Polskie Towarzystwo Statystyczne
Oddział we Wrocławiu

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Silesian Statistical Review

Nr 9 (15)

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2011

RADA NAUKOWA
Walenty Ostasiewicz
Tadeusz Bednarski, Luisa Canal, Stanislava Hronová,
Katarzyna Ostasiewicz, Angiola Pollastri, Emilia Zimkova

KOMITET REDAKCYJNY
Walenty Ostasiewicz (redaktor naczelny)
Zofia Rusnak (zastępca redaktora naczelnego)
Edyta Mazurek (sekretarz naukowy)
Tadeusz Borys, Stanisław Heilpern, Janusz Wywiał

ADRES REDAKCJI
Katedra Statystyki
Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel. (71) 36-80-356, tel./fax (71) 36-80-357
e-mail: walenty.ostasiewicz@ue.wroc.pl

Publikacja jest dostępna na stronie www.ibuk.pl
Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities
http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Spis treści

Summaries

21. SCIENTIFIC STATISTICAL SEMINAR
“MARBURG-WROCLAW”,
MARBURG SEPTEMBER 26–29, 2011

EXTENDED SUMMARIES OF SELECTED PAPERS

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 9 (15)

ISSN 1644-6739

A MATRIX REPRESENTATION OF A NET AMOUNT AT RISK ITS APPLICATION IN PREMIUM PARTITION FOR MULTISTATE INSURANCE CONTRACTS

Joanna Dębicka (Wrocław University of Economics)

The multistate methodology is intensively used in calculation of premiums and reserves of different types of insurances like life, disability, sickness, marriage or unemployment insurances. The pair (S, T) is called a *multiple state model*, and describes all possible insured risk events as far as its evolution is concerned (usually up to the end of insurance). That is, at any time the insured risk is in one of a finite number of states belonging to the *state space S*. Each state corresponds to an event which determines the cash flows (premiums and benefits). By T we denote the *set of direct transitions* between states of the state space.

We consider an insurance contract issued at time 0 and terminating at a later time n (n is the term of policy). Let $X(t)$ denote the state of an individual (the policy) at time t. Hence the evolution of the insured risk is given by a discrete-time stochastic process, $\{X(t) : t = 0, 1, 2, ...\}$ with values in the finite set S. We assume that $\{X(t) : t = 0, 1, 2, ...\}$ is a discrete time Markov process. If we look at the evolution of the contract, then both the presence at a given state and the movement from one state to another may have some financial impact. We distinguish between the following types of cash flows related to multistate insurance: $b_j(k)$ – an annuity benefit at time k if $X(k) = j$, $d_j(k)$ – a lump sum at some fixed time k if $X(k) = j$, $c_{ij}(k)$ – a lump sum at time k if a transition occurs from state i

to state j at that time, $\pi_j(k)$ – a premium amount at some fixed time k if $X(k) = j$, $p_j(k)$ – a period premium amount at time k if $X(k) = j$.

We focus on a discrete-time model, which means that insurance payments are made at the ends of time intervals. Practically it means that annuity and insurance benefits are paid immediately before the end of the unite time (for example a year or month). Premiums are paid immediately after the beginning of the unite time.

Looking from the prospective of financial mathematics, future cash flows (which are realized at time k) are discounted to the present (say time t) by some interest rate. This produces the cash value of future payment stream $\Upsilon_t^{\wp,j}(k)$, where $\wp$ denotes one of the type of cash flows ($\wp \in \{p,\pi,b,d,c_1,c_2,\ldots,c_N\}$ and c_i is the benefit paid if process $\{X(t)\}$ leaves state i).

At moment t the sum of cash value of future payment stream is called prospective loss ${}_tL$ of the insurer at time t, so

$$ {}_tL = \sum_{\wp\in\{b,d,c_1,\ldots,c_N\}} \sum_{j\in S} \sum_{k=t+1}^{n} \Upsilon_t^{\wp,j}(k) - \sum_{\wp\in\{p,\pi\}} \sum_{j\in S} \sum_{k=t}^{n-1} \Upsilon_t^{\wp,j}(k), $$

where benefits are an inflow representing an income to loss fund. Premiums represents an outgo from a loss fund of the insurer. Then *prospective reserve* is a conditional expectation of prospective loss under the condition that at time t the insurance contract is at state i

$$ V_i(t) = \mathrm{E}\left({}_tL \mid X(t) = i\right) $$

$$ = \sum_{\wp\in\{b,d,c_1,\ldots,c_N\}} \sum_{j\in S} \sum_{k=t+1}^{n} \mathrm{E}\left(\Upsilon_t^{\wp,j}(k) \mid X(t) = i\right) - \sum_{\wp\in\{p,\pi\}} \sum_{j\in S} \sum_{k=t}^{n-1} \mathrm{E}\left(\Upsilon_t^{\wp,j}(k) \mid X(t) = i\right). $$

