because in such a case the exact value of time of death is known, if it occurs. Other examples are surveys, where respondents are visited and interviewed repeatedly, for example the Labour Force Sample Survey organized by the Czech Statistical Office, where respondents are interviewed in three months intervals. Then unemployment duration is right censored (if the unemployed didn't find a job) or interval censored (if the unemployed found a job at a three-month interval).

Suppose that the estimation of the survival function of T is of interest. Characteristics such as median time to event, residual median time to event etc. can be derived from the estimated survival function.

The estimates are based on a sample of a size n. It means for all $i = 1,.., n$ we have an interval $(L, R\rangle$ and information whether the event occurred in this interval or didn´t occur up to the end of this interval. Datasets of this type are denoted by I in this paper. For this type of data methods developed for interval censored data should be used. But unfortunately these procedures are complicated and they are not well implemented in widely used statistical packages. On the other hand the estimation of survival function from data with exact observations and right censored observations is known to routine users. Procedures developed for right censored data (especially Kaplan–Meier nonparametric estimate) are known and readily applicable with the use of standard statistical software. A widely used approach in the problem of interval censored data is to transform interval censored data to data that contains only exact and right censored observations and then to use known and easy to apply methods for right censored data. Usually centres of intervals $(L, R\rangle$ are used as exact observations for intervals, where the event occurred and right censored observations remain to be right censored at time point R (dataset III). In the paper three more possibilities how to find exact times for the occurrence of the event are random time points with the uniform distribution on the interval (L,R) (dataset IV), value L (dataset II) and value R in the dataset V. These modifications are summarized as

I. All observations censored, interval $(L_i, R_i\rangle$ or right censored at R_i

II. Interval censored observations exact at the time L_i, right censored observations censored at R_i

III. Interval censored exact at the time $(L_i + R_i) / 2$, right censored observations censored at R_i

IV. Interval censored observations exact at the time with distribution Uniform (L_i,R_i), right censored observations censored at R_i

V. Interval censored observations exact at the time R_i, right censored observations censored at R_i.

In the paper properties of these five approaches are illustrated by a simulation in example 1. Data were created from lognormal distribution. This distribution has positive skewness and its hazard function has an extreme – maximum. In an example 2 results for one selected dataset I (and for its modifications II-V) from the simulation are taken and results are shown and discussed in detail. In the simulation study the distribution of time to event variable T is known and parametric methods, that estimate parameters of the distribution, can be used. Otherwise a proper distribution has to be guessed and a good choice is crucial. The impact of bad choice of the theoretical distribution is not of interest in the paper.

Suppose sample sizes n are relatively small (less than 100), where it is not possible to use large sample properties of estimates. Another complication is usually large number of right censored data. But this situation is frequently met in medical research and approximation III is usually used in order to obtain estimates.

2 Methods

Let T be a positive value random variable with continuous distribution. Denote by $F(t)$ distribution function of T and $S(t)$ survival function of T defined as

$$S(t) = P(T>t)=1-F(t). \tag{1}$$

Suppose a sample of n observations is used for the estimation. Kaplan–Meier method was derived for data with exact and right censored values of T. In this case we put for exact values of time to event T censor variable C=1 and for right censored values at time point T censor variable C=0. For each observation we have a pair (T,C). Suppose that events occurred at k distinct ordered time points $t_1 < t_2 < .. < t_k$. Then Kaplan–Meier product–limit estimator of the survival function S for right censored data is defined as

$$\begin{aligned} \hat{S}(t) &= 1 && t < t_1, \\ &= \prod_{t_i \le t} \left(1-\frac{d_i}{Y_i}\right) && t \ge t_1, \end{aligned} \tag{2}$$

where $t > 0$, [1]. In the formula (2) for $j = 1,..,k$

$Y_0 = n$ (number of observations in the sample),

Y_i = number of individuals at risk at time point t_i (it means number of observations with $T \ge t_i$),

d_i = number of events at the time point t_i.

This estimator for right censored data can be modified for interval censored data [1],[3]. In this case it is not possible to construct such simple procedure. For each individual the event of interest occurred in an interval $(L_i, R_i\rangle$ (an observation is interval censored) or it didn´t occur until R_i (an observation is right censored at time point R_i). Right censored observation can be incorporated as interval censored (R_i,∞). Denote by τ_j ($j = 0,.., k$) ordered time points $(0 = \tau_0 < \tau_1 < .. < \tau_k = \infty)$ which include all time points L_i a R_i, $i = 1,..,n$. The estimate $\hat{S}(t)$ of $S(t)$ is supposed to be constant over intervals $(\tau_{j-1}, \tau_j\rangle$. Denote (for $j = 1,.., k$) by p_j a probability that the event occurs in the interval $(\tau_{j-1}, \tau_j\rangle$, this means

$$p_j = P(\tau_{j-1} < T \le \tau_j) = F(\tau_j) - F(\tau_{j-1}), \sum_{j=1}^{k} p_j = 1. \tag{3}$$

The procedure begins with an initial guess (first approximation) $\hat{p}_j$ of values p_j or with first approximation of $S(t)$

$$\hat{S}(t) = \sum_{t<\tau_{j-1}} \hat{p}_j. \tag{4}$$

Define (0-1) numbers α_{ij} by the formula

$$\begin{aligned} \alpha_{ij} &= 1 \text{ if } (\tau_{j-1}, \tau_j \rangle \subseteq (L_i, R_i \rangle, \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{5}$$

for $i = 1, ., n$ and $j = 1, ., k$.
Number of events which occur at time point τ_j is estimated by

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} \hat{p}_j}{\sum_{l=1}^{k} \alpha_{il} \hat{p}_l}, \tag{6}$$

and estimated number of individuals at risk is

$$Y_j = \sum_{l=j}^{k} d_l. \tag{7}$$

For values Y_j and d_j the Kaplan–Meier estimator (2) is computed in order to receive new estimate of $\hat{S}(t)$. New values of $\hat{p}_j$ are then calculated. This iterative procedure is repeated until no changes in values $\hat{p}_j$ are observed and so a final solution – nonparametric estimator of $S(t)$ is found.

In the simulation study nonparametric estimates are compared with parametric estimates obtained by maximum likelihood method. In the likelihood function $L(\boldsymbol{\theta})$ ($\boldsymbol{\theta}$ is a vector of parameters of the distribution) censored data are incorporated by the term

$$S(R) = 1 - F(R), \tag{8}$$

and interval censored data by

$$S(L) - S(R) = F(R) - F(L). \tag{9}$$

In our problem likelihood function to be maximized has a formula

$$L(\boldsymbol{\theta}) = \prod_{i\text{- right censored}} S(R_i) \prod_{i\text{- interval censored}} (S(L_i) - S(R_i)), \tag{10}$$

where S is a survival function of the underlying distribution. For all computations in the paper procedures for survival analysis and for nonlinear optimization in SPlus 6.2 were used.

3 Illustrations

Example 1. (Simulation study): In this part the estimates of median time to the event will be compared for all methods. In [2] samples of different sizes were investigated, in this paper only small samples are of interest. 100 samples were drawn from lognormal distribution with parameters $\mu = 4.11$ and $\sigma = 1$. This distribution has the mean 100 ($E(T) = 100$) and the median 61 ($t_{0.5} = 61$). The interval censored data were constructed with the use of sums of independent Poisson variables with means (parameter lambda) 15 and 30. Lambda is then the mean length of intervals (L,R). A sample of n values from mentioned lognormal distribution was generated. Values of L and R were found as nearest smaller (L) and higher (R) than T in consecutive values of sums of Poisson´s variables. In the data no exact values of T are present, proportions of right censored data were selected 0.1 and 0.3 (in the mean) for sample sizes 30 and 50 and 0.3 for the sample size 100.

In the study we know the exact distribution, maximum likelihood fit can be used for the estimation of parameters μ and σ ($\boldsymbol{\theta}' = (\mu, \sigma)$) and then for the estimation of median. For the lognormal distribution formulas (8) and (9) are easily rewritten as

$$1 - \Phi\left(\frac{\ln(R) - 4.11}{1}\right) \text{ and } \Phi\left(\frac{\ln(R) - 4.11}{1}\right) - \Phi\left(\frac{\ln(L) - 4.11}{1}\right), \tag{11}$$

where Φ is a distribution function of standard normal distribution. Results of estimated medians are given in the Table 1. In the table in the column % the mean percentage of right censored observation is given to 10 or 30. Kaplan–Meier estimates of medians are given in the Table 2. For interval censored dataset I iterative procedure mentioned above was used, for modified datasets II–V standard Kaplan–Meier procedure was used. In both tables mean values of estimated medians are shown with sample standard deviations in parenthesis.

From Tables 1 and 2 we can see relatively poor estimation of median. But it has to be expected because of small sample sizes, no proper exact values and large number of right censored data. As it is expected, the smallest medians are from dataset II and the largest from data V. Datasets III (the most frequently used approximation of the occurrence times by the centres of censoring intervals) and IV (random times) give similar results. Medians from almost all samples are overestimated, sometimes strongly. It is with the exception of datasets II with exact values of T equal to the lower end of the censoring interval.

Table 1. Maximum likelihood estimates, lognormal distribution fitted

n	%	λ	mean median (standard error) ($t_{0.5}$= 61)				
			I	II	III	IV	V
30	10	30	64.5 (11.1)	48.7 (11.6)	65.5 (11.4)	64.5 (11.6)	76.7 (11.6)
30	10	15	64.2 (12.5)	33.2 (12.4)	66.4 (12.1)	63.8 (13.0)	86.6 (12.2)
30	30	30	81.3 (13.8)	75.2 (16.6)	85.7 (14.4)	85.7 (14.3)	95.1 (14.4)
30	30	15	72.9 (9.9)	53.9 (15.0)	80.6 (8.7)	79.2 (10.7)	97.8 (9.6)
50	10	30	77.6 (14.7)	70.3 (10.7)	88.0 (15.7)	88.1 (15.9)	97.,2 (15.2)
50	10	15	64.4 (8.5)	32.5 (8.1)	66.8 (8.1)	63.4 (8.9)	87.2 (8.3)
50	30	30	89.3 (9.9)	72.2 (12.5)	83.6 (10.6)	83.5 (10.8)	90.8 (10.8)
50	30	15	85.6 (9.0)	55.4 (12.5)	83.4 (9.4)	81.8 (10.2)	92.8 (9.1)
100	30	30	80.8 (9.5)	73.4 (11.4)	85.1 (10.7)	85.2 (10.6)	88.2 (10.3)
100	30	15	77.7 (8.7)	57.2 (11.8)	85.4 (9.9)	84.3 (10.5)	(92.8 (11.7)

Shorter censoring intervals give slightly better results; the dependence on the percentage of right censored observations is not surprisingly visible. In the study original and correct probability distribution was fitted and according to the theory parametric estimates should be superior to nonparametric estimates. It seems to be true in mean values and even smaller standard errors.

Example 2. For this example one dataset from the simulation in example 1 was used. Number of observations is 50 and mean percentage of right censored data is 30. The lengths of the intervals $(L, R\rangle$ are from 23 to 44 with the mean 31.1. It corresponds to the chosen parameter λ = 30. Data contains 39 out of 50 interval censored observations and 11 (22 percent) right censored observations instead of 15 (30 percent of 50 observations). Table 3 presents maximum likelihood estimates of parameters μ and σ for all datasets I–V and estimated medians $\exp(\hat{\mu})$. In the Figure 1 survival functions with these parameters are shown with the horizontal line with $S(t)$ = 0.5. Sample medians are then on the t–axis. In the Table 4 estimated values of medians from Kaplan–Meier method are given with estimated 95% confidence intervals. Figure 2 shows theoretical survival function and Kaplan–Meier estimate of survival function for interval censored data I. Large confidence intervals for medians are shown in the Table 4. In the Figure 1 similar survival functions from datasets I to IV are shown. Survival function for data V seems to be different (shifted to the right). The coincidence with theoretical curve is good for probabilities approximately from 1

to 0.4, and then values of survival function from censored data are remarkably higher than those for lognormal distribution. Similar situation is in the Figure 2.

Table 2. Kaplan-Meier estimates

n	%	λ	mean median (standard error)				
			I	II	III	IV	V
30	10	30	64.9 (12.0)	60.52 (15.59)	67.85 (15.39)	67.5 (15.21)	74.6 (15.4)
30	10	15	67.2 (14,1)	47.96 (15.72)	62.60 (15.63)	63.1 (14.08)	76.23 (15.87)
30	30	30	88.5 (15.20	77.63 (15.66)	84.27 (16.23)	83.58 (16.04)	87.72 (15.15)
30	30	15	82.4 (12.7)	67.30 (12.38)	77.87 (11.73)	80.3 (13.93)	89.0 (13.46)
50	10	30	84.2 (13.1)	72.16 (14.2)	77.24 (14.06)	77.4 (15.18)	81.58 (12.72)
50	10	15	65 (12.8)	51.29 (13.29)	66.60 (13.09)	67.3 (13.24)	80.5 (12.69)
50	30	30	82.23 (13.4)	78.62 (17.75)	84.64 (17.67)	85.0 (18.56)	88.8 (16.18)
50	30	15	81.00 (10.04)	68.72 (13.70)	79.75 (13.92)	80.6 (13.84)	87.1 (10.41)
100	30	30	83.3 (9.91)	74.68 (10.70)	80.15 (11.37)	79.4 (10.77)	84.4 (10.74)
100	30	15	82.9 (9.94)	74.75 (12.73)	86.55 (12.46)	84.5 (10.14)	93.9 (10.16)

Table 3. Maximum likelihood estimates of the parameters for lognormal distribution and estimated values of medians

I		II		III		IV		V	
$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
4.20	1.2	4.07	1.25	4.25	1.35	4.27	1.34	4.52	1.08
66.68		58.56		70.10		71.52		91.84	

Table 4. Kaplan–Meier estimates of medians with 95% confidence intervals

I	II	III	IV	V
67.15	61	64.6	75.5	90.00
(41,160)	(27,163)	(38,171)	(43,179)	(60,194)

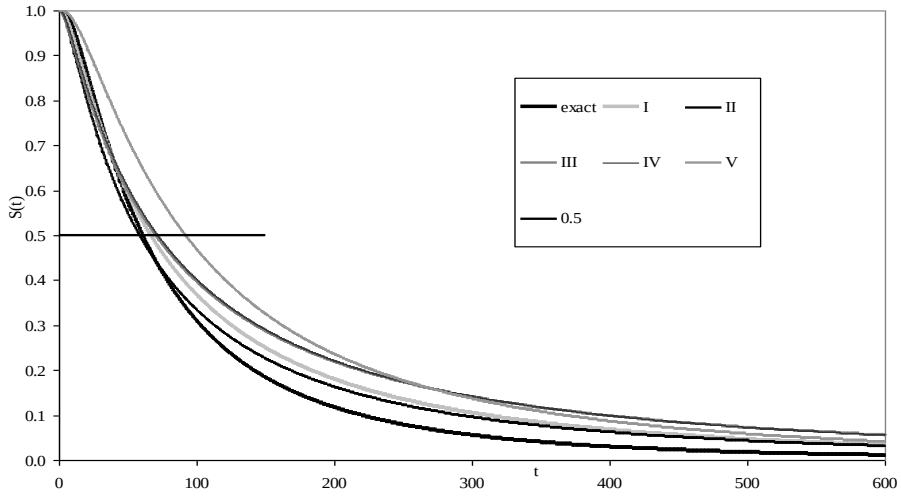


Fig. 1. Parametric estimates of survival function based on parametric maximum likelihood estimates of the parameters

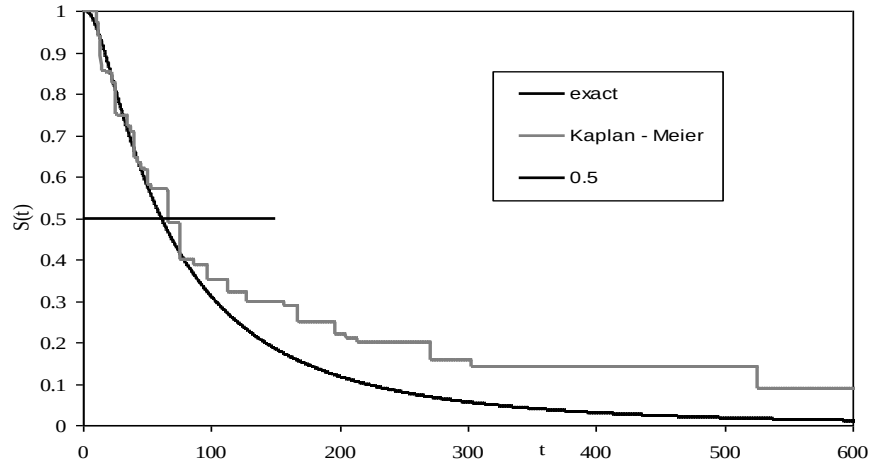


Fig. 2. Survival function of lognormal distribution (exact) and nonparametric estimation of *S* for interval censored data (Kaplan–Meier)

4 Conclusions

In [2] detailed discussion is given for approximation of data I by transformation III for Weibull distribution with respect to different values of its parameters. In the simulation in this paper all estimates for the data of treated types are rather

unsatisfactory and it makes user to be very careful in interpretation. The use of methods for interval censored data should be preferred, if it is possible to perform proper estimation. The approximations used in the paper in order to avoid the use of methods for interval censored data can affect properties of estimates. According to literature (for example [2]) the use of modified data can lead to misleading or even wrong results. But in the case of small sample the use of simpler procedures for modifications III and IV doesn´t result in remarkably worse estimates. The use of random times instead of centres of interval does not seem to give remarkably better results.

References

1. Klein J.P., Moeschberger, M.L.: Survival Analysis, Techniques for Censored and Truncated Data. Springer, (1998). ISBN 387–948295.
2. Odell,P.M., Anderson,K.M., D´Agostino,R.B.: Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull–Based Accelerated Failure Time Model. Biometrics 48, pp. 951--959,(1992)
3. Meel Pyng Ng: A Modification of Peto´s Nonparametric Estimation of Survival Curves for Interval-censored Data. Biometrics 58, pp.439--442, (2003)

Modelling of mortality in Czech Republic[1]

Petr Mazouch

University of Economics, Prague, W. Churchill sq. 4, Prague, Czech Republic
mazouchp@vse.cz

Abstract. Many insurance companies, pension funds or government are interested in mortality progression. They are using several methods how to predict mortality, for example forecasting of life expectancy or any other difficulty models of modelling mortality rates for different age groups. I would like to explain alternative way to modelling and predicting of mortality. This modelling will not be based on predicting mortality rates, but modelling proportion or difference mortality rates of two neighbouring age groups. The proportion is decrease in time to one and the difference decrease to zero. The target of this theory is predict the time, when the proportion will reach the one or difference will reach zero respectively.

Keywords: Mortality, age groups, predictions, life expectancy

1 Introduction

Many insurance companies, pension funds or government are interested in mortality progression. They are using several methods how to predict mortality, for example forecasting of life expectancy or any other difficulty models of modelling mortality rates for different age groups. I would like to explain alternative way to modelling and predicting of mortality. This modelling will not be based on predicting mortality rates, but modelling proportion or difference mortality rates of two neighbouring age groups. The proportion is decrease in time to one and the difference decrease to zero. The target of this theory is predicting the time, when the proportion will reach one or the difference will reach null respectively.

It is very important for the society to know about its mortality. The whole mortality is not so much important as the mortality of each age groups. The development of the mortality gives us information about increasing or decreasing of life condition of these age groups.
Many companies, the same as insurance companies, are interested in researching of mortality of each age group, because the progress of the mortality has impact on their economy. As they have to be prepared for the future development, the companies use

[1] This paper is written under the support of the National Research Programme II of the Ministry of the Education, Young and Sports of the Czech Republic, Project. No 2D06026 „Reproduction of the Human Capital“ and University of economics, Prague, No IGA 27/08

many models for the forecast. Mortality in the time period decrease that means that conditions are improving. This we can also see on the increasing life expectancy. If we focus on mortality of one age group in ling time, we obtain time series. Theoretically we can say that this time series has decreasing trend. Fundamental of the model is to find the trend.
First problem is becoming in the moment, when we really get individually rate of mortality for one age and make time series of them. We find out, that this time series has high volatility and it is not possible to say that mortality in year *t* is every time higher than in *t+1*.

$$m_{t,x} \geq m_{t+1,x} \quad (1)$$

Volatility is cause by uniquely death in young ages (group of dead people is not large, even small deviation can make important change in mortality rate), and in higher ages epidemic of diseases, which are only in some years.
Sometimes we can see the whole time period of constant mortality or slightly increasing. Decreasing trend is therefore very slow and it is not possible to estimate the function. If we estimate this function, than there is one more risk. It is easily possible that the mortality estimated to the future for the age *x* is higher than for the age *x+1*.

$$m_{t,x} \geq m_{t,x+1} \quad (2)$$

This situation may happen (and really happened) in some years, but the development cannot be estimated in long period (we think only about ages 30+).
Mortality is observed in our country almost for one hundred years. For the future forecast is not relevant to use the whole period, because the development in the history was interrupted by some external effects, which can distort the relation of the model.

2 Aim and methodology

As I mention the risks above, it is not possible to make models of mortality for the individual age groups, as it is not obey the situation defined in (2). In case, we express the relationship in form of difference or proportion of mortality of consecutive age groups in one year.

$$m_{t,x+1} - m_{t,x} \quad (3)$$

or

$$\frac{m_{t,x+1}}{m_{t,x}} \quad . \tag{4}$$

Therefore these values are going to 0, resp. 1 in limit in the time.

$$\lim_{t\to\infty}\left(m_{t,x+1} - m_{t,x}\right) = 0 \quad , \tag{5}$$

and

$$\lim_{t\to\infty}\left(\frac{m_{t,x+1}}{m_{t,x}}\right) = 1 \quad . \tag{6}$$

We are going to model difference or portion of two rates of mortality in the time and by keeping the condition (5) and (6) it can not happen situation (2).
This theory is acknowledge already in younger ages (about 30 years), when the mortality is very low and between the ages are not differences (decrease between the ages very slightly).
Theoretically it would be enough to find out rate of convergence of mortality in the age *x+1* to the mortality in the age *x*.
If we look on graphs that show differences and portions between the rates of mortality in the time it is clear that from the point of predictability it is preferable use differences between mortality rates, which is relation (3).

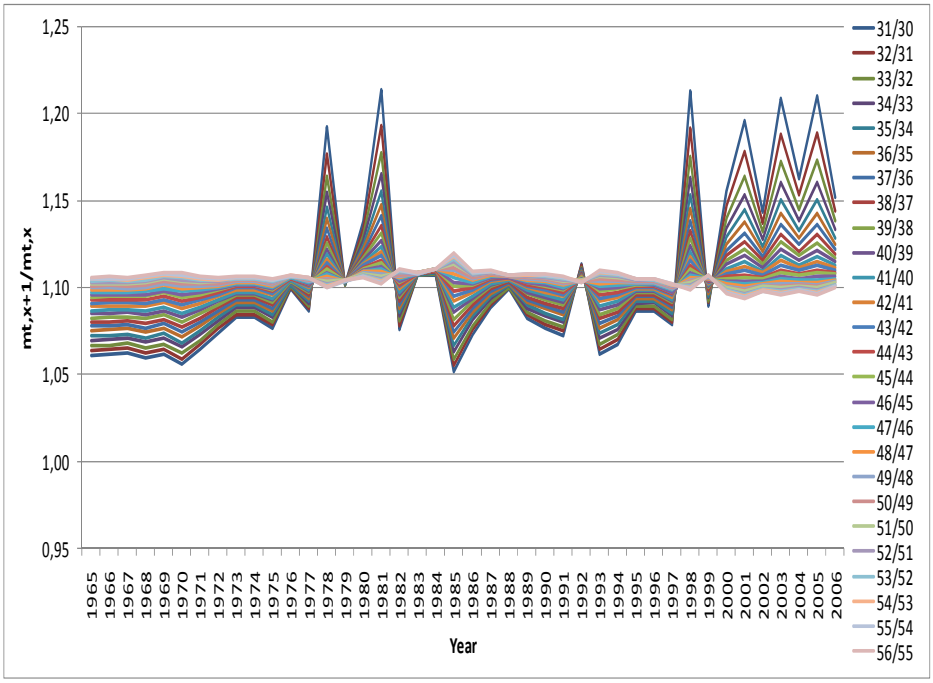


Fig. 1. Ratio of mortality of age x+1 into mortality of age x, women, ages 31 – 56, year 1965 – 2006

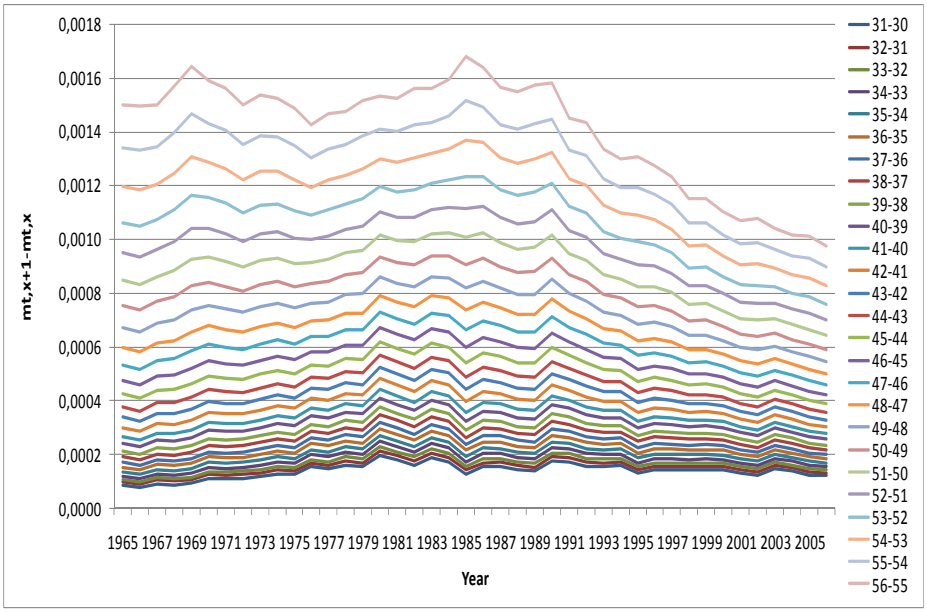


Fig 2. Difference of mortality of age x+1 and mortality of age x, men, ages 31 – 56, year 1965 - 2006

Both graphs show quite high volatility, so it would be good to try to smooth out the function. The best is to use moving averages. These we can use for number of dead people, and then it is necessary to adjust average stage of population or right for rate of mortality. And we also can presume the thing that moving average we can vote in time or in age. In our study we will use second variant. Let's count three-man average without using weight in time.

$$\bar{m}_{t,x} = \frac{m_{t-1,x} + m_{t,x} + m_{t+1,x}}{3}, \qquad (7)$$

or

$$\bar{m}_{t,x} = \frac{m_{t,x-1} + m_{t,x} + m_{t,x+1}}{3}. \qquad (8)$$

3 Results

It is clear from the graphs, that some assumptions, which we made above, weren't so accurate. In short time period, we were focus on, are not limit formula relevant – where difference should go to null and proportion to one. But it is possible to see that it is approximation to the other constant value.

It is also possible to see different volatility at chosen methods of differences and proportions. By the difference the volatility decrease with the age. By the proportion the biggest inconstancy is at the youngest ages.

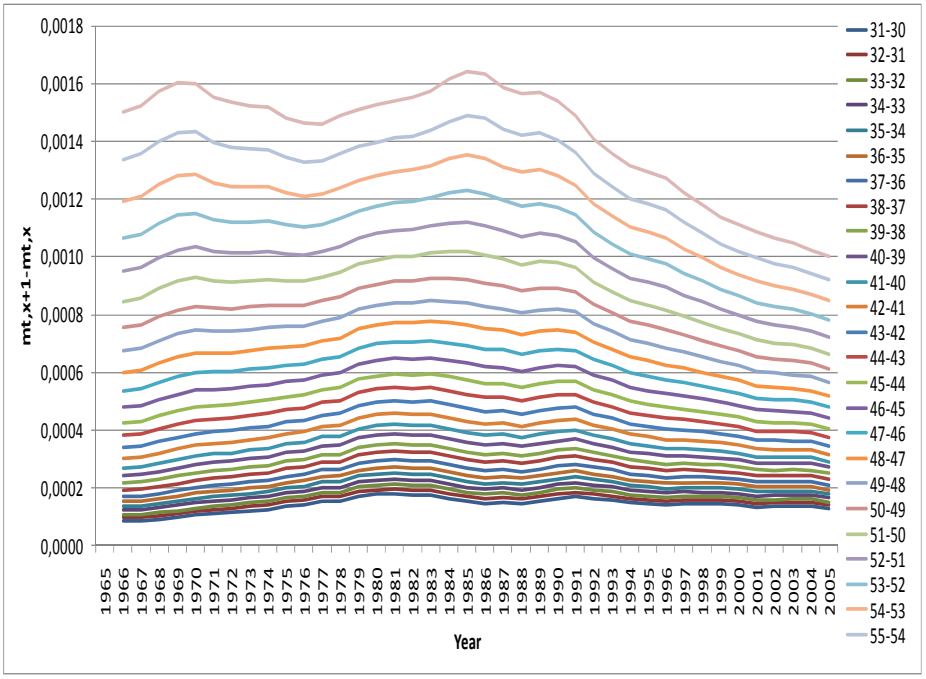


Fig. 3. Difference of mortality of age x+1 and mortality of age x, men, ages 31 – 56, year 1965 – 2006, three-man moving average

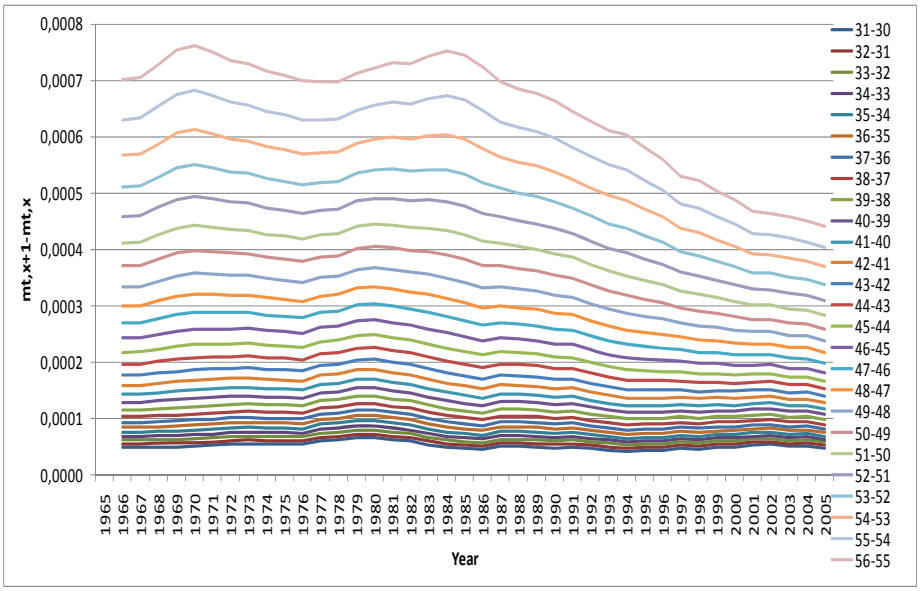


Fig. 4. Difference of mortality of age x+1 and mortality of age x, women, ages 31 – 56, year 1965 – 2006, three-man moving average

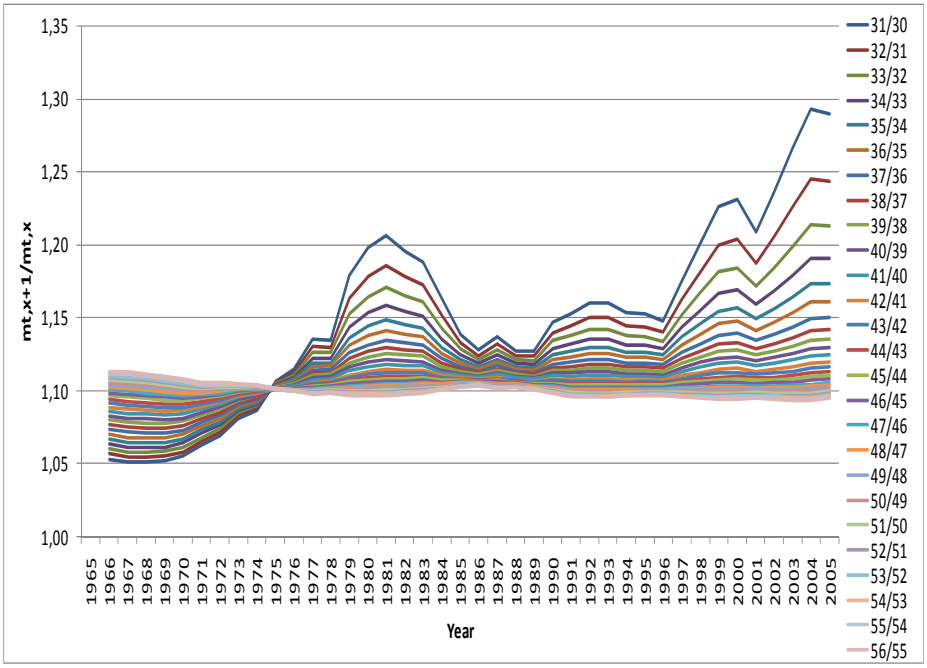


Fig. 5. Ratio of mortality of age x+1 into mortality of age x, men, ages 31 – 56, year 1965 – 2006, three-man moving average

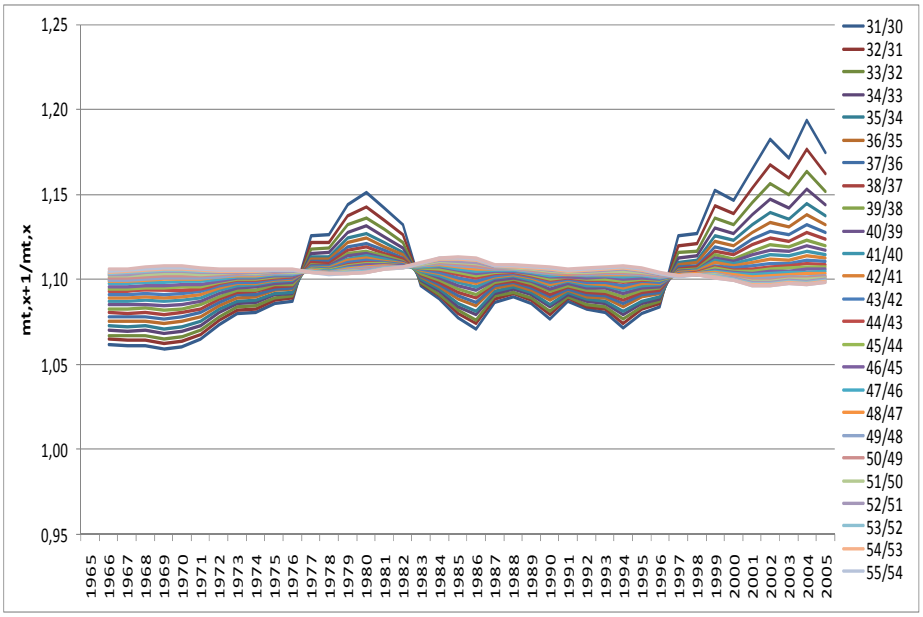


Fig. 6. Ratio of mortality of age x+1 into mortality of age x, women, ages 31 – 56, year 1965 – 2006, three-man moving average

4 Conclusion

From the point of relevance of prediction it is clear, that it is better to use differences between two age groups. Especially in last 20 years the prediction is stabile that is for prediction important. This variant we can use for modeling mortality in age group $x+n$, that means for the higher ages, than the chosen basic group. By the opposite way we are in dangerous, that the mortality for the younger groups after deduction of some value can decrease under the null. From that reason is better to choose model, where we can use proportion of two mortalities. High volatility in young ages is caused by – already mention- small deviation in number of dead people, because in so young age is mortality very low and even small differences can make impression of high decrease. By the sight on older ages we can see that the values are stable on approximately 1,1 . That means that probability of death decrease with the age about 10%. This finding can surely help for better prediction of mortality in higher ages. At these ages we have to make sure that the model works as we think. One more thing is important to say. The volatility in young ages can effect queerly, but because of low values, which mortality rates gain, the value is not important in compares with the old ages, where the mortality rates are several times higher.

References

www.czso.cz

Influence of Child Tax Credit on Inequity of Personal Income Tax in Poland

Edyta Mazurek[1], Marek Kośny[1]

[1] Wrocław University of Economics,
Komandorska Street 118/120, 53-345 Wrocław, Poland
edyta.mazurek@ue.wroc.pl, marek.kosny@ue.wroc.pl

Abstract. Child tax credit (known also as family tax credit) has been introduced in Polish personal income tax system in 2007. This tax credit enables every taxpayer, having dependent children, to subtract the certain amount from his (her) tax duty. As the deducted amount is quite high, analyzed tax credit caused significant change in relative situation of taxpayers having and not having dependent children.

The main aim of the presented paper is the assessment of the influence of this tax credit on tax equity and distribution of tax burden among different family types. Consequences of this change are analyzed on the basis of data from Lower-Silesian tax offices.

Keywords: Redistributive effect, decomposition of redistributive effect, child tax credit.

1 Introduction

Child tax credit has been introduced in Polish personal income tax system for the first time in year 2007. This tax credit enables every taxpayer, having dependent children, to subtract the certain amount from his (her) tax duty. This child tax credit lies in reduction of tax duty by amounts equal to doubled lump sum for each child. It means that in year 2007 Polish taxpayers could deduct from tax duty 1145.08 PLN (about 272.60 EURO) per each dependent child. The deducted amount is quite high, as it is subtracted from tax duty.

Let: y denote gross income and
k – number of dependent children.

Then, according to the current tax system in 2007 (with child tax credit), personal income tax can be calculated in the following way:

if y is less than 43405 PLN
then income tax is equal to $(19\% \cdot y - 572.54 - k \cdot 1145.08)$ PLN;

if $y \in (43405, 85528)$ PLN
then income tax is equal to ($7674.41 + 30\% \cdot$ surplus over $43405 - k \cdot 1145.08$) PLN;

if y is greater than 85528 PLN
then income tax is equal to
$(20311.31 + 40\% \cdot \text{surplus over } 85528 - k \cdot 1145.08)$ PLN.

As can be seen, this tax credit is entitled to every taxpayer (having dependent children), independently of income liable to taxation. The main aim of the presented paper is the assessment of the influence of this tax credit on distribution of income before and after taxation and on efficiency of income redistribution.

Generally, redistribution is defined as difference between concentration coefficients for income before taxation and after taxation, and measures how income tax system reduces inequality in income distribution. Redistributive practices are usually justified as redressing the balance between efficiency and social equity. The basic premise of the redistribution of income is that money should be distributed to benefit the poorer members of society, and that the rich should be obliged to assist the poor.
All analyses are made basing on Polish data from revenue office Wrocław-Fabryczna for fiscal year 2007.

2 Redistribution Measurement and Decomposition of Redistribution Coefficient

The basic index, used in redistribution measurement, is *RE* coefficient, defined as follows (cf. [6]):

$$RE = G_Y - G_{Y-T} \tag{1}$$

where G_Y denotes Gini index for income before taxation and
G_{Y-T} – Gini index for income after taxation.

The value of this coefficient could be interpreted as a percentage of income that is transferred from the richer to the poorer as a result of diversified tax rates. This kind of redistribution does not take the form of direct money transfers. It informs about the extent of such transfers that should be made in case of hypothetical, proportional tax system to get the tax distribution identical to the analyzed one.
RE coefficient could be decomposed. This decomposition aims at answering the question: to what extent is the overall redistribution a consequence of intentional construction of the tax system and to what extent is it restricted by tax inequity? The first component could be interpreted as a measure of actual, theoretical redistributive capacity, while the second reflects undesired – and often unintended – effects of the taxation.

Generally, decomposition of *RE* coefficient could be written as

$$RE = V - H - R \tag{2}$$

where *V* is a measure of vertical effect (decrease in inequality, resulting from the tax system progressivity) and *H* reflects horizontal inequity (unequal treatment of equals). Differences in inequality levels, resulting from changes in orders of taxpayers with respect to income before and after taxation, are captured by component *R*.
Calculation of *RE* coefficient decomposition demands division of the taxpayers' population into groups, distinguished from the point of view of the income level. Let *Y* be a vector of non-decreasing incomes before taxation for *n* taxpayers:

$$Y = (y_1, y_2, ..., y_n), \quad y_1 \le y_2 \le \ldots \le y_n$$

and taxpayers are grouped (with respect to the income) into *k* classes, consisting of $n_1, n_2, \ldots, n_k$ taxpayers respectively. Analogously, *Y-T* would denote incomes after taxation.
There are proposed in the literature several methods of decomposing *RE* coefficient. In this paper we apply decomposition proposed by Aronson, Johnson and Lambert (hereafter AJL) (cf. [1]). AJL method – in the original version – demands groups with exactly equal incomes, but Urban and Lambert (cf. [3]) show the possibility of extension on the groups with similar incomes. They introduce smoothed, linear taxation within each group. The rate of this tax is calculated individually – for each group – as an effective tax rate. Such neutral tax wipes out redistribution within each group. Then AJL decomposition could be given as (cf. [2]):

$$RE = V^{AJL} - H^{AJL} - R^{AJL}, \tag{3}$$

where:

$$V^{AJL} = \left(G_Y^B - G_{Y-T}^B\right) - \left(G_Y^{SW} - G_{Y-T}^W\right),$$
$$H^{AJL} = G_{Y-T}^W - G_{Y-T}^{SW},$$
$$R^{AJL} = G_{Y-T} - G_{Y-T}^{SW} - G_{Y-T}^W.$$

G_Y^B denotes between-group Gini index for income before taxation, where all individual incomes within each group were replaced with average income for a given group. Within-group Gini index (G_Y^W) is given by the formula:

$$G_Y^W = \sum_k \frac{n_k}{n} \cdot \frac{n_k \overline{Y}_k}{n\overline{Y}} \cdot G_{k,Y}, \tag{4}$$

where $G_{k,Y}$ denotes Gini index for *k*-th group, $\overline{Y}_k$ – average income in *k*-th group. Measures concerning income after taxation are denoted by G_{Y-T}^B and G_{Y-T}^W

respectively. Within-group Gini index G_{Y-T}^{SW} is calculated in an analogous way as given by (4), but for the income after taxation and smoothed tax. If $y_1, y_2, \ldots, y_{n_k}$ are the incomes in *k*-th group and $t_1, t_2, \ldots, t_{n_k}$ are respective tax amounts, smoothed tax for *i*-th taxpayer (belonging to *k*-th group) is given by

$$t_i^s = \frac{\sum_{i=1}^{n_k} t_i}{\sum_{i=1}^{n_k} y_i} \cdot y_i = g \cdot y_i \,. \tag{5}$$

Presented decomposition of *RE* requires division of the whole population into groups of taxpayers with similar income. To this end, suitable bandwidth has to be chosen and all the taxpayers have to be assigned to the classes with respect to their income before taxation. In this paper we use the method proposed by Vernizzi and Pellegrino (cf. [5], [6]) for choosing an optimal (suitable) bandwidth.

3 Empirical Analysis

The influence of the child tax credit on the distribution of income before and after taxation was analyzed on the basis of Polish data from Wrocław-Fabryczna tax office for fiscal year 2007. This set of data contains information on income and tax paid for taxpayers that file their tax return in the Municipality of Wrocław, tax office (district identification) Fabryczna. In this analysis households are equated with couples of taxpayers who take advantage of joint taxation. After deleting observations with non-positive gross income, the whole population consists of 23792 households. The analyses were performed by authors' own programmes, written in the "*R*" language.

In order to conduct the analyses, population of 23792 households was divided into subpopulations with respect to the number of dependent children. We introduced the following notation:

- C means family without children,
- C+1 – family with one child,
- C+2 – family with two children,
- C+3 – family with three children,
- C+4+ – family with four or more children.

Results of such a division, in terms of subpopulations' size, are presented in Table 1.

More than a half of the sample was constituted by families without children. Families with one child made around 30 per cent. 16 per cent of families had two children and relatively small share of families had three or more children.

Table 1. Division of the whole sample into five subgroups.

Family type	Numbers of families
C	12711
C+1	6682
C+2	3780
C+3	536
C+4+	83
All	**23792**

Source: own calculations

Table 2 presents average yearly gross income per family and per capita (in EURO) with respect to family type. For families with maximum two children, average income per family increase together with increase in number of children. What is worth observing, average income per family with tree children is almost the same as average income for families with two children. Only average income for families with four or more children is considerably less than for all the others family types. However, analyzing average income per capita, it can be seen that the more children in family, the more difficult is the financial situation.

Table 2. Average yearly gross income (in EURO).

Family type	Average income per family	Average income per capita
C	14858	7429
C+1	18077	6026
C+2	20515	5129
C+3	20302	4060
C+4+	16448	2741
All	16789	6262

Source: own calculations

Table 3. Average tax rates.

Family type	Average income per family	Average tax rate [%] without deduction	Average tax rate [%] actual system
C	14858	9.77	9.77
C+1	18077	10.20	8.73
C+2	20515	11.45	8.90
C+3	20302	12.53	8.77
C+4+	16448	10.77	4.81
All	16789	10.30	9.24

Source: own calculations

Table 3 presents average tax rates for actual Polish tax system in 2007 and for the same system, but without child credit. For actual system, even though average income per family increase, average tax rate decrease. In the system without child credit, average tax rates will be higher for all family types. The biggest difference between average tax rates for actual system and system without child credit is – of course – observed for families with four or more children. Average tax rate for these families would be more than twice as high without child tax credit. In the system without child credit, average tax rates increase together with the increase in average income and the increase in number of children (up to the third child).

Table 4. Redistribution and loss in redistribution for various family types.

Family type	RE	V^{AJL}	H^{AJL}	R^{AJL}
C	1.82%	1.84%	0.01%	0.01%
C+1	2.33%	2.35%	0.01%	0.01%
C+2	3.07%	3.09%	0.01%	0.01%
C+3	3.75%	3.77%	0.01%	0.01%
C+4+	2.29%	2.29%	0.00%	0.00%
All	2.28%	2.30%	0.01%	0.01%

Source: own calculations

Table 4 presents results of the decomposition of redistributive effect for all analyzed types of families. Horizontal and reranking effects are very close to zero. It means that Polish tax system causes almost no reranking in taxpayers' income (within analyzed groups) and no loss in the redistributive effect (meaning unequal treatment of equals). Similar results were received for decomposition of redistributive effect for the whole sample (cf. Table 5). It means that the Polish tax system with child credit (year 2007) did not cause significant loss in redistributive effect of taxation. It is important property of the tax system because it denotes relative lack of "system failures" and compliance with the assumed structure of the system.

Table 5. Results for Polish personal income tax system 2007.

	Present system	**Without children deduction**
RE [%] actual redistribution	2.28	1.90
V^{AJL} [%] potential redistribution	2.30	1.92
V^{AJL}-RE [%] **loss in redistribution [%]**	**0.02**	**0.02**
H^{AJL} [%] horizontal effect	0.01	0.01
R^{AJL} [%] reranking effect	0.01	0.01

Source: own calculations

The personal income tax system reduces inequality maximally in case of families with three children (by 3.75 percentage points), at least (by 1.82 percentage points) for families without children. For the complete sample, tax system with child credit reduces inequality by 2.28 percentage points.

Comparing results of decomposition of redistributive effect for present tax system and the same tax system but without child credit, it can be seen that both tax systems cause almost no reranking in taxpayers' income and no loss in the redistributive effect – accounted for by the unequal treatment of equals. However, present system reduces inequality much stronger. And, as a consequence, child credit increased redistributive effect from 1.90 to 2.28 percentage points.

4 Conclusions

Child credit, introduced in Polish personal income tax system in year 2007, caused significant decrease in average tax rates for families with children and increased redistributive effect from 1.90 to 2.28 percentage points. At the same time, this change had almost no influence on two negative effects of taxation – reranking and horizontal effects.

Taking into account that having children is often considered as one of the main reasons for decreasing the tax duty of families, change in the Polish tax system should be assessed very positively.

References

1. Aronson, J.P., Jenkins, P., Lambert, P.J.: Redistributive effect and unequal income tax treatment. The Economic Journal 104, pp. 262-270 (1994)
2. Kakwani, N.C.: Welfare rankings of income distributions. Advances in Econometrics 3, pp. 191—213 (1984)
3. R Development Core Team: R – A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.Rproject.org, Vienna (2008)
4. Urban, I., Lambert, P.J.: Redistribution, horizontal inequity and reranking: how to measure them properly. Public Finance Review 36(5), pp.563-587 (2008)
5. Vernizzi, A., Pellegrino, S.: On the Aronson-Johnson-Lambert decomposition of the redistributive effect. DEAS, Universita degli Studi di Milano, Working Paper 2007-13 (2007)
6. Vernizzi, A., Pellegrino, S.: On determining "close equals groups" in decomposing redistributive and reranking effects. SIEP, Working Paper 2007-602 (2007)

Dynamic Model of a Small Open Economy Under Fixed Exchange Rates

Petra Medveďová

Department of Quantitative Methods and Informatic Systems, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica, Slovakia
petra.medvedova@umb.sk

Abstract. In the paper a three dimensional dynamic model of a small open economy, describing the development of net real national income, real physical capital stock and nominal money stock is analyzed under fixed exchange rates. Sufficient conditions for the existence of a pair of purely imaginary eigenvalues with the third one negative in the linear approximation matrix of the model are found. Formulae for the calculation of the bifurcation coefficients in the bifurcation equation of the model are derived. Theorem on the existence of business cycles is presented.

Keywords: dynamical model, equilibrium point, matrix of linear approximation, eigenvalues, bifurcation equation, business cycle.

1 Introduction

Toichiro Asada formulated in [1] a Kaldorian business cycle model in a small open economy. He studied both the system of fixed exchange rates and that of flexible exchange rates with the possibility of capital mobility. In this article we investigate the Asada's model which was introduced in [1] under fixed exchange rates. In this case the Asada's model has the form

$$\begin{aligned} &\dot{Y} = \alpha[C + I + G + J - Y], \alpha > 0, \\ &\dot{K} = I, \\ &\dot{M} = pA, \end{aligned} \tag{1}$$

where

$$\begin{aligned} &C = c(Y - T) + C_0, 0 < c < 1, C_0 > 0, \\ &T = \tau Y - T_0, 0 < \tau < 1, T_0 > 0, \\ &I = I(Y, K, r), I_Y = \frac{\partial I}{\partial Y} > 0, I_K = \frac{\partial I}{\partial K} < 0, I_r = \frac{\partial I}{\partial r} < 0, \\ &\frac{M}{p} = L(Y, r), L_Y = \frac{\partial L}{\partial Y} > 0, L_r = \frac{\partial L}{\partial r} < 0, \\ &J = J(Y, \pi), J_Y = \frac{\partial J}{\partial Y} < 0, J_\pi = \frac{\partial J}{\partial \pi} > 0, \end{aligned} \tag{2}$$

$$Q = \beta\left[r - r_f - \frac{\pi^e - \pi}{\pi}\right], \beta > 0,$$
$$A = J + Q,$$
$$\pi = \bar{\pi},$$
$$\pi^e = \bar{\pi}.$$

The meanings of the symbols in (1) and (2) are as follows: Y - net real national income, C - real private consumption expenditure, I - net real private investment expenditure on physical capital, G - real government expenditure (fixed), T - real income tax, K - real physical capital stock, M - nominal money stock, p - price level, r - nominal rate of interest of domestic country, r_f - nominal rate of interest of foreign country, π - exchange rate (fixed), π^e - expected exchange rate of near future (fixed), J - balance of current account (net export) in real terms, Q - balance of capital account in real terms, A - total balance of payments in real terms, α - adjustment speed in goods market, β - degree of capital mobility, α, β - positive parameters, and the meanings of other symbols are as follows, $\dot{Y} = \frac{dY}{dt}, \dot{K} = \frac{dK}{dt}, \dot{M} = \frac{dM}{dt}$, t - time.

In the whole paper we assume as well as Asada in [1] fixed price economy so that p is exogenously given and normalized to the value 1. Asada assumes that the equilibrium on the money market $M = L(Y, r)$ is always preserved what enables using Implicit-function theorem to express interest rate r as the function of Y and M

$$r = r(Y, M), r_Y = \frac{\partial r}{\partial Y} > 0, r_M = \frac{\partial r}{\partial M} < 0.$$

Further we suppose that r_f is also given exogenously because of the assumption of a small open economy. Under these assumptions taking into account (2) and the explicit expression for r the model (1) takes the form

$$\begin{aligned} \dot{Y} &= \alpha\left[c(1-\tau)Y + cT_0 + C_0 + I\left(Y, K, r(Y, M)\right) + G + J(Y, \bar{\pi}) - Y\right] \\ \dot{K} &= I\left(Y, K, r(Y, M)\right) \\ \dot{M} &= J(Y, \bar{\pi}) + \beta[r(Y, M) - r_f]. \end{aligned} \tag{3}$$

In the whole paper we suppose that:

(i) the model (3) has a unique equilibrium point $E^* = (Y^*, K^*, M^*), Y^* > 0, K^* > 0, M^* > 0$, to an arbitrary pair of positive parameters (α, β).

(ii) $0 < I_Y + I_r r_Y < [1 - c(1-\tau)] - J_Y$ at the equilibrium point.

(iii) The functions in the model (3) have the following properties: the function J is linear in the variable Y. The function I is linear in the variable K and r. The function r is linear in the variable M. In the variable Y the functions I, r are nonlinear, and have continuous partial derivatives with respect to Y up to the sixth order in a small neighbourhood of the equilibrium point.

In [1] Asada found sufficient conditions for local stability and instability of the equilibrium point. He studied how changes of the parameter β affect the dynamic characteristics of the model. Asada also showed utilizing reached results on the

equilibrium stability the existence at least one parameter β at which the Hopf bifurcation occurs.

In the presented article we analyse the question of the existence of business cycles analytically. Stable business cycles can arise only in the case when the linear approximation matrix of the model (3) has at the equilibrium point a pair of purely imaginary eigenvalues with the third one negative. Theorem 1 gives sufficient conditions for the existence of a pair of purely imaginary eigenvalues with the third one negative. The bifurcation equation of the model (3) is very important for the existence business cycles. Theorem 2 gives the formulae for the calculation of the bifurcation coefficients in the bifurcation equation. About the existence of business cycles in a small neighbourhood of the equilibrium point says Theorem 3.

2 The analysis of the model (3)

Consider an isolated equilibrium point $E^* = (Y^*, K^*, M^*), Y^* > 0, K^* > 0, M^* > 0$ of (3).

After the transformation

$$Y_1 = Y - Y^*,\ K_1 = K - K^*,\ M_1 = M - M^*,$$

the equilibrium point $E^* = (Y^*, K^*, M^*)$ goes into the origin $E_1^* = (Y_1^*, K_1^*, M_1^*) = (0,0,0)$, and the model (3) becomes

$$\begin{aligned} \dot{Y}_1 &= \alpha[c(1-\tau)(Y_1 + Y^*) + cT_0 + C_0 + G + J(Y_1 + Y^*, \bar{\pi}) + \\ &\quad + I\big(Y_1 + Y^*, K_1 + K^*, r(Y_1 + Y^*, M_1 + M^*)\big) - (Y_1 + Y^*)] \\ \dot{K}_1 &= I\big(Y_1 + Y^*, K_1 + K^*, r(Y_1 + Y^*, M_1 + M^*)\big) \\ \dot{M}_1 &= J(Y_1 + Y^*, \bar{\pi}) + \beta[r(Y_1 + Y^*, M_1 + M^*) - r_f]. \end{aligned} \tag{4}$$

Performing the Taylor expansion of the functions on the right-hand side of this system at the equilibrium point $E_1^* = (0,0,0)$ the model (4) obtains the form

$$\begin{aligned} \dot{Y}_1 &= \alpha[c(1-\tau)Y_1 + I_Y Y_1 + I_K K_1 + I_r r_Y Y_1 + I_r r_M M_1 + J_Y Y_1 - Y_1] + \\ &\quad + \tilde{Y}_1(Y_1, K_1, M_1, \alpha) \\ \dot{K}_1 &= I_Y Y_1 + I_K K_1 + I_r r_Y Y_1 + I_r r_M M_1 + \tilde{Y}_2(Y_1, K_1, M_1) \\ \dot{M}_1 &= J_Y Y_1 + \beta r_Y Y_1 + \beta r_M M_1 + \tilde{Y}_3(Y_1, M_1, \beta), \end{aligned} \tag{5}$$

where the letter indices mean partial derivatives with respect to indicated variables calculated at the equilibrium point E^*, and the functions $\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3$ are nonlinear parts of the Taylor expansion.

The linear approximation matrix of the system (5) has the following form

$$A(\alpha, \beta) = \begin{pmatrix} \alpha[c(1-\tau) + I_Y + I_r r_Y + J_Y - 1] & \alpha I_K & \alpha I_r r_M \\ I_Y + I_r r_Y & I_K & I_r r_M \\ J_Y + \beta r_Y & 0 & \beta r_M \end{pmatrix}. \tag{6}$$

The characteristic equation of $A(\alpha, \beta)$ is given by

$$\lambda^3 + a_1(\alpha, \beta)\lambda^2 + a_2(\alpha, \beta)\lambda + a_3(\alpha, \beta) = 0, \quad (7)$$

where

$$a_1 = -\text{trace}A(\alpha, \beta) = -\{\alpha[c(1-\tau) + I_Y + I_r r_Y + J_Y - 1] + I_K + \beta r_M\}$$

$$a_2 = \text{sum of all principal second} - \text{order minors of } A(\alpha, \beta) =$$
$$= \alpha I_K[c(1-\tau) + J_Y - 1] + \beta I_K r_M + \alpha\beta r_M[c(1-\tau) + I_Y + J_Y - 1] - \alpha I_r r_M J_Y$$

$$a_3 = -\det A(\alpha, \beta) = -\alpha\beta I_K r_M[c(1-\tau) + J_Y - 1].$$

As we are interested in the existence and stability of limit cycles we need to find such values of parameters α, β, at which the equation (7) has a pair of purely imaginary eigenvalues and the third one is negative. We will call such values of parameters α, β the critical values of the model (3). We denote these critical values by α_0, β_0. Mentioned types of eigenvalues are ensured by the Liu's conditions [5]

$$a_1 > 0, a_3 > 0, a_1 a_2 - a_3 = 0. \quad (8)$$

The first inequality is satisfied under the condition (ii) for an arbitrary α. The second inequality is satisfied on the base of the properties of the functions in (2). The equation $a_1 a_2 - a_3 = 0$ gives

$$\{\alpha I_K[c(1-\tau) + J_Y - 1] + \beta I_K r_M + \alpha\beta r_M[c(1-\tau) + I_Y + J_Y - 1] - \alpha I_r r_M J_Y\} \times$$

$$\times \{\alpha[c(1-\tau) + I_Y + I_r r_Y + J_Y - 1] + I_K + \beta r_M\} - \alpha\beta I_K r_M[c(1-\tau) + J_Y - 1] = 0.$$

The equation $a_1 a_2 - a_3 = 0$ can be expressed in the form

$f_1(\alpha)\beta^2 + f_2(\alpha)\beta + f_3(\alpha) = 0$, where

$$f_1(\alpha) = r_M^2\{I_K + \alpha[c(1-\tau) + I_Y + J_Y - 1]\}$$

$$f_2(\alpha) = \alpha I_K r_M\{2[c(1-\tau) + I_Y + J_Y - 1] + I_r r_Y\} + I_K^2 r_M - \alpha r_M^2 I_r J_Y +$$
$$+\alpha^2 r_M[c(1-\tau) + I_Y + I_r r_Y + J_Y - 1][c(1-\tau) + I_Y + J_Y - 1]$$

$$f_3(\alpha) = \alpha^2 I_K[c(1-\tau) + I_Y + I_r r_Y + J_Y - 1][c(1-\tau) + J_Y - 1] - \alpha I_r r_M J_Y I_K -$$
$$-\alpha^2 I_r r_M J_Y[c(1-\tau) + I_Y + I_r r_Y + J_Y - 1] + \alpha I_K^2[c(1-\tau) + J_Y - 1].$$

We see that for an arbitrary β exists $\hat{\alpha}$ such that $f_1(\hat{\alpha}) = 0$, $\hat{\alpha} > 0$.
Put $F(\alpha, \beta) = f_1(\alpha) + f_2(\alpha)\frac{1}{\beta} + f_3(\alpha)\frac{1}{\beta^2} = 0$.
Instead of β introduce $\gamma = \frac{1}{\beta}$ and consider an equation $\Phi(\alpha, \gamma) = 0$, when

$$\Phi(\alpha,\gamma)=\begin{cases} f_1(\alpha), & \gamma=0 \\ f_1(\alpha)+f_2(\alpha)\gamma+f_3(\alpha)\gamma^2, & \gamma\neq 0. \end{cases}$$

We see that $\Phi(\alpha,\gamma)=0$ is equivalent to $F(\alpha,\beta)=0$. Analyse $\Phi(\alpha,\gamma)=0$.