Net amount at risk $nar_j(t)$ for state j at (t+1)-th unit time has the following form

$$ nar_j(t) = \begin{cases} \left(V_j(t+1) + \sum\limits_{\wp\in\{b,d,c_1\}} \wp_j(t+1)\right) - \left(V_1(t+1) + \sum\limits_{\wp\in\{b,d,c_1\}} \wp_1(t+1)\right) & \text{if } q_{1j}(t) > 0 \\ 0 & \text{if } q_{1j}(t) = 0 \end{cases}, $$

where $q_{ij}(t) = P(X(t + 1) = j \mid X(t) = 1)$.

It appears that it is possible to describe the net amount at risk in whole insurance period in a matrix form. To do this we have to introduce the modified multistate model (S*, T*) and describe matrices related with: modified multistate model and its probabilistic structure ($\mathbf{P}(0)$ – vector of initial distribution and $\{\mathbf{Q}^*(k)\}_{k\,=\,0,\,1,\,2,\,\ldots}$ sequence of matrices transition of the process $\{X(t)\}$), cash flows ($\mathbf{C}_{in}$ consists only of an income to a particular fund, $\mathbf{C}_{out}$ consists only of an outgo from a fund and $\mathbf{C}_{in}$ + $\mathbf{C}_{out}$ = $\mathbf{C}$) and discount function ($\boldsymbol{\Lambda}$ consists of discounted and accumulated functions for a process of interest rate $\{Y(t)$: $t \geq 0)\}$)). For modified multistate model it is useful to define matrix

$$\mathbf{Nar} = \left[nar_j(t)\right]_{\substack{j=1,2,\ldots,,N^* \\ t=0,1,2,\ldots,n^*}},$$

where

$$nar_j(t) = \begin{cases} \left(\mathbf{J}_j^T - \mathbf{J}_1^T\right)(\mathbf{V}^T + \mathbf{C}_{in})\mathbf{I}_{t+2} & for \quad \mathbf{J}_1^T\mathbf{Q}^*(t)\mathbf{J}_j > 0 \\ 0 & for \quad \mathbf{J}_1^T\mathbf{Q}^*(t)\mathbf{J}_j = 0 \end{cases}$$

and $\mathbf{V}$ is a matrix of prospective reserves for the whole insurance period, $\mathbf{I}_{t+2} \in \boldsymbol{R}^{\boldsymbol{n}^*+\mathbf{1}}$ is a vector which consists zeros apart from 1 at $t+2$ coordinate, $\mathbf{J}_j \in R^{N^*}$ is a vector which consists of zeros apart from 1 at j coordinate.

Let $p_1(t)$ be a period premium amount payable at time t if the insured is healthy. Moreover, let $p_1^s(t)$ be a saving premium and $p_1^{r_j}(t)$ be a risk premium for the state j. Saving and risk premiums depend on benefits payable at time (t + 1)-th unit time and prospective reserves at time (t + 1)-th unit time.

Theorem (matrix representation of premium partition)

For the insurance contract described by extended multistate model $\left(S^*, T^*\right)$, if $Y(t)$ is stochastic process with stationary increments and net period premiums are paid when $X^*(t) = 1$, then net period premium can be presented as follows

$$p_1(t) = p_1^s(t) + \sum_{j \in S^* \setminus \{1\}} p_1^{r_j}(t)$$

where

$$\begin{aligned} p_1^s(t) &= \mathbf{J}_1^T\left(\left(\mathbf{V}^T+\mathbf{C}_{in}^T\right)\mathbf{I}_{t+2}\mathbf{I}_{t+2}^T\mathbf{\Lambda}^T-\mathbf{V}^T\right)\mathbf{I}_{t+1} \\ p_1^{r_j}(t) &= \mathbf{J}_1^T\mathbf{Q}^*(t)\mathbf{J}_j\mathbf{J}_j^T\mathbf{Nar}^T\mathbf{I}_{t+1}\mathbf{I}_{t+1}^T\mathbf{\Lambda}\mathbf{I}_{t+2} \end{aligned}$$

Matrix notation makes the formulas for saving and risk premiums immediately applicable for numerical calculations and can be used to construct untraditional insurance products.

Matrix approach enables us to give a flexible tool not only for numerical calculations but also for the analysis of gross premiums, emerging costs and profit testing and helps in analysing both a single policy and a portfolio of policies.

As a numerical illustration, a health insurance contract was considered for which saving and risk premiums in whole insurance period were calculated using the matrix notation introduced above.

A PRIVACY-PROTECTING SURVEY DESIGN FOR MULTICHOTOMOUS SENSITIVE VARIABLES

Heiko Groenitz (University of Marburg)

Sensitive variables often appear in surveys. For instance, the interviewer could ask: “How much do you earn?” or “Have you ever evaded taxes?” If such sensitive questions are asked, some interviewees will refuse to respond or will give an untruthful answer. To estimate the distribution of sensitive variables, many randomized response (RR) models were developed since the paper by Warner in [4].

What the RR models have in common is that every respondent is supplied with a randomization device (RD). A RD is an instrument (e.g. coin, deck of cards) used by the interviewee to conduct a random experiment where the experiment has – for a fixed respondent – at least two results. The outcome of the experiment influences the answer. A different approach can be found in the publications of Tian et al. [3], Yu et al. [5], Tan et al. [1] and Tang et al. [2]. These authors proposed some nonrandomized response (NRR) models. That is, their models do not require any RD and thus reduce both the survey complexity and the study costs.

The NRR models of the previously mentioned authors are not applicable to multichotomous sensitive variables like income where all values are sensitive. To overcome this problem, we have developed the diagonal model (DM). Let us consider a sensitive variable $K^* \in \{1, ..., k\}$, $k \geq 2$. The diagonal model requires the choice of a non-sensitive auxiliary variable $W^* \in \{1, ..., k\}$ with known distribution in such a way that K^* and W^* are independent. $W^* = j$ could describe that the respondent is born in period j of the year (where the year is partitioned in k periods). Every respondent is introduced to give a privacy-protecting answer $A^* \in \{1, ..., k\}$ depending on his or her values of K^* and W^*.

The answer pattern can be described with a special table where the replies $A^* = j$ are arranged on certain diagonals. The illustration with diagonals allows an easy presentation of the answer formula to the interviewees. Thus, we conclude that this survey design is clearly understandable and does not demand higher sophistication from the respondents.

We derive the maximum likelihood (ML) estimator for the distribution of K^* where the expectation maximization (EM) algorithm turns out to be beneficial. Further, we calculate standard errors and confidence intervals. Subsequently, we investigate model efficiency and the degree of privacy protection (DPP) depending on the distribution of W^* (denoted with $\mathbf{P}_{W^*}$). We show that there are optimal and non-optimal distributions of W^*. $\mathbb{P}_{W^*}$ is not optimal if the efficiency loss is larger than necessary for a DPP that is provided by $\mathbb{P}_{W^*}$. Of course, it is reasonable to use only optimal distributions $\mathbb{P}_{W^*}$. For these $\mathbb{P}_{W^*}$, the efficiency loss is an increasing function of the DPP. Hence, a decreasing privacy protection is the "price" for increasing efficiency.

In the sequel of the contribution, we extend the diagonal model with covariates. In a survey according to the covariate diagonal model, every interview proceeds as follows: At the beginning, the respondent is asked directly for his or her values of p certain covariates $x^* := (x_1^*, \dots, x_p^*)$, which are presumed to be nonsensitive. Afterwards, a response A^* due to the answer formula of the ordinary diagonal model is demanded.

The probabilities for the categories $K^* = j$ are modeled in dependence of the covariates x^* by a logistic regression model. However, a

ML estimator is not calculable directly, because the respondents give scrambled replies rather than direct replies by construction of the answer formula. Thus, we have to regard the observed frequencies of the given answers and obtain a certain multivariate generalized linear model (GLM). For this GLM, one can compute ML estimates numerically using Fisher scoring. Thereby, in every iteration, a weighted least square estimation is conducted.