It holds:

1. $\Phi(\hat{\alpha},\gamma=0)=0$
2. $\frac{\partial\Phi(\hat{\alpha},\gamma=0)}{\partial\alpha}=\frac{df_1(\hat{\alpha},\gamma=0)}{d\alpha}+\frac{df_2(\hat{\alpha},\gamma=0)}{d\alpha}\gamma+\frac{df_3(\hat{\alpha},\gamma=0)}{d\alpha}\gamma^2=$

$$=r_M^2[I_Y-1+c(1-\tau)+J_Y]\neq 0.$$

By Implicit-function theorem then there exists a function $\alpha=f(\gamma)$ in a small neighbourhood of $\gamma=0$ such that $\hat{\alpha}=f(0)$ and $\Phi(f(\gamma),\gamma)=0$.

We see that to sufficiently large β_0 of parameter β there exists value α_0 of parfameter α such that the pair (α_0,β_0) is the critical pair of the model (3). The following theorem gives sufficient conditions for the existence of a critical pair of the model (3).

Theorem 1. Let the condition (ii) be satisfied. If parameter β is sufficiently large, then there exists a critical pair (α_0,β_0) of the model (3).

3 Existence of limit cycles and their stability

According to the assumption (iii) the model (5) can be itemized in the form

$$\begin{aligned}
\dot{Y}_1 &= \alpha([I_Y+I_r r_Y-\{1-c(1-\tau)-J_Y\}]Y_1+I_K K_1+I_r r_M M_1)+\\
&\quad+\tfrac{1}{2}\alpha I_Y^{(2)}Y_1^2+\tfrac{1}{6}\alpha I_Y^{(3)}Y_1^3+\tfrac{1}{24}\alpha I_Y^{(4)}Y_1^4+O(|Y_1|^5)\\
\dot{K}_1 &= (I_Y+I_r r_Y)Y_1+I_K K_1+I_r r_M M_1+\tfrac{1}{2}I_Y^{(2)}Y_1^2+\tfrac{1}{6}I_Y^{(3)}Y_1^3+\\
&\quad+\tfrac{1}{24}I_Y^{(4)}Y_1^4+O(|Y_1|^5) \qquad (9)\\
\dot{M}_1 &= (\beta r_Y+J_Y)Y_1+\beta r_M M_1+\tfrac{1}{2}\beta r_Y^{(2)}Y_1^2+\\
&\quad+\tfrac{1}{6}\beta r_Y^{(3)}Y_1^3+\tfrac{1}{24}\beta r_Y^{(4)}Y_1^4+O(|Y_1|^5),
\end{aligned}$$

where $I_Y^{(2)}=\frac{\partial^2 I(E^*)}{\partial Y^2}, I_Y^{(3)}=\frac{\partial^3 I(E^*)}{\partial Y^3}, I_Y^{(4)}=\frac{\partial^4 I(E^*)}{\partial Y^4}$.

Consider a critical pair (α_0,β_0) of the model (3). Let as investigate the behavior of Y_1, K_1 and M_1 around the equilibrium $E_1^*=(0,0,0)$ with respect to the parameter $\alpha, \alpha\in(\alpha_0-\varepsilon,\alpha_0+\varepsilon), \varepsilon>0,$ and the fixed parameter $\beta=\beta_0$.

After the shifting α_0 into the origin by relation $\alpha_1=\alpha-\alpha_0$, the model (9) becomes

$$
\begin{aligned}
\dot{Y}_1 &= \alpha_0[I_Y + I_r r_Y - \{1 - c(1-\tau) - J_Y\}]Y_1 + \alpha_0 I_K K_1 + \alpha_0 I_r r_M M_1 + \\
&+[I_Y + I_r r_Y - \{1 - c(1-\tau) - J_Y\}]Y_1\alpha_1 + I_K K_1\alpha_1 + I_r r_M M_1\alpha_1 + \\
&+\tfrac{1}{2}\alpha_0 I_Y^{(2)} Y_1^2 + \tfrac{1}{2} I_Y^{(2)} Y_1^2\alpha_1 + \tfrac{1}{6}\alpha_0 I_Y^{(3)} Y_1^3 + \tfrac{1}{6} I_Y^{(3)} Y_1^3\alpha_1 + \\
&+\tfrac{1}{24}\alpha_0 I_Y^{(4)} Y_1^4 + O_5(Y_1, \alpha_1) \qquad (10)\\
\dot{K}_1 &= (I_Y + I_r r_Y)Y_1 + I_K K_1 + I_r r_M M_1 + \tfrac{1}{2} I_Y^{(2)} Y_1^2 + \tfrac{1}{6} I_Y^{(3)} Y_1^3 + \\
&+\tfrac{1}{24} I_Y^{(4)} Y_1^4 + O_5(Y_1) \\
\dot{M}_1 &= (\beta_0 r_Y + J_Y)Y_1 + \beta_0 r_M M_1 + \tfrac{1}{2}\beta_0 r_Y^{(2)} Y_1^2 + \\
&+\tfrac{1}{6}\beta_0 r_Y^{(3)} Y_1^3 + \tfrac{1}{24}\beta_0 r_Y^{(4)} Y_1^4 + O_5(Y_1).
\end{aligned}
$$

Denote the eigenvalues of (6) as

$$\lambda_1 = \tau(\alpha,\beta) + i\omega(\alpha,\beta), \lambda_2 = \tau(\alpha,\beta) - i\omega(\alpha,\beta), \lambda_3 = \lambda_3(\alpha,\beta)$$

and put $\lambda_1 = i\omega_0, \lambda_2 = -i\omega_0, \omega_0 = \omega(\alpha_0,\beta_0), \lambda_{30} = \lambda_3(\alpha_0,\beta_0)$.
Express the model (10) in the form

$$\dot{x} = A(\alpha_0,\beta_0)x + \bar{Y}(x,\alpha_1),$$

where

$$x = \begin{pmatrix} Y_1 \\ K_1 \\ M_1 \end{pmatrix}, \ \bar{Y} = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \bar{Y}_3 \end{pmatrix}.$$

Consider a matrix $M = (m_{ij}), i,j = 1,2,3$, which transfers the matrix $A(\alpha_0,\beta_0)$ into its Jordan form J, and its inverse matrix $M^{-1} = (m_{ij}^{-1})$.

By the transformation $x = My$, $y = \begin{pmatrix} Y_2 \\ K_2 \\ M_2 \end{pmatrix}$, we obtain

$$\dot{y} = Jy + F(y,\alpha_1), \qquad (11)$$

where $J = \begin{pmatrix} i\omega & 0 & 0 \\ 0 & -i\omega & 0 \\ 0 & 0 & \lambda_{30} \end{pmatrix}$,

$F(y,\alpha_1) = \begin{pmatrix} F_1(y,\alpha_1) \\ F_2(y,\alpha_1) \\ F_3(y,\alpha_1) \end{pmatrix} = \begin{pmatrix} m_{11}^{-1}\bar{Y}_1 + m_{12}^{-1}\bar{Y}_2 + m_{13}^{-1}\bar{Y}_3 \\ m_{21}^{-1}\bar{Y}_1 + m_{22}^{-1}\bar{Y}_2 + m_{23}^{-1}\bar{Y}_3 \\ m_{31}^{-1}\bar{Y}_1 + m_{32}^{-1}\bar{Y}_2 + m_{33}^{-1}\bar{Y}_3 \end{pmatrix}$, $K_2 = \bar{Y}_2, F_2 = \bar{F}_1$, and F_3

is real (the symbol " $-$ " means complex conjugate expression in the whole paper).

Further, we perform a polynomial transformation

$$y = u + h(Y_3, K_3, \alpha_1), \tag{12}$$

$u = \begin{pmatrix} Y_3 \\ K_3 \\ M_3 \end{pmatrix}$, $h = \begin{pmatrix} h_1(Y_3, K_3, \alpha_1) \\ h_2(Y_3, K_3, \alpha_1) \\ h_3(Y_3, K_3, \alpha_1) \end{pmatrix}$, where $h_j(Y_3, K_3, \alpha_1)$, $j = 1,2,3$, are nonlinear polynomials with constant coefficients of the kind

$$h_j(Y_3, K_3, \alpha_1) = \sum_{m_1+m_2+m_3 \geq 2, m_3 \in \{0,1\}}^{4-2m_3} h_j^{(m_1,m_2,m_3)} Y_3^{m_1} K_3^{m_2} \alpha_1^{m_3},$$
$$j = 1,2,3, h_2 = \bar{h}_1.$$

Theorem 2. There exists a polynomial transformation (12) which transforms the model (11) into the form

$$\begin{aligned} \dot{Y}_3 &= i\omega_0 Y_3 + \delta_1 Y_3 \alpha_1 + \delta_2 Y_3^2 K_3 + U^\circ(Y_3, K_3, M_3, \alpha_1) + U^*(Y_3, K_3, M_3, \alpha_1) \\ \dot{K}_3 &= -i\omega_0 K_3 + \bar{\delta}_1 K_3 \alpha_1 + \bar{\delta}_2 Y_3 K_3^2 + \bar{U}^\circ(Y_3, K_3, M_3, \alpha_1) + \bar{U}^*(Y_3, K_3, M_3, \alpha_1) \\ \dot{M}_3 &= \lambda_{30} M_3 + V^\circ(Y_3, K_3, M_3, \alpha_1) + V^*(Y_3, K_3, M_3, \alpha_1), \end{aligned} \tag{13}$$

where $U^\circ(Y_3, K_3, 0, \alpha_1) = V^\circ(Y_3, K_3, 0, \alpha_1) = 0$,
$U^*\left(\sqrt{\alpha_1} Y_3, \sqrt{\alpha_1} K_3, \sqrt{\alpha_1} M_3, \alpha_1\right) = \left(\sqrt{\alpha_1}\right)^5 \tilde{U}(Y_3, K_3, M_3, \alpha_1)$,
$V^*\left(\sqrt{\alpha_1} Y_3, \sqrt{\alpha_1} K_3, \sqrt{\alpha_1} M_3, \alpha_1\right) = \left(\sqrt{\alpha_1}\right)^5 \tilde{V}(Y_3, K_3, M_3, \alpha_1)$,

and $\tilde{U}, \tilde{V}$ are continuous functions.

The resonant coefficients δ_1 and δ_2 in the model (13) are determined by the formulae

$$\begin{aligned} \delta_1 = \delta_1^{(1,0,1)} &= m_{11}^{-1}[(I_Y + I_r r_Y - \{1 - c(1-\tau) - J_Y\})m_{11} + I_K m_{21} + I_r r_M m_{31}] \\ \delta_2 = \delta_2^{(2,1,0)} &= m_{11}^{-1}\Big[\frac{1}{2}\alpha_0 I_Y^{(3)} m_{11}^2 m_{12} + \alpha_0 I_Y^{(2)} m_{11}(m_{11} h_1^{(1,1,0)} + m_{12} h_2^{(1,1,0)} + \\ &+ m_{13} h_3^{(1,1,0)}) + \alpha_0 I_Y^{(2)} m_{12}\left(m_{11} h_1^{(2,0,0)} + m_{12} h_2^{(2,0,0)} + m_{13} h_3^{(2,0,0)}\right)\Big] + \\ &+ m_{12}^{-1}\Big[\frac{1}{2} I_Y^{(3)} m_{11}^2 m_{12} + I_Y^{(2)} m_{11}(m_{11} h_1^{(1,1,0)} + m_{12} h_2^{(1,1,0)} + \\ &+ m_{13} h_3^{(1,1,0)}) + I_Y^{(2)} m_{12}\left(m_{11} h_1^{(2,0,0)} + m_{12} h_2^{(2,0,0)} + m_{13} h_3^{(2,0,0)}\right)\Big] + \\ &+ m_{13}^{-1}\Big[\frac{1}{2}\beta_0 r_Y^{(3)} m_{11}^2 m_{12} + \beta_0 r_Y^{(2)} m_{11}(m_{11} h_1^{(1,1,0)} + m_{12} h_2^{(1,1,0)} + \\ &+ m_{13} h_3^{(1,1,0)}) + \beta_0 r_Y^{(2)} m_{12}\left(m_{11} h_1^{(2,0,0)} + m_{12} h_2^{(2,0,0)} + m_{13} h_3^{(2,0,0)}\right)\Big], \end{aligned}$$

where

$$\begin{aligned} h_1^{(1,1,0)} &= -\frac{1}{i\omega_0} m_{11} m_{12}\left(\alpha_0 I_Y^{(2)} m_{11}^{-1} + I_Y^{(2)} m_{12}^{-1} + \beta_0 r_Y^{(2)} m_{13}^{-1}\right) \\ h_1^{(2,0,0)} &= \frac{1}{2i\omega_0} m_{11}^2\left(\alpha_0 I_Y^{(2)} m_{11}^{-1} + I_Y^{(2)} m_{12}^{-1} + \beta_0 r_Y^{(2)} m_{13}^{-1}\right) \end{aligned}$$

$$h_2^{(1,1,0)} = -\frac{1}{i\omega_0} m_{11} m_{12} \left(\alpha_0 I_Y^{(2)} m_{21}^{-1} + I_Y^{(2)} m_{22}^{-1} + \beta_0 r_Y^{(2)} m_{23}^{-1}\right)$$
$$h_2^{(2,0,0)} = \frac{1}{2i\omega_0} m_{11}^2 \left(\alpha_0 I_Y^{(2)} m_{21}^{-1} + I_Y^{(2)} m_{22}^{-1} + \beta_0 r_Y^{(2)} m_{23}^{-1}\right)$$
$$h_3^{(1,1,0)} = -\frac{1}{i\omega_0} m_{11} m_{12} \left(\alpha_0 I_Y^{(2)} m_{31}^{-1} + I_Y^{(2)} m_{32}^{-1} + \beta_0 r_Y^{(2)} m_{33}^{-1}\right)$$
$$h_3^{(2,0,0)} = \frac{1}{2i\omega_0} m_{11}^2 \left(\alpha_0 I_Y^{(2)} m_{31}^{-1} + I_Y^{(2)} m_{32}^{-1} + \beta_0 r_Y^{(2)} m_{33}^{-1}\right).$$

Proof. The unknown terms $h_j^{(m_1,m_2,m_3)}$, $j = 1,2.3$, and the resonant terms δ_1, δ_2 can be found by the standard procedure which is described for example in [2]. As the whole process of finding them is rather elaborated we do not present it here.

In polar coordinates $Y_3 = re^{i\varphi}$, $K_3 = re^{-i\varphi}$ the model (13) can be written as

$$\begin{aligned} \dot{r} &= r(ar^2 + b\alpha_1) + R^*(r, \varphi, M_3, \alpha_1) + R^*(r, \varphi, M_3, \alpha_1) \\ \dot{\varphi} &= \omega_0 + c\alpha_1 + dr^2 + \frac{1}{r}[\Phi^*(r, \varphi, M_3, \alpha_1) + \Phi^*(r, \varphi, M_3, \alpha_1)] \\ \dot{M}_3 &= \lambda_{30} M_3 + W^*(r, \varphi, M_3, \alpha_1) + W^*(r, \varphi, M_3, \alpha_1), \end{aligned} \tag{14}$$

where $a = Re\delta_2$, $b = Re\delta_1$. The equation

$$ar^2 + b\alpha_1 = 0$$

is the bifurcation equation of the model (14). It determines the behaviour of solutions in a neigbourhood of the equilibrium point of the model (5). Utilizing the results from the bifurcation theory [3] we can formulate the following theorem.

Theorem 3. Let the coefficients a, b in the bifurcation equation exist.

(1) If $a < 0$ then there exists a stable limit cycle for every small enough $\alpha_1 > 0$, if b is positive and for every small enough $\alpha_1 < 0$, if b is negative.
(2) If $a > 0$ then there exists an unstable limit cycle for every small enough $\alpha_1 < 0$, if b is positive and for every small enough $\alpha_1 > 0$, if b is negative.

Acknowledgement. The work was supported by the Slovak grant agency, grant No. 1/4633/07.

References

1. Asada, T.: Kaldorian Dynamics in an Open Economy, Journal of Economics No. 3, 239-269, Springer-Verlag, Printed in Austria (1995)
2. Bibikov, Yu. N.: Local Theory of Nonlinear Analytic Ordinary Differential Equations, Springer-Verlag, Berlin (1979)
3. Bibikov, Yu. N.: Multi-frequency non-linear oscillations and their bifurcations, The Publishing House of the Saint Petersburg University, Saint Petersburg, (in Russian) (1991)
4. Gandolfo, G.: Economic Dynamics, Springer-Verlag, Berlin (1997)

5. Liu, W. M.: Criterion of Hopf Bifurcations without using Eigenvalues, Journal of Mathematical Analysis and Applications, 182, 250-256 (1994)
6. Wiggins, S.: Introduction to Applied Nonlinear Dynamical Systems and Chaos, Springer-Verlag, Berlin (1990)

The Decomposition Methodology of the Annual Change of Cross-Sectional Indicators in the EU-SILC: An Application to the Czech Data

Martina Mysíková[1] and Martin Zelený[2]

[1] Institute of Economic Studies, Charles University in Prague, Opletalova 26, 110 00 Praha 1, Czech Republic
and Institute of Sociology of the Academy of Sciences of the Czech Republic, Prague, Jilská 1, 110 00 Praha 1, Czech Republic
martina.mysikova@centrum.cz
[2] University of Economics, Prague University of Economics, Prague, Department of Economic Statistics (KEST), Nam. W. Churchilla 4, 130 67 Praha 3, Czech Republic
and Czech Statistical Office, Prague, Na padesatem 81, 100 82, Praha 10, Czech Republic
zelenym@vse.cz

Abstract. The aim of this paper is to examine the annual trends of EU-SILC cross-sectional indicators. It reveals the effects that may the applied rotational sample design have on measured and reported annual changes. We suggest and demonstrate a new methodology of decomposition of the actual measured interannual difference of indicators into several effects and study in detail how changing sample (i.e., attrition and incorporating of a new sample replication), changes in responses and different cross-sectional weights affect step by step the indicators. The decomposition methodology is applied on the Czech 2005 to 2006 EU-SILC data.

Keywords: annual trends**,** panel attrition, EU-SILC, rotational panel.

1 Introduction

European Union—Statistics on Income and Living Conditions (EU-SILC) is a harmonised survey compulsory for all EU Member States and EEA countries and it provides data suitable for cross-country comparisons. The survey is designed as a four-year rotational panel in most covered countries.[1] This means that information is collected for four consecutive years for each sub-sample of households (usually called sample replication). Each annual sample consists of four such replications, one of which is dropped every year (households surveyed for the final, fourth time) and a

[1] For more information on the EU-SILC design issues see [4]

one of which are new households sampled for the first interviews. The questionnaire collects information on sampled households (mainly information on living conditions and material deprivation) and on individuals living in these households (individual and job characteristics, wages, income, social allowances etc.).

The EU-SILC provides us with cross-sectional as well as longitudinal datasets. The two subsequent annual cross-sectional samples include an overlapping part of the sample—the sample replications carried forward from the previous survey year. These households which are present in both subsequent cross-sectional samples are further referred to in this paper as core households. This part of the sample is also part of the cross-sectional sample used for the estimation of annual cross-sectional indicators based on the EU-SILC data. The aim of the present paper is to take a closer look at what effects this feature of the annual EU-SILC samples may have on these, in principle cross-sectional, indicators—particularly on their measured and reported annual trends.

Cross-sectional indicators should reflect the annual change of households' situation at a national level. However, the interannual change of cross-sectional indicators' values consists not only of the effect of different conditions in households in terms of changed households' responses but also of several other effects: the composition of the cross-sectional sample changes in two subsequent years in several ways (e.g., the attrition and incorporation of new sample) and the data are weighted by cross-sectional weights assigned to the households, which usually differ each year. The aim of this study is to reveal the components of annual trends of cross-sectional indicators and to decompose the annual change into several effects. We will focus our analysis on trends of selected categorical EU-SILC variables (material deprivation items) and study to what extent they may be affected by the design of the EU-SILC sample.

The rest of this paper is organised as follows. The next section provides a sample design overview and depicts the survey sample changes between two consecutive years. The decomposition methodology of the annual change is outlined in section 3 and decomposition results for annual change between years 2005 and 2006 in the Czech Republic are presented in section 4. The main findings are summarised in Section 5.

2 Sample design[2]

This study focuses mainly on annual trends of indicators in Czech Republic in 2005 to 2006. The Czech rotational panel in this period was designed in such a way that all first-wave households continued in the survey for four consecutive years, i.e. from 2005 to 2008[3]. In 2006 a group of new households was added and remained in the survey until 2009. Therefore, there were two groups of households participating in

[2] For a general introduction to longitudinal sample designs see for example [1] or [2]

[3] Each year one new group of households joins the survey for four consecutive years. The rotational design was fully functional in 2008 when the cross-sectional dataset consisted of four sub-groups of households.

2006, one for the first time and one for the second time. The solid part of **Scheme 1** depicts the sub-samples examined in this study.

Scheme 1. Rotational design in the Czech Republic

2005	2006	2007	2008	2009
1	2	3	4	
	1	2	3	4
		1	2	3
			1	2
				1

Any cross-sectional annual change of indicators is determined mainly by values responded by households continuing from previous year, called core households here, and by new households. However, adding the new households is not the only change of survey sample that can occur between two subsequent years. **Scheme 2** describes all the possible households' statuses that can arise between two consecutive years.

Number "1" indicates the core households that continue in the survey from 2005 to 2006. "2" determines households that responded in 2005 but refused to respond or were not contacted in 2006. Out-of-scope households are labelled "3" and contain households whose all members moved into some institution or out of country, died, households that disappeared by fusion of two sample households, and households that do not contain sample persons anymore. According to the sample design, there are households that are dropped—called "4". In the Czech sample, a sample replication was dropped first in 2009; therefore, households of group "4" do not occur in the present study. However, this group of households has to be mentioned for general purposes and, moreover, a sample replication can be dropped earlier in some countries depending on their sample design for initial years. Therefore, "2" plus "3" plus "4" represent households that dropped out of the 2005 sample.

On the contrary, "5" plus "6" plus "7" denote households added to the sample in year 2006. A split-off household ("5") split from a household that responded in 2005 and so created a new household in the sample. Re-contacted households ("6") are households that were successfully surveyed two years ago but were not contacted last year. Even then such households are kept in the panel and surveyed in current year. Apparently, re-contacted households cannot occur before a third year of the survey. These households thus emerge first in 2007 while they could not exist in 2006 in the Czech survey. Therefore, they are not involved in the present study. Finally, households labelled "7" are new households incorporated to the sample in 2006 (the new sample replication added in 2006).

Scheme 2. Interannual change of sample

2005	2006
1	**1**
Core (continuing households)	**Core** (continuing households)
2	✕
Non-response (refused, non-contacted, lost households)	
3	✕
Out-of-scope (entire household moved into institution or out of country, died, disappeared by a fusion, does not contain sample persons anymore)	
4 (not available in CZ)	✕
Dropped (dropped sample replication)	
	5
	Split-off (households split from 2005 sample households)
	6 (not available in CZ)
	Re-contacted (households that responded in 2004 and non-contacted in 2005)
	7
	New sample (new households)

The analysis will further deal mainly with indicators of material deprivation. The values recorded in 2005 and 2006 will be compared and contribution of particular group of households to the total inter-annual change of the indicator will be analysed.

3 Methodology of decomposition

The measured annual change of cross-sectional indicators' values from rotational design samples consists of many effects: the sample composition changes in two consequent years, actual values reported by respondents in the core overlapping part vary from one year to the next, and different cross-sectional weights are assigned to households in each year. Majority of households surveyed last year forms a core of the sample and continue in the survey in the current year while some households disappear from the sample due to non-response, by getting out of scope of the surveyed population, or by being dropped from the sample. New split-off households appear in the sample, some households from the year before last are re-contacted and a new sub-sample of households is added—as outlined in the interannual change scheme (Scheme 2).

The total annual change of cross-sectional indicators' values can therefore be decomposed into several effects according to these annual changes. A progressive change of an indicator can be followed by altering either the structure of the sample, cross-sectional weights, or recorded values. Scheme 3 depicts the decomposition of cross-sectional change of indicators into these several effects.

The first column of Scheme 3 represents the actual cross-sectional values: the values responded in 2005 by all participating households in the 2005 cross-sectional sample (core households, households that did not respond next year, household that went out of scope in 2006, and households that were dropped in 2006[4]) weighted by 2005 cross-sectional weights. This sample composition corresponds to the actual reported value of cross-sectional indicators from the 2005 survey. The second column covers only values responded in 2005 by core, non-response, and out-of-scope households weighted by 2005 cross-sectional weights. Columns one and two would only vary in excluding responses of households dropped from the sample in 2006. Thus, the difference between the first and second columns in Scheme 3 would represent the effect of dropped households on the annual change of an indicator.[5] However, a sample replication was dropped first in 2009 in the Czech Republic. Therefore, this effect will be excluded in the section 4 describing the decomposition of indicators in the Czech Republic between years 2005 and 2006. Thus, the first and second columns in Scheme 3 would be actually the same in the Czech data in this period.

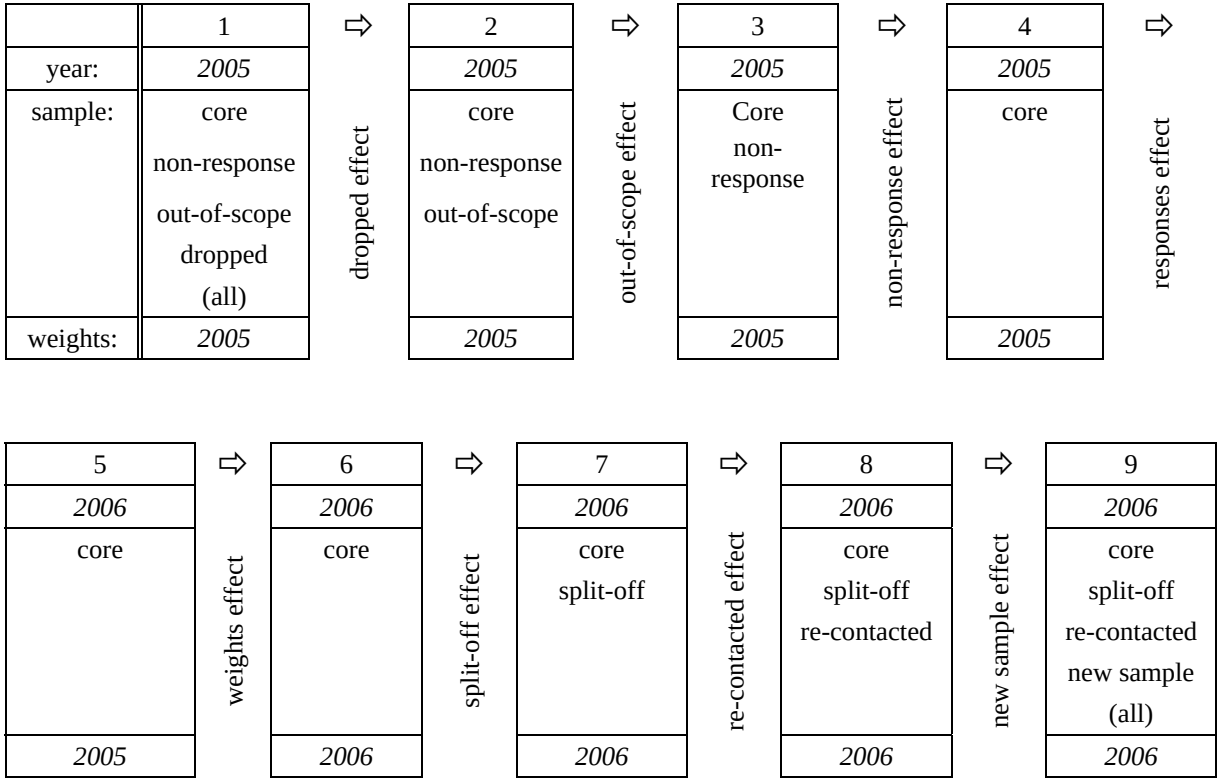
The third column includes only values responded in 2005 by core and non-response households weighted by 2005 cross-sectional weights. Therefore, columns two and three only vary in excluding responses of households that were out of scope in 2006. The difference between the second and third columns in Scheme 3 represents the effect of out-of-scope households on the annual change of an indicator.[6]

[4] Not applicable in the Czech Republic in this period.

[5] More precisely, this effect is in fact "an effect of excluding dropped households".

[6] Again, it would be more precise to call this effect as "an effect of excluding out-of-scope households".

Scheme 3. Decomposition of cross-sectional change of indicators



The fourth column in Scheme 3 compared to column three does not include the values responded by households not responding anymore in 2006 (refusals, non-contacts or other reasons for non-response). The difference in estimates between the fourth and third columns can be attributed to the "non-response effect".[7] The fourth column covers only core households' responded values reported in 2005 and weighted by cross-sectional weights originally assigned to these households.

The comparison of columns five and four depicts the effect of actual change of core households' responses from one year to the next—"responses effect"—since it shows how the core households responded in 2006 compared to 2005 using the same cross-sectional weights. Or, more precisely, it indicates what would have been the result in the sample of core households in 2005 if these households had responded like they later did in 2006.

Subsequently, weights are changed in such a way that from column six onwards cross-sectional weights of 2006 values are applied. Hence, the difference between columns six and five in Scheme 3 stands for the effect of weights caused by the fact that the cross-sectional weights for the core households are generally not identical in years 2005 and 2006.

In the three last columns in Scheme 3, the sample is extended with the remaining groups of households—those, who were not part of the cross-sectional sample in 2005. Moving from column six to column seven, the sample includes also split-off households newly created in 2006. Therefore, the "split-off effect" is captured.

Re-contacted households (households that would have responded in the year before 2005 but would not have in 2005 and would have re-joined the sample in year 2006) are added in column eight. The difference between columns eight and seven then shows the "re-contacted effect". Re-contacted households cannot appear before third year of the survey. In 2006, there was only second year of the EU-SILC survey in the Czech Republic. Therefore, the re-contacted effect could not exist in the applied period and thus will be excluded in section 4, which depicts the decomposition of indicators in the Czech Republic between years 2005 and 2006. Thus, the seventh and eighth columns in Scheme 3 would be actually the same in the Czech data in this period.

Finally, last column includes also responses of households newly added in 2006—the new sample replication surveyed for the first time in 2006. Their inclusion into the sample changes the sample composition and the estimates—here, the "new sample effect" is captured (the difference between columns nine and eight). This means that the last column—column nine—finally contains actual cross-sectional sample in 2006 since all 2006-households are included, their reported values in the 2006-survey questionnaire are used and the 2006-cross-sectional weights are used. The indicator value arrives at its final destination—the actual value estimated from the 2006 cross-sectional sample. The difference between columns nine and one corresponds to the actual annual change in the indicator's value—the difference between the 2006 and 2005 cross-sectional indicators.

[7] Similarly to the previous effects, the effect of non-response could be called "an effect of excluding non-response households".

4 Decomposition results 2005 versus 2006

The Czech survey sample contained 4351 successfully interviewed households in 2005 and 7483 successfully interviewed households in 2006. Figure 1 shows the structure of 2005 versus 2006 sample. 87.6% of the 2005 households were the core households those included also in the 2006 sample. 11.2% of households did not respond next year and 1.2% of the households were those that went out of scope. The core households created only one half (50.9%) of households included in 2006. The split-off households represent only 0.5%. There was a substantial part of new sample replication households in 2006 (48.5%).

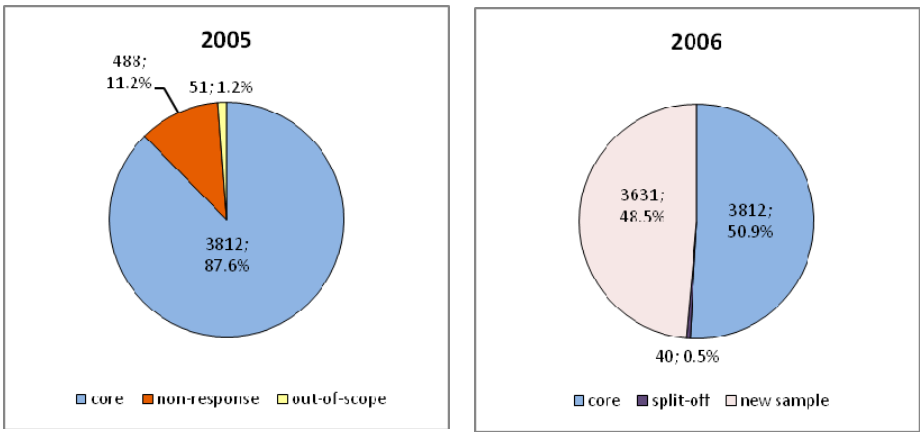


Fig. 1. Structure of 2005 versus 2006 sample
Source: National EU-SILC 2005 and 2006.

This sub-section describes first the annual cross-sectional trend decomposition using several examples of categorical indicators of material deprivation. The indicators are the percentage shares of households having particular problem. Here, the core households may switch between the categories “yes, we have this problem” and “no, we do not have this problem” in their responses between the two years.

The actual published cross-sectional indicator of “damp flat” (EU-SILC variable HH040[8]) in 2005 in the Czech Republic is depicted by first column in Figure 2—estimated 19.91% of households had a problem with various kinds of technical conditions of their flat. Starting at this point the indicator followed its imaginary path described in previous section to its 2006 cross-sectional published value—20.19%, the value represented by last column in Figure 2. Therefore, the total annual change of this indicator is an increase of the share of households suffering from living in a damp flat by approximately 0.3 percentage points.

Excluding the out-of-scope households caused a growth from 19.91% to 20.00% of this indicator (see second column in Figure 2), the result of the fact that within the

[8] The recommended question wording for this variable is: Do you have any of the following problems with your dwelling / accommodation? - a leaking roof, - damp walls /floors/ foundation, - rot in window frames or floor.

excluded group of out-of-scope households a significantly lower share of households responded "yes" to this material deprivation question (11.99%). However, the number of out-of-scope households is very low and therefore there was only moderate influence of their responses on the indicator's value. Exclusion of the non-response households showed a similar effect a further increase from 20.00% to 20.15% was caused by a lower share of households with the problem within the now excluded group of non-respondents (their share of "yes" responses to this material deprivation question was 18.93%).

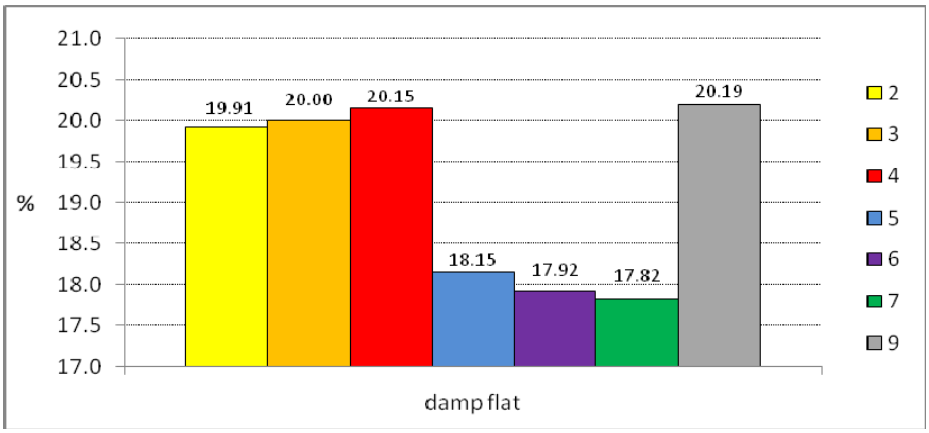


Fig. 2. Decomposition—HH040: Leaking roof, damp walls/floors/foundation, or rot in window frames or floor (households with the problem as a % of total households), 2005x2006
Source: National EU-SILC 2005 and 2006.
Note: For caption, see Scheme 3. The question was whether, in the judgement of the household respondent, the dwelling has a problem with a leaking roof, damp ceilings, dampness in the walls, floors or foundation or rot in window frames and doors.

The effect of responses showed interesting results, it caused a rather significant drop of the indicator from 20.15% to 18.15%. Thus, there occurred a noticeable change of responses from "yes" to "no" within the core households between the two years. Actually, there were considerable shifts in both directions: 314 (8.2%) core households changed the response from "yes" to "no" while 226 (5.9%) of them changed the response in the other direction—from "no" to "yes". 85.5% of core households provided the same response in both years: 465 (12.2%) core households had the problem in both years while 2807 (73.6%) had the problem neither in 2005 nor in 2006. Only 59.7% of core households that had this problem in 2005 confirmed to have this problem also in 2006. 7.5% of core households that had not this problem in 2005 reported newly the problem in 2006.[9]

The effect of weights further lowered the indicator to 17.92%. Also adding the split-of-households caused a slight decrease to 17.82%—despite their low number,

[9] The percentages of core households changing or sustaining their responses stated in the whole text are unweighted.

their influence is strengthened by the very low value of the indicator in this specific sub-group (only 8.55% responded a problem with a damp flat).

Finally, the decreasing trend was overbalanced by the effect of new households and the indicator rose from 17.82% to 20.19%. Therefore, the total annual trend was increasing. 22.74% of new households responded to have a problem with a damp flat compared to just 17.92% of core households in 2006. It seems that households included for the second time in the survey are more optimistic while the new households exhibit more often a negative opinion. The question for further research would be whether the optimism of households rises also in the third year of their participation in the survey.

Similar "negativism" of new households is also present in the indicator of problems with "a dark dwelling" (EU-SILC variable HS160). The effect of new households worked towards an increasing trend although in this case not sufficiently strongly to overbalance the previous rather decreasing effects (see Figure 3). 5.26% of new households themselves responded a problem with a dark dwelling compared to just 4.15% of core households in 2006.

The out-of-scope and non-response effects lowered the indicator. 10.88% households within the group of out-of-scope households had a problem with a dark flat, within the non-response households it was 7.37%. Therefore, exclusion of these groups of households from the sample lowered the indicator.

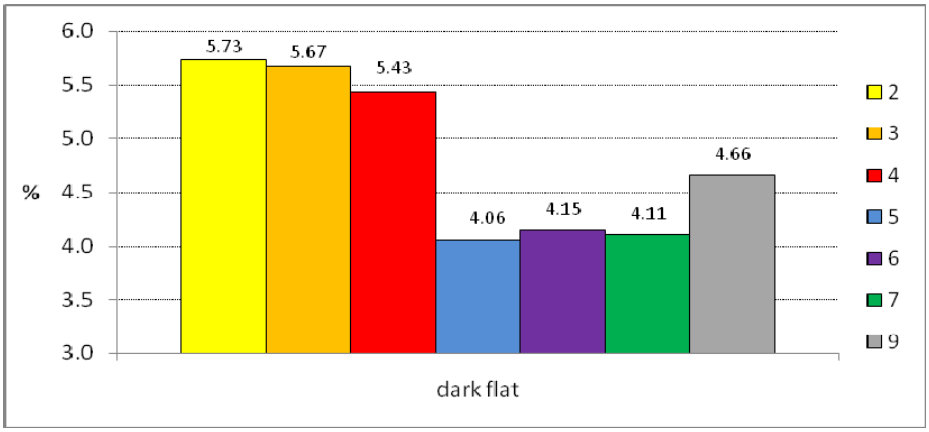


Fig. 3. Decomposition—HS160: Problems with the dwelling: too dark, not enough light (households with the problem as a % of total households), 2005x2006
Source: National EU-SILC 2005 and 2006.
Note: For caption, see Scheme 3. The question was whether the respondent feels "the dwelling to dark, not enough light" to be a problem for the household.

The effect of responses significantly decreased the indicator. The results for core households were 5.43% in 2005 and 4.06% in 2006 (see third and fourth columns in Figure 3). Therefore, there was a shift from "yes" to "no" responses within the core sample households continuing in the survey between the two years also for this material deprivation indicator. Despite relatively small share of core households suffering from living in a dark flat 3.1% of core households changed their responses

between the two years—2.1% from “yes” to “no” and 1.0% from “no” to “yes”. Only 58.0% of core households having this problem in 2005 confirmed to have the problem also in 2006.

The annual trend of an indicator of “ability to keep home adequately warm” (EU-SILC variable HH050) indicates similar main features: the overall trend shows a decrease, the effect of responses lowers the indicator substantially, and the effect of new households increases it (see Figure 4). As far as the stability of core households’ responses between the two years is concerned, 8.4% changed the response (4.8% from “unable” to “able” and 3.6% from “able” to “unable”) and 55.9% of core households that was unable to keep home warm in 2005 reported being unable to manage that also in 2006.

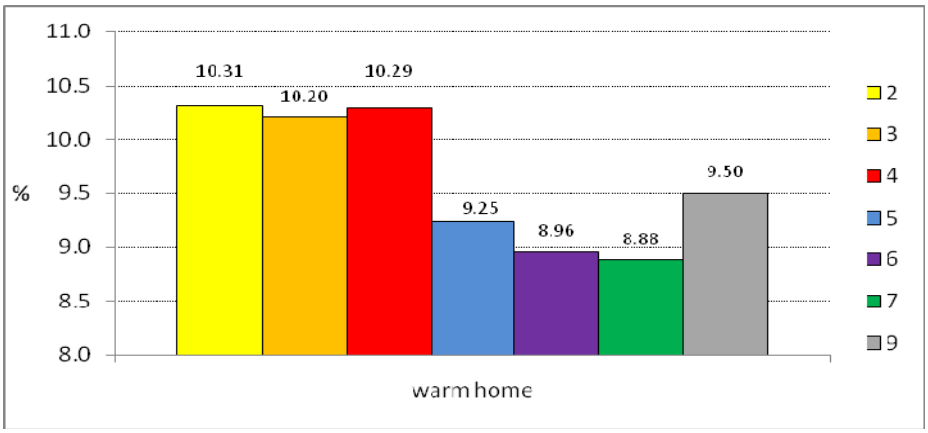


Fig. 4. Decomposition—HH050: Ability to keep home adequately warm (not able households as a % of total households), 2005x2006
Source: National EU-SILC 2005 and 2006.
Note: For caption, see Scheme 3. The question was about ability to pay to keep the home adequately warm.

The indicator of “capacity to face unexpected financial expenses” (EU-SILC variable HS060) shows decreasing effects until the new households are added. However, their slight growth effect from 40.84% to 41.29% does not override the overall annual decrease of the indicator (see Figure 5—last column of “unexpected expenses”). The effect of responses displayed a sizeable drop of the indicator since there was a general shift in responses from “incapable” to “capable”. However, there was again a considerable instability of responses between the two years: 11.6% of core households changed the response from “incapable” to “capable” and 9.2% from “capable” to “incapable”. Only 79.2% of core households responded same answer in both years (32.3% incapable and 47.0% capable) and only 73.6% of core households that were incapable to face unexpected expenses in 2005 were incapable to face them also in 2006.

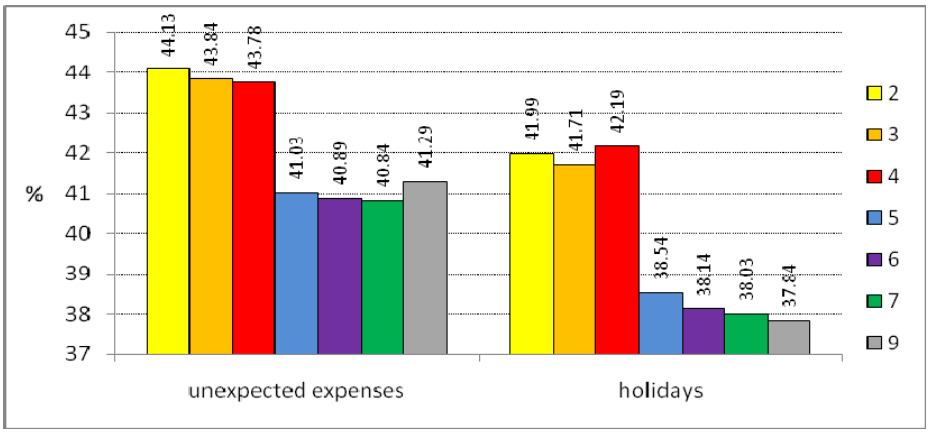


Fig. 5. Decomposition—HS060: Capacity to face unexpected financial expenses and HS040: Capacity to afford paying for one week annual holiday away from home (incapable households as a % of total households), 2005x2006
Source: National EU-SILC 2005 and 2006.
Note: For caption, see Scheme 3. The question on unexpected expenses was whether, according to the household respondent, the household can face itself unexpected financial expenses. The question on holidays was about ability to pay, regardless of whether the household actually want the item. The answer was "yes" if, according to the household respondent, the household can afford to pay for a week's annual holiday away from home.

The annual trend of the indicator of "capacity to afford paying for one week annual holiday away from home" (EU-SILC variable HS040) was also decreasing (see the right side in Figure 5). The out-of-scope effect, i.e. excluding the out-of-scope households, caused a slight decline from 41.99% to 41.71%. The non-response effect caused a growth of this indicator (from 41.71% to 42.19%) since the indicator within the non-response group stood only at 38.35%.

The effect of changed responses of the core sample was very strong—the indicator dropped from 42.19% to 38.54%. Core households that changed the response from "cannot afford" to "can afford" represented 9.6% while 6.0% of core households changed the response in the opposite direction. The share of households that provided the same response in both years was rather low (84.4% of core households); 33.6% of them could not afford a holiday in both years. 77.9% of core households that could not afford a holiday in 2005 could not afford it also in 2006.

The weights and split-off effects further slightly lowered the indicator. Interesting is that also the effect of new households was moderately declining. Adding the group of new households in the sample lowered the indicator from 38.03% to 37.84%. Among the new households themselves the share of households that cannot afford to pay for holiday was even slightly lower (37.64%) than among the core households (38.14%) in 2006—a different pattern than for previous indicators.

All the above-mentioned indicators were more of a subjective nature, largely based on household respondents' opinion and/or feelings. The next group of indicators consist of items of material deprivation that are based on possession or enforced lack of some household amenities and durables rather than on judgement of households' situation.

Figure 6 depicts examples of such indicators based on the possession of basic household amenities. The shares of households living without a bath/shower and without a toilet (EU-SILC variables HH080 and HH090, respectively) were very low and further declined between the two years. The particular effects were almost negligible. The responses of core households were, quite naturally, very stable. The most significant drop occurred due to adding the new households in the sample as shown in last columns in Figure 6. The shares of households living without a bath/shower and without a toilet were only 0.9% within the new household sample.

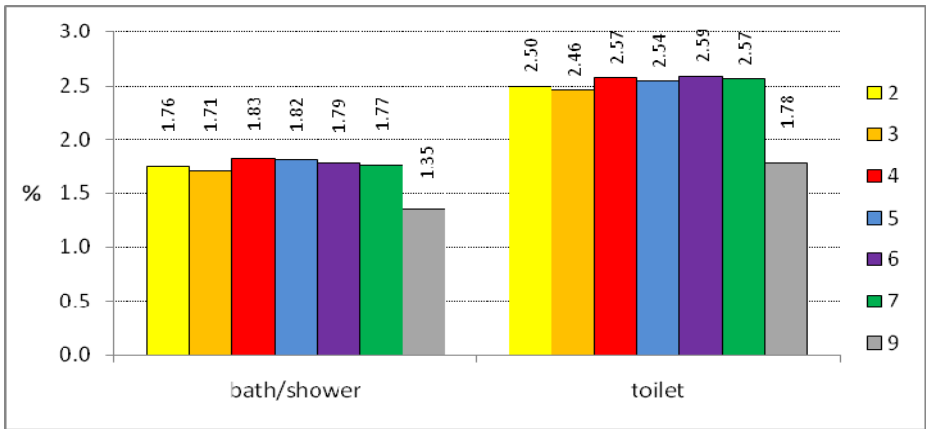


Fig. 6. Decomposition—HH080: Bath or shower in dwelling and HH090: Indoor flushing toilet for sole use of household (households "without" as a % of total households), 2005x2006
Source: National EU-SILC 2005 and 2006.
Note: For caption, see Scheme 3. The question on bath was whether the dwelling has proper room with a bath or a shower.

The indicators of enforced lack of telephone, TV and washing machine (EU-SILC variables HS070, HS080 and HS100, respectively) are similar examples of decreasing annual trends with declining effects of new households (see Figure 7). The effect of responses in case of washing machine worth mentioning since it caused an increase of this indicator. Although the share of households that cannot afford a washing machine is very low approximately one percent of core households changed their response (22 core households (0.6%), could afford/possessed a washing machine in 2005 but could not afford it one year later, while 17 core households (0.5%), could not afford it in 2005 but could afford it/possessed it in 2006).

The most apparent differences between the "subjective indicators" and "indicators of enforced lack" are: (i) the effect of responses does not change the indicators to such a large extent and (ii) the effect of new households lowered the indicators in case of the "indicators of enforced lack".

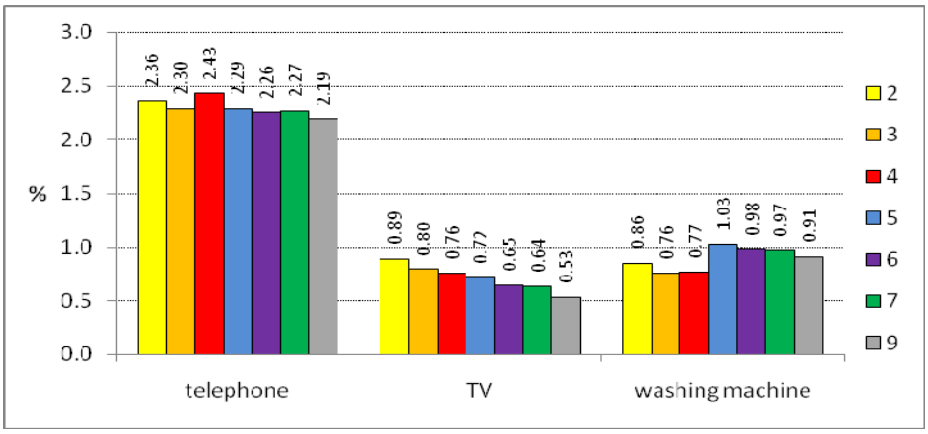


Fig. 7. Decomposition of indicators of enforced lack of durables, based on: HS070: Do you have a telephone (including mobile phone)?; HS080: Do you have a colour TV?; and HS100: Do you have a washing machine? (households that do not possess the item and cannot afford it as a % of total households), 2005x2006
Source: National EU-SILC 2005 and 2006.
Note: For caption, see Scheme 3. The question was whether the household have the item (telephone, including mobile phone; TV; washing machine) or whether the household does not have the item because it cannot afford it (enforced lack) or for other reasons. The percentage of households that do not possess the item and cannot afford it is shown.

5 Conclusion

This study focused on annual cross-sectional trends of indicators in EU-SILC survey. It aimed to reveal what is behind the annual change of indicators: what is caused by actual change of responses of surveyed households and to what extent are the results influenced by the different structure of the sample and weights in each year. We proposed a decomposition method that shows how the value of an indicator in year *N-1* moves to its value in year *N* and follows its imaginary path through various effects. We suggested several effects that consist of the impact of change of reported values by households present in the sample in both subsequent years and also of the impact of the change in sample composition between years *N-1* and *N* (e.g., attrition and incorporation of a new sample replication) and of the impact of different cross-sectional weights. If some extraordinary or unexpected annual change of an indicator occurs this method enables to detect the main source of such a change.

Regarding the results of decomposition conducted with the use of Czech EU-SILC 2005 and 2006, we can conclude some general features. Majority of the indicators of material deprivation based on household respondent's judgement of the households' situation proved that the overall annual trend was decreasing, the effect of responses lowered the indicators the most and the effect of new household increased the indicators in most cases. The most interesting seems to be the fact that without incorporating of the new households in the sample the annual decline would be more apparent. The effect of new households even overbalanced the decreasing trend of all

other previous effect in case of “damp flat” such that the overall trend of this indicator between 2005 and 2006 was increasing.

The indicators of material deprivation that were based on enforced lack of basic amenities and durables rather than on judgement of households’ situation displayed rather different features: (i) the effect of responses did not change the indicators to a great extent and (ii) the effect of new households showed a general tendency to lower the deprivation indicators in case of the “indicators of enforced lack”.

References

1. Kasprzyk D., Duncan G., Kalton G., Singh M.P. , Panel Surveys, John Wiley and Sons, 1990

2. Lynn P., Methodology of Longitudinal Surveys, John Wiley and Sons, 2009

3. Mysíková M., Smetanová T., Tourek Š., Zelený M., Effects of Rotational Design and Attrition in Czech EU-SILC, ESRA 2007 Conference, Prague June 24th-29th 2007

4. Sampling , Eurostat Document E2-EU-SILC 51/01, 2001

Approximation of the Stop-loss Premium for the Compound Poisson Distribution

Anna Nikodem

University of Economics, ul. Komandorska 118/120, 53-345 Wrocław, Poland
anna.nikodem@ue.wroc.pl

Abstract. In this paper the approximation of stop-loss premium for the compound Poisson distribution will be considered. When the claim size distribution is known, we can compute stop-loss premium by using the recursive method. If we don't know this distribution, we can use the approximation of the aggregate loss distribution to compute the stop-loss premium. In literature various approximation are described. In this paper several of them are compared for light tailed and heavy tailed claim size distribution.

Keywords: Compound Poisson distribution, Stop-loss premium.

1 Introduction

In order to protect oneself against large individual claims or against the fluctuation in the number of claims the insurer takes out reinsurance cover for his insurance portfolio. The aggregate claim amount is shared between the cedant and the reinsurer. The expected cost of this insurance is called the net stop-loss premium and is defined

$$\pi(d) = E[(S-d)_+], \tag{1}$$

where S is the aggregate claims amount with support $[0, \infty)$, d is the retention.

The aggregate claims amount of the insurer portfolio is defined by

$$S = X_1 + X_2 + ... + X_N,$$

where the number of claim N is a random variable, the individual claims X_i are independent and identically distributed and the random N and X_i are independent. In this paper we consider the calculation of the stop-loss premium, when the aggregate claims amount has the compound Poisson distribution, i.e. if N is Poisson distributed.

If the aggregate loss distribution is discrete, the stop-loss premium can be calculated

$$\pi(d) = \sum_{x>d} (x-d)\, f_S(s) = \sum_{x>d} [1 - F_S(s)], \tag{2}$$

for continuous distribution the stop-loss premium is given by

$$\pi(d) = \int_d^{\infty} (s-d)\, f_S(s) ds = \int_d^{\infty} [1 - F_S(s)] ds, \tag{3}$$

where $f_S(s)$ is the probability density or the probability function of aggregate claim amount and $F_S(s)$ is the distribution function of *S*.

2 Computation of the Stop-loss Premium

In order to compute the stop-loss premium we can determine the distribution of the aggregate claims amount *S*. When we know the claim size distribution we can use the recursive method. If we don't know this distribution, we can approximate the density $f_S(s)$ by a function that uses the mean, variance, skewness of *S*. For the compound Poisson distribution with parameter λ the first three central moments of *S* are equal to

$$E(S) = E(N) \cdot E(X) = \lambda E(X),$$

$$V(S) = V(N) \cdot \left(E(X)\right)^2 + E(N) \cdot V(X) = \lambda E(X^2),$$

$$E\left[(S - E(S))^3\right] = E\left[(N - E(N))^3\right]\left(E(X)\right)^3 +$$
$$+3V(N)E(X)V(X) + E(N)E\left[(X - E(X))^3\right] = \lambda E(X^3).$$

These parameters can be calculated without knowing the probability density function.

Application the Panjer's Recursion

If the individual claim X_i is a discrete arithmetic random variable with probabilities $p(x)$ and *N* has Poisson distribution with parameter λ, then

$$f_S(0) = e^{-\lambda(1-p(0))}, \tag{4}$$

$$f_S(s) = \frac{1}{s}\sum_{h=1}^{s} \lambda h p(h) f(s-h). \quad (5)$$

Applying the Panjer's recursion, the stop-loss premium can be calculated recursively (see [1], [2], [3]). For integer retention d we have

$$\pi(d) = \pi(d-1) - [1 - F_S(d-1)]. \quad (6)$$

Application of the Normal Approximation

If the skewness of the aggregate claims amount is equal to zero then the compound Poisson distribution can be approximated by a normal distribution with mean $E(S)$ and standard deviation $D(S)$, i.e. $S \sim N(E(S), D(S))$. If $T \sim N(0,1)$ then $S = D(S)T + E(S)$ and the stop-loss premium with the retention d is equal to

$$\pi(d) = E[(S-d)_+] = E[(D(S)T + E(S) - d)_+] = \\ = D(S)E\left[\left(T - \frac{d - E(S)}{D(S)}\right)_+\right].$$

Since for $T \sim N(0,1)$ we have

$$E[(T-t)_+] = \phi(t) - t[1 - \phi(t)],$$

hence the stop-loss premium approximated by normal distribution is equal to (see [3])

$$\pi(d) = D(S)\phi\left(\frac{d - E(S)}{D(S)}\right) - (d - E(S))\left[1 - \Phi\left(\frac{d - E(S)}{D(S)}\right)\right], \quad (7)$$

where $\Phi(x)$ is the standard normal distribution function and $\phi(x)$ is the standard normal density function.

Application NP-approximation

If the skewness of the aggregate claims amount S is small, i.e. $0 \le \gamma_S \le 1$ we can use the NP- approximation. This approximation states that

$$P\left(\frac{S-E(S)}{D(S)} \le x\right) \approx \Phi\left(\sqrt{\frac{9}{\gamma_S^2}+\frac{6x}{\gamma_S}+1}-\frac{3}{\gamma_S}\right) \text{ for } x \ge 1.$$

Since the random variable S is approximated by normal distribution, we have

$$\pi(d) = E[(S-d)_+] = E[(D(S)T + E(S) - d)_+] = \\ = D(S)E\left[\left(T - \frac{d-E(S)}{D(S)}\right)_+\right],$$

where $E(T)=0$, $V(T)=1$, $\gamma > 0$. For random variable T and

$$w(k) = \sqrt{\frac{9}{\gamma^2}+\frac{6k}{\gamma}+1}-\frac{3}{\gamma} \text{ for } k \ge 1,$$

we obtain

$$E\left[\left(T-k\right)_+\right] = \phi(w(k)) + \frac{\gamma}{6}w(k)\phi(w(k)) - d\left[1-\Phi(w(k))\right].$$

Finally, the stop-loss premium for the NP-approximation is equal to (see [1], [3])

$$\pi(d) = D(S)\left[\phi(w(d)) + \frac{\gamma_S}{6}w(d)\phi(w(d)) - d[1-\Phi(w(d))]\right],$$

where $w(d) = \sqrt{\frac{9}{\gamma_S^2}+\frac{6(d-E(S))}{D(S)\gamma_S}+1}-\frac{3}{\gamma_S}$.

Application of the Translated Gamma Approximation

Let the aggregate claim distribution have a form of $S = Z + x_0$, where Z has the gamma distribution with, i.e. $Z \sim gamma(\alpha, \beta)$. Hence, the cumulative distribution function $F_S(s)$ can be approximated by the gamma cumulative distribution function $G(s-x_0;\alpha,\beta)$. The parameters α, β and x_0 are chosen in such way that the first three moments of S and $Z + x_0$ agree. Hence

$$\alpha = \frac{4}{\gamma_S^2}, \quad \beta = \frac{2}{\gamma_S D(S)}, \quad x_0 = E(S) - \frac{2D(S)}{\gamma_S},$$

where γ_S is the skewness of *S*. For the translated gamma distribution the stop-loss premium has a simple expression (see [1], [3])

$$\pi(d) = \frac{\alpha}{\beta}[1 - G(d - x_0, \alpha + 1, \beta)] - (d - x_0)[1 - G(d - x_0, \alpha, \beta)],$$

where $G(x, \alpha, \beta)$ is the gamma cumulative distribution function

Application of the Gaussian Exponential Approximation

For a random variable *S,* when taking values in $[0, \infty)$, with the continuous distribution $F_S(s)$ and a finite mean $E(S)$, the failure rate is defined by

$$h(s) = -\frac{d}{ds} \ln \bar{F}_S(s)), \tag{8}$$

and the mean residual life function

$$m(s) = E[S - s \mid S > s] = \frac{\pi(s)}{\bar{F}_S(s)},$$

where $\bar{F}_S(s) = 1 - F_S(s)$ is the survival function. These two function are related by

$$h(s) = \frac{1 + m'(s)}{m(s)}$$

and $m'(s)$ is the derivative of the mean residual life function. The cumulative failure rate is given by

$$H(s) = \int_0^s h(t)dt = \int_0^s \frac{dt}{m(t)} + \ln\left(\frac{m(s)}{m(0)}\right). \tag{9}$$

Derived from 8 we have

$$\bar{F}_S(s) = \exp\left\{-\int_0^s h(u)du\right\} = \exp\{-H(s)\}.$$

To make use of 9 the survival function is equal to

$$\overline{F}_S(s) = \frac{m(0)}{m(s)} \cdot \exp\left(-\int_0^s \frac{dt}{m(t)} \right), \tag{10}$$

where $m(0) = E(S)$. For a discrete arithmetic compound Poisson distribution, the reciprocal of $m(s)$ is approximately linear and for high values of s the $m(s)$ is constant. Hence (see [2])

$$m(s) = E[S - s \mid S > s] = \begin{cases} \dfrac{1}{\alpha + \beta s}, & 0 \le s \le s_0, \\ \dfrac{1}{\alpha + \beta s_0}, & s \ge s_0, \end{cases} \tag{11}$$

where $0 < \beta \le \alpha^2$. To make use of 10 and 11 we finally obtain that the survival function of S is equals

$$\overline{F}_S(s) = \begin{cases} \left(\alpha + \left(\dfrac{\alpha}{\eta} \right)^2 \cdot \dfrac{s}{E(S)} \right) \cdot \exp\left(-\alpha \cdot \dfrac{s}{E(S)} - \dfrac{1}{2} \left(\dfrac{\alpha}{\eta} \cdot \dfrac{s}{E(S)} \right)^2 \right), & 0 \le s \le s_0, \\ \overline{F}_S(s_0) \cdot \exp\left(+ \left(\dfrac{s - s_0}{m(s_0)} \right) \right), & s \ge s_{0,} \end{cases}$$

where $\alpha = \overline{F}_S(0) = 1 - e^{-\lambda}$, $\beta = \left(\dfrac{\alpha}{\eta} \right)^2$, $\eta \ge 1$. Since the mean residual life function

$$m(s) = \frac{\pi(s)}{\overline{F}_S(s)},$$

then the stop-loss premium has a following form (see [2])

$$\pi(s) = \begin{cases} E(S) \cdot \exp\left(-\alpha \cdot \dfrac{s}{E(S)} - \dfrac{1}{2} \left(\dfrac{\alpha}{\eta} \cdot \dfrac{s}{E(S)} \right)^2 \right), & 0 \le s \le s_0, \\ \pi(s_0) \cdot \exp\left(-\left(\dfrac{s - s_0}{m(s_0)} \right) \right), & s \ge s_0, \end{cases}$$

or, equivalently

$$\pi(s)=\begin{cases} E(S)\cdot\dfrac{\phi\left(\eta+\dfrac{\alpha}{\eta}\cdot\dfrac{s}{E(S)}\right)}{\phi(\eta)}, & 0\le s\le s_0, \\ \pi(s_0)\cdot\exp\left(-\left(\dfrac{s-s_0}{m(s_0)}\right)\right), & s\ge s_0, \end{cases}$$

where $\phi(x)$ is the standard normal density function. The unknown parameters η and s_0 are related to $1+v^2$, where v is the coefficient of variation, in such a way that

$$E(X^2)=\left(1+\left(\frac{D(S)}{E(S)}\right)^2\right).$$

Since the second moment of S equals $\pi^2(0)$, we obtain (see [2])

$$2\frac{\eta}{\alpha\phi(\eta)}\left\{\Phi\left(\eta+\frac{\alpha}{\eta}\cdot z\right)-\Phi(\eta)+\frac{\phi\left(\eta+\dfrac{\alpha}{\eta}\cdot z\right)}{\eta+\dfrac{\alpha}{\eta}\cdot z}\right\}=1+\left(\frac{D(S)}{E(S)}\right)^2,$$

where $z=E(S)\cdot s_0$. To calculate the unknown parameters η and z, first we can find the parameter η_0 by using (see [2])

$$2\frac{\eta_0}{\alpha}\cdot\frac{\bar{\Phi}(\eta_0)}{\phi(\eta_0)}=1+\left(\frac{D(S)}{E(S)}\right)^2,$$

and then we can calculate parameter z

$$z=\frac{1}{2}\left(1+\sqrt{1+4\frac{\eta_0}{\alpha}}\right),$$

where $\alpha=1-e^{-\lambda}$.

3 Examples and Comparison

In this section, the approximation of stop-loss premium are compared for light tailed and heavy tailed claim size distribution. The approximated stop-loss premium will be compared with the stop-loss premium calculated by using the exact method,

i.e. Panjer recursion. For an integer-valued S we can use the Panjer recursion to calculate the probabilities $f_S(s)$. The recursive method can be used also after discretization of the probability density function of the individual claim size X_i. The relative difference between the approximated stop-loss premium and the exact stop-loss premium are shown on the graphs in this chapter.

Light Tailed Distribution

In Fig. 1, the skewness of the aggregate claims amount is equal to 1,06. In this case, the translated gamma approximation and the Gaussian exponential approximation is the most accurate. The largest relative error for the normal and NP approximations, especially for the large retention.

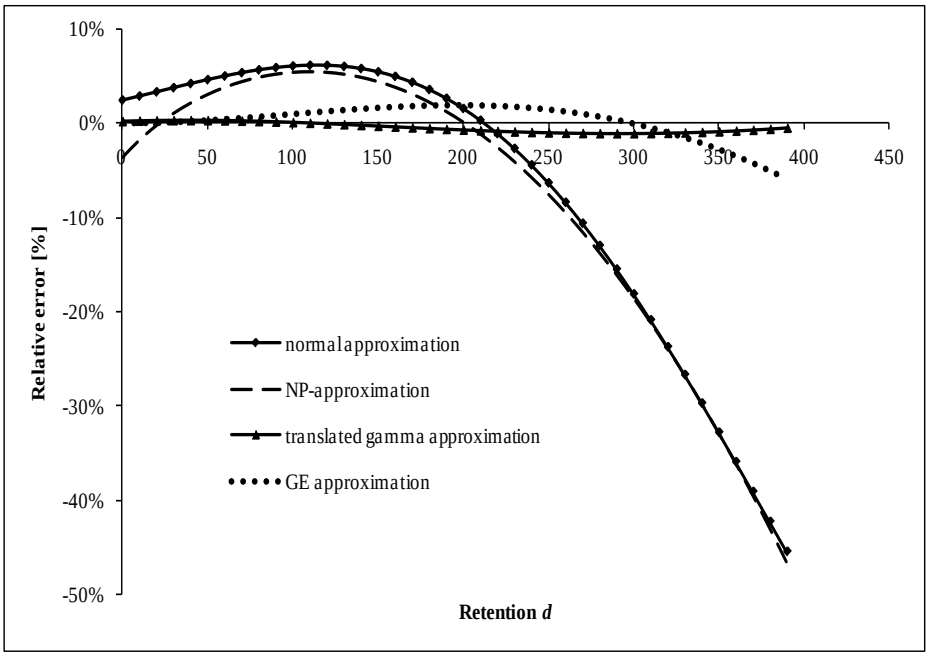


Fig. 1. $N \sim Poisson(4),\ X \sim Exp(0,02),\ E(S) = 200,\ V(S) = 20\,000,\ \gamma_S = 1,06.$

In Fig. 2, the aggregate claims distribution has skewness equals to 4.24 which is greater than 1,06 value that could be found in Fig. 1. In this case, the stop-loss premiums calculated by normal and NP-approximation are the largest. The Gaussian exponential approximation is the best solution here.

Comparing the approximations of stop-loss premium we can see that the Gaussian exponential approximation is best when the Poisson parameter is small, even though the skewness of the aggregate claims amount is large.

Heavy Tailed Distribution

In Fig. 3, the individual claim size has the Weibull distribution. The skewness of the aggregate claims distribution is equal to 2,74.

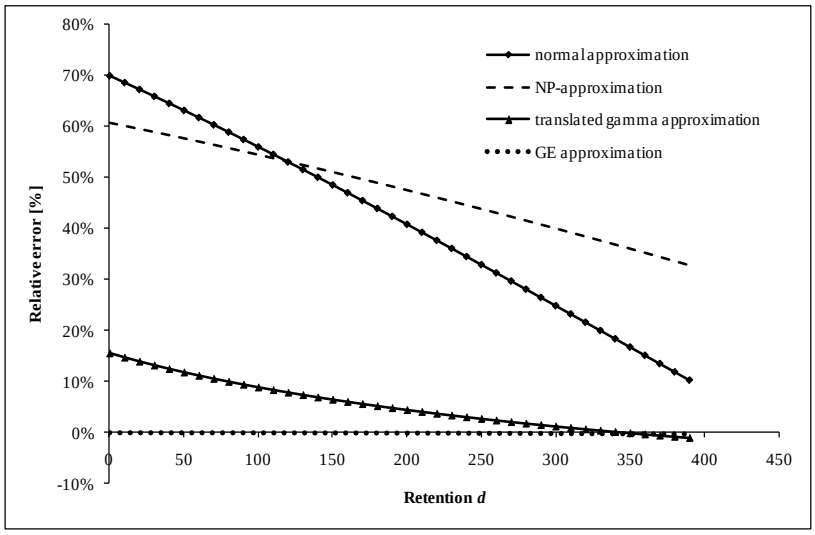


Fig. 2. $N \sim Poisson(0,25),\ X \sim Exp(0,00125),\ E(S) = 200,\ V(S) = 320\,000,\ \gamma_S = 4,24.$

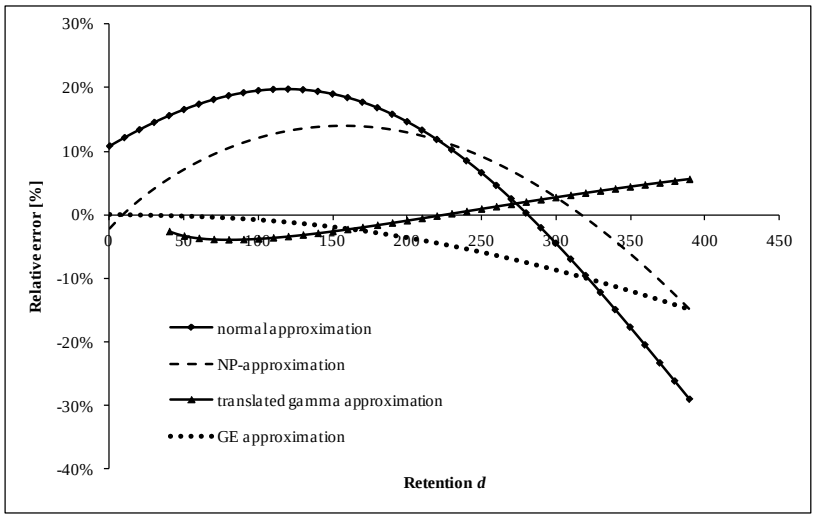


Fig. 3. $N \sim Poisson(5),\ X \sim Weibull(0,5;\ 20),\ E(S) = 200,\ V(S) = 48\,000,\ \gamma_S = 2,74.$

In Fig 4, the skewness of the aggregate claims amount is equal to 1,22. Since the skewness of S is smaller than in Fig. 3, the translated gamma approximation gives more accurate values.

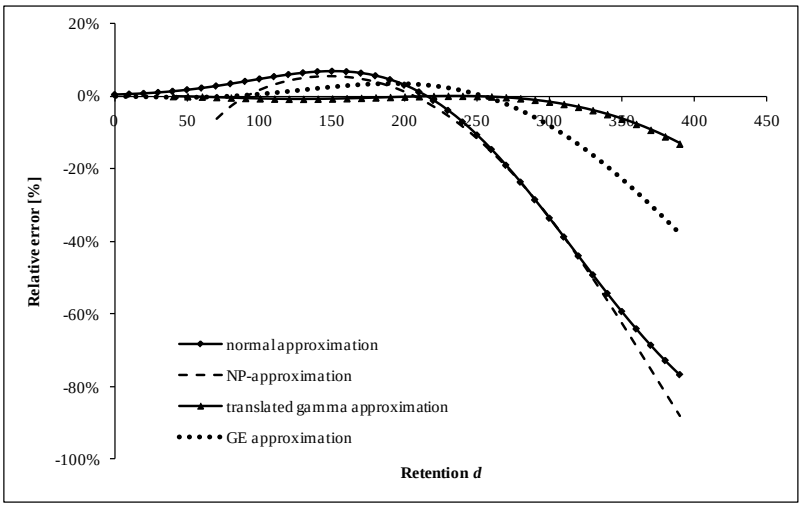


Fig. 4. $N \sim Poisson(25)$, $X \sim Weibull(0{,}5; 4)$, $E(S) = 200$, $V(S) = 9\,600$, $\gamma_S = 1{,}22$.

Conclusion

In case of the light tailed and the heavy tailed claim size distribution, accuracy of the stop-loss premium calculated with the use of the normal approximation and NP-approximation is dependent on the value of the variance. The translated gamma approximation and Gaussian exponential approximation are the best one but they are better for the light tailed claim size distribution than for the heavy tailed claim size distribution.

References

1. Daykin C.D., Pentikainen T., Pesonen M.: Practical risk theory for actuaries. Chapman & Hall, London (1994).
   ISBN 0-412-42850-4.
2. Hurlimann W.: A Gaussian exponential approximation to some compound Poisson distributions. ASTIN Bulletin 33, 41--55 (2003).
3. Kaas R., Goovaerts M., Dhaene J., Denuit M.: Modern actuarial risk theory. Kluwer Academic Publishers, Boston (2001).
   ISBN 0-7923-7636-6.
4. Reijnen R., Alberts W., Kallenberg W.C.M.: Approximation for stop-loss reinsurance premiums. Insurance: Mathematics and Economics 36, 237--250 (2005).

Application of density mixture in the probability model construction of wage distributions

Roman Pavelka[1]

[1] doctoral student of KSTP FIS, University of Economics in Prague,
pavelka@trexima.cz

Abstract. A practice showed, that efforts to model any wage distribution by probability distributions (normal, log-normal etc.) ordinarily in statistics applied are not too successful. Therefore it was used in this contribution as a wage distribution model as a model in shape of distributional mixtures. In accordance with empirical studies the wage distribution is composed from two or more log-normal distributions. Each log-normal distribution defines wage homogeneous subgroup of employees. The statistical model of a wage distribution can be formulated as a mixture (a linear combination) of log-normal distributions. Constituent elements of this model are weighted the element relative share weights of probability distribution the whole model.

Keywords: empirical distribution, wage distribution, normal distribution, mixture of distributions, random variable.

1 Introduction

The knowledge of employee wage distributions, resp. population income distributions is a significant precondition for the appraisal of living conditions, a quality of social security and a social fairness measure in material value distribution created by this society. The employee wage distribution formulation in the analytical functional form and individual parameter estimates of this form enables simple, but as well an apposite wage distribution description. Thereby conditions arise for wage distribution comparisons in individual countries and for different periods of time.

A practice showed, that effort to model any wage distribution by probability distributions (normal, log-normal etc.) ordinarily in statistics applied are not too successful. A classical stochastic approach to the probability modeling according to [6] does not enable to model the whole forms of empirical wage distributions in current heterogeneous societies. Not various transformations of distributional functions or density functions not to allow accommodate a desired theoretical shape often a complicated empirical distribution.

The empirical wage distribution is characterized by great non-uniformity and high variability. Therefore on the basis of the study [4] can be a distribution, from which reviewed a wage random sample originated, consider as a distribution mixture of several homogenous subpopulations. A suitable distributional model of this homogenous wage file by [1] and [3] is a log-normal distribution. Thereby the probability

model of a wage distribution can be defined as a mixture (a linear combination) of log-normal distributions. Individual model components are weighted weights corresponding to component shares the probability distribution of the whole model.

2 Mixture of distributions as a probability model

Probability models for empirical distributions can be divided in general to two main groups, namely parametric models and non-parametric models.

Parametric probability models are characterized by probability functions, resp. density functions $f(x; \boldsymbol{\theta})$, which depend on some value of $P \geq 1$ unknown parameters and they are defined for a value space $\Re_x$ of the random variable X. The set of possible values of the unknown parameter $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$ be called the parametric space $\boldsymbol{\Theta}$. By the function $f(x; \boldsymbol{\theta})$ is specified the whole distribution system and for each parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ it belongs one distribution from this system. Parametric probability models are in accordance with [7] defined in the form

$$\boldsymbol{\Psi} = \left[f\left(x; \theta_1, \theta_2 \ldots \theta_p\right), \theta_j \in \boldsymbol{\Theta}, x \in \Re_x \right], \qquad j = 1, 2, \ldots, P. \tag{1}$$

Non-parametric probability models are represented by models with comparatively few preconditions for the model function form. In accordance with [7] as non-parametric models can be considered statistical models, which a their probability component $\boldsymbol{\Psi}$ is characterized as

$$\boldsymbol{\Psi} = \left[f\left(x\right) \in \boldsymbol{\Omega}_F \subset \boldsymbol{\Omega}, x \in \Re_x \right]. \tag{2}$$

The symbol $\boldsymbol{\Omega}$, resp. $\boldsymbol{\Omega}_F$ denotes a set of all possible, resp. suitable distributions. The function $f(x)$ is not defined in terms some specific distribution system, but with the aid of preconditions referring to distribution properties, for example a value space $\Re_x$ of random variable X, an existence of moments and a distribution smoothness (continuity, differentiability, etc.). Most in use non-parametric estimations of probability distributions are estimations in the form so-called kernel density defined in [8].

Between these main groups of probability models lay so-called mixed models. In compliance with empirical studies [1] and [3] is a wage distribution constituted from two or more log-normal distributions. Each log-normal distribution defines an earnings homogenous subset of employees. Therefore a probability model of any wage distribution can be expressed as a mixture (a linear combination) of log-normal distributions, i.e. as a formulation

$$f_{LN}\left(y\right) = \sum_{j=1}^{K} \pi_j LN\left(y; \mu_j, \sigma_j^2\right). \tag{3}$$

The variable y represents a wage and $f_{LN}(y)$ is the density function in the point y. The label $LN\left(y;\mu_j,\sigma_j^2\right)$ describes j-th partial log-normal density of the wage distribution model (3), $j = 1, \ldots, K$, each with parameters μ_j,σ_j^2. The symbol K stands for finite number of wage model components. The number of model components K is usually in advance not known and it must be estimated under uncertainty conditions about the model structure. A criterion of an optimal density number in the statistical model is estimated by minimization so-called informational criteria by [2]. Each j-th model component is weighted by a weight π_j corresponding to a relative component share in the probability distribution of the whole model, where $\sum_{j=1}^{K}\pi_j = 1$ for $\pi_j > 0$.

3 Parameter estimates of wage distribution mixture models

Parameter estimates of wage distribution probability models compounded from some partial distributions does not usually realize on the basis of directly observed wage values. The reduction of a computing intensity and process time achieves with the aid of the logarithmic transformation of wage values. Thereby the wage population model does not to be estimated as a log-normal density mixture via (3), but after this logarithmic transformation will be as a normal density mixture $N\left(y;\mu_j,\sigma_j^2\right)$ with parameters μ_j,σ_j^2, $j = 1, \ldots, K$. The wage distribution model (3) so comes into a form

$$f_N(y) = \sum_{j=1}^{K}\pi_j N\left(y;\mu_j,\sigma_j^2\right). \qquad (4)$$

In parameter estimates of models given via (4) is first of all needed to find out from which partial distributions the resulting distributional mixture folds up. Further it needs to estimate model parameters, partly partial distribution parameters, partly partial distribution weights in the distributional mixture. Since „classical“ parameter estimate techniques (moment estimations, estimations by minimization a deviation of some model measure from data, for ex. residual squares) of the model (4) are not suitable in accordance with [9], will be parameter estimates of wage distribution models as a mixture realized the maximum likelihood method, i.e. by the expression

$$l\left(\hat{\boldsymbol{\theta}}\right) = \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \qquad (5)$$

The function $l(\boldsymbol{\theta})$ is a logarithmic likelihood function with a parametric vector $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}]^T$, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]^T$, $\boldsymbol{\sigma}^2 = [\sigma_1^2, \ldots, \sigma_K^2]^T$, a $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^T$. The symbol $\hat{\boldsymbol{\theta}} = \left[\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\pi}}\right]^T$ stands for a vector of estimated parameters. The logarithmic likelihood function $l(\boldsymbol{\theta})$ is a function defined as a formulation

$$l(\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f_N(y_i) = \sum_{i=1}^{n} \ln f_N(y_i). \qquad (6)$$

After a putting of the expression for a statistical mixture model (4) into the expression (6) the logarithmic likelihood function $l(\boldsymbol{\theta})$ is formulated as a functional relation

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left[\sum_{j=1}^{K} \pi_j N\left(y_i; \mu_j, \sigma_j^2\right)\right]. \qquad (7)$$

Parameters maximizing the logarithmic likelihood function (7) are estimate by comparison of the likelihood equation system (given as derivations of (7) for logarithmic likelihood function by the all searched parameters) with zero, tj.

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} \ln\left[\sum_{j=1}^{K} \pi_{ij} N\left(y_i; \mu_j, \sigma_j^2\right)\right] = 0. \qquad (8)$$

The likelihood equation system (8) has not according to [5] an explicit solution. For that reason it has to be used for mixture model parameter estimates numerical optimization methods. Since the logarithmic likelihood function (7) need not be in general concave on account of estimated parameters $\hat{\boldsymbol{\theta}} = \left[\hat{\mu}, \hat{\sigma}^2, \hat{\pi}\right]^T$, it is according to [5] preferential to use so-called an Expectation-Maximization (*E-M* for short) algorithm for parameter estimates.

The *E-M* algorithm application for maximum likelihood parameter estimates divides according to [9] into a phase of expectation and phase of maximization. For each observed wage logarithm in a phase of expectation of the *E-M* algorithm it computes the probability p_{ij}, that *i*-th observation, $i = 1, \ldots, n$, comes from *j*-th partial distribution of the statistical model, $j = 1, \ldots, K$, i.e. as a formulation

$$p_{ij} = \frac{\pi_j N\left(y_i; \mu_j, \sigma_j^2\right)}{\sum_{j=1}^{K} \pi_j N\left(y_i; \mu_j, \sigma_j^2\right)}, \qquad i = 1, 2, \ldots, n. \qquad (9)$$

In the maximizing section *E-M* algorithm such estimates the parameter vector, that maximizing the logarithmic likelihood function (7) value. That is a computation of parameter estimations

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} p_{ij} y_i}{\sum_{i=1}^{n} p_{ij}}, \qquad j = 1, 2, \ldots, K, \qquad (10)$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} p_{ij}\left[y_i - \hat{\mu}_j\right]^2}{\sum_{i=1}^{n} p_{ij}}, \qquad j = 1, 2, \ldots, K, \tag{11}$$

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} p_{ij}, \qquad j = 1, 2, \ldots, K. \tag{12}$$

4 Empirical wage distribution modeling as a density mixture

The modeled empirical wage distribution was drawn up as a random sample from wage and personal data of the year 2008. This is a random sample, which is created by 1 000 wage values. Although a representation of the high income employee category is markedly small (their share was 2.95%), their wages evidently influence on the statistical accounted average wage level and particularly moment characteristics (variability, skewness and kurtosis) of the whole sample. The employee share, which wages do not reach the average wage level, is 63.66%. A view of wage logarithm distribution offers Figure 1. A wage logarithm distribution behavior is depicted with the aid of a kernel density and a histogram.

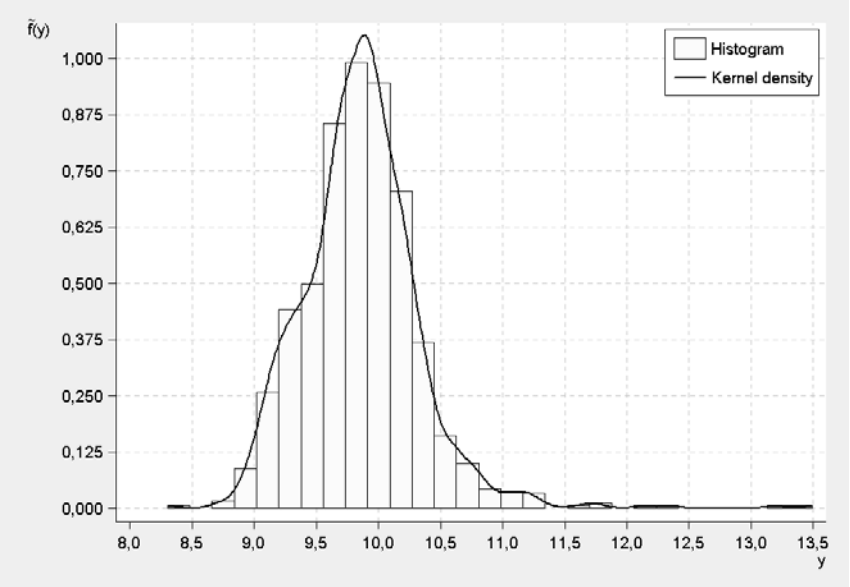


Fig. 1. The kernel density and the histogram of the distribution of empirical wage logarithms.

Homogenous employee groups can not divide by some simple categorization of some group for example according to sex, degrees of education or to the kind of per-

formed work, possibly by the other. The impossibility of classification of the whole wage sample by homogenous wage groups by some simple categorization is produced by the fact, that this group distribution is not wage homogenous. The example of this may be some group of workers, where the most of these workers receive smaller wages. But in this worker category occur some experts, which wages can exceed a wage level of middle level managers. The opposite case arises, if in the managerial group some manager is rewarded wages about some worker level. The same situation exists also in the group of men, of women, and in groups divided according to degrees of education or practice, etc.

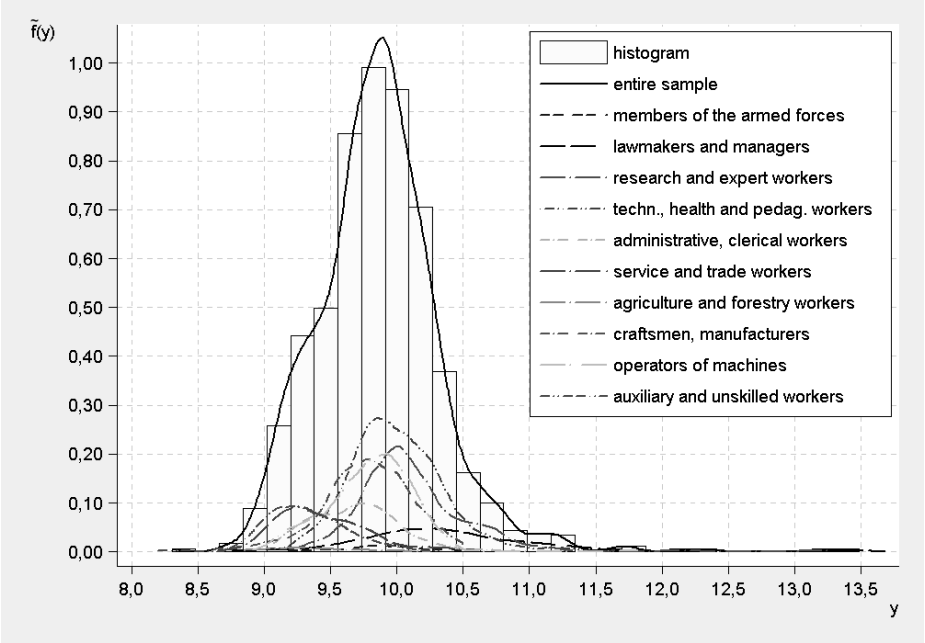


Fig. 2. The mixture of empirical densities in the sample sorted according to professions.

The illustration of empirical density mixtures, that each of them describes the wage logarithm distribution in the sorted sample according to the most important professional categories, it can serve the Figure 2. From this figure is apparent, that wage receivers from individual professional categories cover a wide value spectrum and that not a single professional category homogenous from the point of view of wages.

To findings, how many partial distributions (and thus wage homogenous subsets) the resulting distributional mixture consists, it was executed estimations of model parameters including from 1 to 8 partial distributional components. After the executed estimation, it was discovered, that the optimal model of the wage logarithm distribution is the statistical model consisting from 4 densities of the normal distribution. The criterion to choice of the optimal model was become Akaike's informational criterion value according to [2], that achieved for the model with 4 partial distributions the minimal value. This model as much as possible comes close to the kernel density of

the empirical distribution, which is evident optically from Figure 3. The choice optimality 4 statistical model components for given empirical distribution confirmed one-sample Kolmogorov-Smirnov test according to [2]. Informational criterion values and test statistics (marked as K-S) for model with various item numbers are in the Table 1.

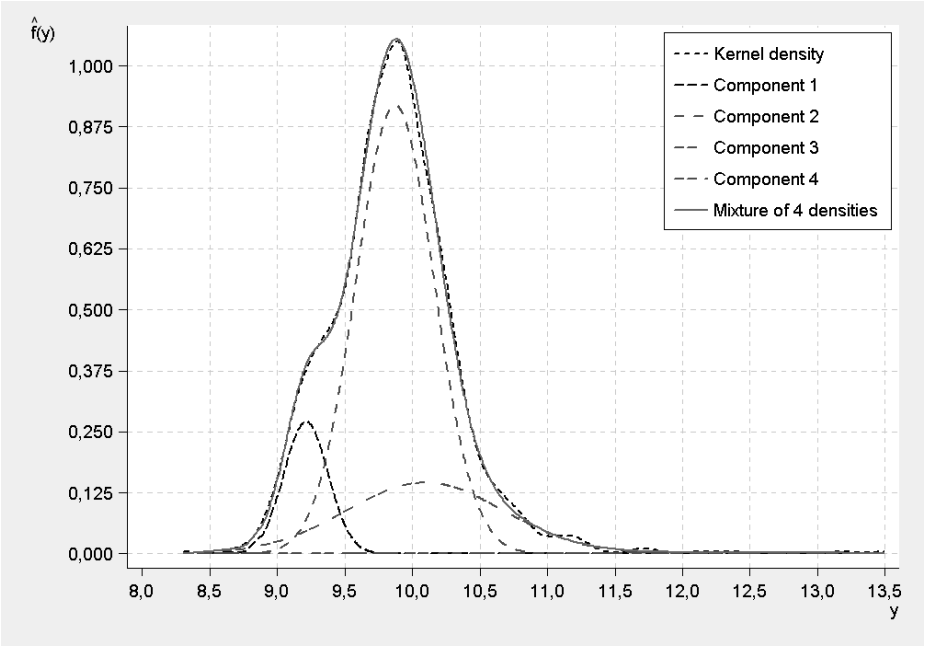


Fig. 3. The mixture of normal densities as an optimal model of the wage logarithm distribution.

Table 1. The optimal estimation of the partial distribution number of the model as a mixture.

Number of model components	1	2	3	4	5	6	7	8
Number of values	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000
Number of parameters	2	5	8	11	14	17	20	23
Log. likelihood	-677	-603	-593	-586	-584	-584	-583	-581
Akaike's inf. criterion	1 358	1 215	1 201	1 195	1 197	1 202	1 205	1 208
K-S statistics	0.4417	0.3328	0.0739	0.0271	0.0314	0.0352	0.0395	0.0419

Distributional mixtures as models with various numbers of partial distributions are displayed graphically in the Figure 4. These models are compared with the kernel density of the empirical wage logarithm distribution. From introduced graphical comparisons it clears, that at minimum suitable model is the model created only from one normal distribution. With increasing the number of model components over the optimal number comes up to move away of the model from the kernel density of the wage logarithm distribution.

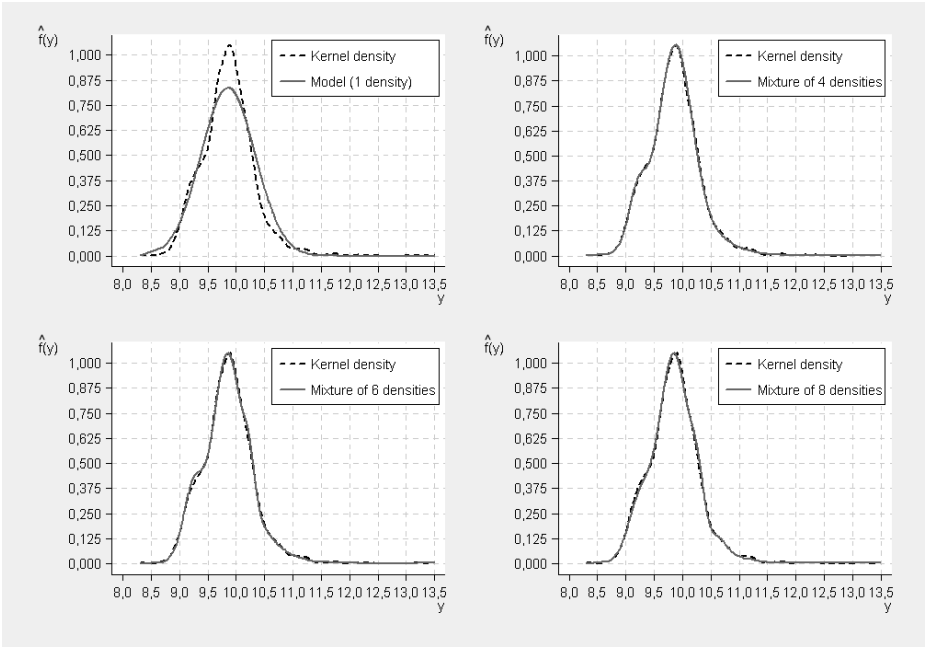


Fig. 4. The estimate of optimal partial distribution number of the statistical model as a mixture.

Estimated parameter values $\hat{\boldsymbol{\theta}} = \left[\hat{\mu}, \hat{\sigma}^2, \hat{\pi}\right]^T$ of the optimal model of the empirical wage logarithm distribution, i.e. a distributional mixture with 4 partial distributions are recorded in the Table 2. Beside estimated parameter values are in the Table 2 described also their standard errors and confidential interval sizes on the 95% level.

Table 2. The estimate of model parameters with the optimal number of partial distributions.

Parameter name	Value	Std. error	95% conf. interval	
mi1	9.2129250	0.0354342	9.1434760	9.2823750
sigma1	0.1497303	0.0251752	0.1003879	0.1990727
pi1	0.1116982	0.0273049	0.0581816	0.1652148
mi2	9.8720880	0.0206752	9.8315650	9.9126110
sigma2	0.2826429	0.0281633	0,2274440	0.3378419
pi2	0.6695857	0.1035618	0,4666083	0.8725631
mi3	10.0918600	0.1304661	9,8361490	10.3475700
sigma3	0.5836871	0.0880035	0,4112034	0.7561708
pi3	0.2129156	0.0960030	0,0247531	0.4010781
mi4	12.3526100	1.2137580	9.9736910	14.7315400
sigma4	0.8287803	0.6364783	-0.4186942	2.0762550
pi4	0.0058005	0.0064185	-0.0067795	0.0183806

The statistical model (4) of the wage logarithm distribution as a distributional mixture with parameters estimated by the maximum likelihood method can be for the optimal component number, i.e. component number $K = 4$ expressed as a formulation

$$
\begin{aligned}
f_N\left(y;\hat{\boldsymbol{\theta}}\right) = \quad & 0.1116982 \cdot \frac{1}{0.1497303\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{y-9.2129250}{0.1497303}\right]^2\right\} \\
& + 0.6695857 \cdot \frac{1}{0.2826429\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{y-9.8720880}{0.2826429}\right]^2\right\} \\
& + 0.2129156 \cdot \frac{1}{0.5836871\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{y-10.0918600}{0.5836871}\right]^2\right\} \\
& + 0.0058005 \cdot \frac{1}{0.8287803\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{y-12.3526100}{0.8287803}\right]^2\right\}
\end{aligned}
\qquad \textbf{(13)}
$$

5 Interpretation of results of the model parameter estimates

Realized empirical studies [1] and [3] confirm that the wage distribution is constituted as a distribution mixture of two or more log-normal distributions. Each partial log-normal distribution defines the earnings homogenous group of employees. Therefore the probability model of wage distributions can describe as a mixture (linear combination) of log-normal distributions in accordance with (3). Each distributional component of the model (3) is weighted by some weight corresponding to the item relative proportion in probability distribution of the whole model (3). Established relative weights (proportions) individual distributions for the model with the optimal number of partial distributions (the number of components $K = 4$) are recorded in the Table 3.

Table 3. The relative weight (proportion) estimate of model components with the optimal number of partial distributions, i.e. for number of components $K = 4$.

component number	1	2	3	4
relative weight (prop)	0.1116982	0.6695857	0.2129156	0.0058005

The relative weight (proportion) system of individual partial components of mixture model is the parameter estimation of the multinomial distribution with parameters $n, \pi_1, \pi_2, \pi_3, \pi_4$. The parameter π_j expresses the probability, that wage logarithm value belongs into the j-th distributional component, resp. into the j-th earnings homogenous group, $j = 1, 2, 3, 4$. It means that wage receiver comes with the probability π_1 to the first earnings homogenous group, with the probability π_2 comes to the second earnings homogenous group, etc.

With the aid of wage distribution model as a mixture of log-normal densities (3) can assess also relative proportions (a distribution) for example women or men, eventually proportions according to other explanatory variables in wage homogenous subsets.

6 Conclusions

The classical stochastic approach to the probability wage distribution modeling according to [6] does not enable under current conditions of wage differentials to model the whole shapes of empirical wage distributions. Not various transformations usually to permit some adapting the desired theoretical shape of distributional functions or density functions to the ragged empirical distribution.

The empirical wage distribution modeling with the aid of distributional mixtures gives some possibility to express a wage distribution as a linear combination of two or more log-normal wage distributions, resp. normal wage logarithm distributions. Each this partial log-normal wage distribution, resp. normal wage logarithm distribution defines some wage homogenous subset. The system of relative weights (proportions) of individual partial mixture model components is the parameter estimate of the multinomial distribution with parameters n, $\pi_1, \ldots, \pi_K$, where K expresses the finite component number of the distributional mixture. The parameter π_j represents the probability, that the wage logarithm value belongs to the j-th distributional component, resp. to the j-th earnings homogenous subset, where $j = 1, \ldots, K$.

The encompassment of individual characteristics into the model can estimate relative proportions (a distribution) wage receivers from the point of view of this characteristics. In subsets with the highest earnings are the most included managers. On the other hand the lowest earnings subsets occupied auxiliary and unskilled workers and service and trade workers.

7 References

1. Aitchinson, J., Brown, J., A., C: The Lognormal Distribution. Cambridge University Press, London (1957), ISBN 0521040116
2. Anděl, J.: Statistické metody. MATFYZPRESS, Praha, (1998), ISBN 8085863278
3. Cowell, F.:. Measurement of inequality. In Handbook of Income Distribution, Volume 1, pp. 87–166. A. B. Atkinson and F. Bourguignon (eds), Elsevier Science, (2000), ISBN 0444816313
4. Flachaire, E., Nuňez, O.: Estimation of Income Distribution and Detection of Subpopulations: an Explanatory Model. Working Paper 03-02. In: Statistics and Econometrics Series, 01. January 2003
5. McLachlan, G. J., D. Peel: Finite Mixture Models. New York: Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics Section, (2000), ISBN 0471006262
6. Sipková, Ĺ.: Prehĺad teoretických východísk merania príjmovej nerovnosti. In: Slovenská štatistika a demografia, ŠÚ SR, vol. 14, num. 3, pp. 36–49, (2004)
7. Spanos, A.: Probability Theory and Statistical Inference: Econometric Modeling with Observational Data. Cambridge, Cambridge University Press, (1999) ISBN 0-511-01097-4.
8. Silverman, B., W.: Using Kernel Density to Investigate Multimodality. Journal of the Royal Statistical Society B 43, pp.97-99, (1981)
9. Titterington, D.,M., Smith, A.F.M, Makov, U.E.: Statistical Analysis of Finite Mixture Distributions. New York, Wiley, (1985), ISBN 0471907634

Discrete Choice Models

Iva Pecáková[1] and Ondřej Vojáček[2]

[1] Department of Statistics and Probability,
University of Economics Prague, Czech Republic
pecakova@vse.cz
[2] Department of Environmental Economics,
University of Economics Prague, Czech Republic
ovojacek1@centrum.cz

Abstract. An important application of the models for categorical dependent variables – discrete choice models – can predict a decision made by an individual and based on the utility or relative attractiveness of competing alternatives as a function of any number of variables. These models are recently wide-spread in various applications areas (transport systems, energetic, environmental studies, medicine, voting behavior etc). The paper summarizes multinomial logit model, nested logit model, probit model and random parameters logit model, and also statistical systems applicable for their estimation.

Keywords: Models for categorical dependent variables, statistical systems for discrete choice models (DCM) estimation.

1 Introduction

Discrete choice modeling (DCM) encompasses a variety of data collection procedures and statistical procedures which can be used to describe the choices made by people among a finite set of alternatives. The models predict a decision made by an individual and based on the utility or relative attractiveness of competing alternatives. The choice made of each person is related to the attributes of the person and also to the attributes of the alternatives available to the person. DCM was developed in parallel by economists and cognitive psychologists since nineteen seventies. The most widely known is the works of D. McFadden, the Nobel Prize winner.

A decision-maker chooses among a set of J options. The dependent variable Y, a discrete variable with a countable number of J values, represents the outcome of the decision. The goal of the analysis is to understand what variables and to what extend influence this choice.

In discrete choice models, the linear combination V_j of the H observed (non-random) factors $[X_{j1}, X_{j2}, ..., X_{jH}] = \mathbf{x}'_j$ with the parameters $\boldsymbol{\beta}' = [\beta_0, \beta_1, ..., \beta_H]$, and the unobserved, random factors (ε_j), $j = 1, 2, \ldots, J$, is considered to be the utility of the alternative j. Then, for the decision-maker i, the utility is

$$U_{ij} = V_{ij} + \varepsilon_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}\,,\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, J. \quad (1)$$

If the decision-maker chooses the alternative which brings the greatest utility to him, then the probability of the choice of the alternative j over j',

$$\pi_{ij} = P(V_{ij} + \varepsilon_{ij} > V_{ij'} + \varepsilon_{ij'}) = P(\varepsilon_{ij'} - \varepsilon_{ij} < V_{ij} - V_{ij'}) \quad (2)$$

is the cumulative distribution function of a random variable $\varepsilon_{ij'} - \varepsilon_{ij} = \varepsilon^*_{ijj'}$. Different discrete choice models are obtained from different assumptions about this probability distribution.

The characteristic feature of DCM is two kinds of explanatory variables: not only characteristics *of the chooser* (constant over the alternatives), but also characteristics *of the choices* (different values for each alternative). Particularly these explanatory variables are usually replaced by artificial (dummy) variables; so, the notation of explanatory variable vector as $\mathbf{x}_j$ and parameter vector as $\boldsymbol{\beta}$ is very general. The ordinary logit model (e.g. in [7]) underlies the discrete choice models theory. However, it is only a special case of one type of discrete choice model – conditional multinomial logit model.

2 Multinomial Logit Model

The most widely used discrete choice model, a *multinomial logit model* (MNL model), is derived under the assumption that each ε_{ij} in (1) has so-called Gumbel (or type I extreme value) distribution with the cumulative distribution function

$$F(\varepsilon_{ij}) = \exp[-\exp(-\varepsilon_{ij})]$$

and with the variance of $\lambda^2\pi^2/6$ (π is Ludolph's number here, λ is a scale parameter). If these random variables are distributed identically (λ can be arbitrarily set to 1) and independently (IID) and follow the Gumbel (type I extreme value) distribution, then their difference follows the logistic distribution (e.g. in [1])

$$F(\varepsilon^*_{ijj'}) = \left[1 + \exp(-\varepsilon^*_{ijj'})\right]$$

with the zero mean and with the variance of $\pi^2/3$. As can be shown (e.g. in [8]), the probability of choice of the alternative j by the individual i is then

$$\pi_{ij} = \frac{\exp(V_{ij})}{\sum_j \exp(V_{ij})} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\sum_j \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})} \quad (3)$$

where $\mathbf{x}_{ij}$ denotes the values of all the explanatory variables for subject *i* and response choice *j*. Since the choice probability depends only on the difference in utility, not on the level of utility, the values of any utility can be normalized to zero; e.g., for the first alternative $\mathbf{x}'_{i1}\boldsymbol{\beta} = 0$, for this alternative (3) can be expressed as

$$\pi_{i1} = \frac{1}{\sum_j \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}$$

and then

$$\mathbf{x}'_{ij}\boldsymbol{\beta} = \ln \frac{\pi_{ij}}{\pi_{i1}}, j = 2, \ldots, J. \quad (\mathbf{4})$$

There is a so-called logit on the right-hand side of (4).

The key problem with the MNL model arises from the IID assumption. Then, the odds of choosing an alternative *j* over an alternative *j′* do not depend on the other alternatives in the choice set or on their values of the explanatory variables:

$$\frac{\pi_{ij}}{\pi_{ij'}} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\exp(\mathbf{x}'_{ij'}\boldsymbol{\beta})} = \exp[(\mathbf{x}_{ij} - \mathbf{x}_{ij'})'\boldsymbol{\beta}]. \quad (\mathbf{5})$$

This direct consequence of the IID assumption – independence of irrelevant alternatives (IIA) – is expressed as a *proportionate shift*: an increase in the probability of one alternative reduces the probabilities for all the other alternatives by the same percentage. If the IIA property holds, it is possible, for example, to reduce a number of choice alternatives without influencing the relations among the remaining ones. It is unrealistic in some applications. The key IID assumption is that the errors are independent of each other. However, unobserved factors related to different alternatives might be similar and hence the random component might be dependent. Then the assumption of independence can be inappropriate.

The IIA condition is usually tested with the Hausman-McFadden test (e.g. in [3]). Presently, the hypothesis of this test is commonly specified as constraints on the parameters of the more general model. For the calculation of the test statistic, each alternative is separately excluded from the model, and the parameters for restricted and unrestricted models are estimated as well as their variance–covariance matrices. The test criterion is chi-square distributed with the degrees of freedom given by the number of estimated parameters.

Although the IID/IIA conditions may be worrying, any unrealistic assumption about the error term is likely to be of small consequence if the amount of information in the unobserved component is minimal. The richness of information in V_{ij} captured in attributes depends in particular on the proper implementation of the design and pretest stages of the choice experiment.

To estimate the parameters of the model (by the maximum-likelihood method), a sample of *n* decision-makers is obtained and their choices are surveyed. For the evaluation, how well the model fits the data, the goodness-of-fit statistics on the basis of the log-likelihood function are usually used (e.g.. in [1]). There exist many such

statistics (e.g. in [7]); the one most used in literature on discrete choice modeling is McFadden's statistic

$$D_{MF} = \frac{\ln L_0 - \ln L_E}{\ln L_0}, \tag{6}$$

where L_0 is the likelihood of the intercept-only model and L_E is the likelihood of the estimated model. The interpretation of this statistic is not the same as that of the R-squared statistic in the linear regression and usually its values are low; fortunately, an unambiguous relationship between them exists that provides better interpretation (e.g. in [2]), where pseudo R-squared values between the range of 0,3 and 0,4 can be translated as an R-square of between 0,6 and 0,8 for the equivalent linear model.

For the comparison of models, the log-likelihood ratio statistic D (so-called deviance) is usually used (e.g. Hensher et al. in [3], Agresti in [1]). It is the statistic for testing the null hypothesis that the restricted model (R) holds against the alternative that the more general, unrestricted model (U) holds:

$$D = -2(\ln L_R - \ln L_U) . \tag{7}$$

It has an approximately chi-square distribution with degrees of freedom equal to the difference in the number of parameters between both the compared models. Wald tests are used most commonly for hypotheses about the significance of the single parameters; however, sometimes likelihood ratio tests are recommended instead (e.g. in [4]). The Wald test is known to have low power and it can be biased where there are insufficient data.

3 Nested Logit Model

If the IIA does not hold, it is necessary to consider a choice model that is less restrictive. Recently, much research effort in this area has been concentrated on relaxing the strong IID and IIA assumptions associated with error terms. The *generalized extreme value* (GEV) model allows correlation in unobserved factors over alternatives; the unobserved portions of utility ε_{ij} for all alternatives jointly have a generalized extreme value distribution. The *nested logit model* is the most widely used member of the GEV family of models.

The choice alternatives are structured into several (K) groups (so-called nests) B_1, B_2, …, B_K. IIA holds within each nest, but it does not hold for alternatives among different nests. The vector of unobserved utility $[\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{iJ}]$ has a generalized extreme value distribution with the cumulative distribution function

$$F(\boldsymbol{\varepsilon}_i) = \exp\left[-\sum_{k=1}^{K} \left(\sum_{j \in B_k} \exp(-\varepsilon_{ij} / \lambda_k) \right)^{\lambda_k} \right]. \tag{8}$$

The parameter λ_k is a measure of the degree of independence in unobserved utility among the alternatives in nest *k;* full independence among all the alternatives in all nests ($\lambda_k = 1$) reduces the nested logit model to a multinomial logit model.

The probability of choice for the alternative $j \in B_k$ is now

$$\pi_{ij} = \frac{\exp(V_{ij} / \lambda_k)\left(\sum_{c \in B_k} \exp(V_{ic} / \lambda_k)\right)^{\lambda_k - 1}}{\sum_{k=1}^{K}\left(\sum_{c \in B_k} \exp(V_{ic} / \lambda_k)\right)^{\lambda_k}} . \qquad (\mathbf{9})$$

This probability of choice can be written as a product of two standard logit probabilities: the probability resulting from the choice among nests (the *upper model*) and the conditional probability resulting from the choice among the alternatives within the nest (the *lower model*)

$$\pi_{ij} = \frac{\exp(W_{ik} + \lambda_k I_{ik})}{\sum_{c=1}^{K} \exp(W_{ic} + \lambda_c I_{ic})} \cdot \frac{\exp(Y_{ij} / \lambda_k)}{\sum_{c \in B_k} \exp(Y_{ic} / \lambda_k)} . \qquad (\mathbf{10})$$

The first probability depends on a part of the observed utility invariable for all alternatives within a nest (W_{ik}) and on a part of the utility that varies across alternatives within a nest (Y_{ij}), $V_{ij} = W_{ik} + Y_{ij}$; the second one depends on Y_{ij} only.

The quantity I_k ,

$$I_{ik} = \ln \sum_{c \in B_k} \exp(Y_{ic} / \lambda_k) ,$$

so-called *inclusive value* IV (or *inclusive utility*) of nest B_k, that enters as an explanatory variable into the upper model, brings in the information from the lower model: it is the log of the denominator of the lower model in (10). The term $\lambda_k I_{ik}$ expresses the utility expected from the choice among the alternatives in nest B_k. Its parameter λ_k can be used to test whether the correlation structure of the nested model differs from the multinomial logit model. The significant test justifies nested structures (e.g. in [5]).

In the nested logit model the scale parameter is introduced in the variance of the unobserved effects for each alternative (the variance is an inverse function to the scale), the same in one nest. A number of researchers have independently shown that the parameter of IV for the upper model is the ratio of the scale parameter of the upper model to the scale parameter of the lower model. Nevertheless, the nested logit model cannot be identified without imposing an additional restriction. One possibility is that the researcher constrains the IV parameters to be the same for all (or some) nests, indicating that the correlation is the same in each of these nests (e.g. in [9]). The reasonableness of this constraint can be tested.

The NL model enables to model choices in a hierarchical structure. These are sometimes interpreted as a sequential decision-making process, that is, that the respondents decide first on the nest and then on the particular alternative within the

nest. However, this decision-making process is not necessary for the nested logit model application. All the parameters of a nested model can be estimated by standard maximum likelihood techniques again.

4 Probit model

The probit model provides an alternative way to fix the problem of the limitations of the multinomial logit model, especially regarding the IID and IIA properties.

The basic assumption of the probit model is that the unobserved utility components are joint normally distributed with the density

$$f(\boldsymbol{\varepsilon}_i) = \frac{1}{(2\pi)^{J/2}\left|\boldsymbol{\Omega}\right|^{1/2}} \exp\left[-0,5\boldsymbol{\varepsilon}_i'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}_i\right] \tag{11}$$

with a mean vector of zero means and a known covariance matrix $\boldsymbol{\Omega}$. The choice probability of the alternative *j* can be then expressed as

$$\pi_{ij} = F(\boldsymbol{\varepsilon}_i) = \int_{\boldsymbol{\varepsilon}_i} f(\boldsymbol{\varepsilon}_i)\mathrm{d}\boldsymbol{\varepsilon}_i \; ; \tag{12}$$

the *J*-dimensional integral is over the set of error terms that result in the choosing of alternative *j*.

With a full covariance matrix, various patterns of correlation and heteroskedasticity can be accommodated according to need, so that the IID and IIA are relaxed. The linear combination of observed factors – the representative utility – in this model is a probit, i.e., a percentile of the normal distribution. However, the probabilities of choice can be expressed only in the form of integrals and they must be evaluated numerically through simulation. Also, the model interpretation is not as straightforward and intuitive as in the logit models.

5 Random Parameters Logit Model

Further model used for discrete choice data analysis (*mixed logit model*, too) also overcomes the two major limitations of the MNL model, i.e., the IIA property and the limited ability of previous models to explicitly account for heterogeneity in data. It is not restricted to normal distributions like the probit; nevertheless, it is more flexible in the treatment of the variances and correlations of the random component. To be able to take into account correlations among the error components of different choice alternatives, the model introduces into the utility function an additional stochastic element that may be heteroskedastic and correlated across alternatives.

The utility of the decision-maker *i* from the alternative *j* is specified in the mixed logit model as

$$U_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \varepsilon_{ij} \quad i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, J; \tag{13}$$

here $\mathbf{x}_{ij}$ are observed variables that relate to the alternative j and the decision-maker i, $\boldsymbol{\beta}_i$ is a vector of coefficients of the observed variables for the decision-maker i representing individuals' tastes; ε_{ij} is a random term with an IID extreme value distribution.

In contrast to a standard logit model, the coefficients (of the variables $\mathbf{x}_{ij}$) vary across decision-makers in the population with a density $f(\boldsymbol{\beta}_i)$. They are considered to be random and can be decomposed into their means α and deviations μ_i. Then

$$U_{ij} = \mathbf{x}'_{ij}\boldsymbol{\alpha} + \mathbf{z}'_{ij}\boldsymbol{\mu}_i + \varepsilon_{ij} \quad i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, J; \tag{14}$$

the parameters of $\mathbf{z}_{ij}$ are random with zero mean, $\mathbf{z}'_{ij}\boldsymbol{\mu}_i$ represents the error component ($z_{ij} = x_{ij}$).

The unobserved portion of utility with the error component $\mathbf{z}'_{ij}\boldsymbol{\mu}_i$ can be correlated among alternatives and/or heteroskedastic for each individual (in the case of a zero error component we obtain the standard logit model).

The mixed logit choice probabilities are conceived as a mixture of the logit function evaluated at different values of parameters $\boldsymbol{\beta}$ with $f(\boldsymbol{\beta})$ as the density of the mixed distribution. The density $f(\boldsymbol{\beta})$ is specified as continued and in particular normal, lognormal, uniform, triangular or any other distributions are used. The applicable distribution is given by expectations about decision-makers' behavior in the particular application.

The mixed logit choice probabilities is then expressed as integrals of standard logit probabilities over a density of parameters evaluated at different values of $\boldsymbol{\beta}$ by the density $f(\boldsymbol{\beta})$,

$$\pi_{ij} = \int \frac{\exp\left(\mathbf{x}'_{ij}\boldsymbol{\beta}\right)}{\sum_j \exp\left(\mathbf{x}'_{ij}\boldsymbol{\beta}\right)} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \,. \tag{15}$$

To specify the distribution of the coefficients, an estimate of its parameters is necessary. Because of that, two sets of parameters are used in the mixed logit model: parameters which enter the logit formula, and parameters which describe the density. The first type parameters have an interpretable meaning as representing the tastes of individual decision-makers; the second parameters describe their distribution across decision-makers.

The greatest value of this model can occur in using each parameter with other linked parameter estimates. The mean parameter estimate for a variable, an associated heterogeneity in its parameter and the standard deviation of the parameter estimate represent the utility of this variable associated with a specific alternative and individual.

Random parameter logit models are computationally intensive to estimate and convergence problems do occur in their applications. The estimation problems differ depending on the distribution of the variables used.

6 Statistical Software for DCM

To estimate models with categorical dependent variables, the more known and more widespread systems are SAS, SPSS, Stata and LIMDEP (e.g. in [6]). SAS and SPSS provides several procedures for these models; various procedures support various models, so the models can be estimated by multiple procedures. Stata and LIMDEP has individualized commands for coresponding models.