References

[1] Tan M.T., Tian G.L., Tang M.L. *Sample surveys with sensitive questions: A nonrandomized response approach*, "The American Statistician" 2009, Vol. 63, pp. 9–16.

[2] Tang M.L., Tian G.L., Tang N.S., Liu Z., *A new non-randomized multi-category response model for surveys with a single sensitive question*: *Design and analysis*, "Journal of the Korean Statistical Society" 2009, Vol. 38, pp. 339–349.

[3] Tian G.L., Yu J.W., Tang M.L., Geng Z., *A new non-randomized model for analysing sensitive questions with binary outcomes*, "Statistics in Medicine" 2007, Vol. 26, pp. 4238–4252.

[4] Warner S.L., *Randomized response: A survey technique for eliminating evasive answer bias*, "Journal of the American Statistical Association" 1965, Vol. 60, pp. 63–69.

[5] Yu J.W., Tian G.L., Tang M.L., *Two new models for survey sampling with sensitive characteristic: design and analysis*, "Metrika" 2008, Vol. 67, pp. 251–263.

THE IMPACT OF DEPENDENCES ON THE ANNUITIES

Stanisław Heilpern (Wrocław University of Economics)

1. Introduction

In the paper we study the impact of dependences on the values of annuities. It is based on the paper [1]. The dependent lifelengths of spouses are studied. The authors studied in [1] the situation in Belgium, but we try to apply the methods from this paper in the Polish case.

First, we introduce the general notation and assumption connected with this subject. Let T_x^M (resp. T_y^W) be remaining lifetimes of an x-year-old man (resp. y-year-old woman) taking value in [0, w_x^M] (resp. [0, w_y^W]). The distribution function and survival function of T_x^M is given by ${}_tp_x^M = P(T_x^M > t) = 1 - {}_tq_x^M$. We can derive ${}_tp_x^M$ using a force of mortality: ${}_tp_x^M = exp\left(-\int_0^t \mu_{x+s}^M\, ds\right)$. The joint distribution of the lifelengths can be described by a joint-life status ${}_tp_{xy} = P\left(T_x^M > t, T_y^W > t\right) = 1 - {}_tq_{xy}$ and by a last-survival status ${}_tp_{\overline{xy}} = P\left(\max\{T_x^M, T_y^W\} > t\right) = 1 - {}_tq_{\overline{xy}} = {}_tp_x^M + {}_tp_y^W - {}_tp_{xy}$.

The random variables T_x^M, T_y^W are positive quadrant dependence (PQD) when $P\left(T_x^M > t, T_y^W > s\right) \geq P\left(T_x^M > t\right)P\left(T_y^W > s\right)$. If T_x^M, T_y^W are PQD, then ${}_tp_{xy} \geq {}_tp_x^M\, {}_tp_y^W$.

2. Pensions

Now we present three pensions connected with the insurance of spouses. Let $v = (1 + \xi)^{-1}$ be the discount factor connected with the annual effective rate ξ. First pension is the widow's pension: $a_{x|y} = a_y - a_{xy}$, where $a_y = \sum_{k=1}^{w_y^W} v^k\, {}_kp_y^W, a_y = \sum_{k=1}^{w_x^M \wedge w_y^W} v^k\, {}_kp_{xy}$. The payments start with the husband's death and terminate with the death of his wife in this case. Next pensions are the n-year joint-life and n-year last survival annuities, done by formulas: $a_{xy;\bar{n}|} = \sum_{k=1}^{n} v^k\, {}_kp_{xy}$ and $a_{\overline{xy};\bar{n}|} = \sum_{k=1}^{n} v^k\, {}_kp_{\overline{xy}}$. They pay \$1 at the end of the years as long as both or either spouse survives.

When the lifelengths are independent we denote these pension by symbols: $a_{x|y}^{\perp}$, $a_{xy;\bar{n}|}^{\perp}$, $a_{\overline{xy};\bar{n}|}^{\perp}$. The independence is the classical assumption often used in practice. But in the real life the lifetimes of spouses are often little, but dependent. There are some common factors, risks influenced on both spouses. There is so called "broken heart syndrome". The aim of this paper is to study the impact of such dependences on the value of pensions.

First we study the nonrealistic, extreme cases, when the lifetimes of spouses are positive and negative perfect dependent. In this case we use the lower and upper Fréchet bounds: $\max\{{}_tp_x^M + {}_tp_y^W - 1, 0\} \leq {}_tp_{xy} \leq \min\{{}_tp_x^M, {}_tp_y^W\}$ and the pensions have the following bounds:

widows pension: $a_{x|y}^{\min} \leq a_{x|y} \leq a_{x|y}^{\max}$,

n-year joint-life annuity: $a_{xy;\bar{n}|}^{\min} \leq a_{xy;\bar{n}|} \leq a_{xy;\bar{n}|}^{\max}$,

n-year last survival annuity: $a^{\min}_{xy;\bar{n}|} \leq a_{xy;\bar{n}|} \leq a^{\max}_{xy;\bar{n}|}$,

where $a^{\min}_{x|y} = \sum_{k=1}^{w_y^W} v^k\ {}_kp_y^W - \sum_{k=1}^{w_x^M \wedge w_y^W} v^k \min\{{}_kp_x^M,\ {}_kp_y^W\}$,

$$a^{\max}_{x|y} = \sum_{k=1}^{w_y^W} v^k\ {}_kp_y^W - \sum_{k=1}^{w_x^M \wedge w_y^W} v^k \max\{{}_kp_x^M + {}_kp_y^W - 1,\ 0\},$$
$$a^{\min}_{xy;\bar{n}|} = \sum_{k=1}^{n} v^k \max\{{}_kp_x^M + {}_kp_y^K, 0\},$$
$$a^{\max}_{xy;\bar{n}|} = \sum_{k=1}^{n} v^k \min\{{}_kp_x^M,\ {}_kp_y^K\},$$

$$a^{\min}_{\overline{xy};\bar{n}|} = \sum_{k=1}^{n} v^k \left(1 - \min\{{}_kq_x^M,\ {}_kq_y^K\}\right),$$
$$a^{\max}_{\overline{xy};\bar{n}|} = \sum_{k=1}^{n} v^k \left(1 - \max\{{}_kq_x^M + {}_kq_y^K, 0\}\right).$$

If T_x^M, T_y^W PQD then we obtain the following relation between these pensions with respect to the independent case: $a_{x|y} \leq a^{\perp}_{x|y}$, $a_{\overline{xy};\bar{n}|} \leq a^{\perp}_{\overline{xy};\bar{n}|}$ and $a_{xy;\bar{n}|} \geq a^{\perp}_{xy;\bar{n}|}$.

3. Markov model

Now we study the Markov model based on stationary Markov chain. It is an appreciated tool for the calculation of life contingencies functions and pensions. We have four states and the forces of mortalities μ_{ij} in this case:

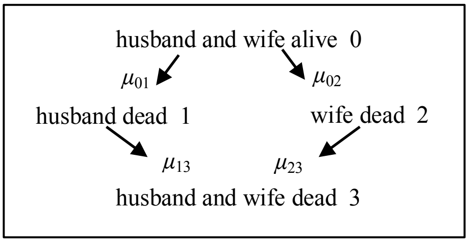

Fig. 1. Markov model

Source: own elaboration based on [1].

We can compute transition probabilities $p_{ij}(t, s)$ using the forces of mortalities in the following way: $p_{00}(t,s) = \exp\left(-\int_t^s \left(\mu_{01}(u) + \mu_{02}(u)\right)du\right)$, $p_{ii}(t,s) = \exp\left(-\int_t^s \mu_{i3}(u)du\right)$ and $p_{0i}(t,s) = \int_t^s p_{00}(t,u)\mu_{oi}(u)p_{ii}(u,s)du$ for $i = 1, 2$, the joint and marginals survival functions:

$$P\left(T_x^M > t, T_y^W > s\right) = \begin{cases} p_{00}(0,s) + p_{00}(0,t)p_{01}(t,s) & 0 \le t \le s \\ p_{00}(0,t) + p_{00}(0,s)p_{02}(s,t) & 0 \le s \le t\text{'} \end{cases}$$

$P(T_x^M > t) = \ p_{00}(0,t) + p_{02}(0,t)$ and $P\left(T_y^W > t\right) = \ p_{00}(0,t) + p_{01}(0,t)$.