SAS is a powerful software package for statistical application. SAS has the CATMOD procedure for the multinomial logit model. The MDC (Multinomial Discrete Choice) module is applicable to perform choice model regression for various discrete models, such as conditional logit, mixed logit, nested logit, and multinomial probit models. However, users without programming experience will spend more time to become familiar with the interface and be able to create codes.

SPSS is a very user-friendly system; it includes procedures for categorical dependent variable models (e.g. the NOMREG command to estimate the multinomial logit model), but its support of DCMs is quite limited. The SPSS COXREG command, which was designed for survival analysis data, makes possible to estimate the conditional logit model. However, the other systems have large potential in these terms.

Stata is capable of logit model regression. On top of that it includes commands for the multinomial logit model (.mlogit) and multinomial probit model (.mprobit – the model took longer time to converge than the multinomial logit model). As to MDC, Stata estimates the conditional multinomial logit (.clogit), nested logit (.nlogit). The Stata command .mixlogit can be used for fitting the mixed logit model.

LIMDEP commands support a variety models with categorical dependent variables. An extension of this very large package, which includes some tools for estimating of discrete choice models, such as multinomial logit, multinomial probit, nested logit and several others is called NLOGIT. Recently NLOGIT has become the premier package for estimation and simulation of multinomial discrete choice models. The latest version of NLOGIT is able to handle heterogeneity in variances of utility functions and mixed logit model.

The disputation about use of logit or probit models remains on. Logit models are simpler, but they have a problem of IIA assumption. Probit models are computationally intensive. There exist papers manifesting that probit will provide more accurate results than logit, some authors contradict.

7 Conclusion and Further Work

The authors of this paper compared the discussed models on data from summer wareside recreation – the preference analysis of the summer holiday-makers on the Macha lake beaches in the summer 2007. For the analysis the NLOGIT4 software is used. The results of our work are very similar for all the models; only the mixed logit model differs slightly from the three other models. The multinomial logit model seems to be quite robust with respect to deviations of the random component distribution from the model. Thus, the presented analysis seems to prove that the

multinomial logit model could be preferred in practice to both the nested logit and the probit models for this type of data even in situations where it does not comply with the basic IID/IIA assumption. Besides, there are three considerable advantages to the multinomial logit model: computational ease, easy-to-obtain probability expression of an individual selecting a given alternative, and straightforward determination and maximization of its likelihood (which reduces possible model estimation difficulties). We are preparing a separate paper with detail results of this analysis.

References

1. Agresti, A.: Categorical Data Analysis, Second edition, John Wiley and Sons, Inc., New Jersey (2002)
2. Domencich T. and McFadden D.: Urban Travel Demand A Behavioural Analysis, American Elsevier Company, New York (1975)
3. Hensher, D.A., Rose, J.M., Greene W.H. : Applied Choice Analysis, Cambridge University Press, New York (2005)
4. Hosmer D. W., Lemeshow S.: Applied Logistic Regression, John Wiley & Sons, New York (2000)
5. Louviere J.J., Hensher D.A., Swait J.: Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Valuation, Cambridge: Cambridge University Press (2000)
6. Park, Hun Myoung: Estimating Regression Models for Categorical Dependent Variables Using SAS, STATA, LIMDEP, and SPSS. Technical working paper, UITS Center for Statistical and Mathematical Computing, Indiana University (2008)
7. Pecáková, I.: Explained Variation Measures for Models with Categorical Responses, in: AMSE 2007 [CD-ROM], Banská Bystrica (2007)
8. Pecáková, I.: Logistická regrese s vícekategoriální vysvětlovanou proměnnou, Acta Oeconomica Pragensia 15/1, pp. 86 – 96 (2007)
9. Train K.: Discrete Choice Methods with Simulation, Cambridge University Press, New York (2003)

Model-Based Curve Clustering Using Unsupervised Learning

Pavel Pešout[1],

[1]University of Economics, KSTP, Nám. W. Churchilla 4,
130 27, Prague, Czech Republic
pavel.pesout@vse.cz

Abstract. Classification is a very common task in information processing and important problem in many areas of science and engineering. In the case of data measured as a function of a dependent variable such as time, the most used algorithms, e.g. *K*-means, may not pattern each of the individual shapes properly. Therefore, in the presented paper we namely focus on some efficient methods of clustering trajectories. Our clustering algorithms are based on a principled method for probabilistic modeling of a set of trajectories as sequences of points generated from a finite mixture model consisting of regression model components. Unsupervised learning is carried out using maximum likehood principles to deal with the hidden data problem – the cluster memberships.

Keywords: classification, probabilistic modeling, density-based methods

1 Introduction

Clustering is the process of grouping data with similar character into the same class or the same cluster. Traditional clustering techniques can be distinguished as two common types which are hierarchical clustering methods and partition-based clustering methods. In hierarchical clustering it is a difficult problem to decide where to cut the formed hierarchical structure – dendrogram – and get the most homogenous clusters. What is more, when two elements are jointed according to the distance measurement, they could not be separated anymore. As a result, it may lead the process to a wrong way.

In partition-based clustering such as *K*-means, the investigator is asked to give the distance measure metric and the number of cluster, which can decrease the accuracy of the result of clustering.

In this paper, on the other hand, we deal with more sophisticated and untraditionally density-based clustering methods using the Maximum Likehood Estimation (*MLE*) to recognize the most homogenous partitioning. In contrast to the hierarchical and the partition-based methods which are focused just on the set of obtained measurements, the density-based methods are able to attend to the whole space among the measurements. For the reason, there are useful above all in the case of time series clustering.

There are two types of the density-based methods which differ in their approach to the cluster memberships. First, we may assume these memberships to be some of the model parameters. These methods are called the Maximum Likehood Approach (*MLA*) methods thorough described by Fraley in (3) and Banfield and Raftery in (3). The classification is two-fold:

- The likehood is analytical or approximative maximized over the jointed parameters.
- Using the estimations is the likehood function criterion maximized over the cluster memberships.

Second way is to assume the cluster memberships to be random variables and hence to use the mixture models. Because of the critical computational cost of finding the global minimum, only the local minimum is looking for with the application of the Expectation-Maximization (EM) algorithm described by Dempster, Laird and Rubin in (1).

The classification is processed by the following manner:

- The cluster memberships are iterative estimated.
- The jointed model parameters are estimated using the membership probabilities.

In the case of the lowly noisy data sets may be useful the probabilistic one-level parametric *EM* model. Gaffney and Smyth are dealing with it in (5). They introduced some advantages of the mixture model clustering:

- The trajectories do not have to be measured in the same data points.
- The measurements belonging to different time series do not have to be the same length.
- The users are not requested to give the number of clusters because of ability to use the Bayesian Information Criterion (*BIC*) proposed by Schwarz in (9).
- The process is iterative which leads to the ability of the choice the best result by initial parameters setting and the rules of committing the individual steps.

The one-level model can be used to effectively account for subpopulations of homogenous behavior. However, more care should be taken when considerable variability exists within each subpopulation or group.

What is needed is the ability to let an individual vary from the template for its group, yet still exhibit the underlying behavior that distinguishes this group from the rest. That is why the random effects regression mixtures are more suitable to highly noisy data sets. A hierarchical model structure is defined with a mixture on parameters at the top level and an individual-specific regression model at the bottom level.

By the general Bayesian context with the application of the Markov Chain Monte Carlo (*MCMC*) techniques are the random effects mixtures introduced by Lenk and DeSarbo in (8) and Jank in (7). Instead, in this paper we focus on the development of *MAP*-based *EM* procedure for parameter inference for the case of polynomial regression models and mixtures of splines made by Gaffney and Smyth in (4) and James and Sugar in (6).

2 Random effects regression mixtures

Suppose we have a set Y of N trajectories generated according to a mixture distribution with K components. For each individual curve we assume measurements $\boldsymbol{y}_i = (y_{i1},\ldots, y_{imi})^{\mathrm{T}}$ in the data points $\boldsymbol{x}_i = (x_{i1},\ldots, x_{imi})^{\mathrm{T}}$. Our aim is to find partitioning into K groups so that objects in the same cluster have high similarity in the form of the shapes and objects in the different clusters have low similarity.

Let the i-th trajectory be generated from the normal regression model

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{r}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{r}_i$ is the p-vector of regression parameters, the m_i x p regression matrix $\boldsymbol{X}_i$ is the Vandermonde matrix evaluated at $\boldsymbol{x}_i$ and $\boldsymbol{\varepsilon}_i$ is the noise m_i -vector. At this bottom level of the hierarchy there is in fact no dependence on the cluster membership and instead, we allow for individual specific heterogeneity.

Let $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ be the parameters at this data-level which allows us to model the individual trajectory behavior. Using them we may assume the i-th conditional distribution taking the form

$$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_i). \tag{2}$$

Furthermore, at the top level of the hierarchy there is a model that describes the distribution of the parameters $\boldsymbol{r}_i$ of each individual. Let $\Phi = \{w_1, \ldots, w_K, \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_K\}$ be the parameters at this level, where w_k is the probability that an observation belongs to the k-th cluster and $\boldsymbol{\varphi}_k$ are the parameters of the distribution on $\boldsymbol{r}_i$ according to the group template

$$f_k(\boldsymbol{r}_i \mid \boldsymbol{\varphi}_k). \tag{3}$$

That is why, the unconditional of class membership prior for $\boldsymbol{r}_i$ is a finite mixture model

$$f(\boldsymbol{r}_i \mid \Phi) = \sum_{k=1}^{K} w_k f_k(\boldsymbol{r}_i \mid \boldsymbol{\varphi}_k). \tag{4}$$

Because of the equality $f(\Theta, \Phi) = f(\Theta \mid \Phi) f(\Phi)$ and knowledge that

$$f(\Theta \mid \Phi) = \prod_{i=1}^{N} \sum_{k=1}^{K} w_k f_k(\boldsymbol{r}_i \mid \boldsymbol{\varphi}_k), \tag{5}$$

we also may, in order to produce the consistent parameter estimates, use the Maximum a Posteriori (*MAP*) - based *EM* algorithm. Let $\boldsymbol{Z}$ be the matrix of memberships z_{ik} where $z_{ik} = 1$ if $\boldsymbol{y}_i$ is a member of the k-th cluster and 0 otherwise. The complete-data *MAP* objective function is for the set R of all $\boldsymbol{r}_i$ given as

$$M_c(\overline{\Theta}, \Phi) = \ln\left[f(Y \mid X, \Theta) f(\Theta, \boldsymbol{Z} \mid \Phi) f(\Phi)\right], \tag{6}$$

where

$$f(\Theta, \boldsymbol{Z} \mid \Phi) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[w_k f_k(\boldsymbol{r}_i \mid \boldsymbol{\varphi}_k) \right]^{z_{ik}}, \overline{\Theta} = \{\Theta - R\}. \quad (7)$$

The *EM* algorithm consists of two *E* and *M*-steps. In the *E*-step, the expected value of the complete-data *MAP* function is taken with respect to the posterior conditional prior of the cluster memberships. We also evaluate the expected value $\boldsymbol{r}_{ik}$ of $\boldsymbol{r}_i$ given $\boldsymbol{y}_i$ and $z_{ik} = 1$ and we set membership probabilities using in the previous *M*-step updated parameters to

$$w_{ik} = \frac{w_k f(\boldsymbol{y}_i \mid \overline{\Theta}^*, \boldsymbol{\varphi}_k^*)}{\sum_{j=1}^{K} w_j f(\boldsymbol{y}_i \mid \overline{\Theta}^*, \boldsymbol{\varphi}_j^*)}. \quad (8)$$

In the *M*-step is this expectation maximized over the parameters Φ and $\overline{\Theta}$. This yields the following form of the *EM* algorithm:

- Randomly initialize the membership probabilities w_{ik}.
- Calculate estimates for $\{\Theta - R\}$ and Φ.
- Make w_{ik} and $\boldsymbol{r}_{ik}$ contemporary.
- Loop to step 2 until the expected value of the complete-data *MAP* function stabilizes.

3 Normal regression model

Suppose the *p*-th order polynomial regression relationship between $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ with an additive Gaussian error term, $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$, and thus

$$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_i) = N(\boldsymbol{y}_i \mid \boldsymbol{X}_i \boldsymbol{r}_i, \sigma^2 \boldsymbol{I}). \quad (9)$$

With the top level parameters $\boldsymbol{\varphi}_k = \{\boldsymbol{\mu}_k, \boldsymbol{Q}_k\}$ we also have

$$f_k(\boldsymbol{r}_i \mid \boldsymbol{\varphi}_k) = N(\boldsymbol{r}_i \mid \boldsymbol{\mu}_k, \boldsymbol{Q}_k), f(\boldsymbol{r}_i \mid \Phi) = \sum_{k=1}^{K} w_k N(\boldsymbol{r}_i \mid \boldsymbol{\mu}_k, \boldsymbol{Q}_k). \quad (10)$$

One problematic issue is that *K* distinct covariance matrices must be estimated. Therefore it is necessary to define hyperpriors for $\boldsymbol{Q}_k$ and w_k. The standard conjugate priors are the multivariate Wishart density $W(\boldsymbol{Q}_k^{-1} \mid \nu_0, \boldsymbol{Q}_0)$ and the multivariate Dirichlet density $D(\mathbf{w} \mid \boldsymbol{\eta}_0)$, where $\mathbf{w} = (w_1, \ldots, w_K)^{\mathrm{T}}$ and $\boldsymbol{\eta}_0 = (\eta_{01}, \ldots, \eta_{0K})^{\mathrm{T}}$.

Note that with the bottom level parameters $\Theta = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K, \sigma^2\}$ we have

$$f(\Phi) = D(\boldsymbol{w} \mid \boldsymbol{\eta}_0) \prod_{k=1}^{K} W(\boldsymbol{Q}_k^{-1} \mid v_0, \boldsymbol{Q}_0), \tag{11}$$

$$f(\Theta, \boldsymbol{Z} \mid \Phi) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[w_k N(\boldsymbol{r}_i \mid \boldsymbol{\mu}_k, \boldsymbol{Q}_k) \right]^{z_{ik}}. \tag{12}$$

In the case of the same covariance matrices $\boldsymbol{Q}_1 = \ldots = \boldsymbol{Q}_K = \boldsymbol{Q}$ we may considerable simplify the complete-data *MAP* function. The hyperpriors for $\boldsymbol{Q}_k$ and w_k do not have to be defined, instead it is assumed $\boldsymbol{r}_i$ given $z_{ik} = 1$ as

$$\boldsymbol{r}_i = \boldsymbol{\gamma}_i + \boldsymbol{\mu}_k \tag{13}$$

and

$$\boldsymbol{\mu}_k = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\beta}_k, \tag{14}$$

where $\boldsymbol{\lambda}_0$ is the p-th vector, $\boldsymbol{\beta}_k$ is the h-th vector and $\boldsymbol{\Lambda}$ is the p x h matrix with h ≤ min(p,K-1). As a result, $\boldsymbol{X}_i\boldsymbol{\lambda}_0$ may be interpreted as the overall mean curve and $\boldsymbol{\Lambda}\boldsymbol{\beta}_k$ as the cluster variable increment.

Thus, with regard to $\boldsymbol{\gamma}_i \sim N(\boldsymbol{0}, \boldsymbol{Q})$, it is possible to rewrite the complete-data *MAP* function into the form

$$\begin{aligned} M_c(\overline{\Theta}, \Phi) = & -\frac{1}{2} \sum_{i=1}^{N} (\log|\boldsymbol{Q}| + \boldsymbol{\gamma}_i^T \boldsymbol{Q}^{-1} \boldsymbol{\gamma}_i) + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log w_k - \\ & - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left[m_i \log \sigma^2 + \frac{1}{\sigma^2} (\boldsymbol{y}_i - \boldsymbol{X}_i(\boldsymbol{\mu}_k + \boldsymbol{\gamma}_i))^T (\boldsymbol{y}_i - \boldsymbol{X}_i(\boldsymbol{\mu}_k + \boldsymbol{\gamma}_i)) \right]. \end{aligned} \tag{16}$$

Note that if no constraints are imposed, $\boldsymbol{\lambda}_0$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\beta}_k$ are confounded. Therefore it is required to set

$$\sum_{k=1}^{K} \boldsymbol{\beta}_k = \boldsymbol{0}, \tag{17}$$

$$\boldsymbol{\Lambda}^T \boldsymbol{X}^T \left(\sigma^2 \boldsymbol{I} + \boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^T \right) \boldsymbol{X}\boldsymbol{\Lambda} = \boldsymbol{I}, \tag{18}$$

where $\boldsymbol{X}$ is a matrix that contains the full range of the data.

4 Experimental results

In this section, we report experimental results with simulated data that show the efficiency of curve modeling techniques when clustering sets of curves. Both

introduced algorithms are implemented in the *Mathematica 4.0* software. Our experiments are carried out as follows:

- A three component mixture model is chosen by sampling 50 individual uncorrelated coefficient vectors from a normal distribution with standard deviation 0,01 centered with the cluster probabilities 12/50, 22/50, 16/50 around $\boldsymbol{\mu}_1$=(140; 3,2; -0,25; -0,04)′, $\boldsymbol{\mu}_2$=(150; -1; -0,7; 0,03)′ and $\boldsymbol{\mu}_3$ = (148; -3,4; -0,8; 0,06)′.
- The sequence of x-points is linearly spaced from 1 to 12 and training set of curves is generated by drawing from normal distribution with Vandermonde matrices $\boldsymbol{X}_i$ and standard deviation 10.

The plot of Fig. 1 shows the generated curve data with classification labels.

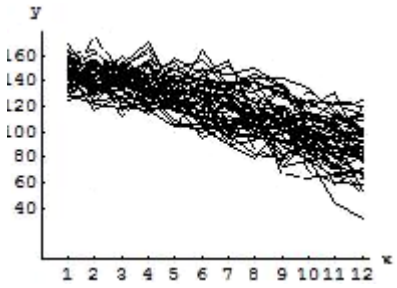


Fig. 1. MAP – curve data.

Let first suppose that there are different covariance matrices $\boldsymbol{Q}_1,\ldots,\boldsymbol{Q}_K$. Thus, there is a issue of setting the hyperparameters for the hyperprior of $W(\boldsymbol{Q}_k^{-1}|\ \nu_0,\ \boldsymbol{Q}_0)$ and $D(\mathbf{w}\ |\boldsymbol{\eta}_0)$. It is profitable to set $\nu_0 = p+1$, $\eta_{01} =\ldots= \eta_{0K} = 1$ and $\boldsymbol{Q}_0 = \boldsymbol{I}/c$ for some positive c that may influence heterogeneous behavior of clusters. We set c = 0,01 and initialize

$$\boldsymbol{r}_{ik} = (\boldsymbol{X}_i^{\boldsymbol{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^{\boldsymbol{T}}\boldsymbol{y}_i. \qquad (\mathbf{19})$$

After five random initialization and iteration processes searching for the local maximum of the complete-data *MAP* function, only four curves are misclassified. The resulting clusters are shown in Fig. 3, Fig. 4 and Fig. 5.

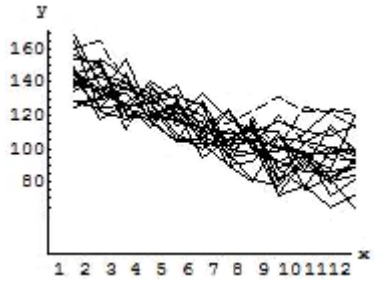


Fig. 2. MAP 1 – cluster 1.

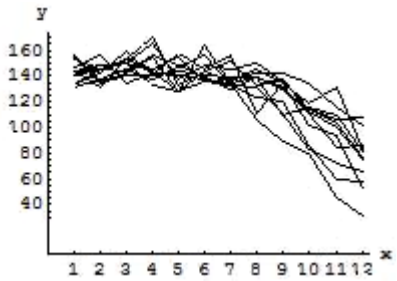


Fig. 3. MAP 1 – cluster 2.

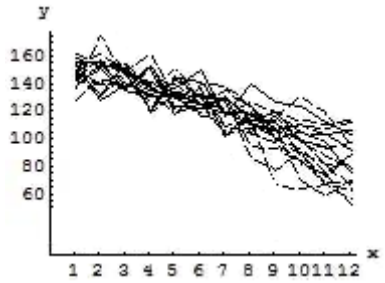


Fig. 4. MAP 1 – cluster 3.

When supposing $\boldsymbol{Q}_1=\ldots=\boldsymbol{Q}_K=\boldsymbol{Q}$ we set $h = 1$ such that the means lie in a restricted subspace and initialize

$$\boldsymbol{\gamma}_{ik} = \boldsymbol{0}, \tag{20}$$

$$\boldsymbol{Q} = \frac{1}{N}\sum_{\boldsymbol{i}=1}^{N}\sum_{k=1}^{K} w_{ik}\left(\frac{\boldsymbol{X}_i\boldsymbol{X}_i^{\boldsymbol{T}}}{\sigma^2}\right)^{-1} \tag{21}$$

with

$$\sigma^2 = \frac{1}{N}\sum_{\boldsymbol{i}=1}^{N}\frac{\left(\boldsymbol{y}_i - \boldsymbol{X}_i\left(\boldsymbol{X}_i\boldsymbol{X}_i^{\boldsymbol{T}}\right)^{-1}\boldsymbol{X}_i^{\boldsymbol{T}}\boldsymbol{y}_i\right)^{T}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\left(\boldsymbol{X}_i\boldsymbol{X}_i^{\boldsymbol{T}}\right)^{-1}\boldsymbol{X}_i^{\boldsymbol{T}}\boldsymbol{y}_i\right)}{m_i\text{-}(p+1)}. \tag{22}$$

Since all three parts of the complete-data *MAP* function involve separate parameters they can be maximized independently of each other. However, the optimizing process consists of two parts because maximizing the excepted value of the third part implies an iterative stand-alone procedure where $\boldsymbol{\lambda}_0$ then $\boldsymbol{\beta}_k$ and finally the column $\boldsymbol{\lambda}_l$ of $\boldsymbol{\Lambda}$ are repeatedly optimized while holding all other parameters fixed.

For the reason we initialize

$$\boldsymbol{\beta}_k = \boldsymbol{0}, \tag{23}$$

$$\lambda_0 = \left(\sum_{\boldsymbol{i}=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^{\boldsymbol{T}}\right)^{\boldsymbol{-1}}\sum_{\boldsymbol{i}=1}^{N}\boldsymbol{X}_i^{\boldsymbol{T}}\boldsymbol{y}_i \tag{24}$$

and moreover

$$\lambda_{11} = \frac{1}{\sqrt{\left(\boldsymbol{X}^{\boldsymbol{T}}\left(\sigma^{\boldsymbol{2}}\boldsymbol{I} + \boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^{\boldsymbol{T}}\right)\boldsymbol{X}\right)_{11}}}. \tag{25}$$

We iterate until all parameters have converged which occurs rapidly. As can be seen in Fig. 5., Fig. 6. and Fig. 7., this approach leads to quite perfect classification with respect to the true cluster memberships.

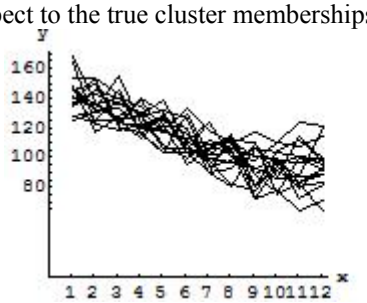


Fig. 5. MAP 2 – cluster 1.

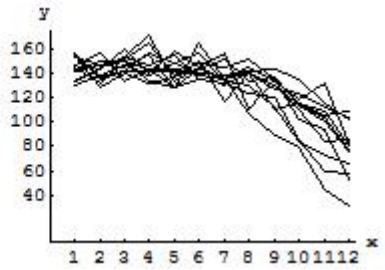


Fig. 6. MAP 2 – cluster 2.

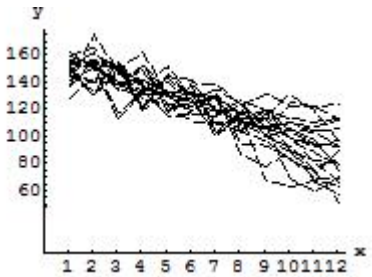


Fig. 7. MAP 2 – cluster 3.

5 Conclusion

This paper was concerned with the extension to the common clustering methods. We focused on the probabilistic density-based two-level hierarchical clustering methods using *MLE* and the efficient *MAP*-based *EM* algorithm for the estimation of both the individual and cluster-specific variability. Specially, we were dealing with an assumption of the same covariance matrices within all clusters. Finally, we reported extensive simulated data experiment that demonstrates that these techniques account for the inherent smoothness information in trajectories and may efficiently handle irregular sampled curves with significant within-cluster variability that is well described through the use of top-level distributions on individual-specific regression parameters.

References

1. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likehood from incomplete data via EM algorithm. Journal of the Royal Statistical Society, No. 39, pp. 1-38 (1977)
2. Banfield, J. D., Raftery, A. E.: Model-Based Gaussian and Non-Gaussian Clustering. Biometrics 49, 803-821 (1993)

3. Fraley, C.: Algorithms for Model-Based Gaussian Hierarchical Clustering. SIAM Journal on Scientific Computing, 20, 270-281 (1998)
4. Gaffney, S., Smyth, P.: Curve Clustering with Random Effects Regression Mixtures. Ninth Inter. Workshop on Artificial Intelligence and Statistics, Key West, FL, January 3-6 (2003)
5. Gaffney, S., Smyth, P.: Trajectory Clustering with Mixtures of Regression Models. Fifth ACM SIGKDD International Conference on Knowledge discovery- Data Mining NY. ACM Press, 63-72 (1999)
6. James, G. M., Sugar, C. A.: Clustering for sparsely sampled functional data. Journal of the American Statistical Association, Vol. 98, 397-408 (2003)
7. Jank, W.: Ascent EM for Efficient Curve-Clustering in Large Online Auction Databases. Robert H. Smith School Research Paper No. RHS-06-008 November (2004)
8. Lenk, P. J., DeSarbo, W. S.: Bayesian inference for finite mixtures of generalized linear models with random effects. Psychometrika, Vol. 65, 93-119 (2000)
9. Schwarz, G.: Estimating the Dimension of a Model. The Annals of Statistics, Vol. 6, 461-464 (1978)

Categorical Data Analysis in R

Martin Prokop
College of Polytechnics Jihlava, Tolstého 16, 586 01, Jihlava, Czech Republic
martin.prokop@vspj.cz

Abstract. The aim of this contribution is to show some possibilities of statistical software R in categorical data analysis. The procedures especially from contingency tables analysis are selected and they are illustrated on examples. Further the study includes one example from my working field, which describes the results of the broad survey about the knowledge of students of the European Union. This survey was carried out in the final grades of the secondary schools in the Vysočina Region aimed at the knowledge and opinions of final year students of the European Union issues. The contingency tables analysis and corresponding tests were selected to process the data.

Keywords: Contingency tables, Pearson´s chi- square test, non-parametric rank methods, Kruskal-Wallis test

1 Introduction

In the questionnaire survey analysis we get categorical data and easy way to illustrate the data are contingency tables. This contribution includes three examples from this field, two examples including the program code to show working with R software and one example from my working field.

According to the character of the data we use suitable tests of the independence. According to [5] in the case of contingency table of the type $r \times c$ (r is the number of rows, c is the number of columns) we usually use statistics:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (1)$$

Alternatively:

$$G^2 = \sum_i \sum_j n_{ij} \ln \frac{n_{ij}}{e_{ij}} \quad (2)$$

e_{ij} is an expected and n_{ij} real frequency. We use the statistic χ^2 in Pearson´s chi-square test, G^2 in likelihood-ratio test. These two statistics have asymptotically $\chi^2_{(r-1)(c-1)}$ distribution with the presumption of the independence.

Previous tests can be used in the case of high frequencies in the contingency table. If frequencies are too small, we can use Fisher´s exact test or we can calculate simulated p-value of χ^2 statistic.

In R there are the procedures *chisq.test* for Pearson´s chi-square test, *simulate.p.value* for small frequencies in the cells and *fisher.test*.

Statistic χ^2 is not standardized, hence it is not suitable for investigation of the dependence intensity. Instead of these statistics we can use contingency coefficients: Pearson´s contingency coefficient,

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{3}$$

Cramer´s contingency coefficient,

$$C = \sqrt{\frac{\chi^2}{n \min(r-1, c-1)}} \tag{4}$$

and phi contingency coefficient.

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{5}$$

Pearson´s contingency coefficient takes values from the interval $[0;1)$, Cramer´s contingency coefficient from $[0;1]$ and phi-contingency coefficient can be used for 4-cells tables.

In R there is the library *vcd* and the procedure *assoc.stats* to calculate these coefficients and values of statistics $\chi^2$and G^2.

Sometimes we get zero in some cell of contingency table. It can be random zero in the case of small range of sampling, but it can be also structural zero in the case of impossible situation. This problem can be solved by iterative proportional fitting procedure, in R with the function *loglin* and the first table for iteration includes numbers 0 in the case of structural zero and 1 otherwise. We test so called quasi-independence.

2 Examples

2.1 Chlamidia

To show the calculations with contingency tables I selected the dataset from [6] with name Chlamidia with the data about chlamidia infections according to the age, race and gender of diseased people.
Following part contains the program code to make contingency table from the data, to calculate marginal and relative frequencies and to process Pearson´s chi-square test of the independence. I selected the dependence of the gender and race of diseased people. Further there are results. The commands are denoted by the symbol `>`.

Input commands
Data input from file

```
> Dataset <- sqlQuery(channel = 1, select * from
[Sheet1$])
> chlam<-Dataset[-101,]
> attach(chlam)
```

Contingency table

```
> kt1<-tapply(Count,list(Race,Gender),sum)
```

Marginal frequencies

```
> marg1<-addmargins(kt1)
```

Relative frequencies in rows

```
> r1kt1<-prop.table(kt1,1)
```

Relative frequencies in columns

```
> r2kt1<-prop.table(kt1,2)
```

Total relative frequencies

```
> rkt1<-prop.table(kt1)
```

Pearson´s chi- square test

```
> chisq.test(kt1)
```

Results (output)
Contingency table with the dependence of the race and gender

```
> kt1
                   Female  Male
American Indian      7600  1388
Asian/Pacific Island  5571  1044
Black               126988 29178
Hispanic             58830 10999
White               107119 17186
```

Contingency table including marginal frequencies

```
> marg1
                     Female  Male    Sum
American Indian        7600  1388   8988
Asian/Pacific Island   5571  1044   6615
Black                126988 29178 156166
Hispanic              58830 10999  69829
White                107119 17186 124305
Sum                  306108 59795 365903
```

Table of relative frequencies in the rows

```
> r1kt1
                        Female      Male
American Indian      0.8455719 0.1544281
Asian/Pacific Island 0.8421769 0.1578231
Black                0.8131604 0.1868396
Hispanic             0.8424866 0.1575134
White                0.8617433 0.1382567
```

Table of relative frequencies in the columns

```
> r2kt1
                         Female       Male
American Indian      0.02482784 0.02321264
Asian/Pacific Island 0.01819946 0.01745965
Black                0.41484705 0.48796722
Hispanic             0.19218707 0.18394515
White                0.34993858 0.28741534
```

Table of total relative frequencies

```
> rkt1
                         Female        Male
American Indian      0.02077053 0.003793355
Asian/Pacific Island 0.01522535 0.002853215
Black                0.34705373 0.079742445
Hispanic             0.16078032 0.030059879
White                0.29275245 0.046968732
```

Pearson´s chi- square test describing the dependence between the race and gender

```
> chisq.test(kt1)

      Pearson's Chi-squared test

data:  kt1
X-squared = 1226.904, df = 4, p-value < 2.2e-16
```

The result is significant. So there exists the dependence between variables gender and race of diseased people.

2.2 Cancer

In the book [2] there is the example and the data about new occurrence of the cancer in the state NY within the year 2003. It is sure, empty cells are structural zeros, because some kind of cancer can not occur in the case of man or woman.

Table 1. Counts of ill people.

Cancer	Men	Women
Lung	7355	4831
Melanoma	1104	964
Ovarian	0	1563
Prostate	9986	0
Stomach	1014	619

We study the dependence between the variables kind of cancer and gender of ill people.

Input commands

Using the function loglin

```
> data<- matrix
(c(7355,4831,1104,964,0,1563,9986,0,1014,618),byrow=T,nro
w=5)
> start<-matrix(c(1,1,1,1,0,1,1,0,1,1), byrow=T,nrow=5)
> dataloglin<-loglin(data,list(1,2),start=start,fit=T)
```

Commands for the calculation of statistics χ^2 and G^2

```
> dataloglin$lrt
[1] 39.99608
> dataloglin$pearson
[1] 40.41931
```

We got the results $\chi^2 = 40.41931$ and $G^2 = 39.99608$ and corresponding p-values are:

```
> pchisq(40.41931,2,lower.tail=F)
[1] 1.671315e-09
> pchisq(39.99608,2,lower.tail=F)
[1] 2.065197e-09
```

Both results are significant. So there exists the dependence between the variables kind of cancer and gender of ill people.

2.3 Example from my Working Field: The Influence of the Languages Knowledge Level on Knowledge of the European Union Issues among Students of the Final Grade of Secondary Schools (Joint Work with Jana Borůvková, Stanislava Dvořáková and Bohumil Minařík)

Background

At the start of the year 2009, a broad survey was carried out in the final grades of the secondary schools in the Vysočina Region aimed at knowledge and opinions of final year students of the European Union issues. The survey was held by electronic method using the ReLa web application (DMB FEM Mendel University in Brno) and it was organized by teachers of Mendel University in Brno and Polytechnic University in Jihlava. Help was offered by the Regional Authority of Vysočina with informing and activating the secondary schools of the region.

The electronic questionnaire was extensive - it contained 69 items. Identification questions enable to structure the file of respondents into lots of segments. For the purpose of this article, the key criterion of segmentation was the level of languages knowledge.

During statistical processing of the obtained information special attention was paid to the identification mark the level of languages knowledge.

Table 2. Structure of respondents according to the level of languages knowledge.

Group	Category	Frequency	Percentage
1	very low	99	6.28173
2	low	450	28.55330
3	high	901	57.17005
4	very high	126	7.99492

From the point of view of statistics, when assessing the differences between the groups of respondents, it is necessary to take several factors into account: above all the discontinuous character of the data, their significant asymmetry, and heterogenity of the dispersion of the groups. This is why non-parametric rank methods were used. Relatively high number of coincidences of order/rank was resolved as follows: uncorrelated random error with normal distribution, zero standard value and minimum dispersion, which does not influence the overall rank of the respondents, was added to the original data.

The set of questions included 40 questions comprising eight thematic areas with five questions in each topic:

. geography of the EU,

. history of the EU,

. institutions, policy and law of the EU,

. economy and finance of the EU,

. culture, education and sport in the EU,

. the Czech Republic in the EU,

. the "seamy side" of the EU

. superlatives in the EU

Kruskal-Wallis test was used to evaluate the differences between groups; the test verifies the hypothesis that more than two independent samplings come from the same continuous partings. The above mentioned processing of the input data enabled unambiguous attachment of the ordinal numbers. Under these circumstances the mean of the order numbers is $\frac{n+1}{2}$ and dispersion $\frac{n^2+1}{12}$.

The test criterion is based on the assessment of variability of mean ranks in individual groups. If the tested hypothesis is true and the range of choice of asymptotically distributed χ^2 is big enough then the number of degrees of freedom equals to the number of groups less one. The test is single-sided. Among several methods available for subsequent testing, which allow different sizes of the groups, Dunn criterion was chosen. It is one of methods for multiple comparison using 95 % Dunn interval.

On the level of single questions a contingency table with two lines and four columns was used which verified the significance of the calculated square contingence.

Results and Discussion

At the beginning we carried out global evaluation of correctness in the knowledge section of the questionnaire to get a general overview. Only the total final percent of correct answers with each respondent was taken into account.

Table 3. Division of respondents of individual groups according to the number of correct answers.

Group	very low	low	high	very high
0 to 10	22.22%	15.33%	8.99%	7.94%
11 to 20	65.66%	66.00%	63.15%	57.14%
21 to 30	12.12%	15.56%	23.64%	30.95%
31 to 40	0.00%	3.11%	4.22%	3.97%

Table 4. Division of respondents of individual groups in the interval of the quantiles.

Group	very low	low	high	very high
0% to 25%	39.39%	31.78%	20.20%	23.81%
25% to 50%	31.31%	25.11%	25.75%	15.87%
50% to 75%	17.17%	24.00%	26.08%	25.40%
75% to 100%	12.12%	19.11%	27.97%	34.92%

For the whole knowledge section of the questionnaire (40 questions) the result of Kruskal-Wallis test is presented in Table 5 and the subsequent testing (Dunn) in the tables 6 and 7.

Table 5. Result of Kruskal-Wallis test (level of languages knowledge).

Level	Count	Sum of rank	Mean rank
very low	99	56448	570.1818
low	450	317556	705.6800
high	901	755540	838.5572
very high	126	113132	897.8730
Total	1576	1242676	788.5000

concordance correction = 0, statistics chi-square = 55.86
degrees of freedom = 3, right probability = 0

Table 6. Subsequent testing of groups according to the level of languages knowledge (Dunn), all questions.

Homogenous subgroups		
Group 1:	very low	
Group 2:	low	
Group 3:	high	very high

Table 7. Statistical significance of differences of ranking for individual thematic groups of questions.

Homogenous subgroups	1. Geography of the EU
Group 1:	very low low high very high
Homogenous subgroups	2. History of the EU
Group 1:	very low low
Group 2:	low high
Group 3:	high very high
Homogenous subgroups	3. Institutions, policy and law of the EU
Group 1:	very low low
Group 2:	low very high high
Homogenous subgroups	4. Economy and finance of the EU
Group 1:	very low low
Group 2:	low high very high
Homogenous subgroups	5. Culture, education and sport in the EU
Group 1:	very low
Group 2:	low
Group 3:	high very high
Homogenous subgroups	6. The Czech Republic in the EU
Group 1:	very low low very high
Group 2:	low very high high
Homogenous subgroups	7. The "seamy side" of the EU
Group 1:	very low
Group 2:	low
Group 3:	high very high
Homogenous subgroups	8. Curiosities and superlatives in the EU
Group 1:	very low low high
Group 2:	low high very high

Table 8. Mean ranking in thematic groups of questions.

Level	Count	A1	A2	A3	A4	A5	A6	A7	A8	Total
very low	417	702.18	681.37	634.80	661.48	586.60	673.26	577.07	685.26	570.18
low	790	739.65	742.43	764.50	758.81	732.82	765.91	713.92	750.56	705.68
high	318	815.62	810.51	814.95	810.85	827.85	811.19	832.47	808.17	838.56
very high	51	836.84	879.85	805.89	834.49	864.58	797.49	906.60	864.45	897.87

Final Results
Mean ranking in thematic groups of questions is corresponding to stated level of the languages knowledge. With higher level of the knowledge the mean rank is better. There are only two exceptions. In the groups 3 and 6 (Table 8) there was better mean ranking in the case of high level knowledge students than in the case of students with very high level of knowledge. In the case of all questions statistical dependence was not proved (p=0.07392 in Pearson´s chi-square test), but in separated groups there was quite strong dependence with p-values less than 0.05 except of the group 8 (p-value nearly over 5%) and group 6. In the case of separated questions half of the results were significant.

3 Conclusions

The aim of this contribution was to show some possibilities of statistical software R in the contingency tables analysis. It was shown, how to realise basic calculations connected with the contingency tables and corresponding tests of independence. The contribution includes one example from my working field and my colleagues discussing the results of questionnaire survey.

References

1. Everitt, B.: An R and S-PLUS companion to multivariete analysis. Springer-Verlag, London (2005). ISBN 1-85233-882-2.
2. Simonoff, J.S.: Analyzing categorical data. Springer, New York (2003). ISBN: 0387007490.
3. Řezanková, H.: Analýza dat z dotazníkových šetření. Professional Publishing, Praha (2007). ISBN-978-80-86946-49-8.
4. Matějů, P., Straková, J.: (Ne)rovné šance na vzdělání: vzdělanostní nerovnosti v české republice, Academia (2006). ISBN-80-200-1400-4.
5. Anděl, J.: Základy matematické statistiky. Matfyzpress, Praha (2005). ISBN 80-7378-001-1.
6. http://pages.stern.nyu.edu/~jsimonof/AnalCatData

An Empirical Analysis of the Permanent Income Hypothesis in the Czech Republic

Lenka V. Půlpánová[1]

[1] Department of Economic Statistics, Faculty of Informatics and Statistics, University of Economics, 130 67 Prague 3, W. Churchill Sq. 4, Czech Republic
lenka.pulpanova@vse.cz

Abstract. This paper re-examines the validity of the permanent income hypothesis under the rational expectations assumption in the Czech economy. The relationship between consumption and income is analysed by means of cointegration analysis in a single-equation model. As consumption and income proved to be co-integrated, the error correction model has been built in order to estimate a long-run relationship between the two variables. The analysis takes advantage of the recent availability of more detailed breakdowns of consumption data and methodologically consistent quarterly time series for the last decade. The results of the permanent income hypothesis test show mixed results, depending on the choice of variables.

Keywords.: Consumption, Income, Rational expectations, Permanent income hypothesis, Co-integration, Error correction model.

1 Introduction

The permanent income hypothesis is a theory of consumption. The hypothesis states that consumption patterns are determined by longer-term expectations of income, therefore transitory short-term changes in income have small impact on consumer spending behaviour.

Much effort has been made and a vast literature exists on whether the response of consumption to income is consistent with the permanent income hypothesis. Since the seminal work of Hall (1978) [10], who explained the stochastic implications of the permanent income hypothesis, manifold tests of the permanent income hypothesis have been developed. Some of the tests were performed in the framework of the structural econometric consumption function e.g. Flavin [9] and some of them in other environments, as shown in the papers of Mankiw and Shapiro (1985) [14], Stock and West (1987) [16], Campbell and Mankiw (1990) [4], and others.

Each empirical test has its pros and cons, as the tools and methods employed often relaxed certain theoretical assumptions. The subsequent research has shown that some of the tests were inappropriate and sometimes highly biased. They did not tackle appropriately problems of spurious regressions, ignored unit roots issues, used de-trended time series and thus concentrated only on a short run relationship between consumption and income etc. The empirical research in this field has always been

closely related to new findings and developments in statistics and econometrics. In this respect, work on co-integration and error correction models, firstly suggested by Granger and further elaborated by Engle and Granger (1987) [6], is perceived as one of the major contributions for the empirical tests of permanent income hypothesis. However, the international empirical evidence does not offer conclusive results and supports the hypothesis only in a few countries [15].

As regards the Czech Republic, several papers on various issues regarding the relation of consumption and income have been published. The research concentrated mostly on the analysis of the consumption function, e.g. Hušek (1999) [12] and Filáček (1999) [8] modelled household consumption and later on Mandel and Tomšík (2003) [13] developed an enlarged version of the consumption function for the Czech Republic and tested the Ricardian equivalence hypothesis in a small open economy.

There are few working papers addressing directly the permanent income hypothesis issue in the Czech Republic. One of them is a specific test of permanent income hypothesis on the Czech voucher privatization by Hanousek and Tůma (1997) [11]. The explicit test of permanent income hypothesis was suggested by Arlt, Čutková and Radkovský (2001) [2]. They developed an econometric model of consumption based on disposable and labour income and rejected the permanent income in the Czech Republic at that time.

This paper aims at re-examining the validity of the permanent income hypothesis in the Czech economy. The test of the permanent income hypothesis employed does not require construction of a structural consumption function model, but might have implications for development of such a model. It also does not elaborate on the consumption behaviour of the Czech households in detail. The main goal of the paper is to analyse relationship between consumption and income while using the tools of co-integration analysis.

The permanent income hypothesis theory implies unit co-integration vector of disposable income and consumption, therefore the empirical strategy is as follows: If co-integration of the variables concerned is confirmed, then both short- and long-run relationships can be estimated by an error correction model. The estimated co-integration vector is then compared with the theoretical one.

2 Permanent income theory and its stochastic implications

The rational expectations theory, suggested in the 1960s, became an integral part of both new classical and new Keynesian macroeconomics. The concept states that outcomes do not vary systematically or predictably from what people expected. Rational expectations assume that people behave rationally, i.e. take into account all available information in order to maximize their utility (e.g. consumption). Linking the rational expectations theory with consumption theories then lead to the permanent income theory under rational expectations.

As explained e.g. in [5], under the life-cycle or permanent income hypotheses, individual consumption does not depend on current income alone, but also on prospects of income in the future. The consumption is determined by the value

of lifetime resources or eventually by permanent income, usually defined as average or expected income. Since the permanent income is assumed to be an annuity of lifetime resources, the two theories are very close. People consume out of their permanent income, which equals the level of consumption that can be sustained while leaving wealth intact. In defining wealth, and measure of human wealth namely, the present value of people's expectations of future labour income is considered. Since the permanent income depends on lifetime resources, it is unlikely to fluctuate much in response to short-term fluctuations in income. Permanent income is therefore smoother than actual current income, so the theory is consistent with the observation that consumption is smoother than income. Over a span of many years, permanent income aligns on average with the measured income, so that in the long run, consumption is proportional to income.

Consumption theory under rational expectations hypothesis was elaborated by Hall (1978) [10]. Based on his theorem, suggested in the same paper, he showed that under certain assumptions, marginal utility of consumption follows a random walk with drift and, with reasonable approximation, the consumption itself should evolve in the same way. Such a conclusion has strong stochastic implications. Consumption growth lagged more than one period has no additional predictive power for current consumption growth. Very rigorous testable implication of the random walk hypothesis is that any other economic lagged variable has no predictive power with respect to current consumption growth including for example lagged income, changes in stock prices etc. Hall explains that permanent income theory of consumptions assumes an intelligent forward-looking consumer (actually rational one), so if the previous value of consumption incorporated all information available at that time, then lagged values of actual income should have no explanatory power once lagged consumption is included, i.e. the best prediction of future consumption is the present level of consumption. Hall used in fact some kind of regression analysis in order to test the permanent income hypothesis and did not explicitly mention the issue of unit roots for the time series used in his analyses. This problem was attacked in successive work of Stock and West (1987) [16]. They clearly addressed the issue that estimation and testing procedures in the presence of unit roots might be highly biased and pointed out that co-integration properties of the regressors (lagged income and other variables) are of high importance.

In this paper we apply the modification of Flavin´s (1981) [9] permanent income hypothesis model. In short it may be described in a following way. Consumption C_t equals permanent income Y_t^P, the annuity value of the sum of human wealth H_t and non-human wealth W_t, r is real interest rate, Y_t is labour income and E_t denotes expectations conditional on the consumers' information set as of time t:

$$C_t = rW_t + r(1+r)^{-1}H_t . \quad (1)$$

$$H_t = E_t \sum_{j=0}^{\infty}(1+r)^{-j} Y_{t+j} . \quad (2)$$

$$W_t = (1+r)W_{t-1} + Y_{t-1} - C_{t-1} . \quad (3)$$

Thus, consumption is proportional to the sum of human wealth, which is the expected present value of future labour income and accumulated savings. Under the permanent income hypothesis, the change in consumption equals the unpredictable change in annuity value of labour income, stemming from information shock:

$$\Delta C_t = C_t - C_{t-1} = r(H_t - E_{t-1}H_t) = \varepsilon_t \ . \quad (4)$$

This result is consistent with the above mentioned conclusion that consumption follows a random walk, where random information shock is not correlated to the expectations in the previous period, $E_{t-1}\varepsilon_t = 0$. As emphasized by Campbell (1987) [3], proposing an alternative test of permanent income hypothesis based on saving rates, also current savings can be obtained from the equations (1) to (3). Let's denote disposable income Y_t^D. It can be expressed as a sum of labour income and accumulated savings.

$$Y_t^D = Y_t + rW_t \ . \quad (5)$$

Then, relationship between disposable income Y_t^D, consumption C_t and current savings S_t is obtained by substituting (2) into (1) and assuming (5), where Δ denotes a standard backward difference:

$$Y_t^D - C_t = Y_t + rW_t - C_t = -\sum_{j=1}^{\infty} E_t(1+r)^{-j}\Delta Y_{t+j} \ . \quad (6)$$

The difference between disposable income and consumption equals to the expected present value of future declines in labour income. This is the so called "saving for a rainy day" feature of the permanent income hypothesis model. It is assumed that the term $\sum_{j=1}^{\infty} E_t(1+r)^{-j}\Delta Y_{t+j}$ is stationary, a saving is a discounted present value of expected changes in labour income, these changes are stationary, so saving is also. This assumption has significant implications. First, a linear combination of consumption C_t and disposable income Y_t^D, which can be thought as saving, is stationary in its level even if both time series are non-stationary I(1). Those two variables are co-integrated in the sense of Engel and Granger [6]. Second, the equation (3) suggests that consumption C_t, labour income Y_t and nonhuman wealth W_t should be co-integrated in the same sense.

If the permanent income hypothesis holds, C_t equals permanent income Y_t^P and C_t-Y_t^D=ε_t, while ε_t is I(0), consumption and disposable income are co-integrated with unit co-integration parameters (1,-1). The same line of reasoning applies when co-integration of labour income, nonhuman wealth and consumption is tested. Therefore, labour income alone should not be co-integrated with consumption at all or with different co-integration parameters, as the non-human wealth must be considered as well. In addition, when consumption and disposable income are co-integrated with

vector (1,-1), then we can derive from the equations (3) and (5) that non-human wealth W_t is also non-stationary I(1):

$$Y_t^D - C_t = W_{t+1} - W_t = \Delta W_{t+1}. \quad (7)$$

In case that consumption and disposable income are co-integrated with vector (1,-1), then labour income might be either stationary I(0) or non-stationary I(1). In both cases the nonhuman wealth W_t is again non-stationary I(1).

3 Sources and data treatment

As regards consumption, economists extended the permanent income hypothesis model in order to allow for factors such as habit persistence in consumption, liquidity constraints on households and the range of durability length of various consumption goods. Consumption can be broken down to durable and non-durable components, while usually only consumption expenditure on non-durable goods and services is considered in empirical analyses. The idea behind is that durable part of consumption is already included in stock of non-human wealth and therefore already in some way in the data. However, some authors e.g. [17] claim that this issue may be irrelevant, as no theory of consumer behaviour needs to be involved in order to derive equations in Section 2. The similar issue of a proper definition for empirical analysis pertains to labour income. Wages and salaries are usually used as a good approximation of labour income.

In estimations two data source were employed, both available at the Czech Statistical Office web pages [19]. The first set of data comes from the quarterly National Accounts (ESA 95 methodology) and the second one from the monthly Population Statistics of the Czech Republic. The Population Statistics provide quarterly data used for per capita calculations.

Household consumption and labour income (approximated by wages and salaries) data come from the National Accounts, the quarterly GDP estimation [18]. Detailed breakdown of nominal and real consumption data is available in the data set on Household final consumption expenditure by durability, where household consumption in domestic concept is broken down to durable goods, semi-durable goods, and non-durable goods and services. The latter data set became available only recently and belongs to the GDP expenditure approach estimation. Quarterly current wages and salaries are available in data on GDP income approach. Current disposable income, which is a balancing item, is taken from the table Transactions in products and distributive transactions for the household sector. Unfortunately, data on nonhuman wealth, which are in national accounts terminology represented by the net wealth of households sector, is not available on a quarterly basis and no approximation, using e.g. the net financial wealth, has been used.

All time series in current prices were deflated by the total household expenditure deflator in order to obtain real household total consumption, real non-durable goods and services consumption, real wages and salaries and real households disposable

income. Finally, all time series were divided by population to obtain data per capita, and subsequently seasonally adjusted by X12-ARIMA method.

The real consumption data, consumption by durability and real wages and salaries are available since the first quarter of 1996, real disposable income from the first quarter 1999 up to the first quarter of 2009. The data used in the empirical analyses therefore count last ten years and one quarter (41 observations). Development of real total consumption, real consumption on non-durables and services, real labour income and real disposable income per capita is shown in the Figure 1, while the Figure 2 describes the same variables, but seasonally adjusted by X12-ARIMA method.

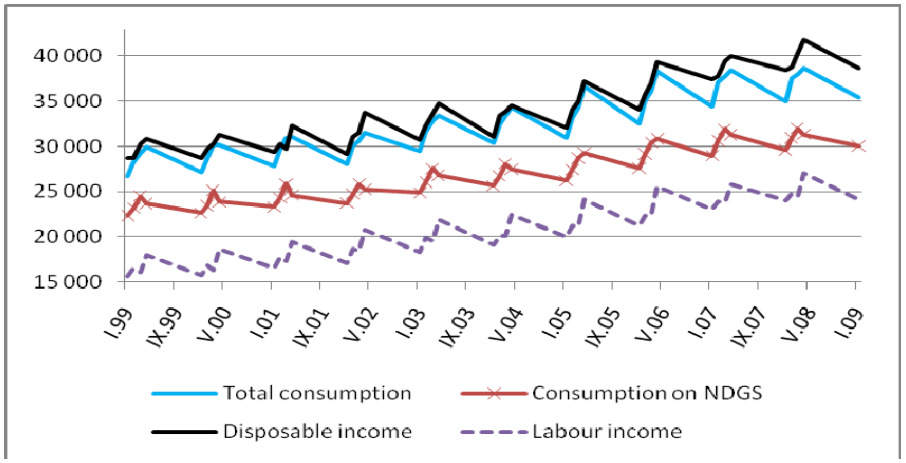


Fig. 1. Total consumption, consumption on nondurable goods and services (NDGS), disposable and labour income (real per capita) in CZK.

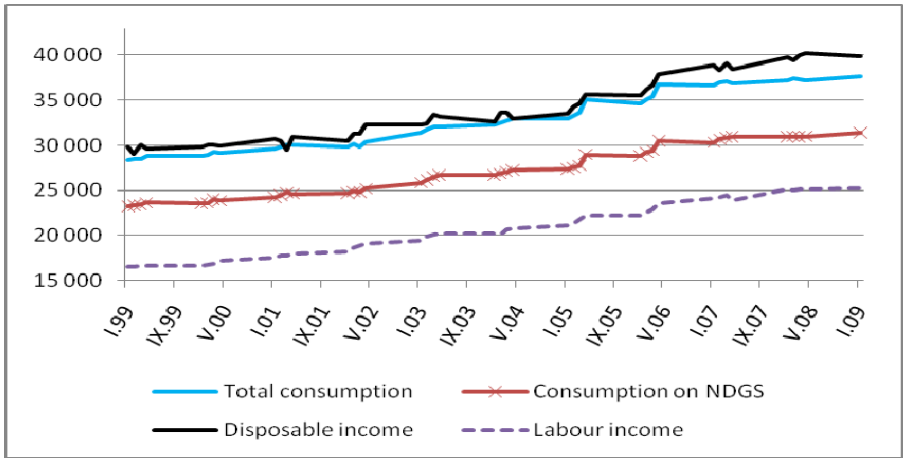


Fig. 2. Total consumption, consumption on nondurable goods and services (NDGS), disposable and labour income (real per capita) in CZK seasonally adjusted by X□12ARIMA.

4 Co-integration of consumption and income in an error correction model

Firm linkage between co-integration and error correction models comes from the Granger representation theorem, stating that two or more co-integrated time series that have an error correction representation, and two or more time series that are error-correcting are co-integrated [6]. However, this holds only for integrated time series processes. An error correction (EC) model is powerful in distinguishing between short-run and long-run relationships. The starting point for the derivation of a single equation EC model may be the autoregressive distributed lag model (ADL).

The derivation of an EC model is shown on an example of consumption C_t and disposable income Y_t^D. Both time series are assumed to be integrated I(1). The static regression of consumption C_t on disposable income Y_t^D can be performed first in order to check in advance whether the two time series are co-integrated.

$$C_t = c + \beta Y_t^D + u_t \,. \tag{8}$$

The non-stationary error terms u_t from the static regression indicates a problem of spurious regression. The co-integration between consumption and income can be debated only if the errors u_t from the static regression are stationary I(0) and two situations can be distinguished. First, if the errors are not autocorrelated, then there is only a long-run relationship between the two variables concerned and the parameter β can be interpreted as a long-run multiplier. Second, if the errors are autocorrelated, then including lagged variables in the static regression may solve the problem of autocorrelation and the equation has a form of e.g. ADL (1,1) model, where ε_t is assumed to be an i.i.d. random variable:

$$C_t = \alpha_0 + \alpha_1 C_{t-1} + \beta_0 Y_t^D + \beta_1 Y_{t-1}^D + \varepsilon_t \,. \tag{9}$$

The ADL model, however, does not explicitly show the long-run multiplier between consumption and income. The ADL model can be transformed into a generalized version of an error correction model. Assuming the ADL(1,1) model, the GECM model takes the following form:

$$\Delta C_t = \alpha_0 + (\alpha_1 - 1)(C_{t-1} - Y_{t-1}^D) + \beta_{02} Y_t^D + (\beta_0 + \beta_1 + \alpha_1 - 1) Y_{t-1}^D + \varepsilon_t \,. \tag{10}$$

The equation (10) explains how quickly the system reacts to any disequilibrium and the coefficient (α_1-1) can be interpreted as the speed at which consumption adjusts to any discrepancy between consumption and the disposable income in the previous period. The equation (10) is further transformed into the EC model:

$$\Delta C_t = \alpha_0 + \beta_0 \Delta Y_t^D + (\alpha_1 - 1)\left[C_{t-1} - \frac{\beta_0+\beta_1}{1-\alpha_1} Y_{t-1}^D\right] + \varepsilon_t \,. \quad (11)$$

The EC model at the same time describes short-run relationships between the variables (the parameter β_0 belonging to the differentiated Y_t^D) and long-run relationships (parameter $\frac{\beta_0+\beta_1}{1-\alpha_1}$ before the non-differentiated term). The co-integration vector is (1, $-\frac{\beta_0+\beta_1}{1-\alpha_1}$) and thus if consumption and income are co-integrated with vector (1,-1), than the second co-integration parameter, which is the long run multiplier $k=\frac{\beta_0+\beta_1}{1-\alpha_1} = 1$ and $\beta_0 + \beta_0 + \alpha_1 = 1$. The permanent income hypothesis might be then regarded as confirmed. It also implies that the average propensity to consume (APC) is constant.

5 Empirical test of the permanent income hypothesis

The empirical test of the permanent income hypothesis, as suggested in the previous sections, is tested in the EC model framework. The analysis takes advantage of recent availability of more detailed breakdowns of household consumption data, which are generally believed to more closely fit to the theoretical assumptions of the permanent income hypothesis. The test is simultaneously performed for the total household consumption data as some other authors [17] deem that there is no justification for distinguishing between consumption of durable and non-durable goods and the results can be compared to the analysis carried out by Arlt et al. (2001) [2].

The permanent income hypothesis representation illustrated above is investigated by means of co-integration analysis of the following pairs of variables expressed in real per capita terms:

1. Consumption on non-durable goods and services CND_t and disposable income Y_t^D,
2. Consumption on non-durable goods and services CND_t and labour income Y_t,
3. Total consumption C_t and disposable income Y_t^D,
4. Total consumption C_t and labour income Y_t.

Autocorrelations of total consumption C_t, consumption of non-durable goods and services CND_t, disposable income Y_t^D and labour income Y_t begin close to one and decline rather slightly. The partial autocorrelations of all variables decline close to zero after the first one, so the time series seem to have properties of a random walk. Also the unit root test, the Augmentd Dickey-Fuller test (ADF), suggests that all time series have a unit root, i.e. are non-stationary.