The Norberg showed in [5] that the lifetimes T_x^M, T_y^W are independent iff $\mu_{01}(t) = \mu_{23}(t)$, $\mu_{02}(t) = \mu_{13}(t)$ and if $\mu_{01}(t) < \mu_{23}(t)$, $\mu_{02}(t) < \mu_{13}(t)$, then they are PQD. In our paper we use the following assumption done by Denuit et al. in [1]:

$$\mu_{01}(t) = (1 - \alpha_{01})\,\mu_{x+t}^M \qquad \mu_{23}(t) = (1 + \alpha_{23})\,\mu_{x+t}^M,$$

$$\mu_{02}(t) = (1 - \alpha_{02})\,\mu_{y+t}^W \qquad \mu_{23}(t) = (1 + \alpha_{13})\,\mu_{y+t}^W.$$

These assumptions link the Markov forces of mortality $\mu_{ij}(t)$ and the marginal lifetime forces of mortality μ_{x+t}^M, μ_{y+t}^W done by the constants α_{ij}. The joint survival function takes the following form in this case;

$${}_tp_{xy} = p_{00}(0, \mathrm{t}) = \exp\left(-\int_0^t\left(\mu_{01}(u) + \mu_{02}(u)\right)du\right) = \left({}_tp_x^M\right)^{1-\alpha_{01}}\left({}_tp_y^W\right)^{1-\alpha_{02}}.$$

We obtain the marginal survival functions ${}_tp_x^M$, ${}_tp_y^W$ from the life tables. The parameters α_{01}, α_{02} must be estimated. We may estimate these parameters using the estimator

$$\hat{\alpha}_{ij} = \arg\min \sum_{k=1}^{n}\left(\Delta\hat{\Omega}_{ij} - \int_0^1 \mu_{ij}(k+t)dt\right)^2$$

based on the increments of the transition function ${}_{ij}(t) = \int_0^t \mu_{ij}(s)ds$. These increments can be estimated by the Nelson-Aalen estimator. For instance we have

$$\hat{\alpha}_{01} = 1 + \frac{\sum_{k=1}^n \Delta\hat{\Omega}_{01} \ln {}_1p_{x+k}^M}{\sum_{k=1}^n (\ln {}_1p_{x+k}^M)^2} \quad \text{and} \quad \Delta\hat{\Omega}_{01} = \frac{L_{0:1}(k)}{L_0(k+1) - L_0(k)}\left(\ln L_0(k+1) - \ln L_0(k)\right),$$

where $L_{0:1}(k)$ is a number of k-year-old husbands dying during 2002, $L_0(k)$ is a number of k-year-old husbands in 2002 and $L_0\ (k + 1)$ is a number of $(k + 1)$-year-old husbands in 2003.

We use the data from Polish Central Statistical Office from 2002 and 2003. There was the Polish General Census in 2002 and the data are more detailed in this year. The effective rate $\xi = 0.03$. We obtain the following values of these parameters:

$\alpha_{01} = 0{,}0706 \qquad \alpha_{02} = 0{,}1155 \quad \alpha_{13} = -0{,}0212 \qquad \alpha_{23} = 0{,}2817.$

The lifetimes T_x^M, T_y^W are PQD. In Table 1 we have the relative values of the widow's pension $a_{x|y}$ when the spouse is of the same age, i.e. $x = y$, for minimum, independent and maximum cases towards Markov model. For Markov case we have one for every age x.

Table 1. The relative values of widow pension towards Markov model

x	30	40	50	60	70	80	90
min	0.724	0.714	0.700	0.671	0.567	0.368	0.078
indep.	1.095	1.094	1.093	1.092	1.098	1.110	1.129
max	1.317	1.331	1.352	1.381	1.442	1.529	1.637

Source: own elaboration.