To see whether the prerequisites for co-integration of variables are fulfilled, the four OLS static regressions were calculated for the four pairs of variables. The residuals from the regressions were then checked for stationarity I(0), while using again the ADF test, but the test results were compared with the stricter MacKinnon's critical values [7], taking into account the number of time series in the static

regression. The ADF test suggests that residuals are all stationary. Therefore, the co□integration between the indicated pairs of time series can be considered. The equation (9) was estimated by the OLS and then transformed in the EC model in the form of equation (11).

5.1 Relationship between consumption on nondurable goods and services CND_t and disposable income Y_t^D

The ADL(1,1) model in the form of the equation (9) for consumption on non-durable goods and services and disposable income with standard errors in brackets was estimated first. * denotes statistical significance for the respective parameter at * 10%, ** at 5%, and *** at 1% level. The parameters α_0 and β_1, were statistically insignificant and therefore excluded, the ADL(1,0) model without a constant then takes this form:

$$CND_t = \underset{(0.10)***}{0.82CND_{t-1}} + \underset{(0.07)**}{0.15Y_t^D} . \quad (12)$$

The EC model derived from the ADL(1,0) model:

$$\Delta CND_t = 0.15\Delta Y_t^D - 0.18[CND_{t-1} - 0.82Y_{t-1}^D] . \quad (13)$$

The estimate of the long term multiplier *k=0.82* and significantly differs from 1. It implies that estimate of the co-integration vector is *(1,-0.82)*, and such a result does not support the propositions of the permanent income hypothesis.

5.2 Relationship between consumption on nondurable goods and services CND_t and labour income Y_t

The EC model was also estimated for consumption on nondurable goods and services and labour income. Nonetheless, with respect to non-validity of the permanent income hypothesis, this estimate has rather an illustrative purpose. The estimate of the EC model was derived from the ADL(1,0) model with a constant, all parameters are significant at 1% level:

$$\Delta CND_t = 2644.02 + 0.44\Delta Y_t - 0.46[CND_{t-1} - 0.94Y_{t-1}] . \quad (14)$$

The long-term multiplier was estimated *k=0.94,* the co-integration vector is *(1,□0.94)*, which shows a stable co-integration relationship between consumption on nondurable goods and labour income.

5.3 Relationship between total consumption C_t and disposable income Y_t^D

Similar ADL (1,1) model in the form of the equation (9) for total consumption and disposable income (with standard errors in brackets) was estimated in order to obtain comparable results with research carried out in 2001 [2]. The parameters α_0 and β_1, were statistically insignificant and therefore excluded from the model, which then takes form of the ADL(1,0) without a constant:

$$C_t = \underset{(0.09)***}{0.80C_{t-1}} + \underset{(0.08)**}{0.20Y_t^D}. \quad (15)$$

The EC model derived from ADL(1,0) model:

$$\Delta C_t = 0.20\Delta Y_t^D - 0.20[C_{t-1} - Y_{t-1}^D]. \quad (16)$$

The estimate of the long term multiplier *k=0.991*. The estimate of the co-integration vector is then *(1,-1)* and the co-integration vector therefore equals to the theoretical co-integration vector implied by the permanent income hypothesis. As a result, the permanent income hypothesis was confirmed for the total consumption and disposable income. This finding differs from permanent income hypothesis test from 2001 [2], as at that time the hypothesis was rejected. There may be several reasons for it e.g. the time series involved were even shorter and before the major methodological revision of national accounts etc.

5.4 Relationship between total consumption C_t and labour income Y_t

Section 2 suggested that if the permanent hypothesis is valid, then consumption and labour income should not be co-integrated with the same co-integration vector as consumption and disposable income. This additional examination also uses the EC model, which was derived from the ADL(1,0) model with a constant, where all parameters are significant at 1% level:

$$\Delta C_t = 5194.93 + 0,53\Delta Y_t - 0.49[C_{t-1} - 1.09Y_{t-1}]. \quad (17)$$

The long-term multiplier was estimated *k=1.09,* the co-integration vector is then *(1,1.09)* and this result is consistent with the permanent income hypothesis.

6 Results and conclusions

The permanent income hypothesis under rational expectations for the Czech economy was tested in order to explore the long run relationship between consumption and income by means of co-integration analysis. The results of the test offer mixed results. Considering the permanent income hypothesis in the form described in the section 2 of the paper, the test for consumption on non-durable goods and services and disposable income rejected the hypothesis. On the contrary, test for total consumption and disposable income does no reject the permanent income hypothesis and the results seems to be encouraging. The international empirical evidence only rarely confirms the permanent income hypothesis, in this context the Czech Republic might be seen rather as an exception.

It needs to be stressed that the outcomes of the test may be considerably dependent on several factors, which should be further analyzed. The results are rather sensitive in respect to the choice and the length of the data series, methods of seasonal adjustments and other data transformations, as well as lag structures of the variables and the estimation methods. In particular in the Czech Republic, the methodologically consistent time series are still relatively short for this kind of analysis and therefore the results should be taken with a caution.

There might be a number of issues, which were not studied in detail, but may have an influence on the test results. The strict assumptions of the permanent income hypothesis were not examined and may not hold. Further analysis is needed in order to reaffirm the theoretical groundings of the permanent income hypothesis and the empirical findings in the Czech Republic.

References

1. Arlt, J.: Kointegrace v jednorovnicových modelech, Journal of Political Economy No 5, VŠE Praha (1997)
2. Arlt, J., Čutková, J., Radkovský , Š.: Analýza spotření funkce v podmínkách ČR, ČNB working paper No. 34, Praha (2001)
3. Campbell, J.Y.: Does saving anticipate declining labor income? An alternative test of the permanent income hypothesis, Econometrical, Vol. 55, No. 6 (1987)
4. Campbell, J.Y., Mankiw, N.G.: Permanent income, current income, and consumption, Journal of Business & Economic Statistics, Vol. 8, No. 3 (1990)
5. Deaton, A.: Understanding consumption, Oxford University Press (1992)
6. Engle, R.F, Granger, W.J.: Co-integration and error correction: representation, estimation, and testing, Econometrica, Vol. 55, No. 2 (1987)
7. Engle, R.F, Granger, W.J.: Long-run relationships, Oxford University Press (1991)
8. Filáček, J.: Model spotřeby domácností v letech 1994-1998, Finance a úvěr No. 49 (1999)
9. Flavin, A.M.: The adjustment of consumption to changing expectations about future income, Journal of Political Economy, Vol. 89, No. 51 (1981)
10. Hall, R.E.,: Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence, Journal of Political Economy, Vol. 86, No. 6 (1978)
11. Hanousek, J., Tůma, Z.: A test of the permanent income hypothesis on Czech voucher privatisation, The William Davidson Institute Working Paper No. 75 (1997)

12. Hušek, R.: Econometric Analysis of the CR Consumption Function, Prague Economic Papers, Vol. 1996, No.1 (1996)
13. Mandel, M., Tomšík, V.: Spotřební funce a Ricardiánská ekvivalence v malé otevřené ekonomice, Journal of Political Economy, VŠE, No 4 (2003)
14. Mankiw, N.G, Shapiro, M.D.: Trends, random walks, and test of the permanent income hypothesis, Journal of Monetary Economics 16 (1985)
15. Slačálek, J.: International Evidence on cointegration between consumption, income, and wealth, Johns Hopkins University, Department of Economics (2004)
16. Stock, J.H., West, K.D.: Integrated regressors and test of the permanent income hypothesis, NBER working paper 2359 (1987)
17. Rudd, J., Whelan K.: A Note on the cointegration of consumption, Income, and Wealth, FEDS working paper No. 53, Board of Governors (2002)
18. Quarterly National Accounts Inventories, Czech Statistical Office (2008)
19. Czech Statistical Office, http://www.czso.cz

Dimensionality Reduction for Ordinal Data

Michaela Ryšánková and Hana Řezanková

University of Economics, Prague, Dept. of Statistics and Probability,
nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
{michaela.rysankova, hana.rezankova}@vse.cz

Abstract. We often obtain ordinal data as the results of questionnaire surveys. The special techniques should be applied in this case. The paper focuses on methods for discovery of groups of similar variables, as cluster analysis, multidimensional scaling (with special similarity measures), and optimal scaling which can help in the process of the dimensionality reduction. This reduction is useful for application of classification methods when objects are classified into groups. The data from questionnaires were analyzed by these methods in the SPSS Statistics and STATISTICA systems.

Keywords: Ordinal variables, cluster analysis, multidimensional scaling, optimal scaling.

1 Introduction

When objects are classified into groups by means of methods of statistical analysis, the choice of variables is an important task. The objects can be characterized by high-dimensional vectors of variable values. For simplification of analysis, dimensionality reduction is needed. In this process, the discovery of the relationships between variables is a very useful tool. Several variables can be replaces by only one of them or by a linear combination of these variables. In this paper we focus on the methods of graphical representation of the relationships – hierarchical clustering with its dendrogram, multidimensional scaling, and optimal scaling.

The first two methods are based on the proximity matrix which contains evaluation of relations in all pairs of variables; the use of these techniques is also possible for ordinal variables. The special similarity measures should be applied in this case. Optimal scaling is the method proposed for nominal and ordinal variables.

This means that there are several different techniques and their variants for graphical representation of variables relationships. Further, there are several similarity measures for ordinal variables. The problem is that we can get the different groups of variables by different methods. By analyzing two different data files we show that a combination of several methods is needed for the data structure discovery.

For the following analyses we use two well-known statistical packages – SPSS (now IBM SPSS) and STATISTICA. For more details concerning possibilities of these systems in the area of searching relationships between variables, see [4].

2 Similarity Measures for Ordinal Variables

In the process of searching groups of similar variables, coefficients of dependency are usually applied as similarity measures, see [3]. Dependency of the ordinal variables is denoted as a rank correlation and their intensity is expressed by correlation coefficients. The best known among them is Spearman's correlation coefficient. Let us have the n x p data matrix $\mathbf{X}$ with the elements x_{ij} where n is a number of objects and p is a number of variables. If investigated ordinal variables $\mathbf{X}_g$ and $\mathbf{X}_h$ express the unambiguous rank, the following formula can be used for *Spearman's correlation coefficient*:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} (x_{ig} - x_{ih})^2}{n(n^2 - 1)}. \quad (1)$$

If this assumption is not satisfied, the process described in [2] must be applied.

Other measures investigate pairs of objects. If, in a pair of objects, the values of both investigated variables are greater (less) for one of these objects, this pair is denoted as concordant. If for one variable the value is greater and for the second one it is less, then the pair is denoted as discordant. In other cases (the same values for both objects exist for at least one variable), the pairs are tied. For the sake of simplification, we will use the following symbols:

Γ – a number of concordant pairs,

Δ – a number of discordant pairs,

Ψ_g – a number of pairs with the same values of variable $\mathbf{X}_g$ but distinct values of variable $\mathbf{X}_h$,

Ψ_h – a number of pairs with the same values of variable $\mathbf{X}_h$ but distinct values of variable $\mathbf{X}_g$.

On these numbers of pairs, *Goodman and Kruskal's gamma* is based for example. Similarly as Spearman's correlation coefficient, it is a symmetric measure. It is expressed as

$$\gamma = \frac{\Gamma - \Delta}{\Gamma + \Delta}. \quad (2)$$

Another symmetric measure is *Kendall's tau-b* (Kendall's coefficient of the rank correlation). It is expressed as

$$\tau_b = \frac{\Gamma - \Delta}{\sqrt{(\Gamma + \Delta + \Psi_g)(\Gamma + \Delta + \Psi_h)}}. \quad (3)$$

Another correlation coefficient is the tau-c coefficient, which is denoted either *Kendall's tau-c* (SPSS) or *Stuart's tau-c* (SYSTAT, SAS). The formula is following:

$$\tau_c = \frac{2q(\Gamma - \Delta)}{n^2(q-1)} \quad (4)$$

where q is the minimum of numbers of categories of the variables $\mathbf{X}_g$ and $\mathbf{X}_h$.

Furthermore, *Somers' d* is used. Both symmetric and asymmetric types of this measure exist. The asymmetric one is expressed as

$$d_{\mathbf{x}_h|\mathbf{x}_g} = \frac{\Gamma - \Delta}{\Gamma + \Delta + \Psi_h}. \quad (\mathbf{5})$$

The symmetric measure is calculated as a harmonic mean of both asymmetric measures, i.e., the final formula is

$$d_{\text{sym}} = \frac{2 \cdot (\Gamma - \Delta)}{2 \cdot (\Gamma + \Delta) + \Psi_g + \Psi_h}. \quad (\mathbf{6})$$

Features of measures mentioned in this section are described in [2].

3 Methods for Searching Groups of Similar Variables

There are several types of techniques which are assigned to dimensionality reduction methods. Some of them are based of the projection of a high-dimensional space into a low-dimensional space. Usually, an object characterized by the vector of values of variables is plotted as a point in two-dimensional space where two of the found dimensions are used. Similarly, values of components or new dimensions can be calculated for variables which can then be plotted in two-dimensional space by means of a dot graph. If groups of variables exist, it can be seen in this graph. Another way how to graphically show relationships between variables is a dendrogram as the result of hierarchical cluster analysis.

We applied a *hierarchical agglomerative algorithm* which starts with each variable in a group of its own; then it merges clusters until only one large cluster remains, see [5]. It is based on the proximity matrix which involves dissimilarities for all pairs of variables. These dissimilarities can be derived on the basis of similarities. Once several variables have been linked together, we need a linkage or amalgamation rule to determine when two clusters are sufficiently similar to be linked together. The *complete linkage method* and *single linkage method* are used for the further analysis. In the former, the dissimilarity between two clusters is determined by the greatest dissimilarity between two variables from these clusters. In the latter, the dissimilarity between two clusters is determined by the lowest dissimilarity between two variables from these clusters.

Multidimensional scaling is also based on the proximity matrix. Non-metric multidimensional scaling is used for the further analysis. It both finds a non-parametric monotonic relationship between the dissimilarities in the proximity matrix and the Euclidean distance between variables, and the location of each variable in the low-dimensional space. The user must pre-specify number of dimensions.

Optimal scaling quantifies categorical variables, resulting in optimal principal components for the transformed variables. The variables can be given mixed optimal scaling levels and no distributional assumptions about the variables are made. Optimal scaling can easily deal with nonlinear relationships between the variables to be analyzed.

4 Applications to Real Data Files

In this section, the analysis of two data files will be described. The first data file is from the research "Males and females with university diploma", No. 0136, Institute of Sociology of the Academy of Sciences of the Czech Republic. The author of this research is the Gender in Sociology Team; the data collection was performed by Sofres-Factum (Praha, 1998). This file contains answers from 1 908 respondents.

The second file is the result of the research "Active lifestyle of university students" which was realized at the Faculty of Electrical Engineering of the Czech Technical University in Prague by dr. Z. Valjent from the Institute of Physical Education and Sports of this university in 2008, see [6], [7]. This file contains answers from 1 453 respondents.

4.1 Males and Females with University Diploma

For the purpose of investigation of relationships between variables, 13 variables expressing a satisfaction concerning a respondent's job from different points of view were analyzed. Respondents evaluated their satisfaction on the scale from 1 (very satisfied) to 4 (very dissatisfied). The similarity matrix based on Kendall's tau-*b* was created in the SPSS system. This matrix was transformed to the dissimilarity matrix by subtraction of the values from 1 in Microsoft Excel. The transformed matrix was analyzed by the complete linkage method of hierarchical cluster analysis in the STATISTICA system (for the reason of better quality of graphs). The resulting dendrogram is shown in Fig. 1.

If we do a cut in the distance 0.85 in the dendrogram, we obtain two larger clusters; with cutting in the distance 0.6 in the dendrogram we obtain six smaller clusters. The first lager cluster represents two groups: 1) satisfaction with salary, remuneration and evaluation of working enforcement; 2) satisfaction with perspective of the company and possibility of promotion. The further smaller clusters represent the following groups of variables: satisfaction with relationships in the company and relationships between males and females, satisfaction with scope of employment and use of respondent degree of education, satisfaction with management of the company, respondent supervisor and possibility to express own opinion, and the alone variable expressing a satisfaction with working burden.

Then the transformed matrix described above was analyzed by non-metric multidimensional scaling (MDS) in the STATISTICA system. The resulting dot plot for the 1st and 2nd dimensions is in Fig. 2. We can see that the relations between variables differ a little. While variables *p1a* and *p2b* (satisfaction with salary and remuneration) are the most similar according to cluster analysis, with the use of MDS the distance between variables *p1f* and *p2g* (satisfaction with scope of employment and use of respondent degree of education) are the smallest. MDS can provide different views on the data. The variables can be expressed by more than two dimensions and for each combination of them a special view on variables is provided.

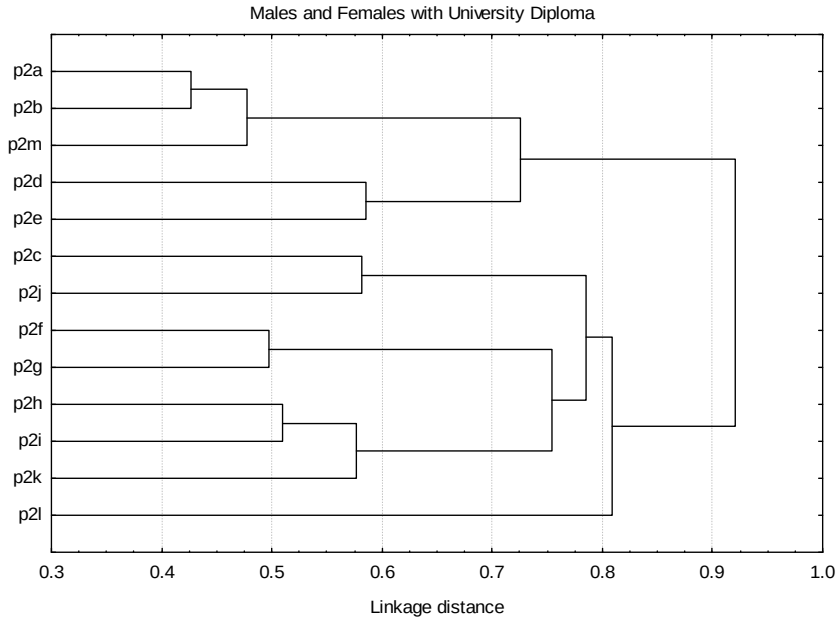


Fig. 1. Dendrogram of relationships between variables based on complete linkage.

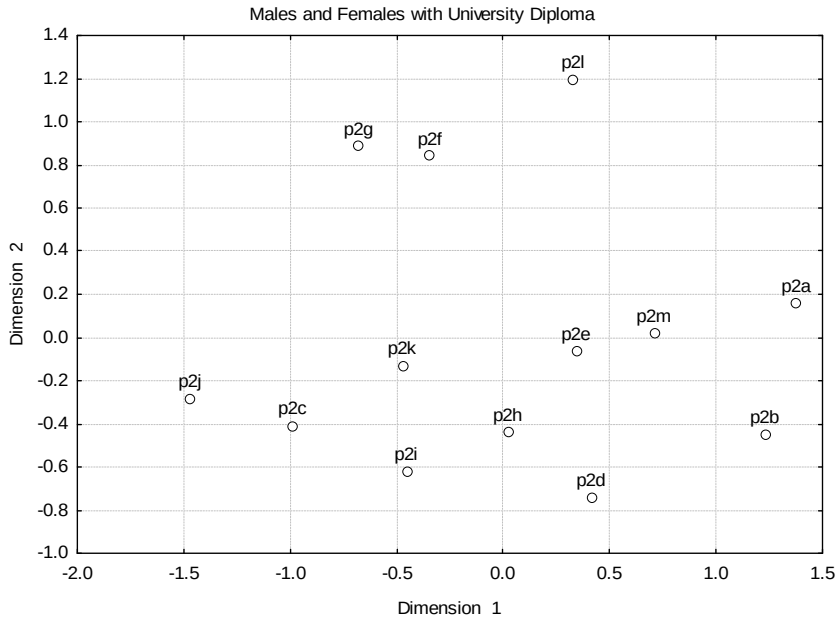


Fig. 2. Dot plot of relationships between variables obtained by multidimensional scaling.

In the end, we applied the CATPCA (principal component analysis for the categorical data) procedure as the method of optimal scaling in the SPSS system. We found some values of component loadings greater than 0.6 in three dimensions. In the resulting graph for the first two dimensions, we can distinguish three pairs and one triplet of variables which are very close. We considered that each variable is an object characterized by three values of component loadings (corresponding to three dimensions). Then the relationships between variables can be plotted by a dendrogram. With the use of the single linkage method and Euclidean distance we obtained variables *p1f* and *p2g* well separated, see Fig. 3, similarly as in Fig. 2.

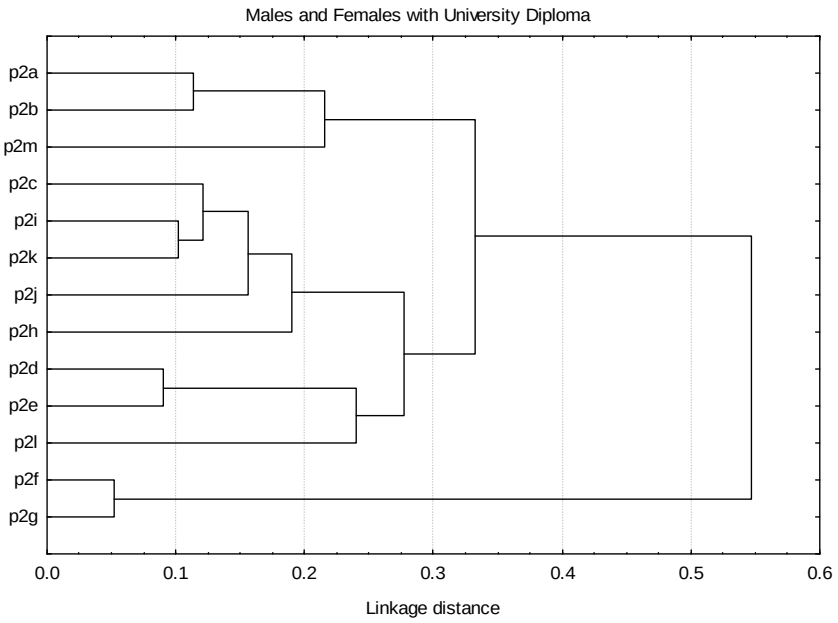


Fig. 3. Dendrogram of relationships between variables based on optimal scaling.

4.2 Influence of University Environment on Active Life Style of University Students

For the purpose of investigation of relationships between variables, 15 variables expressing a satisfaction concerning different points of view of the students' life were analyzed. Respondents evaluated their satisfaction on the scale from 1 (no satisfaction) to 7 (very satisfied). The similarity matrix based on Kendall's tau-*b* was created in the SPSS system. This matrix was transformed in the same way as was described in Section 4.1. The transformed matrix was analyzed by the complete linkage method of hierarchical cluster analysis in the STATISTICA system. The resulting dendrogram is shown in Fig. 4.

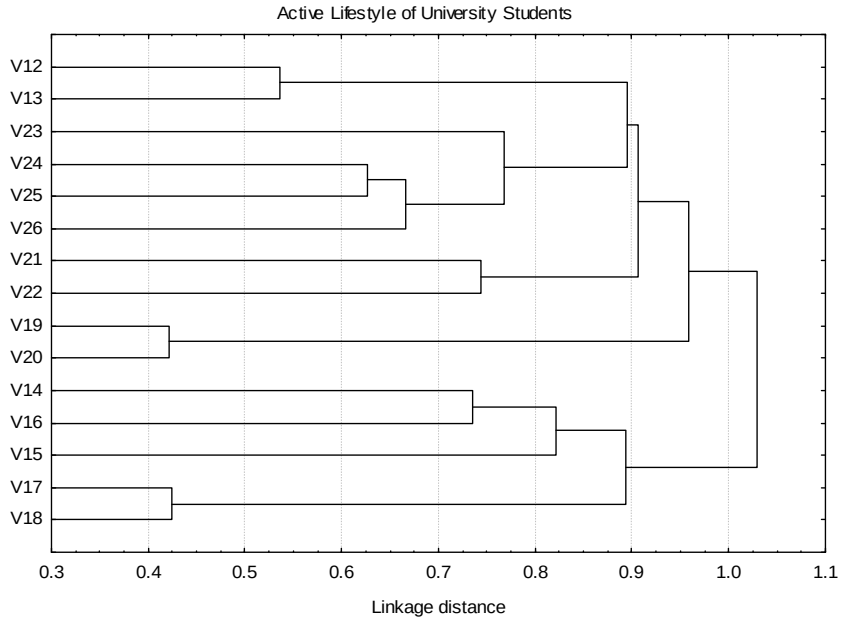


Fig. 4. Dendrogram of relationships between variables based on complete linkage.

If we do a cut in the distance 1 in the dendrogram, we obtain two larger clusters; with cutting in the distance 0.85 in the dendrogram we obtain six smaller clusters. The most similar answers in a pair are satisfaction with study generally and studying results (V19 and V20), and satisfaction with partner in life and sex life (V17 and V18). With the latter groups the following answers are merged into the second large cluster: satisfaction with success and acknowledgments from others (V14), with friends (V16) and family life (V15). The first large cluster further represents satisfaction with visage and body weight (V12 and V13), with leisure time, health state, fitness, and total quality of life (V23 to V26), with financial situation and housing (V21 and V22).

As in the previous example, the transformed matrix was analyzed by non-metric multidimensional scaling in the STATISTICA system. The resulting dot plot for the 1st and 2nd dimensions is in Fig. 5. In this case we can see that the pair formed by variables V17 and V18 (satisfaction with partner in life and sex life) is distinguished as well as in cluster analysis. With vertical splitting the graph according to the value –0.2 we obtain two groups of variables corresponding to two clusters in Fig. 4.

In the end, we applied again the CATPCA procedure in the SPSS system. We found some values of component loadings greater than 0.6 in three dimensions. In the resulting graph for the first two dimensions, the pair formed by variables V17 and V18 were separated distinctively from other variables. With the application of the single linkage method to the component loadings for three dimensions, we obtain again variables V17 and V18 well separated in a dendrogram, see Fig. 6.

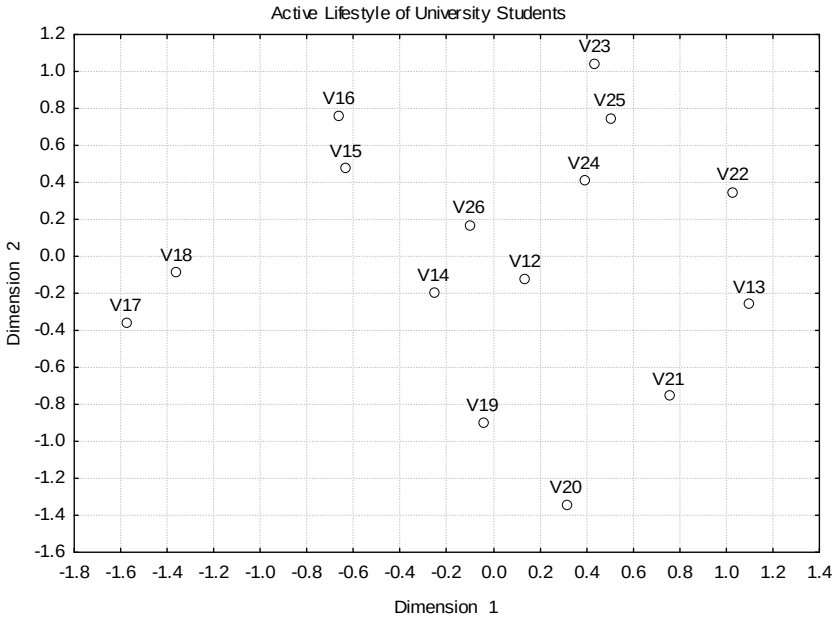


Fig. 5. Dot plot of relationships between variables obtained by multidimensional scaling.

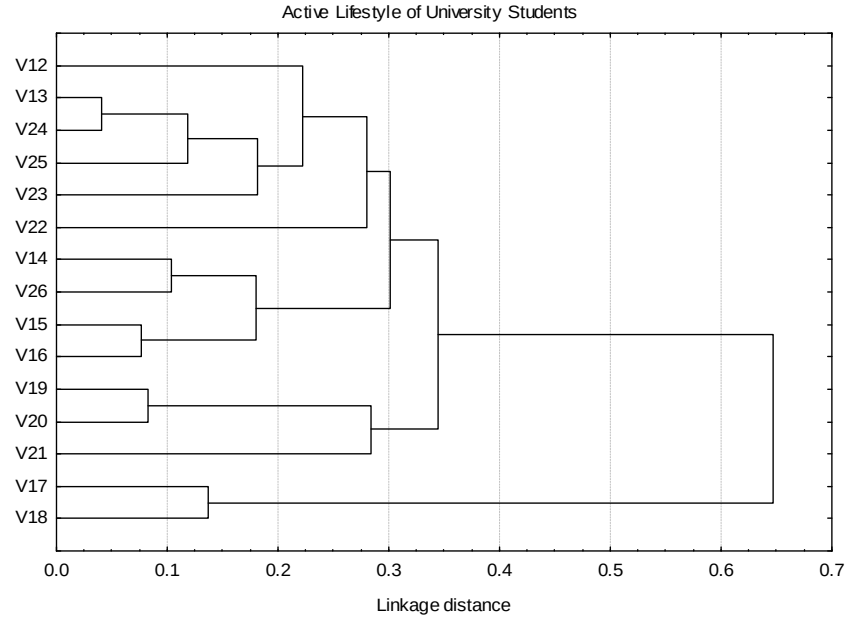


Fig. 6. Dendrogram of relationships between variables based on optimal scaling.

Moreover, in this case we obtained a distinct view on the relationships between variables. The most similar variables are V13 and V24 (body weight and health state) which are merge with V25 (fitness), V23 (leisure time) V12 (visage) and V22 (housing) into one from four clusters. The very close variables V15 and V16 (family life and friends) are merger into the second clusters with the pair of variables V14 and V26 (success and acknowledgments from others and total quality of life). The third cluster is formed by the pair of variables V19 and V20 (study generally and studying results) and the variableV21 (financial situation).

5 Conclusion

We presented only selected problems concerning methods for the dimensionality reduction in the data file with ordinal variables. Although there is a lot of ways how to determine groups of similar variables which are ordinal, in some papers only traditional factor analysis is applied. For example in [1] the data file concerning the features of typical and ideal policeman was analyzed. In each of these two cases, 24 variables were investigated. Respondents evaluated the features on the scale from 1 (positive evaluation) to 7 (negative evaluation). Factor analysis was used in this case.

The important task in the process of searching groups of similar variables is a determination of the number of these groups (clusters or dimensions). In the case of cluster analysis, several coefficients for cluster number determination exist, see [5]. The problem is that the user can obtain different results by different coefficients. The searching disjunctive groups of similar variables can be difficult in cases when groups of variables are overlapping. In this case, fuzzy cluster analysis could be used.

In our further research we plan to propose a technique for combination of results obtained by different methods with the aim to describe the data structure in an aggregate form. However, in this process the interpretation of groups of variables as underlying factors should be considered. For this reason the combination of results can be only a helpful tool for the finally description of the structure presented by an analyst.

Acknowledgments. This work was partially supported by project MSM6138439910.

References

1. Moulisová, M.: Výzkum percepce policisty. Kriminalistika, 42(1), 56--71 (2009)
2. Řezanková, H.: Analýza dat z dotazníkových šetření. Professional Publishing, Praha (2007)
3. Řezanková, H.: Cluster Analysis and Categorical Data. Statistika, 89(3), 216--232 (2009)
4. Řezanková, H., Húsek, D.: Comparison of the SAS, SPSS and STATISTICA Systems in the Area of Clustering Variables. Computational Statistics & Data Analysis, 41(2) 331--339 (2002)
5. Řezanková, H., Húsek, D., Snášel, V.: Shluková analýza dat. 2nd. ed. Professional Publishing, Praha (2009)
6. Valjent, Z.: Aktivní životní styl vysokoškoláků. Doctoral thesis, FTVS UK, before defense
7. Valjent, Z., Flemr, L.: Kvalita života studentů technické university. Acta (2009), in press

New Developments in Fuzzy Cluster Analysis

Hana Řezanková[1] and Dušan Húsek[2]

[1] University of Economics, Prague, Dept. of Statistics and Probability,
nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
hana.rezankova@vse.cz
[2] Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod vodárenskou věží 2, 182 07 Praha 8, Czech Republic
dusan.husek@cs.cas.cz

Abstract. The paper deals with a special class of cluster analysis methods where a membership degree is calculated for each object and each cluster. These methods are investigated under the name fuzzy cluster analysis. We present some emerging topics in this area, such as relation fuzzy clustering, soft clusters ensembles, similarity of fuzzy clusters, visualization of clustering results, simultaneous clustering and feature discrimination, and techniques for cluster number determination. Some tasks are illustrated by clustering of binary variables.

Keywords: Fuzzy cluster analysis, ensembles of fuzzy clustering, relationships between clusters and variables, cluster number determination.

1 Introduction

Cluster analysis has become a widely accepted synonym for a broad array of activities of exploratory data analysis and model development in science, engineering, life sciences, business and economics. Clustering (or segmentation) is the process that groups similar objects together and forms clusters. It is an important tool of exploratory data analysis. It is applied in areas such as data mining, text mining, image analysis, and pattern recognition. Clusters are groups of objects homogeneous within and desirably heterogeneous in between. The rationale of intra-group homogeneity is that objects with similar attributes are likely to respond in a similar manner to a given action. Identification of groups of people with similar opinions or customers with similar demands applications can be mentioned as an example. In the above-mentioned cases the overlapping groups can exist. Fuzzy cluster analysis is usually applied to the assignment of such objects to the groups As result, we obtain not only information about object membership but also even more membership degrees for each object and each cluster. In such a way the object can be assigned to more than one cluster with a given level of assignment.

For clustering, each object is represented by a set of features. It means that the variables involved in the clustering process are defined in advance. There can be different types of these variables. In the following text, we will consider only

quantitative and binary ones. For example, if an object is characterized by two quantitative variables, it can be represented as a point in the two-dimensional space, if it is characterized by two binary variables, it can be represented as vertex of a unit square.

The basic concept in cluster analysis is similarity. It is investigated for two objects or two clusters. However, algorithms of cluster analysis are often based on distances expressing space relationships of objects or clusters. Distance and similarity measures are widely used not only in clustering but also in machine learning and other fields. An attempt to formalize similarity measure and relation between similarity and distance is given in [4]. Let $\mathbf{x}_i$ be a vector of feature values, which characterizes the *i*th object. Then the distance between the *i*th and *j*th objects can be calculated as Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ for example (in the following text we will consider an object and a representing vector as synonyms), i.e.

$$d_{\mathrm{E}}(\mathbf{x}_i,\mathbf{x}_j) = d_{ij} = \sqrt{\sum_{l=1}^{m}(x_{il} - x_{jl})^2} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\| \qquad (\mathbf{1})$$

here *m* is a number of variables.

We suppose that the data set consisting of *n* objects, i.e. $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ should be partitioned into *k* clusters $C_1, \ldots, C_k$. In some algorithms, the representative of each cluster is determined. This can be either one object from a cluster (so called *medoid*) or a new vector characterizing the center of a cluster (*centroid*). In the latter case, the *centroid* is usually created by average values of individual variables. Further on, the centroid of the *h*th cluster will be denoted as $\overline{\mathbf{x}}_h$. Then the distance between the *i*th object and the corresponding centroid can be expressed as

$$d_{\mathrm{E}}(\mathbf{x}_i,\overline{\mathbf{x}}_h) = d_{ih} = \left\|\mathbf{x}_i - \overline{\mathbf{x}}_h\right\|. \qquad (\mathbf{2})$$

In the following, the individual algorithms applications are illustrated using the real data set. This comes from the 1984 United States Congressional Voting Records Database (http://mlearn.ics.uci.edu/databases/voting-records). It includes 16 variables (voting on selected topics) and 435 cases but only 230 cases without missing values were included in the analysis. Variables are binary with the value 1 in case of the answer "yes" and value 0 in case of the answer "no". In this task, the clustering of variables instead of the clustering of cases is considered. This means that we look for groups of topics with similar answers.

2 The Basic Algorithms and Visualization of Clustering Results

Fuzzy clustering has been studied very intensively in the past decades. A lot of papers have been published in journals, conference proceedings and in some monographs, e.g. [2] and [7]. There are many different algorithms used for fuzzy (soft) cluster

analysis. Fuzzy *k*-means is one of them, see e.g. [11]. It is based on a generalization of the classical (hard) *k-means* (also HCM – *hard c-means*) algorithm which analyzes the data set of n object, i.e. $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, with the aim to minimize the objective function

$$J_{\mathrm{HCM}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih} d_{ih}^2 \tag{3}$$

where k is a number of clusters, the elements $u_{ih} \in \{0, 1\}$ indicate the assignment of data to clusters (1 means the assignment) and d_{ih} is the distance between the j-th object and the center of the h-th cluster. Further, the following conditions have to be satisfied: $\sum_{h=1}^{k} u_{ih} = 1$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} u_{ih} > 0$ for $h = 1, \ldots, k$.

The *fuzzy k-means* (frequently FCM – *fuzzy c-means*) algorithm minimizes the objective function

$$J_{\mathrm{FCM}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih}^q d_{ih}^2 \tag{4}$$

where elements $u_{ih} \in \langle 0, 1 \rangle$ are membership degrees, and the parameter q ($q > 1$) is called the fuzzifier or weighting exponent (usually $q = 2$ is chosen). Furthermore, the same conditions as in the previous case have to be satisfied:

$\sum_{h=1}^{k} u_{ih} = 1$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} u_{ih} > 0$ for $h = 1, \ldots, k$.

Another technique is based on the hard *k-medoids* (HCMdd) algorithm. The function

$$J_{\mathrm{HCMdd}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih} \left\| \mathbf{x}_i - \mathbf{m}_h \right\| \tag{5}$$

is minimized here under the same assumption as for the function (3) where $\mathbf{m}_h$ is a medoid in the hth cluster.

The *fuzzy k-medoids* (FCMdd) algorithm minimizes the objective function

$$J_{\mathrm{FCMdd}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih}^q \left\| \mathbf{x}_i - \mathbf{m}_h \right\|^2 \tag{6}$$

under the same assumption as for the function (4).

In some cases, only the proximity matrix is known, instead of the original vectors of feature values. In this matrix, each pair of objects is numerically described by a real-valued relation. We distinguish between similarity and dissimilarity relations. An example of the former is a number of common elements of objects in the pairs. Cluster analysis based on the proximity matrix is sometimes called *relation clustering*, see [14]. For example, two basic algorithms of relation fuzzy clustering can be used. They can be also applied to a set of original vectors of feature values from which the proximity matrix is calculated.

The first algorithm is *relation fuzzy k-means* (RFCM – *relation fuzzy c-means*). For $q = 2$ it is called FANNY, see [10], and it is implemented in the statistical software system S-PLUS. In this algorithm, the objective function

$$J_{\text{FANNY}} = \sum_{h=1}^{k} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ih}^2 u_{jh}^2 d_{ij}}{2\sum_{j=1}^{n} u_{ih}^2} \tag{7}$$

is minimized. In this function, d_{ij} is the known distance between the ith and jth objects, and u_{ih} and u_{jh} are unknown. It is a special case of the function (4) when $\bar{\mathbf{x}}_h = \sum_{i=1}^{n} u_{ih}^2 \mathbf{x}_i \Big/ \sum_{i=1}^{n} u_{ih}^2$ see [15] p. 232.

The second algorithm is *relational fuzzy k-medoids* (RFCMdd – *relation fuzzy c-medoids*). It is based on the fuzzy *k*-medoids algorithm, see above.

For a membership representation, traditional graphs for disjunctive clustering can be used, and also some special graphic techniques, see [16]. An example of a traditional graph in the S-PLUS system is a *silhouette plot*. The width and the direction of the rectangles for the ith object is determined by the value

$$\psi_i = \frac{\eta_i - \mu_i}{\max\{\eta_i, \mu_i\}} \text{ where } \eta_i = \sum_{j \in C_g} d_{ij} \Big/ (n_g - 1),\ \mu_i = \min_{h \neq g}\left(\sum_{j \in C_h} d_{ij} \Big/ n_h \right), \tag{8}$$

and n_g is a number of objects in the gth cluster. It is supposed that the object is assigned only to one cluster (C_g) according to the highest value of a membership degree. It means that $u_{ig} = \max_h(u_{ih})$.

2.1 Example – Voting Records Database

In this example, we applied the FANNY algorithm in the S-PLUS system to the 1984 United States Congressional Voting Records Database. We clustered variables (voting on selected topics) into three clusters. First, we applied fuzzy cluster analysis to the original data sets directly (for the transposed data matrix), and we obtained the

membership degrees less than 0.6 for all the variables and all the clusters. For this reason, we applied factor analysis first and then we analyzed the first two vectors of factor loadings (the further vectors of factor loadings explain only a small amount of variability), see [9]. In Table 1, there are the values of membership degrees both for original data set and for factor loadings (the values greater than 0.48 are highlighted).

Table 1. Membership degrees for three clusters of variables obtained by algorithm FANNY (for original data set and for factor loadings)

Variable	Original data set			Factor loadings		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
1	0.389	0.389	0.222	**0.528**	0.366	0.106
2	0.343	0.343	0.314	0.127	**0.730**	0.143
3	0.411	0.411	0.178	**0.848**	0.107	0.045
4	0.224	0.224	**0.552**	0.044	0.064	**0.892**
5	0.201	0.201	**0.598**	0.049	0.077	**0.874**
6	0.258	0.258	**0.484**	0.072	0.127	**0.801**
7	0.408	0.408	0.184	**0.835**	0.109	0.056
8	0.418	0.418	0.164	**0.857**	0.097	0.046
9	0.414	0.414	0.172	**0.870**	0.087	0.043
10	0.358	0.358	0.284	*0.382*	*0.322*	*0.296*
11	0.372	0.372	0.258	0.100	**0.837**	0.063
12	0.226	0.226	**0.548**	0.031	0.047	**0.922**
13	0.247	0.247	**0.506**	0.065	0.112	**0.823**
14	0.243	0.243	**0.514**	0.053	0.076	**0.871**
15	0.396	0.396	0.208	**0.789**	0.152	0.059
16	0.382	0.382	0.236	**0.732**	0.182	0.086

A *silhouette graph* is part of the procedure for fuzzy cluster analysis in S-PLUS. Although it has been proposed for disjunctive clustering, it reflects membership degrees relatively well. In the third cluster, the order of rectangle widths corresponds with the order of membership degree values in the individual clusters, see Fig. 1.

3 Ensembles of Fuzzy Clustering and Similarity of Fuzzy Clusters

Sometimes, the user has at his disposal results of clustering (assignments of individual objects to the certain number of clusters) obtained by different ways, and he has no access to the original features of the objects. For example, in marketing research customers are segmented in multiple ways based on different criteria (need-based, demographic, etc.). The user can be interested in obtaining a single, unified segmentation.

Combining clustering is more difficult than combining the results of multiple classifiers. Before combining the clustering, one has to identify which clusters from different clusterings correspond to each other. Moreover, the number of clusters in individual solutions might vary, see [12].

For results of hard clustering, graph-theoretic approaches have been proposed in the literature. They are based on the hypergraph representation of clustering, see

Table 2. In this table, $C_h^{(q)}$ denotes the hth cluster in the qth clustering and $u_{ih}^{(q)}$ represents the membership degree of the ith object to the hth cluster in the qth clustering. In case of hard clustering $u_{ih}^{(q)} \in \{0, 1\}$.

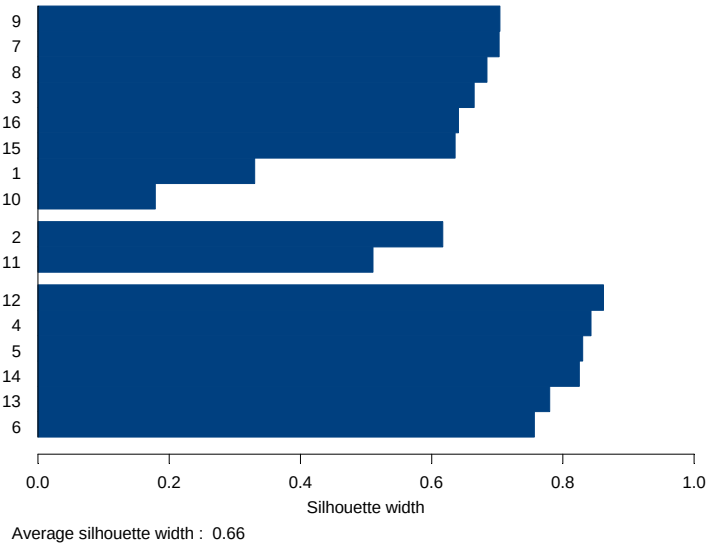


Fig. 1. Silhouette plot for three clusters of variables (algorithm FANNY for factor loadings)

Table 2. Hypergraph representation of clustering.

cluster / object	$C_1^{(1)}$	$C_2^{(1)}$	…	$C_{k^{(1)}}^{(1)}$	…	$C_1^{(r)}$	$C_2^{(r)}$	…	$C_{k^{(r)}}^{(r)}$
$\mathbf{x}_1$	$u_{11}^{(1)}$	$u_{12}^{(1)}$	…	$u_{1k^{(1)}}^{(1)}$	…	$u_{11}^{(r)}$	$u_{12}^{(r)}$	…	$u_{1k^{(r)}}^{(r)}$
$\mathbf{x}_2$	$u_{21}^{(1)}$	$u_{12}^{(1)}$	…	$u_{1k^{(1)}}^{(1)}$	…	$u_{21}^{(r)}$	$u_{12}^{(r)}$	…	$u_{1k^{(r)}}^{(r)}$
…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	$u_{n1}^{(1)}$	$u_{n2}^{(1)}$	…	$u_{1k^{(1)}}^{(1)}$	…	$u_{n1}^{(r)}$	$u_{n2}^{(r)}$	…	$u_{1k^{(r)}}^{(r)}$

Cluster-based similarity partitioning algorithm (CSPA) is an example of the graph-theoretic approach. In the CSPA technique, a similarity matrix is computed as $\mathbf{W} = (1/r)\mathbf{U}\mathbf{U}^{\mathrm{T}}$, where r is a number of clusterings. Then a clustering algorithm based on a proximity matrix can be used.

Another technique is *hypergraph partitioning algorithm* (HGPA), which seeks directly to partition the hypergraph defined in Table 2 by eliminating the minimal number of hyperedges.

The third technique is *meta-clustering algorithm* (MCLA). First, it seeks to solve the cluster correspondence problem and then uses voting to place objects into final consensus clusters. In MCLA, the similarity of clusters C_h and C_g is computed, based on the number of objects that are clustered into both of them. When using Jaccard measure,the similarity is computed according to the formula (9)

$$sim(C_h, C_g) = \frac{\left|C_h \cap C_g\right|}{\left|C_h \cup C_g\right|} \tag{9}$$

where the numerator expresses the number of objects which belong to both clusters and the denominator expresses the total number of objects in both clusters together. The similarity matrix for clusters is partitioned into *meta-clusters*. Then, each object is assigned to the meta-cluster it is most associated to.

Finally, we have to mention *hybrid bipartite graph formulation* (HBGF). This method models the objects and clusters simultaneously in a graph. The CSPA algorithm models the ensemble as a graph with the vertices representing objects, while the MCLA algorithm models the ensemble as a graph of clusters. The HBGF technique combines these two ideas and represents the ensemble by a bipartite graph in which the individual data points and the clusters of the constituent clustering are both vertices. This graph is partitioned into *k* parts. The method yields a co-clustering solution.

For the results of fuzzy clustering, several approaches have been proposed in the literature. One of them is solving soft ensembles with *information-theoretic k-means* (ITK). The ITK algorithm is very similar to the *k*-means algorithm, differing only in the fact that it uses the KL-divergence (Kullback-Leiber) instead of Euclidean distance as a distance measure. Each object in a soft ensemble is represented by a concatenation of *r* posterior membership probability distributions obtained from the clustering algorithms. In Table 1, $u_{ih}^{(q)}$ represents the membership degree of the *i*th object to the *h*th cluster in the *q*th clustering. In case of fuzzy clustering $u_{ih}^{(q)} \in \langle 0, 1 \rangle$. The distance measure between two objects is defined using KL-divergence, which calculates the "distance" between two probability distributions. For objects $\mathbf{x}_i$ and $\mathbf{x}_j$ it can be calculated as

$$d_{\mathrm{KL}}(\mathbf{x}_i, \mathbf{x}_j) = -\sum_{q=1}^{r} w^{(q)} \sum_{h=1}^{k^{(q)}} u_{ih}^{(q)} \ln \frac{u_{ih}^{(q)}}{u_{jh}^{(q)}} \tag{10}$$

where $w^{(q)}$ are clustering specific weights, such that $\sum_{q=1}^{r} w^{(q)} = 1$. On the basis of the matrix $\mathbf{U}$ and KL-divergence the objects can be clustered into *k* clusters.

Another technique is based on CSPA. It is a *soft version of CSPA* (sCSPA). One can use either the $\mathbf{U}\mathbf{U}^{\mathrm{T}}$ matrix or the similarity matrix created on the basis of Euclidean distance. In the latter, the distance between objects $\mathbf{x}_i$ and $\mathbf{x}_j$ is calculated as

$$d_{\mathrm{E}}(\mathbf{x}_i,\mathbf{x}_j)=\sqrt{\sum_{q=1}^{r}\sum_{h=1}^{k^{(q)}}\left(u_{ih}^{(q)}-u_{jh}^{(q)}\right)^2}\ . \tag{11}$$

On the basis of the proximity matrix the objects can be clustered into *k* clusters.

Further techniques are a *soft version of MCLA* (sMCLA) and a *soft version of HBGF* (sHBGF). In the former, Euclidean distance can be used as a measure of the difference of two clusters.

In the previous text, Euclidean distance was suggested to be used as a *measure of the difference of two fuzzy clusters*. However, some other approaches have been proposed for aggregation of fuzzy clusters, see [1]. The expression (9) is a base. Each pair of clusters can be compared to each other in the following ways:

$$sim(C_h,C_g)=\frac{\sum_{i=1}^{n}\min\{u_{ih},u_{ig}\}}{\sum_{i=1}^{n}\max\{u_{ih},u_{ig}\}} \tag{12}$$

or

$$sim(C_h,C_g)=\frac{\sum_{i=1}^{n}u_{ih}u_{ig}}{\sum_{i=1}^{n}(u_{ih}+u_{ig}-u_{ih}u_{ig})}\ . \tag{13}$$

3.1 Example – Voting Records Database

We combined results of two fuzzy clusterings – partitioning of original data set and partitioning of factor loadings, see Table 1. It is easy to create both the $\mathbf{UU}^{\mathrm{T}}$ matrix and the similarity matrix created on the basis of Euclidean distance. In both cases, we applied FANNY and PAM (Partitioning Around Medoids) algorithms. In all cases, the 10th variable was added to the clusters formed of the 2nd and 11th variables when three clusters were required. However, it is evident that assignment of this variable is problematic, as we can see in Fig. 2 (a negative value for the rectangle width). It corresponds to the low values of membership degrees in Table 1. A comparison with some other graphic representations (factor analysis, multidimensional scaling) is described in [13].

Further, we applied the FANNY algorithm to create 7 clusters. Then, we calculated a proximity matrix for clusters according to the formula (12), and applied hierarchical cluster analysis (complete linkage). The resultant dendrogram obtained by the STATISTICA system is in Fig. 3. The joined clusters 1 and 3 contain higher values of membership degrees (from 0.45 to 0.83 in some of these two clusters) for variables 1, 3, 7, 8, 9, 15, and 16, which correspond to the first cluster in Fig. 2. The cluster 2 (contains variables 2 and 11) is merged with the cluster 7 (contains only the 10th variable), what corresponds to the second cluster in Fig. 2. The joined clusters 4, 5, and 6 contain variables corresponding to the third cluster in Fig. 2. This particular technique can serve for cluster number determination, see Section 5.

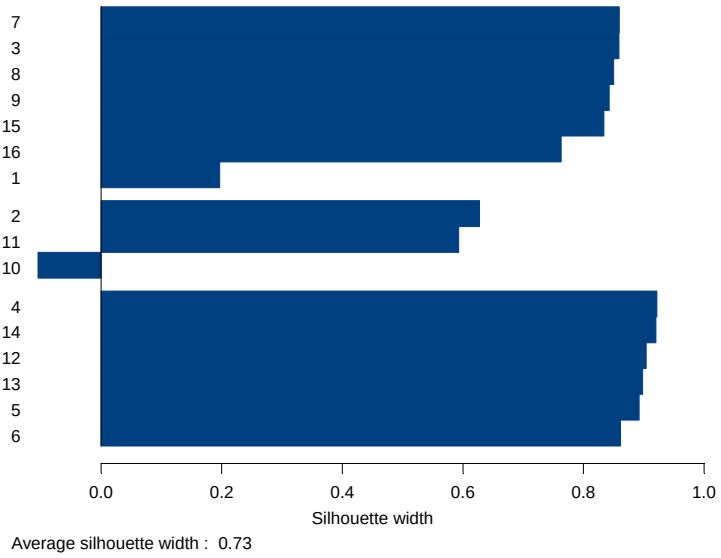


Fig. 2. Silhouette plot for three clusters of variables (FANNY algorithm for the combination of clustering results with Euclidean distance for relationships between variables)

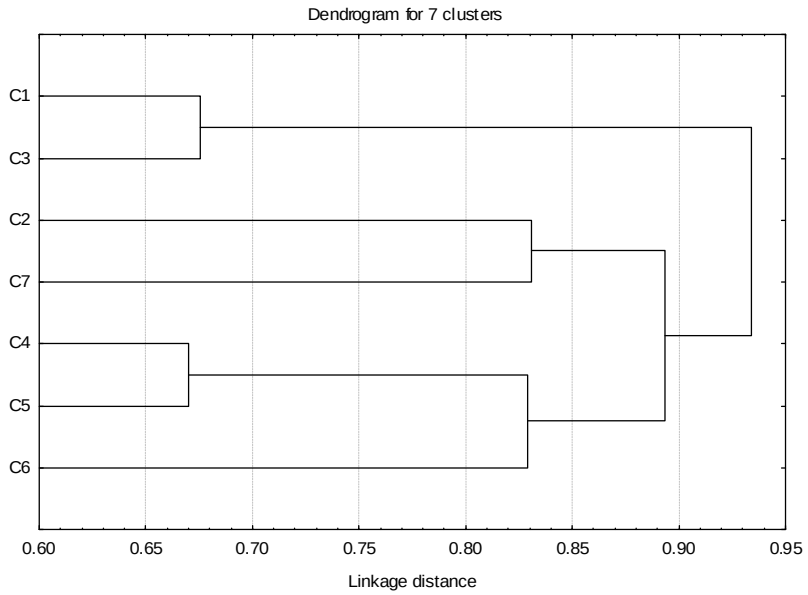


Fig. 3. Dendrogram for clustering of fuzzy clusters (complete linkage)

4 Relationships between Clusters and Variables

The problem of selecting or weighting the best subset of features is an important part of the design of algorithms for real world tasks of data analysis. In [5], two versions of simultaneous clustering and attribute discrimination (SCAD) are proposed. The SCAD-1 algorithm minimizes the following objective function:

$$J_{\text{SCAD-1}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih}^{q} \sum_{l=1}^{m} v_{hl} d_{ihl}^{2} + \sum_{h=1}^{k} \delta_h \sum_{l=1}^{m} v_{hl}^{2} \qquad (\mathbf{14})$$

where $v_{hl} \in \langle 0, 1 \rangle$ are feature weights (v_{hl} represents the weight of the lth variable in hth cluster), $\sum_{l=1}^{m} v_{hl} = 1$ for $h = 1, \ldots, k$, and d_{ihl} is given by $d_{ihl} = x_{il} - \bar{x}_{hl}$.

The SCAD-2 algorithm omits the second term from its objective function by incorporating a discriminant exponent r, and minimizing

$$J_{\text{SCAD-2}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih}^{q} \sum_{l=1}^{m} v_{hl}^{r} d_{ihl}^{2} \,. \qquad (\mathbf{15})$$

Further, the author considers the subset feature weighting. A set of features is partitioned into several subsets, and weights for these feature subsets are calculated.

Another way of expressing the relationship between the clusters and the variables is *fuzzy cluster loading model*, see [15]. It is close to the problem of fuzzy clustering for the weights determination in the weighted regression model. It has been proposed in order to obtain the interpretation of the result fuzzy clusters. The model is defined as

$$u_{ih} = \sum_{l=1}^{m} x_{il} z_{lh} + \varepsilon_{ih} \,, \; i = 1, \ldots, n, \; h = 1, \ldots, k \qquad (\mathbf{16})$$

where z_{lh} is called *fuzzy cluster loading* and ε_{ih} is an error. A fuzzy cluster loading shows the fuzzy degree that represents the amount of loading of the hth cluster to the lth variable, i.e. it shows how each cluster can be explained by each variable.

4.1 Example – Voting Records Database

We solved the modified model (16) – linear regression model with the constant element. The input matrix **X** consisted of two vectors of factor loadings as variables. The **U** matrix contained membership degrees obtained by FANNY algorithm for three clusters. The result-estimated parameters are shown in Table 3.

Table 3. Estimates of the z_{lh} parameters of the modified model (16).

cluster parameter	C_1	C_2	C_3
z_{1h}	–0.473	–0.048	0.521
z_{2h}	–0.312	0.591	–0.278

The values of factor loadings are both negative and positive. The first cluster was created on the basis of negative values in the first variable (the absolute values were higher than 0.5); the values of the second variable were either negative or positive near to zero. The second cluster was created on the basis of high positive values in the second variable (higher than 0.5), and the third cluster was created on the basis of high positive values in the first variable (higher than 0.7).

5 Cluster Number Determination

Many various coefficients (indices) have been proposed in the literature for the cluster number determination and evaluation of a resulting clusters quality; see [2], [3], and [6]. The partition coefficient (PC) and partition entropy (PE) are the most popular. The former is also known as Dunn's coefficient, and it is implemented in the S-PLUS system, including the normalized form. It is calculated according to the formula

$$I_{\mathrm{PC}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{k} u_{ih}^2 \qquad (\mathbf{17})$$

with the values from $1/k$ to 1. We can distinguish two extreme situations:
completely fuzzy clustering: all $u_{ih} = 1/k \Rightarrow I_{\mathrm{PC}} = 1/k$,
hard (*disjunctive*) *clustering*:
for one u_{ih}: $u_{ih} = 1$, and for all others: $u_{ih} = 0 \Rightarrow I_{\mathrm{PC}} = 1$.

We can compare the values of this index for a different number of clusters (k) – the higher value is better. However, the upper limit of values of this coefficient is dependent on the number of clusters. For this reason, the normalized coefficient

$$I_{\mathrm{PC}}^{*} = \frac{I_{\mathrm{PC}} - 1/k}{1 - 1/k} = \frac{kI_{\mathrm{PC}} - 1}{k - 1} \qquad (\mathbf{18})$$

is used in the S-PLUS system. In this way, we obtain the values from the interval $\langle 0, 1 \rangle$, see [10].