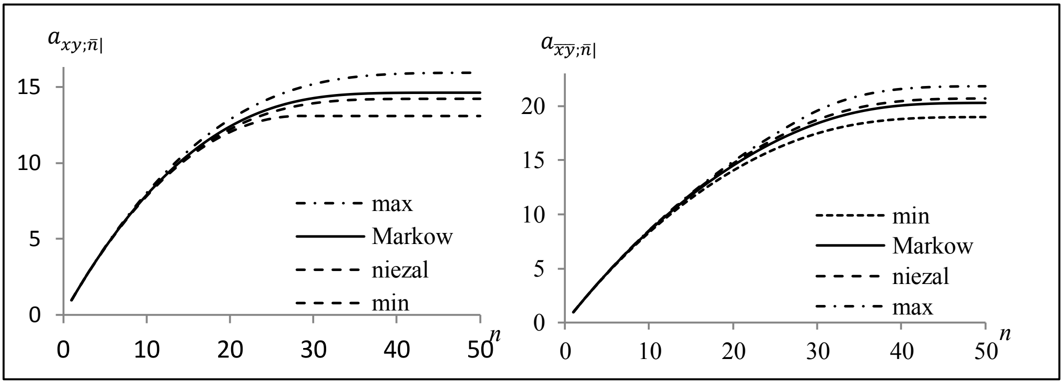

Fig. 2. The values of pensions $a_{xy;\bar{n}|}$ and $a_{\overline{xy};\bar{n}|}$.

Source: own elaboration.

We see that if the Markov model is truth, then the window's pension when we assume independent lifetimes is overestimated. This overestimate is equal to 10% on the average and it increases with the age x. We obtain the similar situation for Frechet bounds, but the errors are bigger, particularly for lower bound.

In Figure 2 the graph of values of pensions $a_{xy;\overline{n}|}$ and $a_{\overline{xy};\overline{n}|}$, when $x = y = 50$ is given. We see that the classical, independent case underestimates the n-year join-life annuity for 50 years old spouses. But this underestimation is smaller than in the previous case, smaller than 3%. For age less than 20 years, the differences are not essential and after 30 years they stabilize. Similar situation we have for n-year last-survival annuity. But the independent case overestimates the true pension.

4. Copula model

Now we assume that the survival function of lifetimes is described by the copula C, the link between the joint and marginal distributions:

$$P(T_0^M > t,\ T_0^W > s) = \mathrm{C}\big(P(T_0^M > t),\ P(\ T_0^W > s)\big).$$

The conditional survival probability connected with joint lifetimes takes the following form.

$${}_tp_{xy} = P\big(T_x^M > t, T_y^W > t\big) = \frac{C(\ {}_{x+t}p_0^M,\ {}_{y+t}p_0^W)}{C(\ {}_xp_0^M,\ {}_yp_0^W)}.$$

We derive ${}_xp_0^M$, ${}_yp_0^W$ using the empirical distributions of T_0^M, T_0^W. The raw date $n = 360$ become from two cemeteries in Wrocław.

Now we present the procedure of chose of copula [2, 3]. For simplicity, we will investigate the simpler Archimedean copula induced by the decreasing generator φ: $\{0, 1] \to \mathrm{R}_+$ only: $C_\alpha(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$ [4, 3]. First we choose four families of copulas: Clayton, Gumbel, Frank and AMH. Second, we estimate Kendall's coefficient of rank correlation $\hat{\tau} = 0.156$ and we choose from each family representative with the theoretical Kendall coefficient $\tau = \hat{\tau}$. This theoretical Kendall coefficient takes the form $\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi\ (t)}dt$. Next, we select from these representatives the "best" copula using criterion based on Kendall's function: $S_n = \int_0^1 \left|\sqrt{n}(K_n(t) - K_C(t))\right|^2 dK_C(t)$, where $K_C(t) = P(F(T_0^M, T_0^K) \leq t) = t - \frac{\varphi_C(t)}{\varphi\ _C(t)}$ is theoretical Kendall's function based on C and $K_n(t)$ is empirical Kendall's function. We obtain the smallest value of this criterion for AMH copula $C_\alpha(u,v) = \frac{uv}{1-\alpha(1-u)(1-v)}$, for $\alpha = 0{,}5879$.

In Table 2 the relative values of the widow's pension $a_{x|y}$ when the spouse are in the same age towards AMH case are given.

Table 2. The values of widow's pension

	40	50	60	70	80	90
min	0.713	0.708	0.688	0.585	0.368	0.072
indep.	1.093	1.106	1.120	1.131	1.109	1.038
max	1.330	1.368	1.416	1.486	1.527	1.506

Source: own elaboration.