Another index is based on the variability measurement by means of entropy. It is called *PE* (*Partition Entropy*) *index*. It can be expressed as

$$I_{\mathrm{PE}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{k} u_{ih}\ln(u_{ih}) \tag{19}$$

with values from interval ⟨0, ln*k*⟩. The lower value is better. However, the upper limit of values of this coefficient depends on the number of clusters. For this reason, the application of the normalized coefficient is a preferable solution.

In the following example, we used an analogy to the mutability measure proposed by C. Gini (Gini's index) for the suitable number of cluster identification in our application. Whereas the PC coefficient measures concentration in clusters, the proposed GC coefficient measures variability, i.e.

$$I_{\mathrm{GC}} = 1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{k} u_{ih}^2 \ . \tag{20}$$

The normalized form of this coefficient is then

$$I_{\mathrm{GC}}^{*} = \frac{kI_{\mathrm{GC}}}{k-1} \ . \tag{21}$$

Further, we used the PE index (19) and the normalized form of this coefficient in the form

$$I_{\mathrm{PE}}^{*} = \frac{I_{\mathrm{PE}}}{\ln k} \ . \tag{22}$$

In both case (21) and case (22), the minimum of these coefficients means the best partitioning.

For the reason that we obtain disjunctive assignment objects into clusters by the FANNY algorithm we can also use for final evaluation techniques for hard clustering. In the S-PLUS system, the average silhouette width is a part of the silhouette plot, see Fig. 1 and Fig. 2. It is an average of the silhouette widths, see (8), and it will be denoted by the symbol $\overline{\psi}$. The higher value represents better partitioning objects to clusters.

5.1 Example – Voting Records Database

We analyzed the data matrix presented in Table 1. The resulting graph for three clusters is shown in Fig. 2. We calculated coefficients (20), (21), (19), (22) for 2, 3, 4, and 5 clusters. The values of these coefficients are presented in Table 4 as well as average silhouette widths.

Table 4. The evaluation of variables partitioning into different number of clusters.

coefficient / number of clusters	I_{GC}	I^*_{GC}	I_{PE}	I^*_{PE}	$\overline{\psi}$
2	0.183	0.367	0.308	0.445	0.69
3	0.206	**0.309**	0.409	**0.373**	0.73
4	0.247	0.329	0.525	0.379	**0.78**
5	0.277	0.346	0.606	0.376	0.70

We can see that partitioning to three clusters was evaluated as optimal by both the normalized Gini index and the normalized PE index. With the use of an average silhouette width, the partitioning into four clusters is optimal (two small clusters were created, one with the 1st and 10th variables, and the second with the 2nd and 11th variables).

Another example of cluster number determination with the use of the coefficients mentioned above is in [8]. Further, we can start from the higher number of clusters and join these clusters into the smaller number of them, see Section 3.

6 Conclusion

Fuzzy cluster analysis is a very useful tool for revealing a data structure. In this paper, we presented one special task – fuzzy clustering of binary variables. Clustering of cases is mainly considered in the algorithms proposed in the literature. However, some algorithms are based on the proximity matrix (relational fuzzy clustering). In this case, also variables can be clustered without problems. Moreover, we suggest to apply factor analysis first, and then to cluster vectors of factor loadings.

Furthermore, we show that the silhouette plot proposed for visualization of the results of disjunctive clustering can also represent a result of fuzzy clustering very well.

A lot of coefficients (indices) have been proposed in the literature for cluster number determination. However, their values usually depend on the actual investigated number of clusters. We show that only normalized coefficients can give a suitable evaluation of individual partitioning.

Acknowledgments. This work was partially supported by projects AV0Z10300504, 205/09/1079, MSM6138439910, and 1M0567.

References

1. Abonyi, J., Feil, B.: Aggregation and Visualization on Fuzzy Clusters Based on Fuzzy Similarity Measures. In: Oliveira, J.V., Pedrycz, W. (eds.) Advances in Fuzzy Clustering and its Applications, pp. 93--121. John Wiley & Sons, Chichester (2007)
2. Abonyi, J., Feil, B.: Cluster Analysis for Data Mining and System Identification. Birkhäuser Verlag AG, Berlin (2007)

3. Bezdek, J.C., Pal, N.R.: Some New Indexes of Cluster Validity. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 28(3), 301--315 (1998)
4. Chen, S., Ma, B., Zhang, K.: On the similarity metric and the distance metric. In: Formal Languages and Applications: A Collection of Papers in Honor of Sheng Yu, Theoretical Computer Science, vol. 410(24-25), pp. 2365--2376 (2009) DOI: 10.1016/j.tcs.2009.02.023, http://www.sciencedirect.com/science/article/B6V1G-4VR9FHM-2/2/6aa0974302d6e3f1e4066171d1692be3
5. Frigui, H.: Simultaneous Clustering and Feature Discrimination with Applications. In: Oliveira, J.V., Pedrycz, W. (eds.) Advances in Fuzzy Clustering and its Applications, pp. 285--312. John Wiley & Sons, Chichester (2007)
6. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM, Philadelphia (2007)
7. Höppner, F., Klawon, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis. Methods for Classification, Data Analysis and Image Recognition. John Wiley & Sons, New York (2000)
8. Húsek, D., Řezanková, H., Dvorský, J. Social Group Identification and Clustering. In: International Conference on Computational Aspects of Social Networks, pp. 73--79. IEEE Computer Society, Danvers (2009)
9. Húsek, D., Řezanková, H., Frolov, A.A.: Overlapping Clustering of Binary Variables. In: Knowledge Extraction and Modelling [CD-ROM]. TILAPIA Edizioni, Italy (2006)
10. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, Hoboken (2005)
11. Kruse, R., Döring C., Lesot, M.-J.: Fundamentals of Fuzzy Clustering. In: Oliveira, J.V., Pedrycz, W. (eds.) Advances in Fuzzy Clustering and its Applications, pp. 3--30. John Wiley & Sons, Chichester (2007)
12. Punera, K., Ghosh, J.: Soft Cluster Ensembles. In: Oliveira, J.V., Pedrycz, W. (eds.) Advances in Fuzzy Clustering and its Applications, pp. 69--91. John Wiley & Sons, Chichester (2007)
13. Řezanková, H., Húsek, D., Snášel, V.: Clusters Number Determination and Statistical Software Packages. In: Database and Expert Systems Application, pp. 549--553. IEEE Computer Society, Turin (2008)
14. Runkler, T.A.: Relational Fuzzy Clustering. In: Oliveira, J.V., Pedrycz, W. (eds.) Advances in Fuzzy Clustering and its Applications, pp. 31--51. John Wiley & Sons, Chichester (2007)
15. Sato-Ilic, M.: Fuzzy Regression Clustering. In: Oliveira, J.V., Pedrycz, W. (eds.) Advances in Fuzzy Clustering and its Applications, pp. 229--246. John Wiley & Sons, Chichester (2007)
16. Wiswedel, B., Patterson, D.E., Berthold, M.R.: Interactive Exploration of Fuzzy Clusters. In: Oliveira, J.V., Pedrycz, W. (eds.) Advances in Fuzzy Clustering and its Applications, pp. 123--136. John Wiley & Sons, Chichester (2007)

Utility Function in an Insurance

Jana Špirková

Department of Quantitative Methods
and Information Systems,
Faculty of Economics,
Matej Bel University,
Tajovského 10,
975 90 Banská Bystrica, Slovakia
jana.spirkova@umb.sk

Abstract. Utility theory belongs to very interesting part of modern decision making theory. We develop the basic concept of utility theory on a determination of a gross annual premium in a non-life insurance. We introduce specific values of a gross annual premium on the basis person's utility function wich was founded empirically by short personally interviewing.

Keywords: Utility function. Expected utility. Insurance.

1 Introduction

A more modern approach of the utility theory was advanced by John von Neumann and Oskar Morgenstern in 1947 in their book Theory of Games and Economic Behavior [3]. There, they proposed that a utility function may be tailored for any individual, provided certain assumptions about the individual's preferences hold. These assumptions provide several valid, basic shapes for the utility function. In 2007 was published 60th-anniversary edition of this book [4]. Lapin in [2] describes and explains very interestingly employment of the utility function in decision making.
Modern approach to the utility function is also described in [5].
This paper was inspired by the book Modern Actuarial Risk Theory [1] and by my students, who want to know more information about a generation of person's utility function, which they know until now only on the theoretical level.
This paper is organized as follows. In Section 2 we recall basic properties of the utility functions and their applications in the insurance. In Section 3 we describe concrete person's utility function of our respondent, who responds in our short interview. Moreover, we calculate maximal gross premium for an insurance policy.
Finally, in Section 4, some conclusions and indications of our next investigation about mentioned topic are included.

2 The Utility Function

2.1 Basic Properties

The utility function may be used as the basis for describing an individulal's approaches to the risk. Three basic approaches have been characterized.
The opposite cases are the *risk averse*, who will accept only favorable gambles, and the *risk seeker* or by other words *risk loving*, who will pay a premium for the privilege of participing in a gamble. Between these two extremes lies the *risk-neutral*, who considers the face value of money to be its true worth.
Throughout most of their lives, people are typically risk averse. Only gambles with high expected payoff will be attractive to them. The risk averse's marginal utility diminishes as the rewards increase, so that the risk averse's utility function exhibits a decreasing positive slope as the level of monetary payoff becomes larger.
Such a function is concave, see Figure 1.

The risk loving's behavior is the opposite of the risk averse's behavior.
The risk loving will prefer some gambles with negative expected monetary payoffs.
The risk loving is typically self-insured, believing that the risk is superior to forgoing money spent on premiums.
The risk loving's marginal utility is increasing. Each additional euro provides a disproportionately greater sense of well-being. Thus, the slope of the risk loving's utility function increases as the monetary change improves. This function is convex what also viewed from Figure 2.

The utility function for a risk neutral's is a straight line. The utility is equal to the utility for the expected value. Risk-neutral individuals buy no casually insurance, since the premium charge is greater than the expected loss. Risk-neutral behavior is typical by persons who are enormously wealthy. Many people may be both risk averses and lovings, depending on the range of monetary values being considered.

2.2 Utility Function in an Insurance

The fundamental proposition of the modern approach of the utility is that it is possible to obtain a numerical expression for an individual's preferences. Because people usually have different approaches to the risk, two persons faced with an identical decision may actually prefer different courses of action.
In this section we will discuss the utility as an alternative expression of payoff that reflects a person's approaches.

Suppose that our respondent owns a capital w and that he values welth by the utility function u.
The next Theorem 1. describes properties of the utility function and its expected value [2], (see also Figure 2).
It can be written as follows.

Theorem 1. (Jensen's inequality)
If $u(x)$ is a convex function and X is a random variable, then

$$E[u(X)] \geq u(E[X]) \tag{1}$$

with equality if and only if $u(x)$ is linear on the support of X or $\text{var}(X) = 0$.

From the inequality (1) and also from Figure 1 it follows that for a concave utility function holds

$$E[u(w - X)] \leq u(E[w - X]) = u(w - E[X]). \tag{2}$$

In this case decision maker is rightly called *risk averse*. He prefers to pay a fixed amount $E[X]$ instead a risk amount X.

Fig. 1. Concave utility function - risk averse approach

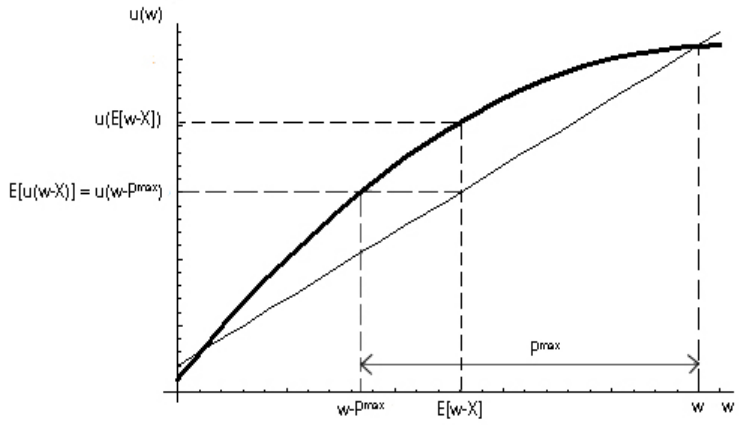
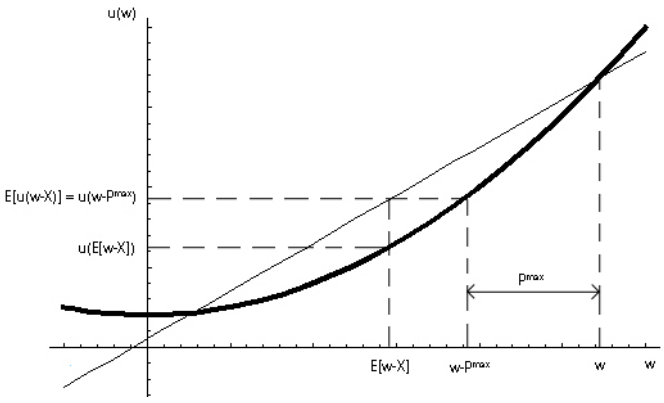


Fig. 2. Convex utility function - risk loving approach



In this part we illustrate by evaluating an individual's decision of whether or not to buy insurance. Now, suppose that our respondent has two alternatives - to purchase an insurance or not to do so. Assuming he is insured against a loss X for a premium P. If he is insured that means certain alternative. This decision take us utility value $u(w-P)$. If he is not insured that means uncertain alternative.

In this case is expected utility $E[u(w-X)]$.

From Jensen's inequality (2) we get inequality

$$E[u(w-X)] \le u(E[w-X]) = u(w-E[X]) = u(w-P). \tag{3}$$

Since utility function u is a non-decreasing continuous function, this is equivalent to $P \le P^{\max}$, where $P^{\max}$ denotes the maximum premium to be paid. This so-called *zero utility premium* is the solution to the following utility equilibrium equation

$$E[u(w-X)] = u(w-P^{\max}). \tag{4}$$

The difference $(w-P^{\max})$ is also called *certainty equivalent - CE*.

The insurer with the utility function $U(x)$ and a capital W, with insurance the loss X for a premium P if

$$E[U(W+P-X)] \ge U(W) \tag{5}$$

hence $P \geq P^{\min}$, where $P^{\min}$ denotes the minimum premium to be asked. This premium from solving the utility equilibrium equation reflecting the insurer's position:

$$U(W) = E\left[U\left(W + P^{\min} - X\right)\right]. \tag{6}$$

From a theoretical point of view, insurers are often considered to be risk neutral. So for any risk X, disregarding additional costs, a premium $E[X]$ is sufficient.
Therefore,

$$U(W) = E\left[U\left(W + E[X] - X\right)\right] \tag{7}$$

for any risk X.
The shape of a person's utility function fundamentally affects the relationships between utilities, expected payoffs, certainty equivalents, and maximal premiums.

Remark 1. We recall that the expected utility we calculate by well-known formula

$$E[u(X)] = \sum_{i=1}^{n} u(x_i) \cdot p_i, \tag{8}$$

where $X = (x_1, x_2, ..., x_n)$ is a vector of the possible alternatives and p_i, for $i = 1,2,...,n$, appropriate probabilities.

Remark 2. When possible monetary outcomes fall in the decision maker's range of risk averse, the following properties hold:

1. Expected payoffs $EP = E[w - X]$ are greater then their counterpart certainty equivalent $CE = w - P^{\max}$.
2. Expected utilities $E\left[u\left(w - X\right)\right]$ will be less than the utility of the respective expected monetary payoff $u\left(w - P^{\max}\right)$.
3. Risk premiums $RP = EP - CE$ are positive.

When possible monetary outcomes fall in the decision maker's range of risk loving, the following properties hold:

1. Expected payoffs $EP = E[w - X]$ are less then their counterpart certainty equivalent $CE = w - P^{\max}$.
2. Expected utilities $E\left[u\left(w - X\right)\right]$ will be greater than the utility of the respective expected monetary payoff $u\left(w - P^{\max}\right)$.
3. Risk premiums $RP = EP - CE$ are negative.

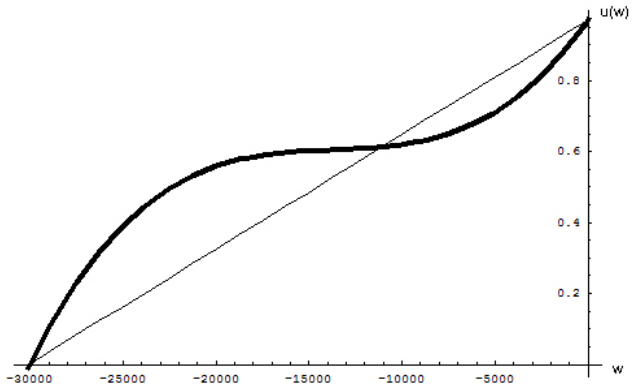
3 The Maximal Premium and Person's Utility Function

In practice, the utility function is found empirically by personal interviewing decision maker. This function is easily constructed from the information gleaned in a short interview. This interview was created by my student in her student scientific activity [6]. Respondent can use this function in any personal decision analysis in which the payoff fall between 0 and 30000 €.

Now we recall the interview, which my student put together [6].
Suppose that we are owner of the investment, which brings us zero payoff now or a loss 30000 €. However, you have a possibility to step aside from this investment at penalty in sequence: A: 1000 €, B: 5000 €, C: 10000 €, D: 15000 €, E: 25000 €.
Your portfolio manager can provide you with information with what probability you can loss 30000 €. Take a think! What would be the biggest probability of the loss, so that you retain mentioned investment?

Only a few well spread out graphed points are required. From the interview we choose the person's data points for individual losses and correspoding probabilities $(0,1)$, $(-1000,0.8)$, $(-5000,0.75)$, $(-10000,0.60)$, $(-15000,0,60)$, $(-20000,0.40)$, $(-30000,0.00)$, and construct appropriate utility function, because this utility function is for larger losses concave, and for smaller losses convex, as shown in Figure 3.

Fig. 3. Utility function and expected utility of our respondent (the function 3. from Table 1)



This curve has the interesting shape, which reflects the underlying approach to the risk of our respondent.
Different person's utility functions for our respondent were created by system SPSS 13.0 for a comparison.

Maximal premium $P^{\max}$ was calculated by inverse function u^{-1} to the utility equilibrium equation (4)

$$P^{\max} = w - u^{-1}\left(E\left[u\left(w - X\right)\right]\right) \tag{9}$$

by system Mathematica 5.
Utility functions are used to compare investments to each other. For this reason, we can scale a utility function by multiplying it by any positive constant and (or) translate it by adding any other constant (positive or negative). This kind of transformation is called a positive affine transformation. All our results would be the same.
Quadratic and cubic utility functions and also their invariant functions created by adding constant 3001 (dislocation for only positive input values), are written in Table 1.
In this table you can see values Adjusted R square, F statistics and significance level for individual functions. On the basis statistical parameters we assume that cubic function is fitting function. Moreover, there are corresponding individual expected utilities in this table expressed as linear functions.

Remark 3. Expected utilities (for the utility functions 1.,3. from Table 1) we can calculate by linear function, which is assigned uniquely by points $\left(-30000, u\left(-30000\right)\right)$ and $\left(0, u\left(0\right)\right)$, or by the formula (8) also.
In both cases we get the same values of the expected utilities. For utility functions 2. and 4. we calculate expected utilities by the analogical way.

Table 1. Utility function and expected utility

	Utility function and expected utility	Adjusted R square	F	Sig.
1.	$u(x) = -2.820 \cdot 10^{-10} x^2 + 1.856 \cdot 10^{-5} x + 0.904$ $E[u(x)] = 27.02 \cdot 10^{-6} x + 0.904$	0.883	23.677	0.006
2.	$u(x) = -2.820 \cdot 10^{-10} x^2 + 3.548 \cdot 10^{-5} x + 0.093$ $E[u(x)] = 27.02 \cdot 10^{-6} x + 0.093$	0.883	23.677	0.006
3.	$u(x) = 1.310 \cdot 10^{-13} x^3 + 5.383 \cdot 10^{-9} x^2 + 7.588 \cdot 10^{-5} x + 0.971$ $E[u(x)] = 32.29 \cdot 10^{-6} x + 0.971$	0.983	117.290	0.001
4.	$u(x) = 1.310 \cdot 10^{-13} x^3 - 6.404 \cdot 10^{-9} x^2 + 1.0652 \cdot 10^{-4} x + 0.003$ $E[u(x)] = 3.23 \cdot 10^{-5} x + 0.003$	0.983	117.290	0.001

From Table 2 you can see that the insured person is willing to pay more than the expected loss to achieve ``piece of mind".
A fresh utility function would be required for evaluating a decision with more extreme payoffs or if our respondent's attitudes changes because of a new job or lifestyle change. Moreover, the utility function must be revised over time.

Table 2. Expected utility and maximal premium

Probability p	$E[u]$ with respect to quadratic function 1.	$P^{\max}$ (€)	$E[u]$ with respect to cubic function 3.	$P^{\max}$ (€)
0.01	0.9040	0.00	0.9710	0.00
0.05	0.8959	433.57	0.9613	129.10
0.10	0.8635	2114.20	0.9226	669.09
0.20	0.7419	7807.63	0.7773	3237.82
0.30	0.6608	11198.10	0.6804	6032.05
0.40	0.5798	14342.30	0.5835	18180.08
0.50	0.4987	17293.40	0.4867	22796.40
0.60	0.4176	20080.40	0.3898	25032.40
0.70	0.3366	22724.70	0.2929	26643.10
0.80	0.2555	25252.10	0.1960	27938.20
0.90	0.1745	27671.10	0.0992	29036.30
1.00	0.0934	30000.00	0.0023	30000.00

4 Conclusion

We have shown how to construct person's utility function and we have calculated maximum premium for loss 30000 € with respect to person's utility function. Constructing the utility function for an insurer is very difficult, we can say that it is almost impossible. However, we think that person's utility function of individual insured person would be very important for the insurer. On the basis of the person's utility function he knows, what approach to the risk concrete insured has, and so how he will behave to the own wealth. In our next work we want to investigate the insurer's utility function, and to finish whole mentioned model from Section 2.

Acknowledgments: This work was supported by grant VEGA 1/0539/08.

References

1. Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M.: Modern Actuarial Risk Theory, Kluwer Academic Publishers, Boston, (2001)

2. Lapin, L. L.: Quantitative Methods for Business Decisions, 5th edition, San Jose State University, (1975)
3. Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior, Princeton University Press, US, (1947)
4. Neumann, J., Morgenstern, O., Khun, H. W., Rubinstein, A.: Theory of Games and Economic Behavior, Sixtieth-Anniversary Edition, Princeton University Press, US, (2007)
5. Norstad, J.: An Introduction to Utility Theory, http://homepage.mac.com/j.norstad, (2005)
6. Riesová, R.: Utility function in a theory and practice (in Slovak), Student Scientific Activity, Banská Bystrica, (2009)

Predictive Power of Neural Networks in Finance

Jiří Trešl

University of Economics Prague, W.Churchill Square 4,
130 67 Prague, Czech Republic, tresl@vse.cz

Abstract. The possibility of application of neural networks for the prediction of both stock and exchange rate returns was investigated. First, the capability of neural networks to reveal specific underlying process was studied using different simulated time series. Second, actual weekly returns from Czech financial markets were analyzed and predicted. Particularly, the problems connected with capturing of outliers and structural breaks were discussed. The predictive power of neural metworks was investigated both as a function of network architecture and the length of training set.

Keywords: neural networks, financial time series, predictive power

1 Introductory Remarks to Neural Networks

Artificial neural networks (ANN) are now frequently used in many modelling and forecasting problems, mainly thanks to the possibility of the use of computer intensive methods. Recently, they have been increasingly applied in financial time series analysis as well [1], [2]. The main advantage of this tool is the ability to approximate almost any nonlinear function arbitrarily close. Particularly in financial time series with complex nonlinear dynamical relationships, the ANN can provide a better fit compared with parametric linear models. On the other hand, usually it is difficult to interpret the meaning of parameters and ANN are often treated as „black box“ models constructed for the pattern recognition and prediction. Further, excellent in-sample fit does not guarantee satisfactory out-of-sample forecasting.

Generally, the ANN is supposed to consist of several layers. The *input layer* is formed by individual inputs (explanatory variables). These inputs are multiplied by *connection strengths* which are called *weights* in statistical terminology. Further, there is one or more *hidden layers,* each consisting of certain number of *neurons.* In the hidden layer, the linear combinations of inputs are created and transformed by the *activation functions.* Finally, the *output* is obtained as a weighted mean of these transformed values. Usually, this kind of ANN is referred to as *multilayered feedforward network* and we restrict ourselves to the most commonly used models with one or two hidden layers. It is useful to realize, information flows only in one direction here, from inputs to output. In time series problems, variables are measured over a time interval and we suppose the existence of relationships among variables at

successive times. In this case, our objective is to predict future values of a variable at a given time using its lagged values at earlier times. We restrict here to the case, when single numeric variable is observed and its next values is predicted using number of lagged values.

The mathematical representation of the feedforward network with one hidden layer and logsigmoid activation functions is given by the following system [1]

$$\begin{aligned} n_{k,t} &= w_{k,0} + \sum_{i=1}^{I} w_{k,i} x_{i,t} \qquad N_{k,t} = 1/\left[1+\exp\left(-n_{k,t}\right)\right] \\ Y_t &= \gamma_0 + \sum_{k=1}^{K} \gamma_k N_{k,t} + \sum_{i=1}^{I} \beta_i x_{i,t} \end{aligned} \tag{1}$$

The first equation describes the creation of linear combination of input variables x using weights w, whereas second one expresses the creation of neuron N by logsigmoid activation function. The third equation explains that output value Y can be obtained either from neurons N or from inputs x directly. Clearly, if there are no hidden layers, the model reduces to purely linear one.

2 Predictions with Simulated Data

First, the various types of ANN were trained and applied to three kinds of simulated time series. The main aim was to investigate prediction ability with respect to the length of time series (250 or 500 is believed to be enough and to prevent overtraining), the number of lagged explanatory values (10 or 20) and the number of hidden layers (1 or 2). In each case, 10 last values were used for prediction, so that either 240 or 490 values were left as training data. To quantify the prediction ability, the following goodness of fit measures were computed: *Mean Prediction Error* (MPE), *Mean Deviation of Prediction Errors* (MDPE) and *Mean Absolute Prediction Error* (MAPE)

$$\begin{aligned} MPE &= \frac{1}{h}\sum_{j=1}^{h}\left(y_{n+j} - Y_{n+j}\right) = \frac{1}{h}\sum_{j=1}^{h} e_{n+j} \\ MDPE &= \frac{1}{h}\sum_{j=1}^{h}\left|e_{n+j} - \bar{e}\right| \qquad MAPE = \frac{1}{h}\sum_{j=1}^{h}\left|e_{n+j}\right| \end{aligned} \tag{2}$$

In all formulas, n denotes the number of training data and h the prediction length. Further, the following structural notation will be used: *length of time series – number of lagged values – number of neurons in the first hidden layer – number of neurons in the second hidden layer*. For example, the notation *250–10–03–01* specifies 250 data in time series, 10 lagged explanatory values and 3 (resp.1) neurons in the first (resp. second) hidden layer. All computations were performed with the use of *STATISTICA* software, version 7.

2.1 Simulation 1: Deterministic Chaos

Even some simple nonlinear *deterministic* systems can under certain conditions pass to chaotic states due to extremely sensitivity both to initial conditions and control parameters [3]. As an example, let us consider discrete time system described by *logistic difference equation*

$$y_{t+1} = Ay_t\left(1 - y_t\right) \tag{3}$$

with control parameter *A=4,* because the most chaotic behaviour is observed just for this value. Clearly, the values from the interval <0,1> will be mapped again into this interval.

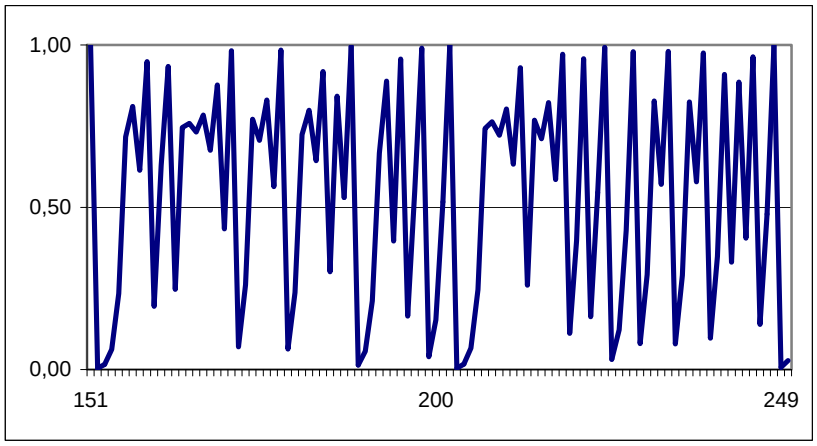


Figure 1. Simulation of Deterministic Chaotic Behaviour (Length 100)

As for modelling, it is obvious from the following table, longer time series provided better results. On the other hand, the number of lagged values and hidden layers are of minor importance here.

Table 1. Results of ANN Modelling: Deterministic Chaos

Network Type	MPE	MDPE	MAPE
500-20-07-00	-0.048	0.100	0.080
500-20-10-08	-0.013	0.220	0.217
500-10-02-00	-0.139	0.203	0.188
500-10-02-02	-0.079	0.121	0.111
250-20-04-00	-0.042	0.325	0.325
250-20-10-10	+0.065	0.252	0.251
250-10-05-00	-0.066	0.295	0.305
250-10-05-04	-0.082	0.234	0.246

2.2 Simulation 2: Bilinear Process

The simplest diagonal form of this process can be written as [4]

$$y_t = \alpha y_{t-1} u_{t-1} + u_t \qquad u_t \approx N\left(0, \sigma^2\right) \tag{4}$$

Clearly, the first term on right hand side leads to process nonlinearity. Further, the condition of stationarity is

$$|\lambda| = |\alpha\sigma| < 1 \tag{5}$$

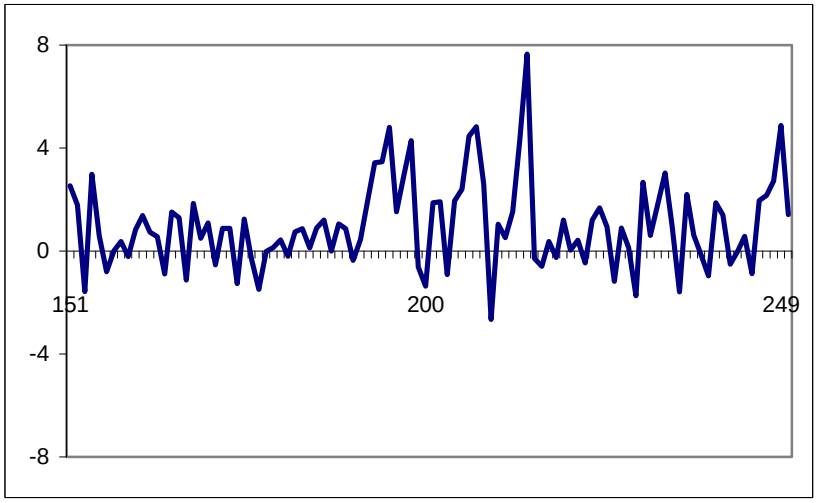


Figure 2. Simulation of Bilinear Process (Length 100)

On the contrary to the previous case, there is no preferred model and the results are comparable.

Table 2. Results of ANN Modelling: Bilinear Process

Network Type	MPE	MDPE	MAPE
500-20-01-00	0.740	1.326	1.431
500-20-02-01	0.428	1.282	1.337
500-10-04-00	0.716	1.257	1.328
500-10-07-02	0.619	1.291	1.346
250-20-01-00	0.746	1.324	1.479
250-20-01-01	0.501	1.348	1.457
250-10-04-00	0.456	1.265	1.320
250-10-05-05	0.409	1.273	1.277

2.3 Simulation 3: AR(1) Kesten Process

This process is a natural generalization of classical AR(1) process to the form [5]

$$y_t = \alpha y_{t-1} + u_t \qquad u_t \approx N\left(0,\sigma^2\right) \qquad \alpha \approx R\left(a,b\right) \tag{6}$$

where *R* denotes regular distribution.

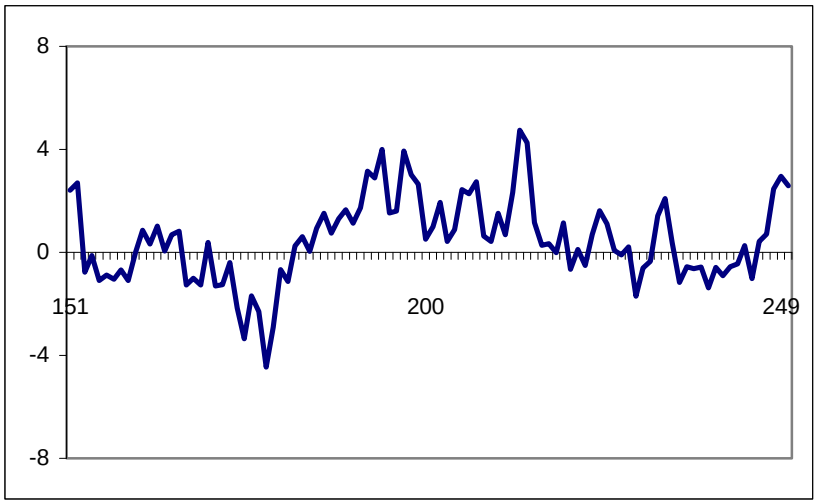


Figure 3. Simulation of Kesten Process with $\alpha \approx R\left(0.3,1.3\right)$ (Length 100)

Again, MDPE and MAPE exhibit relatively slow variations and the best results are achieved with 250-10-03-00 model. As seen from the Fig.5, there is some kind of compensation in the sense that common trend of cumulative predictions is growing. Due to this fact, cumulative actual returns and their predictions are almost the same for 8th week. On the other hand, actual returns are systematically lower in comparison with predictions up to 5th week.

Table 3. Results of ANN Modelling: Kesten Process

Type	MPE	MDPE	MAPE
500-20-03-00	0.511	1.211	1.144
500-20-06-05	0.551	1.279	1.205
500-10-02-00	0.441	1.062	1.022
500-10-03-02	0.363	1.077	1.005
250-20-01-00	0.314	1.243	1.191
250-20-04-03	0.728	1.402	1.337
250-10-03-00	0.215	0.937	0.887
250-10-01-01	0.595	1.131	1.131

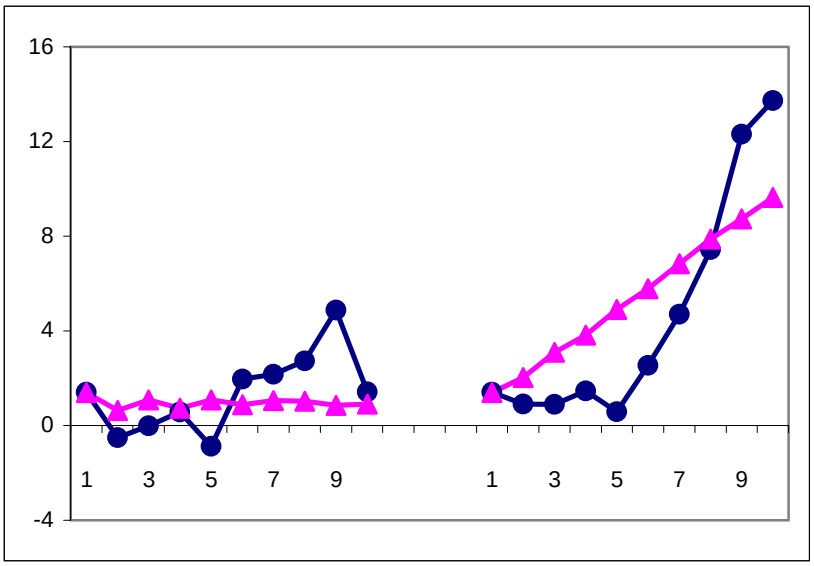


Figure 4. Bilinear Process: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulated Values Model: 250-10-05-05

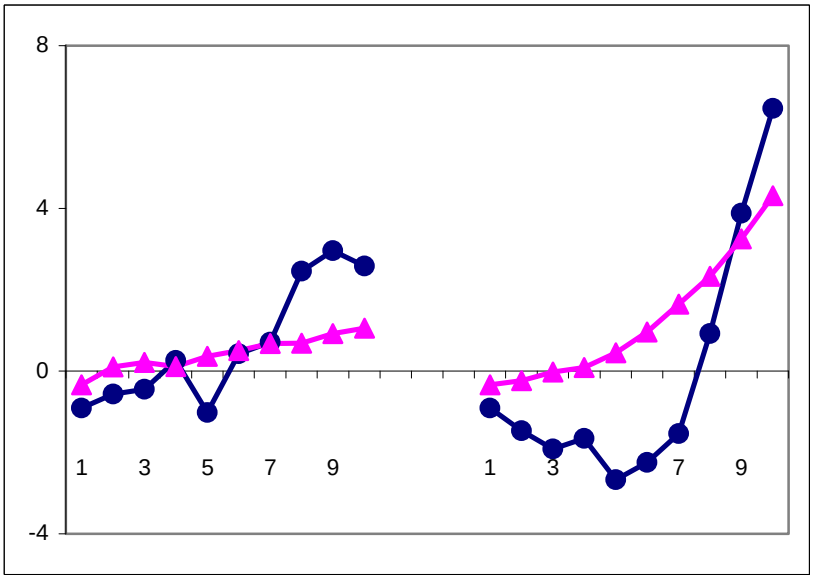


Figure 5. Kesten Process: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulated Values Model: 250-10-03-00

3 Predictions with Financial Data

The main aim was the testing of ANN predictive power in financial applications. We employed weekly logarithmic stock returns (CEZ, KB, TEL,UNIP) and exchange rate returns (CZK/EUR, CZK/GBP, CZK/CHF, CZK/USD) during 2005-2008, i.e. 200 weekly values for each time series. In all cases, 25 preceding values were used and last 12 values were left for prediction testing. Further, both linear models and neural networks with one and two hidden layers were applied.

3.1 Prediction of stock returns

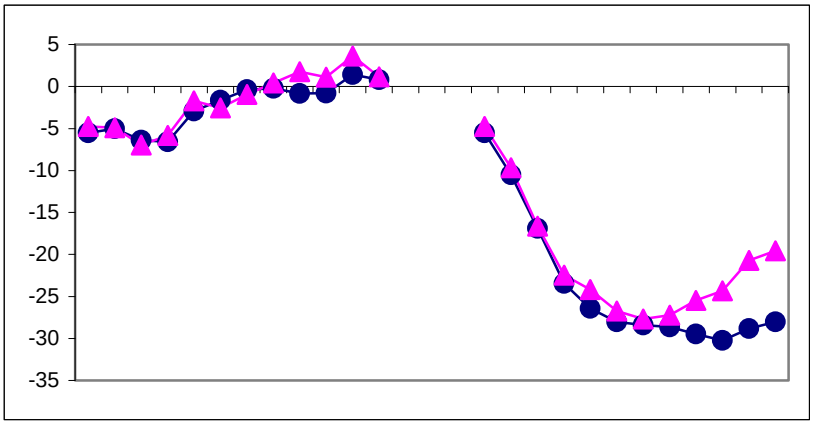


Figure 6. CEZ Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-05-00

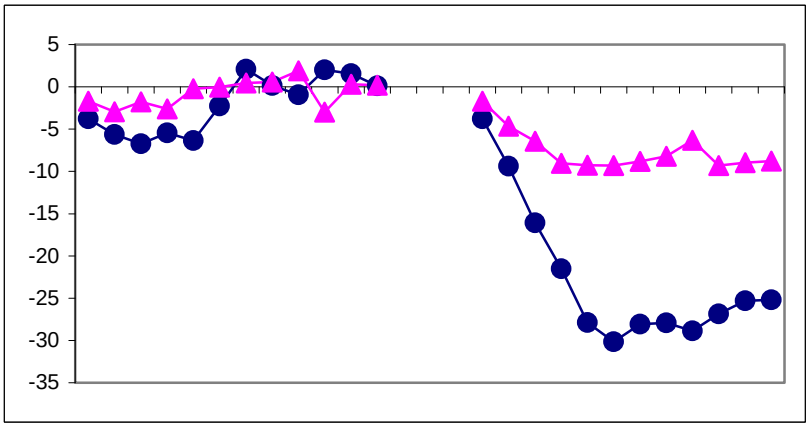


Figure 7. KB Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-12-10

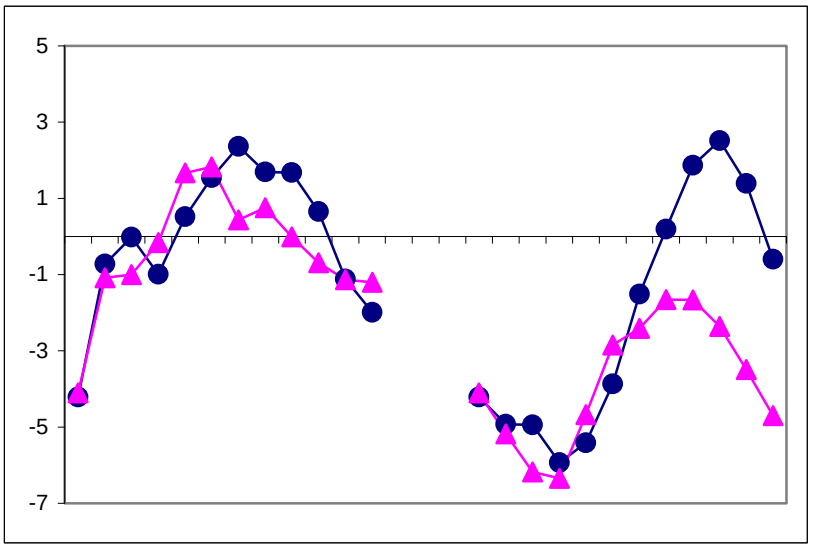


Figure 8. TEL Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-12-06

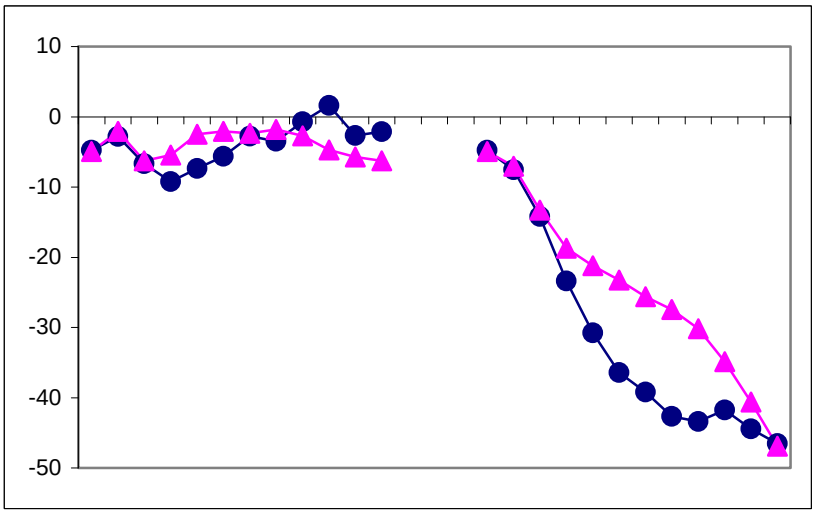


Figure 9. UNIP Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-03-04

3.2 Prediction of exchange rate returns

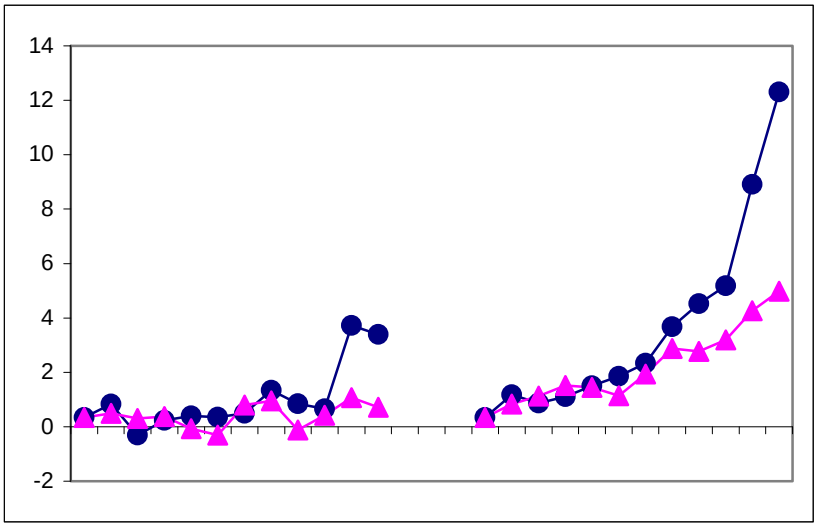


Figure 10. CZK/EUR Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-06-00

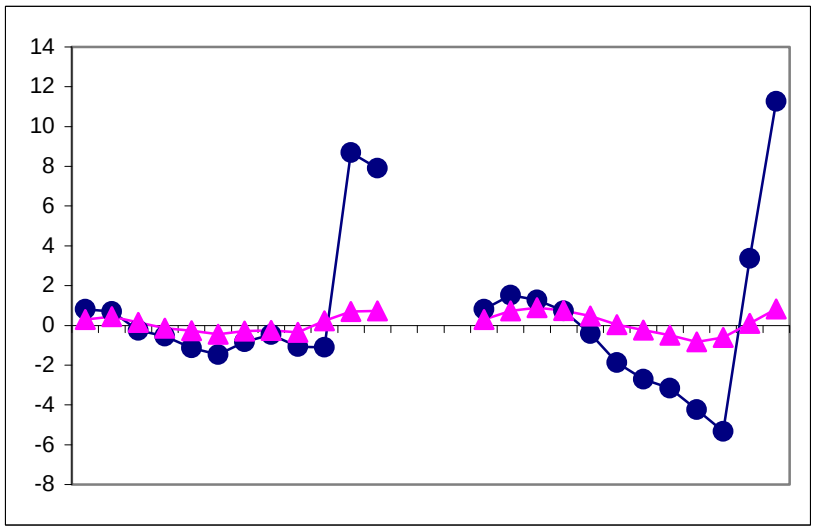


Figure 11. CZK/GBP Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-06-00

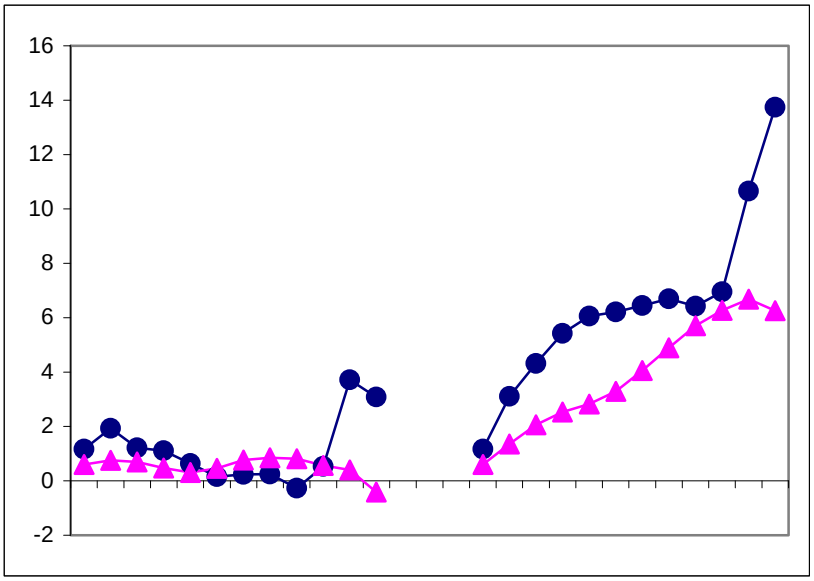


Figure 12. CZK/CHF Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-12-06

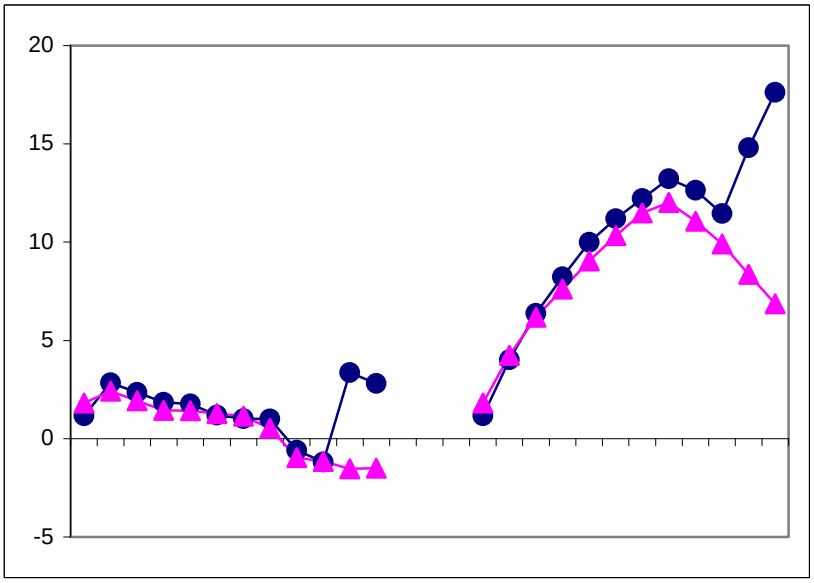


Figure 13. CZK/USD Returns: Actual Values (Circles) versus Predictions (Triangles).
Left: Original Values Right: Cumulative Values Model: 200-25-05-00

Table 4. Results of ANN Modelling: Stock Returns

Stock	Model	Type	MPE	MDPE	MAPE
CEZ	Linear		-1.550	1.227	1.910
	One Layer	200-25-05-00	-0.704	0.869	1.054
	Two Layers	200-25-06-02	-1.565	1.281	2.075
KB	Linear		-2.197	3.211	3.460
	One Layer	200-25-04-00	-1.833	2.964	3.138
	Two Layers	200-25-12-10	-1.368	2.356	2.672
TEL	Linear		+0.906	1.322	1.296
	One Layer	200-25-01-00	+0.527	1.572	1.538
	Two Layers	200-25-12-06	+0.342	0.865	0.868
UNIP	Linear		-2.611	1.892	3.044
	One Layer	200-25-02-00	-0.153	2.897	2.943
	Two Layers	200-25-03-04	+0.028	2.597	2.592

Table 5. Results of ANN Modelling: Exchange Rate Returns

Stock	Model	Type	MPE	MDPE	MAPE
CZK/EUR	Linear		+1.087	0.823	1.121
	One Layer	200-25-06-00	+0.610	0.753	0.786
	Two Layers	200-25-02-01	+1.273	0.830	1.290
CZK/GBP	Linear		+0.730	2.389	2.059
	One Layer	200-25-06-00	+0.870	2.237	1.787
	Two Layers	200-25-04-02	+1.090	2.343	1.801
CZK/CHF	Linear		+1.537	0.959	1.537
	One Layer	200-25-08-00	+0.152	1.293	1.247
	Two Layers	200-25-12-06	+0.623	1.021	1.048
CZK/USD	Linear		+0.913	1.104	1.519
	One Layer	200-25-05-00	+0.896	1.236	1.044
	Two Layers	200-25-03-04	-0.642	1.253	1.395

4 Conclusion

The first group of findings is related to artificial data. Undoubtedly, the best results were obtained for deterministic chaotic process, because there is relatively simple relation between neighbouring values. Second, the process itself is bounded between zero and one and the notion of outliers is meaningless here. On the other hand, the results for both bilinear and Kesten processes are markedly worse due to ability to create sudden random excursions. Further, the results seem to be similar for the lengths of time series used (500 and 250), number of lagged values (20 and 10) and number of hidden layers (1 and 2).

As for stock returns, different kinds of individual behaviour were revealed. Both for CEZ and TEL, there has been good agreement between real data and predictions till 7th week and some deviations occured after this time. On the other hand, predictions of UNIP returns balanced out with respect to their sign, so that corresponding mean predicted error was very small. The worst results were achieved in the case of KB returns, where both actual values and predictions exhibited negative signs up to sixth week, but absolute values of predictions were systematically lower. In most cases, neural networks with two hidden layers turned out to be the best alternative.

Exchange rate returns exhibited similar behaviour, but there were strongly manifested outliers. In all cases, 11th and 12th actual values were strong positive outliers with markedly worse predictions. Thus, the corresponding deviations occured, but the general agreement between actual values and predictions has been observed till 10th week for CZK/USD and CZK/EUR returns, whereas CZK/CHF one manifested some kind of compensation. Further, the signs of actual values and predictions in 11th and 12th weeks were the same for CZK/EUR and CZK/GBP and opposite in the case of CZK/USD. In most cases, neural networks with one hidden layers turned out to be the best alternative.

References

1. McNelis, P.D.: Neural Networks in Finance: Gaining Predictive Edge in the Market. Elsevier Academic Press, Amsterdam (2005). ISBN 978-0124859678.
2. Franses, P.H., van Dijk, D.: Non-Linear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge (2000). ISBN 978-0521779654.
3. Hilborn, R.: Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers. Oxford University Press, Oxford (2001). ISBN 978-0198507239.
4. Tsay, R.S.: Analysis of Financial Time Series. Wiley, New York (2002). ISBN 978-0471415442.
5. Sornette, D.: Critical Phenomena in Natural Sciences. Springer, Berlin (2000). ISBN 978-3540674627.

Synchronisation of Business Cycles - Cross Country Analyses

Vladimír Úradníček, Emília Zimková

Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica
vladimir.uradnicek@umb.sk, emilia.zimkova@umb.sk

Abstract. The main aim of the contribution is to assess the degree of co-movements of the German, French, Italian and Slovak economies over the last 15 years. Attention is focused not only on GDP development but also on the behavior of the demand side components of GDP. Methodology of contribution covers Harding and Pagan approach to measure the degree of synchronisation by the correlation coefficient among their respective reference cycle.

Keywords: Synchronisation of business cycle. Convergence process. Cross-country analyses. Correlation coefficients.

1 Introduction

The conventional wisdom says that the synchronisation of business cycle among member countries is a fundamental prerequisite for an effective Monetary Union. The introduction of the single European currency in Slovakia has broden our interest in business cycle synchronization, which is a part of so called structural convergence. In our previous research we analyzed real economic convergence of the Slovak Republic to Euro zone by so called β convergence, σ convergence and also by cointegration [3]. In our previous research [3] we analyzed behavior of the Slovak gross domestic product (GDP) and the GDP of the Euro area. This time we would like to adopt a different point of view which highlights the role played by GDP components. The paper presents business cycles of demand side GDP components: consumption, investment, import, export and inventory. The research covers three biggest economies of Euro area (Germany, France, Italy) and Slovakia.

The structure of the paper is as follows: section two presents data and methodology of research. The section three gives empirical results of the research and the section four concludes.

The contribution was prepared in the framework of the project VEGA 1/4634/07:

"Variant methods of prediction of small and medium sized enterprises development after introducing single European currency in the Slovak Republic" and project VEGA 1/0229/09 "Analyses of Selected Issues of Financial and Banking Market after the Slovak Republic Accession into EMU".

2 Data and methodology

In this section, after the first look at the statistical properties of the data, we will apply methodology used by Harding and Pagan (2002, 2006) to measure the degree of synchronization by the correlation coefficient among their respective reference cycle.

Statistical properties of the Data

The used data are from Eurostat database quarterly national accounts expressed at current prices. The data cover period since first quarter of 1995 till the first quarter of 2009.

Consumption (C) refers to households expenditures, Investment (I) refers to public and private investment, Import (IMP) (export EXP) refers to import (export) of goods and services to the rest of the world, Inventory (DINV) refers to the contribution of inventory change to GDP growth, net trade reflects net external demand (NED). It is worth to say, that NED express external demand outside of the Euro zone, and it does not take into consideration intra-trade between Euro zone countries. The methodology is taking into consideration fact that foreign trade has an impact on the foreign exchange reserves which influence the level of the exchange rate. Thus the trade within Euro zone countries is not relevant.

Only GDP, Consumption, Investment, Import and Export series are tested, DINV and NED are clearly stationary.

Underlying assumption in the business cycle literature is that any time series X_t can be expressed as the sum of trend (T), cycle (C), seasonal (S) and irregular (I) (unobserved) components as follows: $X_t = T + C + S + I$. The trend includes long-term fluctuations both stochastic as well as deterministic, the cycle refers to medium-term systematic fluctuations, the seasonal components refer to oscillations with period shorter than 1,5 years, while the irregular component is the residual of the decomposition. As the components are not observable, to focus on business cycle fluctuations one needs to assume independence among them and transform the observed series through application of a filter. In our contribution the Hodrick-Prescott Filter was applied.

Table 1. Test for unit root in Level/1st difference

	Level	1st Difference	Order of Integration
Italy	**IT**		
GDP	-0,816	-2,916	1
C	-1,449	-4,771	1
I	-2,571	-5,977	1
IMP	-2,552	-3,733	1
EXP	-1,932	-3,863	1
France	**FR**		
GDP	-0,916	-4,015	1
C	-3,352	-6,651	1
I	-1,484	-3,739	1
IMP	-2,91	-3,765	1
EXP	-2,276	-3,597	1
Germany	**DE**		
GDP	-1,448	-3,641	1
C	-2,398	-4,485	1
I	-2,916	-6,698	1
IMP	-2,634	-3,745	1
EXP	-1,841	-3,347	1
Slovak Republic	**SK**		
GDP	-2,591	-6,398	1
C	-2,835	-4,984	1
I	-2,287	-5,443	1
IMP	-2,842	-4,351	1
EXP	-3,08	-4,427	1

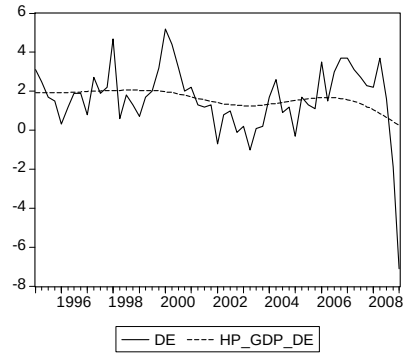


Fig. 1. GDP growth and HP of Germany

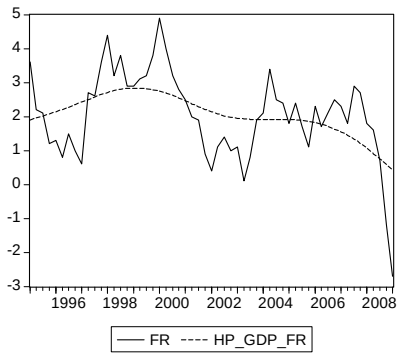


Fig. 2. GDP growth and HP of France

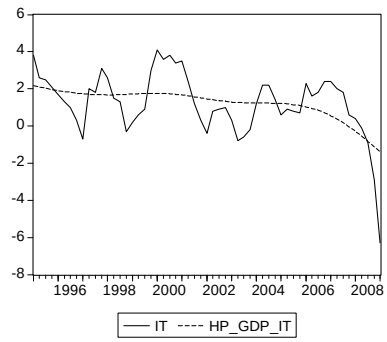


Fig. 3. GDP growth and HP of Italy

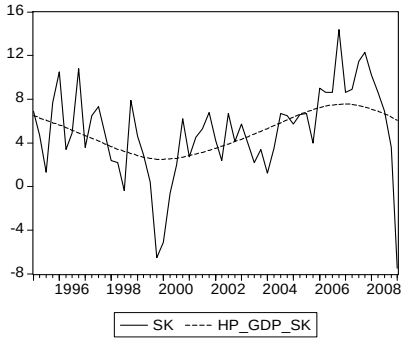


Fig. 4. GDP growth and HP of Slovakia

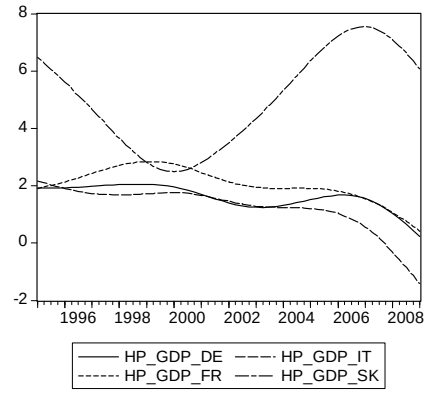


Fig. 5. HP of GDP growth in all analyzed countries

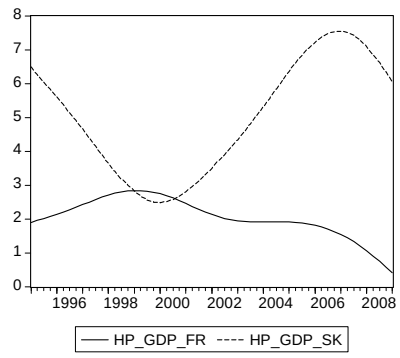


Fig. 6. HP of GDP growth in Slovakia/France

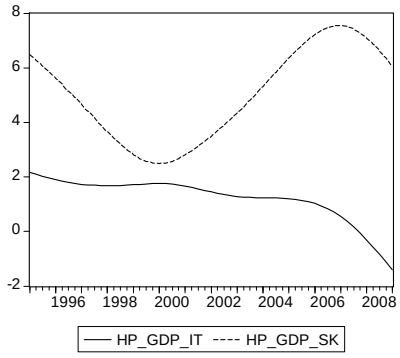


Fig. 7. HP of GDP growth in Slovakia/Italy

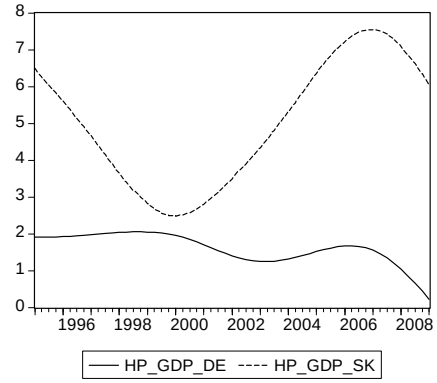


Fig. 8. of GDP growth in Slovakia and Germany

From the figures which reflect the GDP growth filtered by Hodrick-Prescott we can draw several conclusions. There might by evidence of significant convergence of business cycles between analysed countries from the short and medium term. In the long run more significant differences appear. Regarding comparison of the Slovak GDP behavior and the GDP behavior of the largest Euro economies, there is considerable divergence till year 2004. In the period of Slovak economic decline till 2000, analyzed economies experienced either dynamic (France) or modest economic economic growth (Germany and Italy). In the period of robust Slovak economic growth till 2002, analysed countries experienced rather weak economic growth. If we compare three largest economies in the Euro zone we can say that the most robust economic growth was in France.

In our analyses we have applied HP filter also on the components of GDP, e.g. consumption, investment, import, export and inventory over the cycle for Germany, France, Italy and Slovakia.

Measure of synchronization

Regarding methodology in the contribution is applied the Harding and Pagan approach (2002, 2006). Given two variable Xt and Yt Harding and Pagan proposed to measure the degree of synchronisation by the correlation coefficient among their respective reference cycle. The reference cycles are represented by two dummy variables S_t^x and S_t^y, which take values one during expansions (defined as the period between troughs and peaks) and values zero during recessions (defined as the period between peaks and troughs in the filtered series). Having defined the reference cycles, the correlation coefficient can be measured by running the following regression (Harding, Pagan, 2006):

$$s_x^{-1} s_y^{-1} S_y^t = a_1 + \rho_S s_x^{-1} s_y^{-1} S_x^t + u_t \,. \tag{1}$$

where

s_x is the standard deviation of the variable S_x^t, s_y is the standard deviation of the variable S_y^t, ρ_S is the estimated correlation coefficient and a_1 is a regression constant. Tests for synchronisation can be carried out on the parameter ρ_S. A t-test for the null hypothesis that ρ_S quals zero can be performed, provided autocorrelation corrected standard errors are used.

3 Empirical results

Table 2 reports the estimated correlation coefficients obtained by estimating equation (1) for GDP and all demand components for each pair of countries.

Table 2. Correlation between reference cycles.

	GDP	C	I	EXP	IMP	NED	DINV
IT-DE	0,23	0,11	0,26	0,39	0,62	0,16	0,02
IT-FR	0,77	0,61	0,52	0,66	0,66	0,52	0,64
FR-DE	0,18	0,18	0,05	0,44	0,69	-0,56	0,47
SK-DE	0,26	-0,11	0,29	0,44	0,13	-0,84	-0,78
SK-FR	0,31	-0,17	-0,16	0,24	0,25	-0,55	0,32
SK-IT	0,11	0,09	-0,31	0,25	0,23	0,3	-0,24

According our analyses there is evidence of strong correlation between business cycles of France and Italy. Rather weak positive correlation is between the rest of analyzed countries. As we can see from the figures in the contribution, results are influenced by rather weak economic performance in Germany after reunion. Later on, in 2006 and 2007 the Germany and France were economically growing while growth in Italy as was slowing down. Economic performance of all analyzed countries in 2008 and 2009 was hit by current economic crisis.
Slovak business cycle is correlated with German and French ones. The highest correlation is between German and Slovak export. As we know, the German export is the largest over the world (second position has China followed by USA), thus the correlation of Slovak and German export cycle is very positive. Less positive aspect of analyzed export is fact that while Germany is exporting large machinery and sophisticated products the Slovak export includes only limited value added products (Slovak export include mainly cars, TV sets, chemicals).
Consumption growth in business cycle of largest economies of Euro zone was rather modest also due to prudent public consumption which is in frame of the Stability and Growth Pact. On the contrary Slovak consumption was growing at two digit level. In first half of analyzed period there was a strong public consumption which was after the entry into European Union replaced by strong private consumption.
The development of the net external demand in Slovakia and Italy is negative. It is proving that the export of goods and services outside of the Euro zone is in long term smaller that the raw material imported from Russia and goods from outside of Euro zone. Germany and France in long run record a positive performance in the net external demand. That is why the correlation of the NED business cycle between Slovakia and those two countries is negative.

4 Conclusion

Synchronisation of business cycles of three largest Euro zone countries (Germany, France, Italy) and Slovakia was analyzed by Harding and Pagan approach (2002, 2006). In analyzed period there is evidence of strong correlation between business cycles of France and Italy.
Cyclical development of all demand side components of GDP are strongly correlated too.
Rather weak positive correlation of business cycle is between Germany and rest of analyzed countries and Slovak Republic and rest analyzed countries. As to correlation of all demand components of Germany and Slovakia the most positive outcome is the correlation of the export cycle.

5 References

1. Bry, G., Boschan, C.: Cyclical analysis of time series: selected procedures and computer programs, NBER technical paper, N. 20. (1971). Available online at http://www.nber.org/chapters/c2145.pdf (27 July 2009).
2. Bulligan, G.: Synchronisation of cycles: a demand side perspective, In: Convergence or Divergence in Europe? Springer. P. 384. (2007).
ISBN 3-540-32610-3.
3. Gavliak, R., Úradníček, V., Zimková, E.: Error correction models and real convergence – case of Slovakia, In: 11th AMSE conference, Wisla (2008), in print.
4. Harding, D., Pagan, A.: Dissecting the cycle: a methodological investigation. In: Journal of Monetary Economics 49 pp. 365 - 381. (2002). Available online at http://www.sciencedirect.com/science/article/B6VBW-44TVC6K-1/2/be5761f0ac5411002e46fa1ad6058869 (27 July 2009).
ISSN 0304-3932.
5. Harding, D., Pagan, A.: Synchronisation of cycles. In: Journal of Econometrics 132 (2006), pp. 59 – 79. Available online at http://www.sciencedirect.com/science/article/B6VC0-4FMK8MH-4/2/30878812d8beac5668385285cbb3f926 (27 July 2009).
ISSN 0304-4076.
6. Harding, D., Pagan, A.. An Econometric Analysis of Some Models for Constructed Binary Time Series. National Centre for Econometric Research, Working Paper 39, January 2009. Available online at:
http://www.ncer.edu.au/papers/documents/NCER_WpNo39Jan09.pdf (27 July 2009).
7. National Bank of Slovakia: Analyses of convergence of the Slovak economy. (2009).

Clustering Data Based on Directly Unobservable Attributes

Ondřej Vilikus

The University of Economics, Faculty of Informatics and Statistics, Department of Statistics and Probability, Prague, Czech Republic
ondrej.vilikus@vse.cz

Abstract. Traditional data clustering methods rely on distances measured in space defined by directly observable measures. However in some applications of cluster analysis such as market segmentation we would rather obtain clusters that would be based on attributes that are not directly observable. These attributes such as sensitivity to different factors in analysis of preference are obtained as parameters of estimated models and thus cannot be used as inputs for traditional clustering even if there is substantive heterogeneity in the sample.This article explores ways how to handle samples heterogeneous in parameters of estimated models with use of latent variable models. Hierarchical Bayesian models with latent variables seem to be the most effective tool for solving this problem as they incorporate both the possibility to estimate model parameters for each individual and allow use of other information about the individuals to be used to get more accurate and meaningful results.

Keywords: Cluster analysis, market segmentation, latent class models, hierarchical Bayesian models.

1 Introduction

Bayesian theorem as a tool for accounting for uncertainty by combining apriori knowledge and information contained in our data has been around for more than two centuries. But it hasn't been until few tens of years until we have seen dramatic increase in use of Bayesian methods in many applications including marketing.

Until the mid 1980s, Bayesian methods appeared to be impractical because the class of models for which the posterior could be computed was no larger than the class of models for which exact sampling results were available, see [2]. Since simulation methods and in particular Markov Chain Monte Carlo (MCMC) has been adopted for posterior estimation tremendous variety of models could be used even in situations where traditional likelihood based modeling suffered from identification problems. Once free of computational constraints Bayesian models have incurred dramatic growth of popularity.

Since 90s tens of papers illustrating applications and often superiority of Bayesian models have been published in marketing journals and Bayesian methods are already finding its way in specialized software developed for analyzing marketing data.

Purpose of my article is to describe another application of Bayesian approach in dealing with heterogeneity of consumers in marketing models. After reviewing past and present approaches ideas for further research are outlined.

2 Dealing with Heterogeneity of Consumers

Marketing is concerned with understanding and reacting to the behavior of individual consumers. While the decisions are made at the individual level, due to constraints set by the sparsity of individual level data decisions of the customers have traditionally been modeled at the aggregate level. This approach gives reasonable and robust solutions in the situations where consumers are alike in their preferences and choice behavior. (Un)fortunately this is rarely true and aggregate models not allowing for heterogeneity of consumers are likely to produce spurious and in many cases misleading results.

To illustrate possible drawbacks aggregate level model and available remedies that have been developed to solve the problem of heterogeneity lets assume simplified model of consumer behavior described in the next paragraphs. This example represents common multinomial probit model used in choice-based conjoint analysis, however all conclusions are valid for any other model of consumer behavior.

2.1 Modeling Consumer Behavior – Illustrative Example

Choosing a new car is for the consumer usually a decision making process that take a few weeks since the purchase represents a substantial part of household's budget and many parameters of available options need to be considered in the decision making process. Suppose the customer is maximizing his utility (or alternatively highest value obtained as utility divided by cost). When choosing out of J alternatives, i-th customer chooses the j'-th alternative when his utility from this choice is the maximum available

$$y = j' \text{ if } U_{ij'} = \max_j \{U_{ij}\}. \quad (1)$$

We do not observe utilities of the choices but only various attributes of the choices, represented by vector $\boldsymbol{x}_j$. In our example, these can include brand of the car, its gas consumption, power, storage compartment volume, various extra equipment, price and many others. As preferences of each consumer are then represented by vector $\boldsymbol{\beta}_i$, random utility model is in the form

$$y = j' \text{ if } \boldsymbol{x}_{j'}^{\mathrm{T}} \boldsymbol{\beta}_i + \varepsilon_{ij'} = \max_j \{\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_i + \varepsilon_{ij}\}. \quad (2)$$

Aggregate level model assumes all preferences $\boldsymbol{\beta}_i$ to be equal. If there are groups of consumers with different preferences, averaging these distinct subgroups may

cause some attributes to be reported as not influential even though they may be the most important. Suppose β_{i1} and β_{i2} are elements of vector $\boldsymbol{\beta}_i$ corresponding to how strongly consumer prefers a car with more powerful engine over one with low consumption and whether he prefers rather a car that is more spacious or small and easy to park.

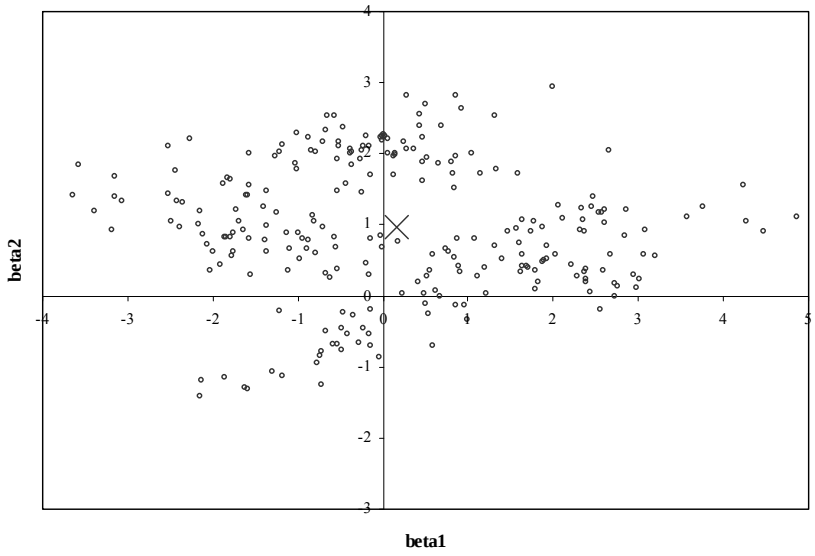


Fig. 1 Illustration of aggregate level model results (marked with a cross) on two parameters with strong heterogeneity of individual preferences (circles).

Results obtained by aggregate level model are illustrated in Fig. 1. Even though both attributes are highly relevant to majority of customers, averaging out their preferences leads to seeming insignificance of β_1 and substantial loss of information with regards to β_1. It seems that the larger the car is the better however for many customers this is not true and modeling their preferences would be highly inaccurate.

2.2 Market Segmentation Approach

One way of dealing with the implicit assumption that consumers differ from each other but only to a certain extent is market segmentation. It could be believed that consumers can be grouped into relatively homogenous groups or segments. Identification of these segments that can be then target of specialized marketing activities is usually referred to as market segmentation.

Wedel and Kamakura in [10] note that: *"Market segmentation is an essential element of marketing in industrialized countries. Goods can no longer be produced and sold without considering customer needs and recognizing the heterogeneity of*

those needs." Since market segments are not usually real entities apparent in the market but rather artificial grouping of the population based on manager's needs and goals that can help to satisfy customers' needs, market segmentation is usually an object of a study on customers' behaviors and attitudes.
Identification of market segments is highly dependent on the bases (variables or criteria) used to defined them so these have to be chosen carefully on the basis of specific purposes of the segmentation study. Frank, Massy and Wind in [4] define classification of segmentation basis as shown in Table 1.

Table 1. Classification of Segmentation Bases.

	General	Product-specific
Observable	Cultural, geographic, demographic and socioeconomic variables	User status, usage frequency, store loyalty and patronage, situations
Unobservable	Psychographics, values, personality, life-style	Psychographics, benefits, perceptions, elasticities, attributes, preferences, intentions

Bases that are observable and that can be directly measured can be used as basis for segmentation disregarding the method used. We are also relatively free to use bases that are unobservable but that can be obtained indirectly by asking consumers in a survey such as personality, life-style or perceptions and intentions. However in many situations the desired basis for market segmentation cannot be obtained from the customer even if asked simply because he or she isn't aware of it or cannot give satisfactory explanations (price-elasticities, attribute importance, part-worth utilities…). Such bases need to be modeled indirectly and cannot be used as an input for market segmentations when traditional methods of cluster analysis are used.

When we are dealing with heterogeneity of the market in directly unobservable measures we have an option to segment the consumers with the use of latent class models for choice behavior.

Latent class models became popular in mid 90s as a tool for analyzing choice-based conjoint data sets. In contrast to aggregate analysis where the model assumes that all the variability between respondents is random latent class attempts to find of respondents who share similar preferences. It then estimates specific parameters for specified number of groups and probability for each respondent of belonging to each via iterative algorithm as EM (expectation-maximization).

It has been found that latent class methods may be successful in recovering known data structure and that segments based on latent class solution provide better fit aggregate model or model where clusters based on cluster analysis on observable measures have been created, see [9].

If we get back to our example outlined in section 2.1 we may generalize the aggregate model to a latent class model. Assume that there is K groups of consumers distinct in preferences. Utilities from each of J alternatives are then given as

$$U_{ij} = \sum_{k=1}^{K} \boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_k \cdot p_{ik} \tag{3}$$

where $\boldsymbol{\beta}_k$ represents preference of *k*-th group and p_{ik} probability that *i*-th consumer belongs to the *k*-th group. Individual preferences can be therefore modeled as linear combinations of group vectors creating a convex shape of possible values in parameter space as seen in Fig. 2. or they can be assigned to the most probable group. Estimates of parameters for each group as well as group sizes give us an easy to understand picture of the heterogeneity of the sample.

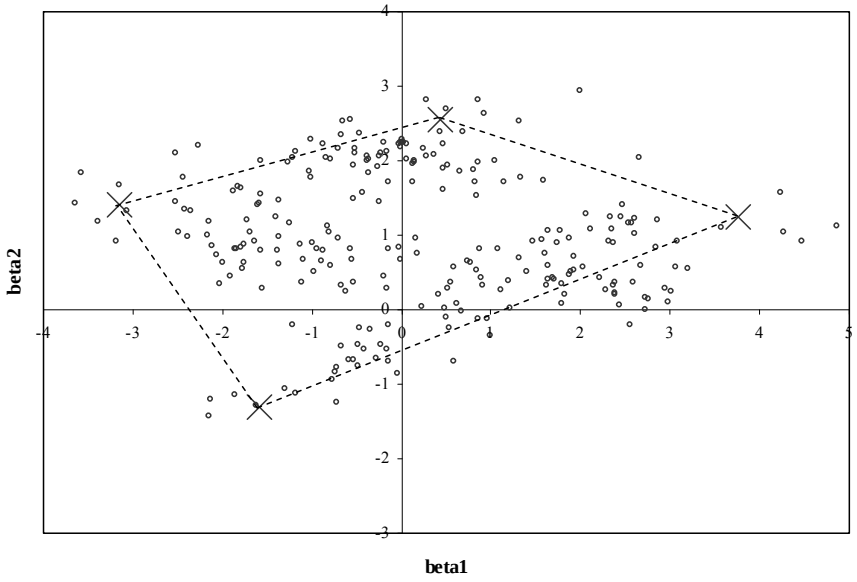


Fig. 2 Solution with four clusters and simplex of attainable individual level solutions.

2.3 Hierarchical Bayesian Models

Different approach for dealing with the heterogeneity of customers is offered by hierarchical Bayesian models. Preferences of customers are supposed to be different but similar in some extent. In hierarchical models estimated parameters for each individual respondent are treated as random variables coming from a common continuous distribution. Individual level parameters and parameters of the upper level distributions are estimated at the same time with use of MCMC based on our apriori assumptions. We are then interested in posterior distribution (or its characteristics such as posterior mean) of the individual level parameters given our observed data.

Hierarchical model consist of the individual level likelihood and two stages of priors:

- Likelihood: $p(\boldsymbol{y}_i, \boldsymbol{x}_{ij} \mid \boldsymbol{\beta}_i)$,

- First-stage prior: $p(\boldsymbol{\beta}_i \mid \bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_\beta)$,
- Second-stage prior: $p(\bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_\beta, \boldsymbol{\tau})$.

Likelihood of our data $\boldsymbol{y}_i, \boldsymbol{x}_i$ has the same form as (3) assuming that vector of choices has multinomial distribution with parameters given by individual utilities based on $\boldsymbol{\beta}_i$ and $\boldsymbol{x}_{ij}$

$$P(y = j') = P\left(\boldsymbol{x}_{j'}^{\mathrm{T}} \boldsymbol{\beta}_i + \varepsilon_{ij'} = \max_j \left\{\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{\beta}_i + \varepsilon_{ij}\right\}\right). \tag{4}$$

The first-stage prior specifies that individual level parameters $\boldsymbol{\beta}_i$ come from a multivariate normal distribution with mean $\bar{\boldsymbol{\beta}}$ and covariance matrix $\boldsymbol{\Sigma}_\beta$

$$\boldsymbol{\beta}_i \sim \mathrm{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_\beta). \tag{5}$$

At last purpose of the second-stage prior is to define the distribution of the parameters $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_\beta$. This can be uniform or other non-informative prior.

Assuming continuous distribution of individual level parameters has several advantages over assumption of discrete grouping of respondents into artificial number of segments. They seem to better represent the heterogeneity in the sample and predict individual choice behavior with greater accuracy, see [3], [6] or [10]. On the other hand discrete representation of heterogeneity is generally more actionable for the managers and might provide better description of the population. For other problems associated with use of hierarchical models see [7].

2.4 Mixture Random Coefficient Models

As both of the two above mentioned approaches have their own advantages, models combining both discrete and continuous heterogeneity have been recently developed by Allenby and Rossi in [1] or Lenk and DeSarbo in [5]. Such models use multinomic prior distribution for defining probability that given consumer belongs to a specific segment while heterogeneity in the sample is further modeled by some continuous prior. This means that instead of statement (5) in the model two statements defining the assignment A_i of *i*-th consumer to the segment and heterogeneity among particular segments are included:

$$P(A_i = k) = P_k \text{ for } k = 1, \ldots, K, \tag{6}$$

$$\boldsymbol{\beta}_k \sim \mathrm{N}(\bar{\boldsymbol{\beta}}_k, \boldsymbol{\Sigma}_\beta) \text{ for } k = 1, \ldots, K. \tag{7}$$

3 Topics For Further Research

Additionally, Bayesian approach might be further employed in other areas where it can offer straight-forward way for dealing with some common problems that arise while modeling consumer preferences. As these areas are object of future research and need to be validated by Monte Carlo simulations as well as on real world data, I describe them in separate section.

3.1 Linking of Heterogeneity in Preferences with Measurable Attributes

Use of exogenous information (outside the information already available in the choice tasks) to the model to improve the estimates of preferences and market predictions has already been suggested by Orme and Howell in [8]. Rather than assume respondents' preferences are drawn from a normal distribution with mean vector $\bar{\boldsymbol{\beta}}$ and covariance matrix $\boldsymbol{\Sigma}_{\beta}$, their HB model assumes that respondents' preferences are related to the covariates represented by additional measurable attributes through a multivariate regression model:

$$\beta_i \sim \mathrm{N}\left(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{z}_i, \boldsymbol{\Sigma}_{\beta}\right). \tag{8}$$

When the effect of covariates is significant it may be a good sign that these variables might be a good choice as a basis for market segmentation.

To obtain clusters that would be more identifiable by measurable attributes, such covariates could be linked to the definition of the segments rather than to the individual-level preferences itself. Vector of probabilities $\boldsymbol{A}$ of assignment i-th consumer to the k-th segment could be modeled by Dirichlet distribution:

$$\boldsymbol{A} \sim \mathrm{Dirichlet}\left(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{z}\right). \tag{9}$$

3.2 Number of Clusters Selection

When doing cluster analysis, common issue is the selection of proper number of clusters. Model-based methods require choosing the number prior to the estimation. Even though various indicators have been developed to choose the number of clusters that is optimal in some way, each requires the model to be estimated separately with different numbers of clusters and the choice of the proper number is then based on some heuristics or the assumption that adding an additional cluster needs to bring significant enhancement in some model quality measure.

As hierarchical models can be used for model selection as well, estimation of number of clusters as a search for model with highest posterior probability among class of models with various numbers of clusters is also possible. Over above mentioned this approach this enables us to support the probability of choosing model with convenient number of clusters (generally 4 to 8 market segments are preferred as

they offer enough differentiation yet still allow for easy targeting) by using suitable prior such as Poisson distribution with prior mean corresponding to our desired number of clusters. When the likelihood won't favor strongly some solution resulting number of clusters with highest posterior probability won't differ much from our desired number.

4 Conclusions

Bayesian methods find their use in many areas of marketing research and we can expect to see their applications in this field more frequently in the future as they will be taken as a primary and not alternative way for analyzing consumer behavior. They offer model-based approach that allows us to creatively assess wide range of phenomena that we are interested in with the possibility to incorporate the prior knowledge we have.

Modeling consumer preferences in conjoint studies is an area where the use of Bayesian methods have already proved its strengths over formerly used methodologies and became popular among researchers. Not only can we estimate more precisely what we had already estimated before but it has been found that we can also incorporate additional pieces in our models that can make them better for our particular application. It is worthwhile to explore more possibilities in our future research.

References

1. Allenby, G.M., Rossi, P.E.: Marketing Models of Heterogeneity. Journal of Econometrics 89 (1999)
2. Allenby, G.M., Rossi, P.E.: Bayesian Statistics and Marketing. Marketing Science 22, (2003)
3. Allenby, G.M., Rossi, P.E.: Perspectives Based on 10 Years of HB in Marketing Research. Sawtooth Software (2003)
4. Frank R.E., Massy, W.F, Wind, Y.: Market Segmentation. Prentice Hall, London (1972) ISBN: 978-0135575796
5. Lenk, P.J., De Sarbo, W.S.: Bayesian Inference for Finite Mixture of Generalized Linear Models With Random Effects. Psychometrica 65 (2000)
6. Orme, B.: Hierarchical Bayes: Why All the Attention? Quirk's Marketing Research Review (2000)
7. Orme, B.: New Advances Shed Light on HB Anomalies. Sawtooth Software Inc. (2003)
8. Orme, B., Howell, J.: Prediction Application of Covariates within Sawtooth Software's CBC/HB Program: Theory and Practical Example. Sawtooth Software Inc. (2009)
9. Sawtooth Software Inc.: The CBC Latent Class Technical Paper (2004)
10. Wedel, M., Kamakura, W.: Market Segmentation – Conceptual and Methodological Foundations. Kluwer Academic Publishers, Massachusetts (2000) ISBN: 978-0792386353

The impact of human capital on productivity in all industries in the Czech Republic

Kristýna Vltavská

Dept. of Economic Statistics, W. Churchill Sq. 4 ,
13067 Prague, Czech Republic
kristyna.vltavska@vse.cz

Abstract. The paper focouses on human capital and its impact on productivity in all industries in the Czech Republic. It concentrates on development in the period between the years 2002 and 2006. The article implicitly examines the importance of investments in human capital for increasing both national and international competitiveness of industries. Employing decomposition of the contribution of labour services into the contribution of hours worked and the contribution of labour composition one can find out the impact of labour skills on the productivity.

Keywords: labour composition, labour services, total factor productivity

1 Introduction

Total factor productivity is one of the most suitable indicators to evaluate economic performance. In measuring productivity the labour input reflects the work time, effort and skills of the workforce. While the data on hours worked capture the time dimension, they do not reflect the skill size. This is achieved by using the labour services as the labour input. Using the decomposition of the labour services into labour composition and hours worked or number of employed one can find out the impact of labour skills and one can divide the industries into low-skilled, medium-skilled and high-skilled ones.

The aim of this paper is to take into account the influence of human capital (measured as the level of education) on total factor productivity in Czech industries in the period of time between 2002 and 2006.

2 Standard computation of TFP

Index of productivity of two factors (capital and labour) by the gross value added ($=A_1/A_0$) is computed by indices of product (Y), capital (K) and labour (L):

$$\frac{Y_1}{Y_0} = \frac{A_1}{A_0}\left(\frac{K_1}{K_0}\right)^{1-\alpha}\left(\frac{L_1}{L_0}\right)^{\alpha} \qquad (1)$$

where Y_1/Y_0 is the index of value added,
K_1/K_0 is the index of fixed assets,
L_1/L_0 is the index of hours worked
α is the average share of compensation of employees on value added.

The analysis uses data from the Czech Statistical Office which involve value added, hours worked and fixed assets in the Czech Republic divided by the industries in the period between 2002 and 2006. As an input of labour the ESA 95 recommended hours worked. The data are divided into 14 industries in this paper. Industry A (agriculture, hunting and forestry) is put together with industry B (fishing) because this industry is too small for an individual analysis. Because of using the multiplicative computation the results do not work out.

Table 1 Calculation of total factor productivity, using hours worked (H) as labour input, total growth from 2002 to 2006 (%)

	VA	**H**	**K**	**TFP**
Total	23.29	1.34	3.61	17.42
A + B Agriculture, forestry, fishing	4.96	-6.77	-1.58	14.39
C Mining and quarrying	-9.95	-4.47	-0.40	-5.36
D Manufacturing	43.77	1.23	7.73	31.82
E Electricity, gas and water supply	25.20	-3.94	-4.69	36.74
F Construction	14.62	2.16	7.66	4.21
G Wholesale and retail trade; repairs	40.63	0.72	7.97	29.32
H Hotels and restaurants	-28.81	2.19	-0.64	-29.89
I Transport, storage and communication	17.85	0.39	5.97	10.78
J Financial intermediation	24.48	-2.27	1.01	26.10
K Real estate, renting and business activities	18.47	4.72	5.02	7.72
L Public administration and defence	0.39	4.14	-0.90	-2.74
M Education	30.14	2.72	0.23	26.41
N Health and social work	-17.44	2.28	4.13	-22.48
O Other community, social and personal service activities	15.35	3.54	13.93	-2.21

Source: Czech Statistical Office, author's computation

Gross value added of the whole economy was growing 23.29% in the period between 2002 and 2006. The main proportion of this was constituted by TFP (17.42%). Fixed

assets and hours worked represent a much lower influence with 3.61% and 1.34% respectively.

3 Alternative computation of TFP

3.1 Data and Methodology

The productivity of various types of labour input is different. This paper distinguishes four levels of education – primary, secondary without A levels, secondary with A levels, tertiary. Standard measures of labour input do not take these differences into account. But it is necessary to measure labour input that takes the diversity of labour force into account when analysing productivity and the contribution of labour to output growth. These measures are called labour services [5], as they allow for differences in the amount of services delivered per unit of labour in the growth accounting approach. It is assumed that the flow of labour services for each labour type is proportional to hours worked, and workers are paid according to their marginal productivities[1]. Then the index of labour services input LS is given by:

$$\Delta \ln LS_t = \sum_l \bar{v}_{l,t} \Delta \ln H_{l,t} \qquad (2)$$

where $\Delta \ln H_{l,t}$ indicates the growth of hours worked by labour type l and weights are given by the average shares of each type in the values of labour compensation. Thus, aggregation takes into account the changing composition of the labour force. A shift in the share of hours worked by low-skilled workers to high-skilled workers will lead to a growth of labour services which is larger than the growth in total hours worked. This difference is called the labour composition effect.

The Czech Statistical Office does not publish the data of hours worked divided by the level of education; that is why the number of employed are used for the further analysis.

Because of using the labour services it is necessary to modify (1) into:

$$\frac{Y_1}{Y_0} = \left(\frac{K_1}{K_0}\right)^{1-\alpha} \left(\frac{LS_1}{LS_0}\right)^{\alpha} \left(\frac{A_1^*}{A_0^*}\right), \qquad (3)$$

where LS_1/LS_0 is the index of labour services.

[1] In the analysis average wages are used instead of marginal productivities because one can assume that the average wages are indicators of marginal productivities.

The index of labour services is divided into the index of number of employed and labour composition:

$$\frac{LS_1}{LS_0} = \left(\frac{L_1}{L_0}\right)\left(\frac{LC_1}{LC_0}\right). \qquad (4)$$

At first, labour services are computed then the development of labour composition as proportion of the development of labour services and the development of number of employed.

3.2 Results

Table 2 Calculation of total factor productivity, using number of employed (L) as labour input, total growth from 2002 to 2006 (%)

	VA	**L**	**K**	**TFP**
Total	23.29	0.41	3.61	18.51
A + B Agriculture, forestry, fishing	4.96	-10.13	-1.58	18.67
C Mining and quarrying	-9.95	-5.91	-0.40	-3.91
D Manufacturing	43.77	1.41	7.73	31.59
E Electricity, gas and water supply	25.20	-2.17	-4.69	34.27
F Construction	14.62	1.15	7.66	5.26
G Wholesale and retail trade; repairs	40.63	0.36	7.97	29.80
H Hotels and restaurants	-28.81	4.80	-0.64	-31.63
I Transport, storage and communication	17.85	-1.37	5.97	12.75
J Financial intermediation	24.48	-3.16	1.01	27.25
K Real estate, renting and business activities	18.47	6.51	5.02	5.91
L Public administration and defence	0.39	-0.26	-0.90	1.55
M Education	30.14	-5.25	0.23	37.05
N Health and social work	-17.44	5.01	4.13	-24.50
O Other community, social and personal service activities	15.35	1.80	13.93	-0.55

Source: Czech Statistical Office, author's computation

After using number of employed as the labour input, the total factor productivity of the whole economy rises up to 18.15% between the years 2002 and 2006. It is caused by the decrease of the index of labour input from 1.34% when using hours worked to 0.41% using number of employed. Again, the multiplicative computation is used; that is why the total sums do not match.

Table 3 Calculation of total factor productivity, using labour services as labour input, total growth from 2002 to 2006 (%)

	VA	LS	L	LC	FA	TFP*
Total	23.29	1.47	0.41	1.07	3.61	17.27
A + B Agriculture, forestry, fishing	4.96	-9.72	-10.13	0.41	-1.58	18.13
C Mining and quarrying	-9.95	-6.12	-5.91	-0.20	-0.40	-3.70
D Manufacturing	43.77	2.78	1.41	1.37	7.73	29.84
E Electricity, gas and water supply	25.20	-1.12	-2.17	1.05	-4.69	32.84
F Construction	14.62	1.67	1.15	0.53	7.66	4.71
G Wholesale and retail trade; repairs	40.63	1.07	0.36	0.71	7.97	28.88
H Hotels and restaurants	-28.81	4.13	4.80	-0.67	-0.64	-31.19
I Transport, storage and communication	17.85	-0.65	-1.37	0.72	5.97	11.94
J Financial intermediation	24.48	-3.50	-3.16	-0.34	1.01	27.70
K Real estate, renting and business activities	18.47	6.95	6.51	0.44	5.02	5.48
L Public administration and defence	0.39	2.55	-0.26	2.81	-0.90	-1.23
M Education	30.14	-3.49	-5.25	1.76	0.23	34.55
N Health and social work	-17.44	8.13	5.01	3.12	4.13	-26.68
O Other community, social and personal service activities	15.35	1.70	1.80	-0.11	13.93	-0.44

LS = contribution of labour services
L = contribution of number of employed
LC = contribution of labour composition
Source: Czech Statistical Office, author's computation

By using the decomposition of contribution of labour services into contribution of number of employed and contribution of labour composition one can find out how a shift in the portion of number of employed by low-skilled workers to high-skilled workers leads to a growth of labour services that is larger than the growth of number of employed, mainly in industries D, E, L, M and N (and vice versa). The labour composition effect caused movement of more educated workforce from less productive industries to more productive ones.

The value of contribution of labour composition shows that during the period between 2002 and 2006 the level of education grows in following industries: N – Health and

social work (3.12%), L - Public administration and defence (2.81%), M – Education (1.76%), D – Manufacturing (1.37%) and E – Electricity, gas and water supply (1.05%). One can see that the growth is achieved in public sector industries. It is caused by the need to increase the qualification level of workforce of the sectors in question. On the other hand, the level of education falls in these industries: O – Other community, social and personal service activities (-0.11%), C – Mining and quarrying (-0.2%), J – Financial intermediation (-0.34%) and H – Hotels and restaurants (-0.67%).

4 Conclusion

There is an obvious influence of human capital (measured as the level of education) on productivity in the Czech industries between the years 2002 and 2006. The decomposition of contribution of labour services on TFP into contribution of number of employed and contribution of labour composition showed that high-skilled workforce shifted to the more productive industries (e.g. Health and social work, Public administration and defence, etc.) in the examined period of time.

References

1. Czech Statistical Office: National Accounts. Prague, 2009.
2. Czech Statistical Office: Labour Force Survey, 2002 – 2006.
3. Jílek, J., Moravová, J.: Ekonomické a sociální indikátory, Futura, Praha 2007
4. Jílek, J., Vojta, M.: K roli produktivity výrobních faktorů v české ekonomice. In: Statistika 2008, 1, pp. 13--29.
5. O'Mahony, M., Timme,r P. M., van Ark, B.: EU KLEMS Growth and Productivity Accounts: Overview November 2007 Release, www.euklems.net

Sampling Distribution of Some Characteristics of Location

Jana Ištvánfyová, Luboš Marek, Libor Svoboda, Michal Vrabec

The University of Economics in Prague, W.Churchilla 4, Prague 3, Czech Republic
istvanfy@vse.cz, libor.svoboda@vse.cz, marek@vse.cz, vrabec@vse.cz

Abstract. Monte Carlo studies play more increasing role in the establishment and verification of results in both theoretical and applied statistics. The present paper is an attempt to find the approximate sampling distributions of some commonly used and some newly proposed measurement of the central tendency, because their exact distributions have so far not been found, and are highly unlikely to be obtained analytically. From our MC-experiments, performed on each of several distribution functions for various values of sample size n, the probability distribution functions of the measures mentioned above were defined empirically. Further, the values of the expectations, standard errors, skewness and kurtosis coefficients of these measures were calculated.

Keywords: Monte Carlo studies, approximate sampling distributions, cauchy distributions, measurement of the central tendency, mean, standard errors, skewness and kurtosis coefficients

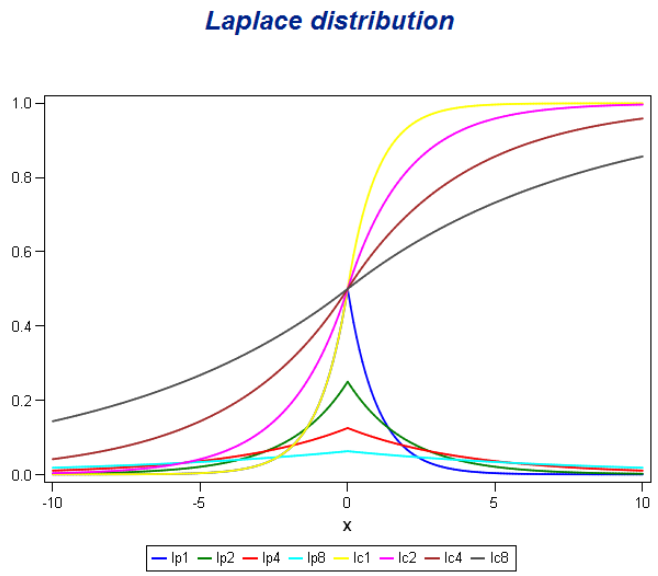
1 Introduction

Though the empirical measures of central tendency such as the median, mid-range (or mid-extreme), interquartile mean, truncated and winsorized mean have long been used in theoretical [4] and applied statistics, no relevant explicit formulae giving the small sample distributions or characteristics thereof have been published, as far as we know. The same is true as far as the empirical measures of relative variability and/or the measures of kurtosis are concerned. It seems that these sampling properties are, in general, in no way so analytically tractable as sample mean, sample variance etc. However, in order that some insight into these properties may be gained, use must be made of the Monte Carlo simulation on a computer, which will yield approximations to these properties probably sufficiently accurate for most operational purposes.

Following the previous paper [2] a [3], we are concerned in the present paper with the evaluation of the numerical values of expectations, variances (standard deviations), coefficients of skewness and coefficients kurtosis for some commonly accepted empirical measures of the central tendency on the assumption that the random sample is generated from some standard parent populations. The comparison

of the expectations of the individual measures will be enable us to establish the rank of these measures from the point of view of their bias, and the comparison of the variances (standard errors) will give us their rank from the point of view of their efficiency.

The sampling properties of every estimator depend upon (1) sample size n, (2) the type of the parent population (distribution function), and (3) the value of the shape parameter. In the present paper, sample sizes n = 3, 7, 15, 23 and 31 were chosen, and the distributions used are some standard probability distributions, partly continuous and partly discrete ones. Six values of the measurement of the central tendency parameter of these distributions were selected such that median, midrange, interquartile (interpentile) range, truncated mean and winsorized mean were used for each distribution. Some of these distributions have no tail, some have a short tail, and some a long tail. In all cases, 5,000 samples were generated and in each sample the values of all the measures investigated were computed. On the one hand, the set of these variate-values was grouped into a frequency distribution, and, on the other hand, the main sample characteristics of these variate-values, i.e. the mean, the standard deviation, skewness and kurtosis characteristics, were computed. Picture 1 shows an example of such empirical density and distribution function (for more scale parameter) a set of values as well as their processing.



Picture 1

2 Theory

Parent populations

In our investigation, use was made of nine parent populations (probability distributions) the desired properties of which were dealt with in the introductory section. Now, we are going to specify these selected populations giving their names.

Continuous distributions:	Cauchy,
	Laplace,
	lognormal,
	normal,
	Weibull.
Discrete distributions:	binomial,
	geometric,
	hypergeometric,
	Poisson.

The measures investigated and their estimators

The measures describing the individual characteristics of the distribution of a random variable X , viz., the measures of location, dispersion and asymmetry (skewness) can be constructed, as it is well-known, on the basis of either the moments or the quantiles, or on the basis of both these quantities.

If central tendency (measure of the "middle" or "expected" value of the data set) of a distribution is measured, several mean may be constructed [1,4,5], such as the well-known **arithmetic mean** (for k=1 and w_i=1/n) , truncated mean (the arithmetic mean of data values after a certain number or proportion of the highest and lowest data values have been discarded), interquartile mean (special case of the truncated mean, using the interquartile range) winsorized mean (similar to the truncated mean, but, rather than deleting the extreme values, they are set equal to the largest and smallest values that remain) and so on. All these characteristics can also weigh.

$$am = \overline{X_w} = \frac{\sum_{i=k}^{n-k+1} w_i X_{(i)}}{\sum_{i=k}^{n-k} w_i}, \; 1 \le k < \frac{n}{2} \tag{1}$$

Median is described as the number separating the higher half of a sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, the median is not unique, so one often takes the mean of the two middle values.

$$P(X \le med) \ge \frac{1}{2} \text{ and } P(X \ge med) \ge \frac{1}{2} \tag{2}$$

Whereas these three coefficients have been used at all extensively in practice, the others are being of theoretical importance only. Nevertheless, we are going to discuss the various measures in turn, including, in addition, new constructions that have recently been presented in statistical literature. Whereas these coefficients have been used at all extensively in practice, the others are being of theoretical importance only. Nevertheless, we are going to discuss the various measures in turn, including, in addition, new constructions that have recently been presented in statistical literature [1,5,7].

Mid-range or mid-extreme

$$mr = \frac{X_{\max} + X_{\min}}{2} \tag{3}$$

where $X_{\max}$ is the maximum (largest value) and $X_{\min}$ is the minimum (smallest value).

Mid-quartile

$$qr = \frac{X_{75} + X_{25}}{2} \tag{4}$$

where $X_{0.25}$ is the lower (first) quartile and $X_{0.75}$ is the upper (third) quartile.

Mid-pentile

$$pr = \frac{X_{80} + X_{20}}{2} \tag{5}$$

where $X_{0.20}$ is the lower (first) pentile and $X_{0.80}$ is the upper (fourth) pentile (any of the four points that divide an ordered distribution into five parts, each containing a fifth (20%) of the population).

Numbers and sizes of samples

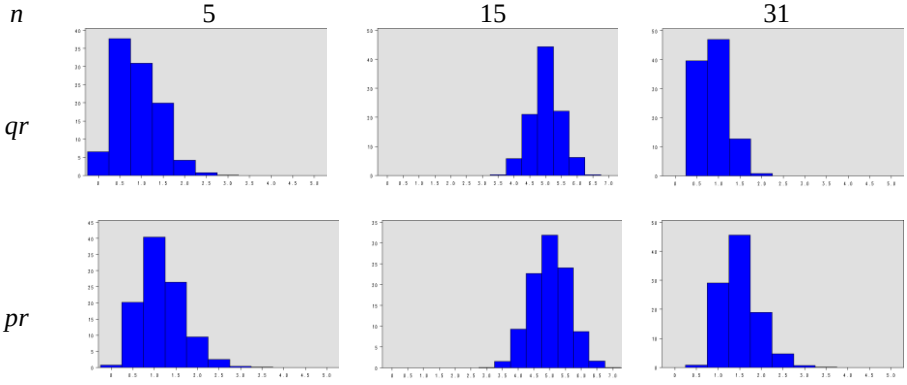
At our experiments were taken 5,000 samples from each distribution (e.g., from the Cauchy distribution with parameter value $\mu = 0$, $\sigma = 1$) and of a given sample size (e.g., $n = 15$). As far as the sample sizes (n) are concerned, they were decided upon, on the one hand, with due account taken of the demand to investigate the sampling properties of each estimator for both small and medium large samples, and, on the other hand, considering the advantage of an unambiguous definition of the quartiles in a finite sample (data ungrouped). The sizes were, therefore, chosen as follows: $n = 3$, 7, 15, and 31 (generally, $n = 2^p - 1$, where $p = 2$, 3, 4, and 5). Later, size $n = 23$ turned out to be useful to add to the foregoing ones. With all these sizes, the quartiles (pentiles) equal to the values on the following units of the sample :

	$n = 3$	$n = 7$	$n = 15$	$n = 23$	$n = 31$
First pentile $\tilde{x}_{20}$	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(5)}$	$x_{(6)}$
First quartile $\tilde{x}_{25}$	$x_{(1)}$	$x_{(2)}$	$x_{(4)}$	$x_{(6)}$	$x_{(8)}$
Second quartile $\tilde{x}_{50}$	$x_{(2)}$	$x_{(4)}$	$x_{(8)}$	$x_{(12)}$	$x_{(16)}$
Third quartile $\tilde{x}_{75}$	$x_{(3)}$	$x_{(6)}$	$x_{(12)}$	$x_{(18)}$	$x_{(24)}$
Fourth pentile $\tilde{x}_{80}$	$x_{(3)}$	$x_{(6)}$	$x_{(13)}$	$x_{(19)}$	$x_{(26)}$

Sampling distributions of estimators

The basic aim of our research was to provide "data" for assessing the sampling distributions of central tendency estimators. Hence, sequences of 5,000 values of each of those statistics were ordered into frequency distributions and the asked sampling distributions were thus defined empirically. Nevertheless, they of course enable us to assess the degree of normality - see the example (binomial distribution) in Table 1.

Table 1. Frequency distributions mid-quartile and mid-pentile (depending on *n*).



The second way of evaluating the provided sequences of estimates consisted in the computation of their arithmetic means, variances and coefficients of skewness and in their use as estimates of the true expected values, $\mathbf{E}(\bullet)$, true variances, $\mathbf{D}^2(\bullet)$, and coefficients of skewness $\alpha(\bullet)$.

For example, when estimator $\hat{\tau}$ was evaluated, the following formulae were used

(a) the estimated bias

$$\hat{\mathbf{B}}\ \hat{\tau}\ = \hat{\mathbf{E}}\ \hat{\tau}\ - \tau = \frac{1}{5000}\sum^{5000}\hat{\tau} - \tau,$$

(b) the estimated variance

$$\hat{\mathbf{D}}^2\ \hat{\tau}\ = \hat{\mathbf{E}}\ \hat{\tau} - \hat{\mathrm{E}}\ \hat{\tau}\ \ ^2 = \frac{1}{4999}\sum^{5000}\left(\hat{\tau} - \frac{1}{5000}\sum^{5000}\hat{\tau}\right)^2,$$

and its square root, i.e. the estimated standard deviation (standard error) $\hat{\mathbf{D}}\ \hat{\tau}$,

(c) classical coefficient skewness (the third moment of normalized variate)

$$\hat{\alpha}\ \hat{\tau}\ = \frac{\hat{\mathrm{E}}\ \hat{\tau} - \hat{\mathrm{E}}\ \hat{\tau}\ ^3}{\hat{\mathrm{D}}^2\ \hat{\tau}\ ^{3/2}} = \frac{\sqrt{n\ \ n-1}}{n-2}\ \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}},$$

(d) classical coefficient kurtosis (the third moment of normalized variate)

$$\hat{\beta}\ \hat{\tau}\ = \frac{m_4}{m_2^2} - 3 = \frac{\hat{\mathrm{E}}\ \hat{\tau} - \hat{\mathrm{E}}\ \hat{\tau}\ ^4}{\hat{\mathrm{D}}^2\ \hat{\tau}\ ^2} - 3 = \frac{(n+1)\,n\,(n-1)}{(n-2)(n-3)}\ \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} - 3\frac{(n-1)^2}{(n-2)(n-3)}.$$

The bias was evaluated both absolutely and with respect to standard error, i.e. as the rate $\hat{\mathrm{B}}\ \hat{\tau}\ /\hat{\mathrm{D}}\ \hat{\tau}\ $. The same characteristics were computed for all the other estimators.

For every set of samples the rank of five estimators of central tendency investigated was determined from all the four points of view. Moreover, each aspect was applied to both the mutual comparison of successive statistics (estimators) and the "dynamic" comparison of one statistic, i.e. how its properties change as sample size *n* grows and parameter(s) of the parent population (distribution function) changes its values.

3 Illustrations - selected results[1]

Cauchy distribution[2]

Density function	$f(x;\mu,\sigma) = \frac{1}{\pi}\left[\frac{\sigma}{(x-\mu)^2+\sigma^2}\right]$
Distribution function	$F(x;\mu,\sigma) = \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right)+\frac{1}{2}$
Characteristic function	$\phi_X(t;\mu,\sigma) = \mathrm{E}(e^{iXt}) = \exp(i\,\mu t - \sigma\,\lvert t\rvert)$
Inverse cumulative distribution function	$F^{-1}(p;\mu,\sigma) = \mu + \sigma\tan\left[\pi\ \ p-\frac{1}{2}\right]$

[1] In this paper only a part of the full MC-results have been presented, and the interested reader is referred to an unpublished manuscript for fuller details.

[2] Without a defined mean, it is impossible to consider the variance or standard deviation of a standard Cauchy distribution, as these are defined with respect to the mean. But the second moment about zero can be considered. It turns out to be infinite.

Table 2. Main Statistical Characteristics (Cauchy distribution)

Mean

n	*am*	*med*	*mr*	*pr*	*qr*	*tav*[3]	*wav*[4]
3	3,178	-0,020	8,767	14,777	4,777	1,552	0,859
7	-1,865	-0,010	-6,537	-6,537	0,037	1,229	0,659
15	0,654	-0,007	4,903	4,903	0,007	1,125	0,451
31	-0,200	0,002	-1,445	0,006	0,004	0,521	0,345

StdDeviation

n	*am*	*med*	*mr*	*pr*	*qr*	*tav*	*wav*
3	151,122	2,081	226,61	216,068	24,406	3,003	0,859
7	58,146	0,774	203,19	203,195	2,929	1,821	0,659
15	46,340	0,447	346,90	46,903	0,828	0,936	0,451
31	145,978	0,292	220,96	0,578	0,473	0,300	0,345

Skewness

n	*am*	*med*	*mr*	*pr*	*qr*	*tav*	*wav*
3	50,74	-7,402	50,76	50,762	50,762	16,500	3,479
7	-31,36	-0,004	-31,49	-31,486	5,028	5,519	1,582
15	6,59	-0,249	6,64	6,640	-0,402	1,600	1,043
31	-39,97	-0,036	-40,08	-0,114	-0,110	1,497	0,794

Kurtosis

n	*am*	*med*	*mr*	*pr*	*qr*	*tav*	*wav*
3	2713,18	192,618	2714,64	2714,63	2714,63	346,763	23,895
7	1210,80	6,553	1217,52	1217,52	126,30	264,453	21,386
15	1089,05	2,329	1097,21	1097,20	9,48	204,728	18,192
31	1991,21	0,633	1996,39	2,43	1,59	156,290	3,849

4 Conclusions

Firstly, it is worth to note that the relations

$$\left|\widehat{\mathbf{B}}\bullet_{31}\right| < \left|\widehat{\mathbf{B}}\bullet_{15}\right| < \left|\widehat{\mathbf{B}}\bullet_{7}\right| < \left|\widehat{\mathbf{B}}\bullet_{3}\right|$$

between all type measures hold in the case of the continuous distributions.

[3] trimmed mean

[4] Winsorized mean

Secondly, estimators, investigated in this study, the following inequalities hold in almost all cases:

for distributions that are have long tail

$$\left|\hat{\mathbf{B}}\ med\right| \le \left|\hat{\mathbf{B}}\ wav\right| \le \left|\hat{\mathbf{B}}\ tav\right| \le \left|\hat{\mathbf{B}}\ qr\right| \le \left|\hat{\mathbf{B}}\ am\right|,$$

for symmetrical distributions

$$\left|\hat{\mathbf{B}}\ med\right| \le \left|\hat{\mathbf{B}}\ am\right| \le \left|\hat{\mathbf{B}}\ wav\right| \le \left|\hat{\mathbf{B}}\ tav\right| \le \left|\hat{\mathbf{B}}\ qr\right|.$$

Secondly, estimators, investigated in this study, the following inequalities hold in almost all cases:

References

1. Coles, S.: An Introduction to Statistical Modeling of Extreme Values,. Springer-Verlag, New York-Heidelberg-Berlin (2001). ISBN 978-1852334598.
2. Čermák, V., Vrabec, M.: Sampling distributions of some measures of skewness: Results of a MC-simulation. Acta Oeconomica Pragensia (1998), vol. 6, pp. 35 – 48. RIV J18/98 311400026.
3. Čermák, V. Lauber, J., Vrabec, M.: Sampling distributions of some measures of rel-variability: Results of a MC-simulation. In Statistical Modelling in Economic Practice. Acta Oeconomica Pragensia (1997), vol. 7, pp. 15 – 55.
4. Embrechts, P., Klüppelberg C., Mikosch T.: Modelling extremal events for insurance and finance. Springer-Verlag, New York-Heidelberg-Berlin (2008), corr. 4th printing. ISBN: 978-3540609315.
5. Fisher, R. A.: Moments and product moments of sampling distributions. Proc. London Math. Soc., (1930); s2-30: 199 - 238.
6. Leadbetter, M.R., Lindgreen, G., Rootzén, H.: Extremes and related properties of random sequences and processes. Springer-Verlag, New York-Heidelberg-Berlin 1984. ISSN 0026-1335.
7. Resnick, S.I.: Extreme values, regular variation and point processes. Springer-Verlag, New York-Heidelberg-Berlin, 1st ed. 1987. 2nd printing 2008, ISBN: 978-0387759524.

The Settlement Process and Its Properties

Pavel Zimmermann[1]

[1] University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3,
Czech Republic
zimmerp@vse.cz

Abstract. This article will studies the so called settlement process which is the process through which claims pass in an insurance company while being handled by the claim adjusters. This process describes how the incurred value of a claim evolves between the reporting and its settlement. In the next section, some modeling ideas are demonstrated on liability data with real market properties. Although the actual real values itself are not published, the techniques used as well as the observed properties are believed to be transferable on other real portfolios.

Keywords: Settlement process, Reserve Risk.

1 Introduction

In the recent years there has been an enormous growth of demand for new modeling techniques to evaluate insurance risk of the insurance company. Most of the current models (e.g. [2]) are based on so called triangle schemes, i.e. based on extrapolating the development of aggregated loss from past claims. These models generally do not allow capturing many individual properties of claims such as the limits. Also, modeling many popular reinsurance programs, such as excess of loss, is difficult. Therefore it seems advantageous to develop a model that would be based on development of the individual losses.

Modeling the so called reserve risk, i.e. basically the insurance risk connected with claims which already occurred in the past but have not yet been settled, based on individual claim level seems to be relatively complicated due to the fact that each modeled claim has its own unique handling history. Therefore it seems to be necessary to understand the process which a claim passes through in the insurance company in between its reporting and settlement. This process will be referred to as the settlement process.

Few theoretic articles have been published on this topic previously. The theoretical background of the individual claim level models was set in the article [5]. This paper is a common reference to most of the consecutive articles. The author considers a fully time-continuous model. He suggests using the marked Poisson process as a model of the claims process. This theoretical framework was followed for example in the article [3] which is slightly more practical and also contains an application. An

interesting method of individual claims modeling was also published in [5]. This method is however specially designed for projecting only a limited number of very large claims. Both of the latter articles recognize that the settlement process is compounded of two groups of random components:

a) Differences of the incurred value over consecutive development periods.
b) Opened / closed claim indicators in consecutive development periods.

An interesting paradox between the theoretical models which are time continuous and the practical models which only work with discrete time periods (usually development years) occurs. The first contribution of this article is that the process is not modeled over consecutive development periods but the so called stages of the process are defined. Stages basically correspond with the period between two incurred value changes. Second contribution of this article is that the claim closure probability is not considered to be a constant, instead it is considered being dependent on the current incurred value of the claim which is quite natural property which can be expected in many practical cases.

We will first try to formulate this process in a correct way. The distribution of the ultimate severity (the ultimate incurred value) will be derived. Later we will investigate some properties of the settlement process based on data with real market properties.

2 The Settlement Process

In this section, we try to introduce some basic terms and define the process in a correct way. When a claim is reported to the insurance company it enters the so called 'settlement process'. At first, the claim adjusters set up a reserve which corresponds to their initial information about the claim. During the settlement process adjusters adjust their view on the severity of the claim as they are collecting new information arriving to the company or as they perform their own investigation. Whenever they feel that the actual reserve is inappropriate, they adjust it. At some point (when both the insurer and insured come to an agreement or a trial result), the claim is settled and paid out. The time between two changes of the reserve will be referred to as the stage of the settlement process. In this article it will be assumed that all claims are paid at once (i.e. as a lump sum) and no re-openings are possible. This obviously excludes annuities which require rather different methods and models which are off the scope of this paper.

Unless stated otherwise, the random variables will be denoted with the capital letters and its values with corresponding lower case letters. The settlement process is a marked process $\{X_s, T_s, A_s s \geq 1\}$, where X_s is the change of reserve in the s-th stage of the settlement process, T_s is the corresponding time point and A_s is dichotomous variable (a 'flag') indicating whether the claim is opened ($A_s = 1$) or closed ($A_s = 0$). Formally we put $X_s = 0$ and $T_s = \infty$ for the closed claims, i.e. for stages s such that $A_s = 0$. The cumulative process

$$Y(t) = \sum_{T_s \le t} X_s \tag{1}$$

will be referred to as the process of the incurred value.

3 The Ultimate Incurred Value

From the above text, it is obvious that the incurred value is constant within each stage of the settlement process, i.e.

$$Y(t) = Y_s, \text{ for } T_s \le t < T_{s+1} \tag{2}$$

where Y_s is defined as

$$Y_s = \sum_{k=1}^{s} X_k \ . \tag{3}$$

where the index *k* corresponds with the time order of the adjustment.

The most important variable for the reserve risk modeling is the severity of the claim which in this set up corresponds with the ultimate incurred value, i.e. $Y(\infty)$. In this section we will investigate the distribution of this variable. It is obvious that there is only a finite number of nonzero adjustments before the claim is settled. This number will be denoted as $\bar{S}$. Since after the settlement the incurred value is not changing anymore the equation $Y(\infty) = Y_{\bar{S}}$ holds. An illustration of the settlement process can be found in the figure 1. This means that there are two sources of randomness in the ultimate incurred value:

1. The number nonzero adjustments until the settlement $\bar{S}$.
2. The actual values of the adjustments X_s, $s = 1,...,\bar{S}$.

We first put some restrictive but in many cases realistic assumptions on the process which allow keeping the complexity of the model at a reasonable level. Namely, we will assume that

1. The probability that a claim will be closed in the *s*-th stage given that it was opened in the (*s-1)*-th stage depends only on the incurred value of the claim in the *s*-th stage.
2. The distribution of the adjustment X_s depends only on the incurred value in the previous stage Y_{s-1}
3. The maximum possible number of nonzero adjustments for each claim is known and will be denoted as *m*. This number is practically of course unknown but can be easily estimated as 'sufficiently' large. In most practical tasks, *m=11* stages should be sufficient.

We will first consider the (ultimate) incurred value to be a discrete variable (table losses) which will make all the following equations more transparent. Under these assumptions we can write for the ultimate incurred value

$$P(Y_{\bar{s}} = y) = \sum_{s=1}^{m} P(\bar{S} = s, Y_s = y) = \sum_{s=1}^{m} P(\bar{S} = s \mid Y_s = y) P(Y_s = y) \tag{4}$$

We can rewrite this in the terms of the adjustments X_s as

$$\begin{aligned} &P(Y_{\bar{s}} = y) = P(\bar{S} = 1 \mid X_1 = y) P(X_1 = y) + \\ &P(\bar{S} = 2 \mid X_1 + X_2 = y) P(X_1 + X_2 = y) + ... + \\ &P(\bar{S} = m \mid X_1 + X_2 + ... + X_m = y) P(X_1 + X_2 + ... + X_m = y). \end{aligned} \tag{5}$$

The distributions of the sums of the adjustments can be further written using the following 'convolution like' formulas and plugged back in the equation (4):

$$\begin{aligned} &P(Y_{\bar{s}} = y) = \\ &= \sum_{s=1}^{m} \sum_{x_1} \sum_{x_2} ... \sum_{x_{s-1}} P(\bar{S} = s \mid X_1 = x_1, X_2 = x_2, ..., X_s = y - x_1 - x_2 - ... - x_{s-1}) \\ &\qquad P(X_1 = x_1, X_2 = x_2, ..., X_s = y - x_1 - x_2 - ... - x_{s-1}). \end{aligned} \tag{6}$$

2.1 The probability of a settlement

In this section we will investigate further the probability that a claim will be settled in a particular stage *s*, i.e. the probability

$$P(\bar{S} = s \mid X_1 = x_1, X_2 = x_2, ..., X_s = y - x_1 - x_2 - ... - x_{s-1}) \tag{7}$$

We can write for this probability using the above mentioned 'flags' A_s :

$$\begin{aligned} &P(\bar{S} = s \mid X_1 = x_1, X_2 = x_2, ..., X_s = y - x_1 - x_2 - ... - x_{s-1}) = \\ &= P(A_1 = 1, A_2 = 1, ..., A_{s-1} = 1, A_s = 0, A_{s+1} = 0, ..., A_m = 0 \mid X_1 = x_1, X_2 = x_2, ... \\ &\quad ..., X_s = y - x_1 - x_2 - ... - x_{s-1}). \end{aligned} \tag{8}$$

Furthermore we can write this equation using the conditional probabilities.

$$\begin{aligned} &P(\bar{S} = s \mid X_1 = x_1, X_2 = x_2, ..., X_s = y - x_1 - x_2 - ... - x_{s-1}) = \\ &= P(A_1 = 1 \mid X_1 = x_1, X_2 = x_2, ..., X_s = y - y_{s-1}) \\ &P(A_2 = 1 \mid X_1 = x_1, X_2 = x_2, ..., X_s = y - y_{s-1}, A_1 = 1) ... \\ &P(A_{s-1} = 1 \mid X_1 = x_1, ..., X_s = y - y_{s-1}, A_1 = 1, A_2 = 1, ..., A_{s-2} = 1) ... \\ &P(A_s = 0, A_{s+1} = 0, ..., A_m = 0 \mid X_1 = x_1, ..., X_s = y - y_{s-1}, A_1 = 1, ..., A_{s-1} = 1) \end{aligned} \tag{9}$$

We can now simplify the notation using:

$$P(A_s = 0, A_{s+1} = 0,..., A_m = 0 \mid X_1 = x_1,..., X_s = x_s, A_1 = 1,..., A_{s-1} = 1) = \\ = q_s(x_1, x_2,..., x_s) \quad (10)$$

and

$$P(A_s = 1 \mid X_1 = x_1,..., X_s = x_s, A_1 = 1,..., A_{s-1} = 1) = z_s(x_1, x_2,..., x_s), \quad (11)$$

According to the assumption 1, the probabilities of settlement (and therefore the probabilities of remaining opened) depend only on the last incurred value (and not on the whole 'trajectory' of the previous adjustments). This means that

$$q_s(x_1, x_2,..., x_s) = q_s(y_s) \quad (12)$$

and

$$z_s(x_1, x_2,..., x_s) = z_s(y_s) \quad (13)$$

where $q_s(y_s)$ is the probability that a claim will get settled in the *s*-th stage given the incurred value in this stage, and $z_s(y_s)$ is the probability that a claim remains opened in the *s*-th stage given the incurred value in this stage. Notice also that

$$z_s(y_s) = 1 - q_s(y_s). \quad (14)$$

If we plug this back in the equation (9), we get

$$P(\overline{S} = s \mid X_1 = x_1, X_2 = x_2,..., X_s = y - x_1 - x_2 - ... - x_{s-1}) = \\ z_1(y_1) z_2(y_2) \cdot ... \cdot z_{s-1}(y_{s-1}) q_s(y_s). \quad (15)$$

2.2 The Distribution of the Adjustments

Similar algebra can be done with the distribution of the adjustments:

$$P(X_1 = x_1, X_2 = x_2,..., X_s = y - y_{s-1}) = \\ P(X_1 = x_1) P(X_2 = x_2,..., X_s = y - y_{s-1} \mid X_1 = x_1) = \\ P(X_1 = x_1) P(X_2 = x_2 \mid X_1 = x_1) P(X_3 = x_3,..., X_s = y - y_{s-1} \mid X_1 = x_1, X_2 = x_2) \quad (16)$$

For this equation we can write using the incurred values Y_s:

$$P(X_1 = x_1, X_2 = x_2,..., X_s = y - y_{s-1}) = \\ P(Y_1 = y_1) P(Y_2 - y_1 = x_2 \mid Y_1 = y_1) P(Y_3 - y_2 = x_3 \mid Y_1 = y_1, Y_2 = y_2)... \\ P(Y_s - y_{s-1} = y - y_{s-1} \mid Y_1 = y_1, Y_2 = y_2,..., Y_{s-1} = y_{s-1}) \quad (17)$$

Using the assumption 2, we can write

$$P(X_1 = x_1, X_2 = x_2,..., X_s = y - y_{s-1}) = \\ P(Y_1 = y_1)P(Y_2 - y_1 = x_2 \mid Y_1 = y_1)P(Y_3 - y_2 = x_3 \mid Y_2 = y_2)... \\ P(Y_s - y_{s-1} = y - y_{s-1} \mid Y_{s-1} = y_{s-1}). \tag{18}$$

We can again simplify the notation using

$$p_1(y_1) = P(Y_1 = y_1) \tag{19}$$

and

$$p_s(y_s \mid y_{s-1}) = P(Y_s = y_s \mid Y_{s-1} = y_{s-1}). \tag{20}$$

If we plug this back in the equation (16), we get

$$P(X_1 = x_1, X_2 = x_2,..., X_s = y - y_{s-1}) = p_1(y_1)p_2(y_2 \mid y_1) \cdot ... \cdot p_s(y \mid y_{s-1}) \tag{21}$$

2.3 The Distribution of the Ultimate Incurred Value

Now we can finally write the formula for the distribution of the ultimate incurred value:

$$P(Y_{\bar{S}} = y) = \sum_{s=1}^{m} \sum_{y_1} z_1(y_1)p_1(y_1) \sum_{y_2} z_2(y_2)p_2(y_2 \mid y_1)... \\ \sum_{y_{s-1}} z_{s-1}(y_{s-1})p_{s-1}(y_{s-1} \mid y_{s-2})q_s(y)p_s(y \mid y_{s-1}). \tag{22}$$

We can use the same algebra to derive a similar formula for the case of the continuous (ultimate) incurred value:

$$f_{\bar{S}}(y) = \sum_{s=1}^{m} \int_{\Omega_1} z_1(y_1)f_1(y_1) \int_{\Omega_2} z_2(y_2)f_2(y_2 \mid y_1)... \\ \int_{\Omega_{s-1}} z_{s-1}(y_{s-1})f_{s-1}(y_{s-1} \mid y_{s-2})q_s(y)f_s(y \mid y_{s-1})dy_{s-1}dy_{s-2}...dy_1 \tag{23}$$

where $f_s(y \mid y_{s-1})$ denotes the corresponding transition density distribution and Ω_s is the corresponding domain of definition of the variable y_s.

Notice that the main reason to derive the formulas (22) and (23) is that that the terms of these formulas are practically estimateable. Unfortunately finding a closed form formula for the distribution (23) will in general case be impossible. A closed form formula can be found only under some relatively restrictive assumptions such as assumption that all the probabilities of a claim remaining opened $z_s(y_s)$ are constants

(i.e. independent of the incurred value y_s). In general, either Monte Carlo simulations will have to be used or it will be needed to perform a discretization and use the formula (22).

3 Observed Properties of the Adjustments

In this section we will have a closer look on the real properties of the adjustments. The data used have real properties (real data with some noise). To avoid handing over information that could be used by the competition, specific values are mostly not published. Instead a graphical representation of the results is presented. We believe that the actual values are not really important. Instead we focus on properties or techniques that could transferable to other portfolios.

On purpose we selected a group of claims that appears to be one of the most complex to model – the liability bodily claims. The maximum possible number of nonzero adjustments for each claim was set to *m = 11*. First we will concentrate on modeling the distribution of the incurred value adjustments $X_s = Y_s - Y_{s-1}$, that is the distribution $f_s(y \mid y_{s-1})$. For this purpose we will use the generalized linear models. Namely we will assume (similarly as in [3]) the logarithmic link function (i.e. the multiplicative structure) with the gamma error function.

The data considered is the value in the stage *s* (Y_s) as the response variable, the value in the stage *s-1* (Y_{s-1}), and the stage *s* in which was the adjustment $X_s = Y_s - Y_{s-1}$ observed as the predictors. Since the main purpose of this article is to reveal the basic properties and the behavior of the variables in the settlement process, we used a similar method as suggested in [1]. Namely we discretized the predictor Y_{s-1} into categories. This way we get in fact a nonparametric estimate of the impact of this predictor on the response variable Y_s. The discretization of Y_{s-1} was performed on the logarithmic scale i.e. the length of the intervals is exponentially growing on the original scale. The (unconditional) distribution of the first adjustment $Y_1 = X_1$ involves only well known statistical methods - there is no "special" technique for this in respect of the settlement process. Therefore we focus in the following text only on modeling the conditional ('transition') distributions starting from *s=2*. That means we first construct a model of the form:

$$E(Y_{si}) = \exp(\eta_{si}) \tag{24}$$

where

$$\eta_{si} = \beta_0 + \sum_j y_{s-1,j,i}\beta_j + \sum_k s_{k,i}\alpha_k + \sum_{j,k} y_{s-1,j,i}s_{k,i}\gamma_{j,k} \tag{25}$$

where β_0 is the intercept, $y_{s-1,j,i}$ is the dummy variable indicating whether the *i*-th claim was in the stage (*s-1*) in the *j*-th category of the incurred value, $s_{k,i}$ is the dummy variable indicating in which stage *s* was the observation observed, the last term incorporates the impact of the interactions and $\beta_j, \alpha_k, \gamma_{j,k}$ are the parameters.

The parameter estimates of the GLM assigned to the dummy variables corresponding with the stages of the process of the incurred value can be found in the figure 1 in the appendix. We can observe that at the first sight the stage itself does not really have that much impact on the conditional mean of Y_s given Y_{s-1}, i.e. on the mean of X_s. The reference category was chosen to be the category '*stage 2*' which denotes the adjustment from the first stage to the second stage. The only significant difference from the reference category was observed for the category '*stage 3*' (significance 0,001) and 'almost' for the '*stage 5*' (significance equals to 0,069). Otherwise no significant differences were measured. Notice that the parameter estimates for the category '*stage 3*' and '*stage 5*' are very similar. This suggests that there are two different groups of stages: *Stage 3* and *Stage 5* in one group and all other stages in another group.

Because of these results we will fit a new model with a predictor '*StageII*' which is a transformed (recoded) original variable '*stage*'. The variable *StageII* is dichotomous only. Namely *StageII =1* for *s=3* and *s=5* and *StageII =0* for all other stages. This time however, the model was constructed with both main effects (Y_{s-1} and *stageII*) as well as the first order interaction. The results of the Wald's test of the model effects are displayed in the table 1.

Table 1. The results of the Wald's test of the model effects.

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	161414,1	1	,000
Y_{s-1}	6219,2	15	,000
StageII	7,0	1	,008
Y_{s-1} * *StageII*	41,8	15	,000

The table suggests that this time all predictors (including the interaction) significantly contributed to the likelihood of the model. The estimate of the parameter assigned to the dummy variable corresponding with '*StageII=1*' was positive and significant (the actual value could not be published) suggesting that in general, the value of Y_s (given the value of Y_{s-1}) is on average higher in the *stages 3* and *5* than in the other stages. An interesting insight in this difference is provided by the estimates of the parameters assigned to the interaction dummy variables. The interaction parameters can be found in the figure 3. As we can see the significant interaction parameter estimates were observed for the (discretized) values of Y_{s-1} approximately between 20 000 and 120 000 Czk, i.e. the medium sized claims. (Notice that the horizontal axis is in the logarithmic scale.) These significant

parameter estimates are negative. Notice, that the general impact of the main effect '*StageII=1*' was positive. This in fact means that this general impact of the main effect is 'reduced' by the negative interaction effect for the medium values of Y_{s-1}. I.e. the resulting estimates of Y_s will be different for the lowest and larger claims for the two '*StageII*' categories. For the medium sized claims, large difference between '*StageII=1*' and '*StageII=0*' is not expected.

The parameter estimates of the dummy variables corresponding with the categories of the incurred value in the stage *s-1* can be found in the figure 4. The results are not really surprising. The fact that Y_{s-1} is a significant factor for estimating Y_s was quite expected. The questions to answer will however be: 'For what size of the Y_{s-1} do the claims tend to increase / decrease?' As we already mentioned in the previous paragraph, this tendency has to be studied separately for the two categories of *StageII*. To uncover these tendencies, the predicted values of Y_s (denoted as $\hat{Y}_s$) for each class (i.e. for each combination of the values of the factors Y_{s-1} and *stageII*) were compared to the corresponding value of Y_{s-1}. Since the variable Y_{s-1} was categorized, we have to pick a value representing each category. The observed mean value of Y_{s-1} (denoted as $\overline{Y}_{s-1}$) was considered for each category of Y_{s-1}. Two charts were constructed for this purpose. On the figure 5 we can find the prediction of Y_s on the vertical axis and the corresponding values of $\overline{Y}_{s-1}$ on the horizontal axis. Both axes are in the logarithmic scale. In the figure 6, similar comparison is performed on the relative numbers. Namely vertical axis contains the predicted values $\hat{Y}_s$ divided by the corresponding value of $\overline{Y}_{s-1}$, and the horizontal axis contains again the values of $\overline{Y}_{s-1}$. The horizontal axis is in the logarithmic scale. The estimated relative changes $\hat{Y}_s / \overline{Y}_{s-1}$ can be thought of as individual development factors of the claims. The resulting charts suggest that:

1. Smallest claims tend to grow (i.e. are under-reserved) especially in the stage 3 and 5 (corresponding with the value *stageII = 1*).
2. We can also observe the result mentioned in the previous paragraph, i.e. that the difference between the stages is negligible for the medium sized claims.
3. Largest claims tend to grow in the stage 3 and 5 (*stageII* = 1) while as they tend to slightly decrease in the other stages.

The above mentioned results only concerned with the conditional expected value of Y_s. As stated before, the multiplicative structure with the gamma distributed error term was considered. Notice that in this setup the variance of the error term is a function of the (square) of the conditional mean (see e.g.[4] for details). That also means that the variance changes with the value of Y_{s-1} because this value determines the response variable. Notice however that a systematic impact of stages on the conditional variance cannot be captured by this model if there is no systematic impact of the stages on Y_s. This aspect could be a matter of the further research since a systematic impact of the stages on the variance would in this case be natural since one

would expect that the adjustments in the later stages should more or less correspond just with some fine tuning of the claim adjusters' estimate while as in the earlier stages, certain large changes can be expected.

4 Observed Properties of the Settlement Probability

In this section we will study the observed properties of the probability of the claim settlement (given that the claim was opened), that is the component $z_s(y_s)$ of the distribution function of the ultimate incurred value $f_{\bar{S}}(y)$. As stated above, it is assumed in this article that this probability depends in a given stage only on the incurred value in this stage. The data available consists of:

1. A dichotomous explanatory variable A_s. $A_s = 0$ denotes claims settled in the *s*-th stage and $A_s = 1$ denotes claims that remained opened in the given stage;
2. The corresponding stage *s;*
3. And the incurred value Y_s in the given stage.

This setup implies that the suitable model will be also generalized linear model, this time with the binomial error term and the logit link function, i.e. the logistic regression. The odds ratio estimates (i.e. the exponentials of the parameter estimates) can be found in the figure 7 and 8. The charts displays the impact of the stage *s* or incurred value Y_s on the odds of a claim staying opened, i.e. on the measure

$$odds_s(y_s) = \frac{z_s(y_s)}{q_s(y_s)}. \tag{26}$$

relatively to the odds of the reference category.

We will first study the impact of the incurred value. The reference category was chosen to be the first category, i.e. claims with the incurred value between 0 and 8 000 Czk. The impact of the last category, i.e. claims higher than 14,6 mil. Czk, could not really be considered as reliable due to lack of data. It also didn't prove to be significant. All other parameter estimates were significantly different from zero therefore we can conclude that the odds systematically differ from the odds in the reference category. The estimates suggest that the probability of a claim remaining opened is systematically growing with the growing incurred value. This quite corresponds with practical expectations since it is quite natural that larger claims tend to pass through more stages of the settlement process as there is much more to investigate than for the small claims (e.g. there are more objects or victims to visit etc.).

The impact of the stage is on the other hand somewhat surprising. The reference category was chosen to be the stage 1. All parameter estimates except for the parameters assigned to the dummy variables corresponding with the stage 10 and 11 are statistically significant. This suggests that the odds are different for different stages. The estimates suggest that at first the relative difference in odds between the stage 1 and stage *s* increase with the increasing *s*. Namely the odds that the claim will

remain opened decrease with the increasing *s*. In later stages, however, the difference is decreasing again and in the stage 10 or eleven the difference between the odds is not significantly different from the stage 1 anymore. That is the odds that the claim will remain opened will again start to increase with the increasing *s*. This could probably be explained by the fact that only the most complicated claims remain opened in the later stages, therefore the probability that they will remain opened tend to grow back up.

Conclusions

In this article the settlement process and the process of the incurred value was firstly defined. Further a distribution of the ultimate incurred value was derived under restrictive but practical (and in theory quite common) assumptions. The derived distribution function contains only practically estimateable components. The statistical properties of these components were studied.

Acknowledgments. This paper was supported by grant number 26/08 of the internal grant agency of VSE (IGA).

References

1. Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., Thandi, N.: A Practitioner's Guide to Generalized Linear Models. (http://www.watsonwyatt.com/europe/news/glmpaper/media/Practitioners_Guide.pdf)
2. England, P. D., Verral, R. J.: Stochastic Claims Reserving in General Insurance. British Actuarial Journal 8, 443–544 (2002)
3. Larsen, C. R.: An individual claims reserving model. ASTIN Bulletin International Actuarial Association, 37, No.1: 113-132, 2007.
4. McCullagh, P., Nelder, J.A.: Generalized Linear Models, Second Edition. Chapman & Hall, London (1989).
5. K. Murphy and A. McLennan. A method for projecting individual large claims. Casualty Actuarial Society Forum, 205 - 236, 2006.
6. Norberg, R. Prediction of outstanding liabilities in non-life insurance. ASTIN Bulletin International Actuarial Association, 23, No.1: 95-115, 1993.

Appendix: Figures

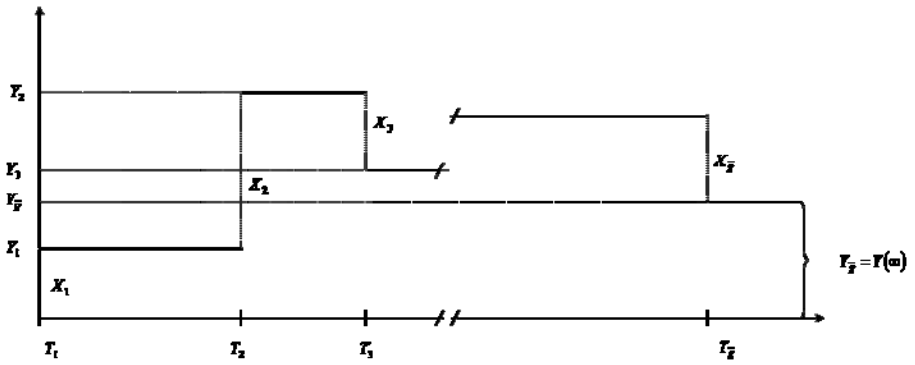


Fig. 1. Illustration of the settlement process.

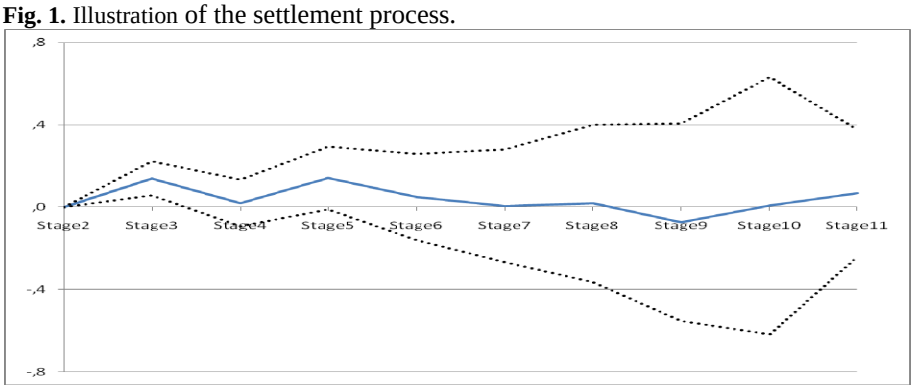


Fig. 2. The estimates of the parameters assigned to the dummy variables corresponding with the stages of the process of the incurred value. The reference category is 'stage 2'. The dotted lines are the 95% confidence intervals.

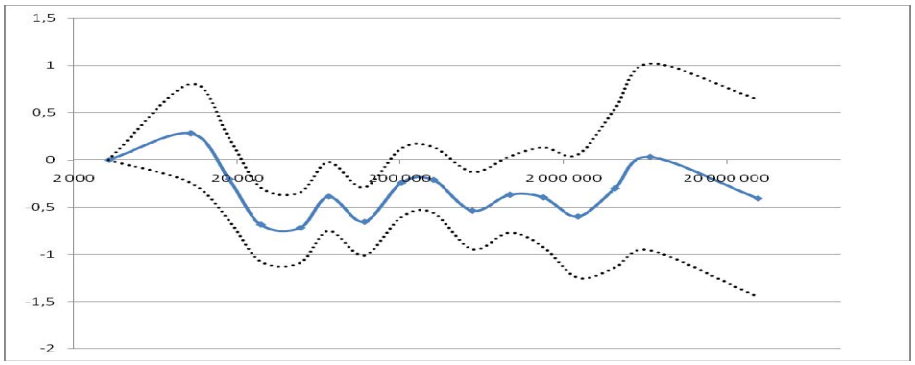


Fig. 3. The estimates of the parameters assigned to the dummy variables corresponding with the interaction of the incurred value in s-1 and StageII. The dotted lines are the 95% confidence intervals. Horizontal axis is in logarithmic scale.

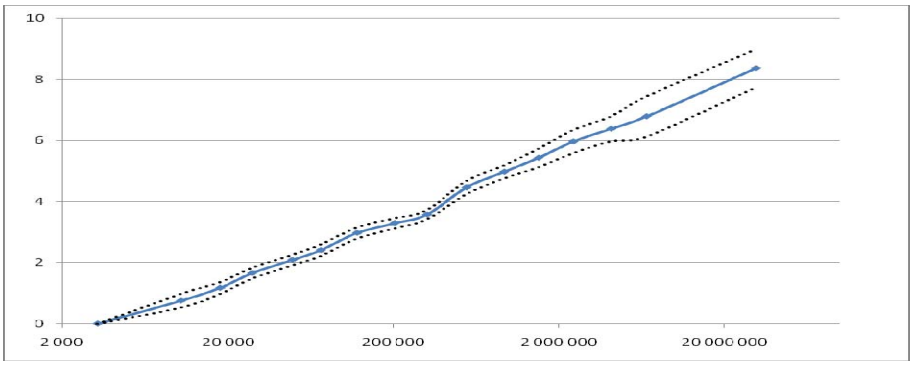


Fig. 4. The estimates of the parameters assigned to the dummy variables corresponding with the intervals of the incurred value in the stage s-1. The dotted lines are the 95% confidence intervals. Horizontal axis is in logarithmic scale.

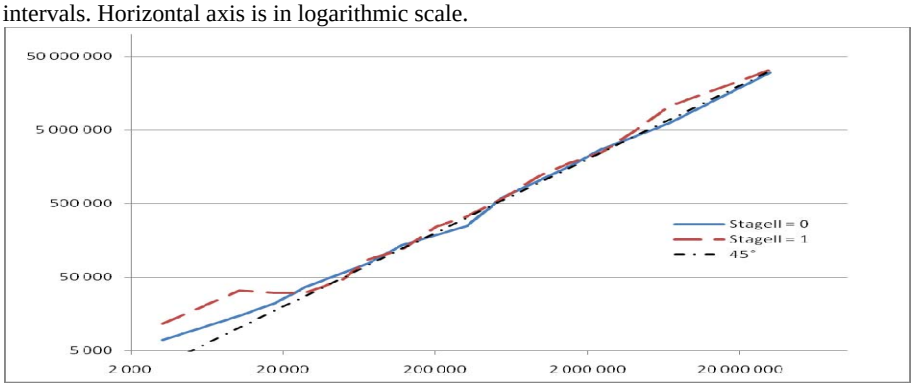


Fig. 5. Predictions of Y_s on the vertical axis and the corresponding means of categories of Y_{s-1} on the horizontal axis. Both axes are in the logarithmic scale.

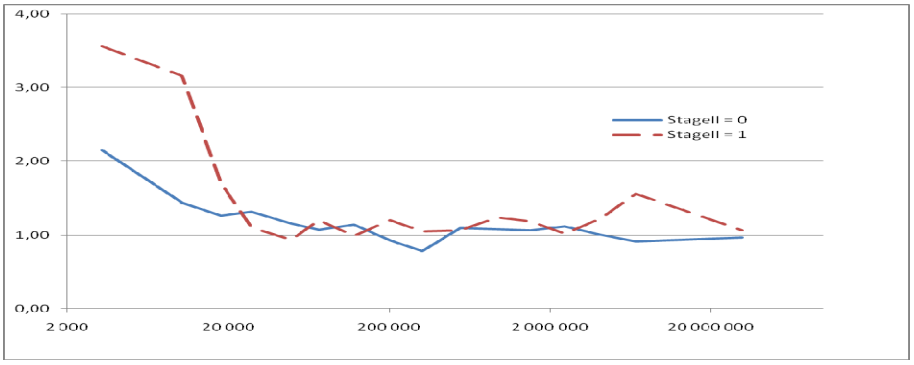


Fig. 6. Predictions of Y_s divided by corresponding means of categories of Y_{s-1} on the vertical axis and the corresponding means of categories of Y_{s-1} on the horizontal axis. Horizontal axis is in logarithmic scale.

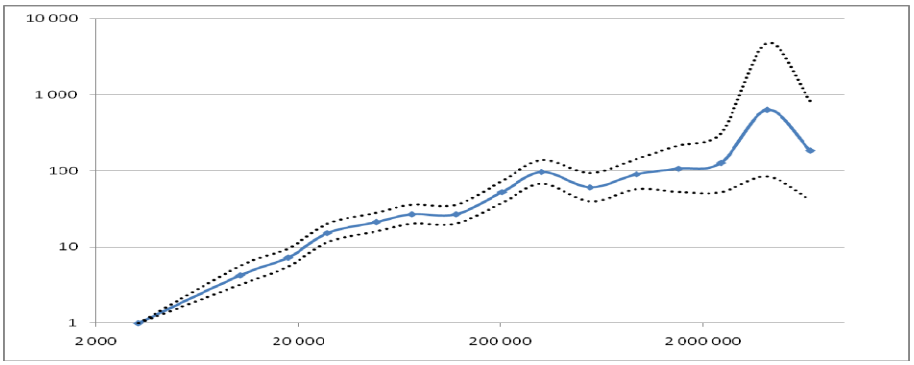


Fig. 7. Estimate of odds ratio (exponentials of the estimated parameters) for different categories of the incurred value Y_s. Both axes are in the logarithmic scale.

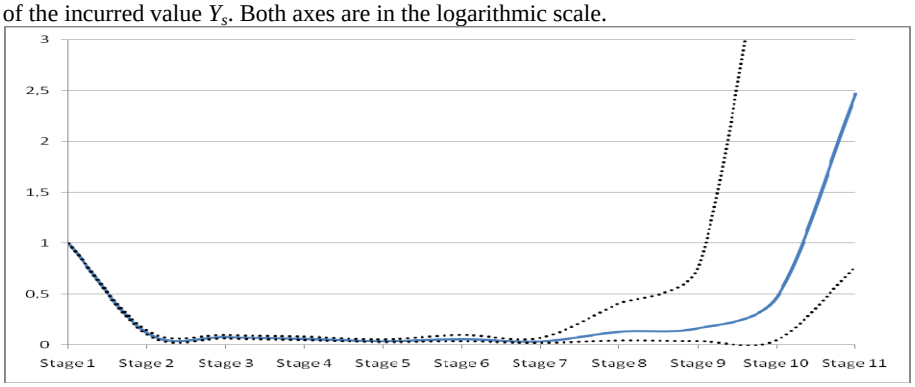


Fig. 8. Estimate of odds ratio (exponentials of the estimated parameters) for different stages. *Stage 1* was selected to be the reference category.

The Mortality Development in the Czech Republic

Pavel Zimmermann[1] , Milan Perina[1]

[1] University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3,
Czech Republic
{zimmerp, perina}@vse.cz

Abstract. In the recent years there has been observed an enormous general decrease of mortality in the Czech Republic. This paper tries to reveal the structure of this development. Logistic regression is applied on the population data for this purpose. The dependence of the one year mortality rate on the age, calendar year of observation and gender is studied.

Keywords: Mortality, Logistic regression

1 Introduction

After the political and economical changes in 1989 connected with the transition from centrally planned economy to capitalistic economy, the mortality has started to decrease with an enormous speed in the Czech Republic. This paper has no ambition to explain the reasons of this evolution. In this paper we try to uncover the nature of this development and describe its impact in different ages and for different genders. We will try to put emphasis on facts that could be important for potential forecasting of the mortality in our future work.

The fact that the mortality has been decreasing in the recent years is now generally known. The purpose of this paper is to answer questions of the following kind:

1. What does the trend in mortality look like? Has the trend change during the past few years?
2. Is the trend similar for each age group or does the trend differ in different age groups?
3. Is the trend similar for both genders?

All these questions will be studied in detail in this article. All tests performed in this article are conducted on a 0.05 significance level. The parameter estimates and all other calculations were performed in SPSS. Some of the charts were also prepared in MATLAB.

2 The Data

The data used in this article was downloaded from Human Life-Table Database webpage [3]. The data range is set to 1989 - 2006 since the mortality data before 1989

is not relevant and the mortality data after 2006 wasn't available at the time the analyses were performed. We analyse ages higher than 15 mainly because of ages below 15 is not interesting in acturial point of view. The highest age considered was 84 years due to instability of the estimates for the higher ages. More details regarding the data preparation and transformation could be found in [3] (section: Methodology Note).

3 Formulation of the Problem

3.1 Logistic Regression

In this section, we will go briefly through the basic principles of the logistic regression method (see e.g. [2] for details). Logistic regression is a special case of generalized linear model, where the link function is a logit function and the error distribution is considered to be binomial. It describes the relation between the probability of observing a 'success', and several explanatory variables - both categorical variables (so called factors) and covariates. More precisely we consider a binomial variable *Y~Bi(n,q)*, where *n* is the number of trials and *q* is the probability of success. The conditional expected value of the variable *Y/n*

$$q = E\left(\frac{Y}{n}\middle| x_1, x_2, \ldots\right) \tag{1}$$

is then modeled with the regression function:

$$Ln\left(\frac{q}{1-q}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \tag{2}$$

where β_j are the regression parameters and x_j are the explanatory variables. The fragment $q/(1-q)$ is so called 'odds' and it expresses the odds of success, i.e. the probability of success divided by the probability of failure. The term on the left side of this equation will be referred to as the 'log odds'.

In our case the binomial variable *Y* denotes the number of persons that die ('the success') in the specified year, *n* is the population size and *q* is the probability that a person dies within a year. This probability will be referred to 'probability of death' in this article. The values of the explanatory variables will be denoted by corresponding indexes. The considered explanatory variables are the calendar year of observation (denoted as *t*), age group (denoted as *age*) and gender of the person (denoted as *sex*).

3.2 Factors or Covariates

Some explanatory variables, such as gender, are discrete by its nature. For some variables (such as age or calendar year), one can chose whether to asses them as (continuous) covariates or again as a (discrete) factors. The main purpose of this article is to describe the relations in the process of changing mortality. In this case, it is better to consider these variables as categorical. This way, the estimate of the impact of the explanatory variables is in fact nonparametric (see [1] for details).

Factor variables are coded as the dummy variables in the regression function (2). The reference category has to be chosen for each variable. That means for the variable *age*, which was discretized in 14 different categories, we have a block of 13 dummy variables (denoted as *age(1)*, *age(2)*, …, *age(13)*). The age group ’80-84’ was chosen to be the reference category. That means for this age group, all these 13 dummy variables equal to 0. For the age group ‘15-19’, the variable *age(1)* equals 1 and all other variables equal to 0 etc.

The role of the variable calendar year of occurrence (i.e. the variable *t*) is slightly different. For the prediction purposes, it will be needed to consider this variable (or its transformation) as a continuous (i.e. covariate). In this article however we try to analyze the impact of the consecutive years separately to uncover the nature of the trend. This variable has 18 numerical values (corresponding with the years 1989 - 2006). The year 2006 (last observed) was chosen to be the reference year.

3.3 Mortality Curve and Log Odds of Death Curve

The age structure of the probabilities of death is usually denoted as the mortality curve. The development of the mortality curve for males and females in the Czech Republic in the years 1989 - 2006 is displayed in figures 6 and 7 in the appendix.

As noted in the section 3.1. the model is constructed for the log of odds. Therefore it will be convenient to work with the log odds of death curves instead of the mortality curves itself. The log odds of death curve will be defined as the age structure of the odds that a person with a given gender will die within a year, i.e. the transformation of the above mentioned mortality curves:

$$u_{x,t}^{(g)} = Ln\left(\frac{q_{x,t}^{(g)}}{1-q_{x,t}^{(g)}}\right) \tag{3}$$

where $q_{x,t}^{(g)}$ denotes the probability of death of a person with gender *g* and age *x* in the calendar year *t*, and $u_{x,t}^{(g)}$ is the corresponding log of odds of death. This curve will be referred to simply as the ‘log odds curve’ in the further text. The development of the log odds curve in the years 1989 - 2006 is displayed in the figures 8 and 9 in the appendix.

Notice that the log odds can be transformed back to probabilities using a simple transformation

$$q_{x,t}^{(g)} = \frac{1}{1+e^{-u_{x,t}^{(g)}}}. \tag{4}$$

This transformation is monotonically increasing function of $u_{x,t}^{(g)}$, therefore any increase of the odds is translated to an increase of probabilities of death $q_{x,t}^{(g)}$ and vice versa.

4 Results & Answers

First we will study the significance of all variables in the model (omnibus test). The significance of the variables in the model is usually tested using the likelihood ratio test. This test uses the *-2 log likelihood* statistic. The variables considered in the model are the main effects as well as the interactions of the first order. The values of this statistic for the model with only the intercept and with the model with all three main effects and all three second order interactions is displayed in the table 1. The likelihood ratio test tests the null hypothesis that adding the explanatory variables and interactions to the model has not significantly reduced the -2 log likelihood criterion (i.e. has not significantly improved the model). The test is based on the χ^2 distribution. Since the null hypothesis is rejected with quite high significance, we can come to a result that the variables in the model do increase the ability to model the probability of death.

Table 1. Omnibus Test of Model Coefficients.

Model	-2 Log likelihood
Only intercept	12 931 402
Variables	9 367 702
Difference	3 563 700
Significance	,000

This result is not really surprising - probably no one will be surprised that e.g. the age influences the probability of death. Therefore we will proceed to some 'more informative' tests - so called Wald's tests.

4.1 Wald's Tests of Factors and Interactions

There are two main options of using the Wald's test in the logistic regression. One possibility is to test the significance of a set of several dummy variables that corresponds with values of a particular factor at once. Or we can test the individual parameters of the dummy variables (similar to the *t*-tests in linear regression) to reveal

'the unique contribution of each predictor, in the context of the other predictors - that is, holding constant the other predictors - that is, eliminating any overlap between predictors.' (Explanation is cited from [4]). In the table 2 we can find the results of the Wald's tests of the parameters of all the factors and first order interactions considered in the model. The degree of freedom ('df') is equal to the number of dummy variables that correspond to the factor (i.e. the number of possible factor values minus one).

Table 2. Wald's Tests of Factors and Interactions.

Parameter	Wald's Test	df	Sig.
age	95 165	13	.000
t	3 369	17	.000
sex	1 851	1	.000
age * t	855	221	.000
age * sex	9 966	13	.000
sex * t	135	17	.000
constant	590	1	.000

All main effects and interactions of the first order proved to be statistically significant. The second order interaction did not prove to contribute significantly to the model, therefore was not considered any further.

Now we will study the individual parameters of the dummy variables for each factor or interaction.

4.2 The Main Effects

First we will study the impact of the main effects. The impact of the main effect is in general not very surprising. Mostly it corresponds with the general knowledge or expected results.

The Impact of the Calendar Year

Firstly we will demonstrate the impact of the calendar year on the odds of probability of death. As noted previously, the general trend is that mortality is decreasing over time. The coefficients for the particular years are displayed in the figure 1.

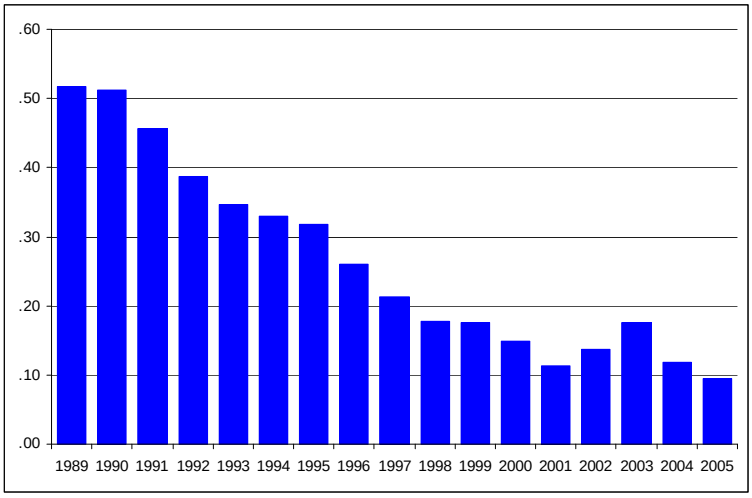


Fig. 1. The parameter estimates of the dummy variables corresponding with the factor '*t*', i.e. corresponding with the particular calendar year. The reference year is 2006.

These parameters basically reflect parallel shifts of the whole log odds curve (i.e. the change of the log odds regardless of the age group). The significance of these estimates was for all dummy variables very high - the corresponding p-values were in all cases lower than 0.001. This result was, as noted above, quite expected since it corresponds with the general knowledge that the mortality has been constantly decreasing over the past years. The estimates suggest however difficulties when forecasting. Quite stable decreasing trend was somehow 'disturbed' in the years 2002 - 2004. It is now difficult to forecast whether the heap in 2003 is only a one time event and the trend will continue or if this means the change of the trend it self.

The Impact of Age

The impact of the age is quite straightforward. It is not a surprising result that the death probability (and therefore also the log odds) increases with the increasing age. The impact of the age is displayed in the figure 2. Again all the parameter estimates are significantly different from 0 with all p-values lower than 0.001.

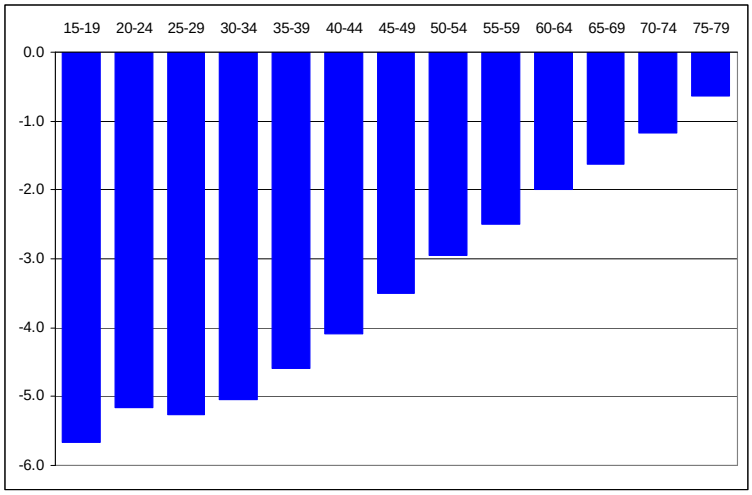


Fig. 2. The parameter estimates of the dummy variables corresponding with the factor '*age*', i.e. corresponding with the particular age group. The base age group is '80-84' years (the last considered).

The Impact of the Gender

Again, it is a common knowledge that gender in general is a factor strongly determining the mortality. In this case, there is only one dummy variable, which equals 1 for females and 0 for males. The parameter estimate corresponding with the dummy variable *sex(1)* (the females) equals to -0.390 and again is highly significant (p-value < 0.001).

4.3 The Interactions

As stated previously, only the first order interactions were considered. The interactions help revealing the structure of the mortality change for specific combinations of the values of more factors (two in the case of first order interactions).

The Interaction of Age and Gender

In the main effect section a separate influence of both the age (regardless of the gender) and the gender (regardless of the age group) were studied. The interaction of age and gender could help answering the following question: 'Is the impact of the gender same over all ages, i.e. are the log odds curves for both genders parallel, or is the impact different for some age groups?'

Since the interaction *age*sex* is proved to carry out a statistically significant information, we can already answer that the log odds curves for both genders are not

parallel. There is some significant extra impact of the gender above the general (=main effect) gender impact considered previously for some specific age groups. The parameter estimates are displayed in the figure 3.

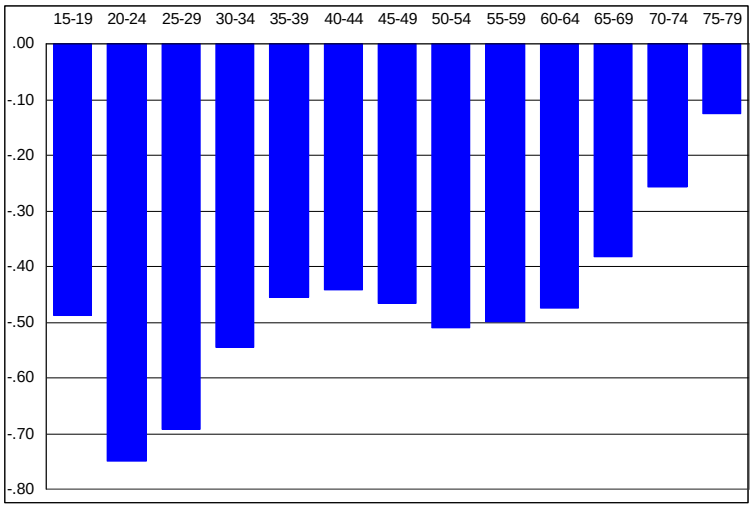


Fig. 3. The parameter estimates of the dummy variables corresponding with the interaction of the factors '*age*' and '*sex*'.

The results suggest that the difference of mortality for both genders is even higher in the younger ages - the general negative impact of the dummy '*sex(1)*' (female gender) is further decreased by the interaction dummy variables of the lower ages. In higher ages the parameter estimates corresponding to this interaction are closer to zero, i.e. the difference between the genders is relatively lower. For this interaction, parameter estimates of all dummy variables were also significantly different from zero.

The Interaction of Calendar Year and Gender

This interaction could help us answering the question whether the evolution of mortality is similar for both genders or if there is or was in the past some difference of the trend of the two genders. The resulting parameter estimates can be found in the figure 4.

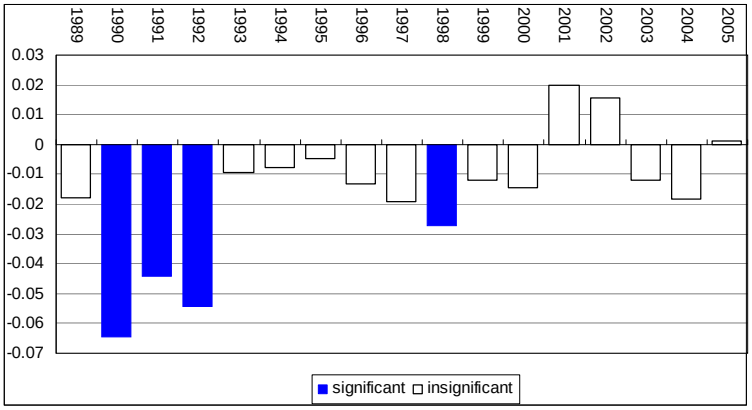


Fig. 4. The parameter estimates of the dummy variables corresponding with the interaction of the factors '*sex*' and '*t*'. Estimates that are significant on the 0.05 level are displayed with filled bars.

Although the whole interaction proved to carry significant information (see the table 2), not all the parameter estimates were statistically significant. In fact, it seems that if the parameter estimate of the dummy variable corresponding with the interaction of the values '*sex(1)*' (female gender) and '*t=1998*' is considered to be an outlier, the only significant difference in the mortality evolution between the genders was at the beginning of the nineties (1990 - 1992). The fact that all these significant parameter estimates are negative suggests that decrease of the mortality was at the beginning faster for female than for male gender - the male gender was at the beginning somehow delayed in the general trend.

In the more recent years this difference disappeared. ('Males caught up females.') It seems that in the case of forecasting the mortality these interactions will not have to be considered in the future years. Nothing suggests that the general impact of the calendar year will be different for different genders in the future.

The Interaction of Calendar Year and Age

The last interaction is the interaction between the calendar year and the age group ('*t***age*'). In the chapter 4.2, we stated that in general, the mortality is decreasing over the calendar years. The estimates of the parameters of the dummy variables corresponding with the main effect '*t*' reflected the parallel shifts of the log odds curve. The parameter estimates of the dummy variables corresponding with this interaction should help answering the question:

'Is the evolution of the mortality similar for each age group or does the mortality evolve significantly different in some age groups?'

The parameter estimates of the dummy variables corresponding with the '*t***age*' interaction can be found in the figure 5. The estimates significantly different from zero (on the 0.05 significance level) are marked with the coloured background.

Due to higher amount of dummy variables corresponding with this interaction, this interaction is probably the most difficult to analyze. The only systematic (and therefore interpretable) 'block' of the significant parameter estimates is for the higher age groups ('65-69', '70-74', '75-79') and the calendar years 1989 - 2002. All this significant values are positive, i.e. it reduces the general decrease of the mortality which is incorporated in the model due to the main effect '*t*'. This suggests that the trends are different for different age groups. The mortality decreases faster for the younger age groups than for the elder ones.

t\age	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
1989	-0.22	-0.23	-0.14	-0.07	-0.11	-0.05	-0.08	-0.14	-0.09	-0.07	0.06	0.05	0.04
1990	-0.08	-0.15	-0.08	0.02	0.03	0.03	0.01	-0.06	-0.04	-0.02	0.09	0.09	0.06
1991	0.06	-0.14	0.03	0.04	0.01	0.04	0.03	-0.08	-0.04	-0.05	0.07	0.07	0.04
1992	0.06	0.00	0.10	0.14	0.09	0.10	0.04	-0.02	0.01	0.01	0.10	0.10	0.06
1993	0.09	0.02	0.09	0.11	0.06	0.03	0.01	-0.07	-0.02	-0.03	0.06	0.11	0.05
1994	0.09	-0.01	0.16	0.08	0.02	0.05	0.01	-0.05	-0.04	-0.04	0.05	0.10	0.04
1995	0.13	-0.05	0.07	0.10	0.06	0.05	-0.02	-0.04	-0.04	-0.04	0.05	0.11	0.06
1996	0.04	-0.13	0.07	0.08	0.06	0.03	0.01	0.01	-0.03	-0.01	0.06	0.11	0.07
1997	0.11	0.04	0.10	0.10	0.12	0.14	0.06	0.04	0.01	0.01	0.09	0.13	0.09
1998	0.15	-0.08	0.06	0.00	0.02	0.10	0.04	0.01	0.02	-0.03	0.08	0.12	0.09
1999	0.06	-0.01	0.11	0.07	-0.03	0.02	0.01	-0.02	0.00	-0.02	0.03	0.09	0.07
2000	0.12	0.03	0.06	0.05	-0.01	0.06	0.06	0.01	0.00	-0.04	0.05	0.07	0.07
2001	0.09	0.06	0.09	0.03	-0.01	0.07	0.01	-0.01	0.02	-0.07	0.03	0.06	0.06
2002	0.07	0.05	0.08	0.02	-0.03	0.05	0.00	-0.03	0.00	-0.08	0.01	0.04	0.04
2003	-0.09	-0.03	0.04	0.05	-0.03	-0.07	-0.05	-0.06	-0.03	-0.10	-0.03	-0.01	0.00
2004	0.03	-0.04	0.04	-0.02	0.01	-0.05	-0.03	-0.05	-0.04	-0.08	-0.02	-0.01	-0.01
2005	-0.04	-0.02	0.07	0.05	-0.05	-0.09	-0.02	-0.07	-0.03	-0.08	-0.04	-0.03	-0.01

Fig. 5. The parameter estimates of the dummy variables corresponding with the '*t*age*' interaction. The estimates significantly different from zero (on the 0.05 significance level) are marked with the coloured background.

Acknowledgments: This article was supported from grant IGA 27/08.

References

1. Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., Thandi, N.: A Practitioner's Guide to Generalized Linear Models. http://www.watsonwyatt.com/europe/news/glmpaper/media/Practitioners_Guide.pdf
2. Kleinbaum, G., David, Klein, Mitchel: Logistic Regression: A Self Learning Text. Second Edition. Springer. New York(2002)
3. The Human Life-Table Databese, http://www.lifetable.de/
4. Wuensch, K.L.: Binary Logistic Regression with SPSS. http://core.ecu.edu/psyc/wuenschk/MV/MultReg/Logistic-SPSS.doc

Appendix: Development of the Mortality Curves

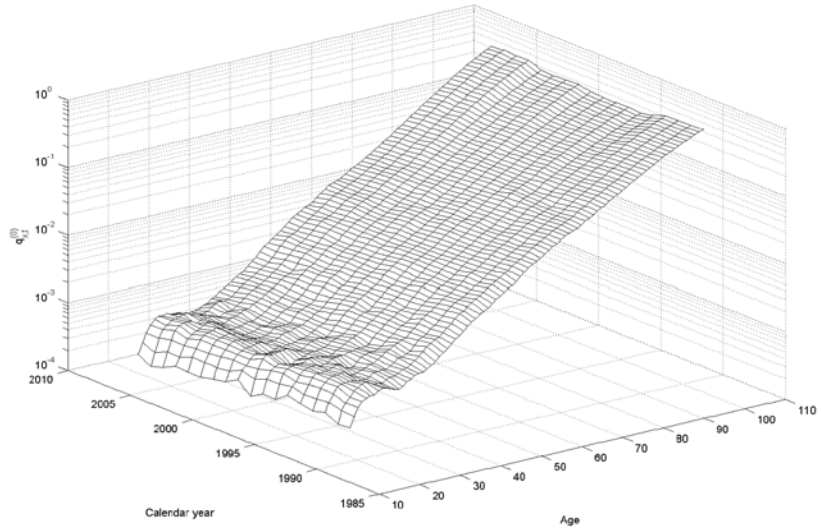


Fig. 6. Development of the mortality curves for males in the years 1989 - 2006 in the Czech Republic. (Logarithmic scale).

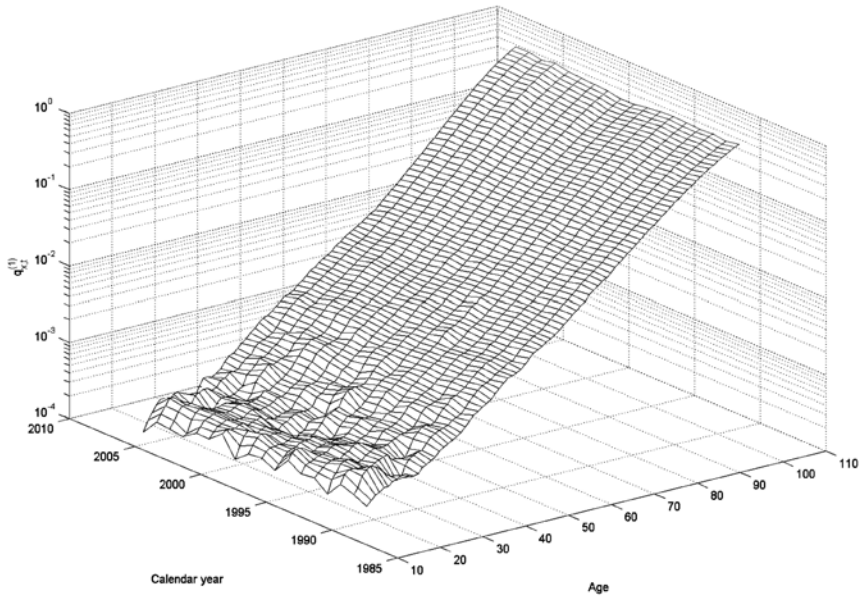


Fig. 7. Development of the mortality curves for females in the years 1989 - 2006 in the Czech Republic. (Logarithmic scale).

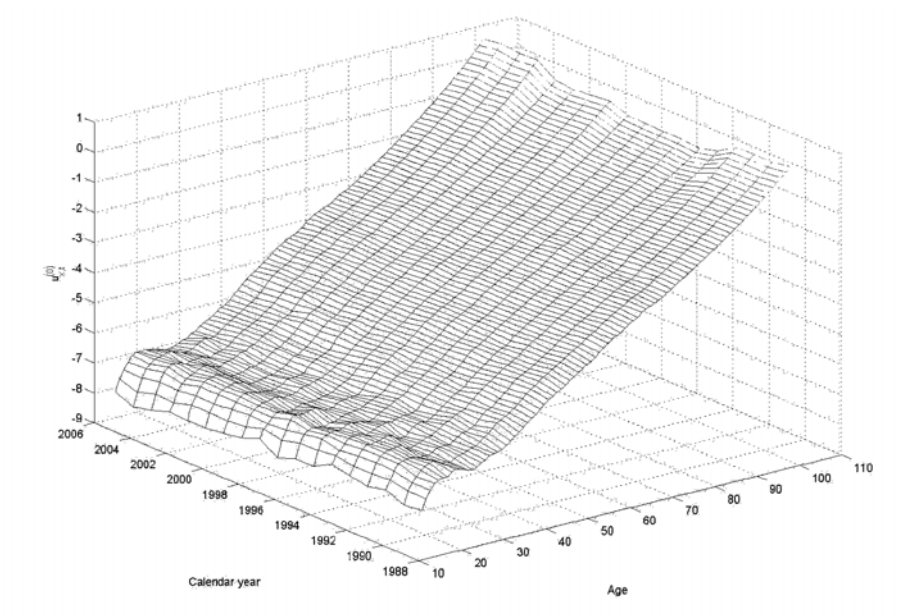


Fig. 8. Development of the log odds curves for males in the years 1989 - 2006 in the Czech Republic.

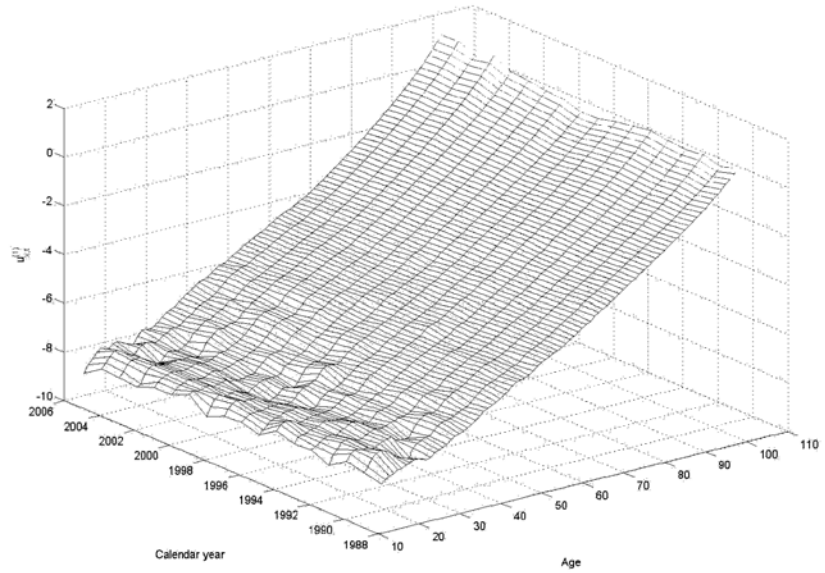


Fig. 9. Development of the log odds curves for females in the years 1989 - 2006 in the Czech Republic

Vydavatel: Vysoká škola ekonomická v Praze
Nakladatelství Oeconomica
Rok vydání: 2009
Tisk: Vysoká škola ekonomická v Praze
Nakladatelství Oeconomica
Sazba: autoři
Tato publikace neprošla redakční úpravou.