We obtain little different situation than in Markov model. The biggest overestimation equal to about 13% for independent case is obtained for the age of husband death equal to 70 years, but for the greater years the overestimation **radically** decreases with the age x. The underestimation for the joint-life annuity $a_{xy;\bar{n}|}$ for the independent case is observed (see Table 3). It is about 3% and it stabilize**s** after 30 years.

Table 3. The n-year joint-life annuity $x = y = 50$

n	10	20	30	40	50
min	0.987	0.941	0.871	0.863	0.862
indep.	0.992	0.976	0.964	0.962	0.962
max	1.035	1.075	1.114	1.133	1.135

Source: own elaboration.

Figure 3 represents the values of the widow's pension for the different copulas, Markov model and independent case. We see that all cases except independency are essentially different after 60 years, but the model based on the Gumbel copula is radically different.

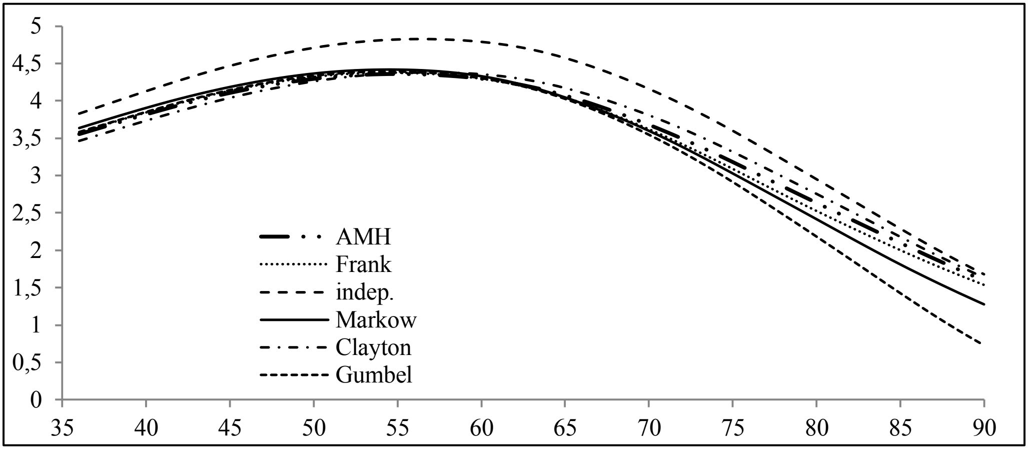

Fig. 3. The values of the widow's pension for the different copulas, Markov model and independent case

Source: own elaboration.

References

[1] Denuit M., Dhaene J., Le Bailly de Tilleghem C., Teghem S., *Measuring the impact of a dependence among insured lifelengths*, "Belgian Actuarial Bulletin" 2001, Vol. 1 (1), (), pp. 18–39.

[2] Genest C., Rivest L.-P., *Statistical inference procedures for bivariate Archimedean copulas*, "JASA" 1993, Vol. 88, pp. 1034–1043.

[3] Heilpern S., *Funkcje łaczące*, AE, Wrocław 2007.

[4] Nelsen R. B., *An Introduction to Copulas*, Springer, New York 1999.

[5] Norberg R., *Actuarial analysis of dependent lives*, "Bulletin de l'Association Suisse des Actuaries" 1989, Vol. 40, pp. 243–254.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 9 (15)

LOGISTIC REGRESSION MODELS IN POVERTY ANALYSES

Zofia Rusnak (Wrocław University of Economics)

1. Introduction

The main aim of this work is the attempt at applying the logistic regression model in order to establish poverty determinants as well as at indicating which of the proposed factors influence the probability of a certain type of household's falling into the sphere of relative poverty.

The analysis of the poverty sphere requires establishing the poverty line. In this paper the analysis is concerned with relative poverty which is a relative lack of funds for maintaining a household. The expenditure of households have been used as the indicator of household's wealth; an original equivalence OECD 0.7/0.5 scale[1] has been employed in order to calculate equivalent expenditure and make it possible to compare the households of different size and demographic composition. Half of the average equivalent expenditure calculated for the collective of all households studied in 2008 BBGD households' budget research has been used as the relative poverty line.[2]

The basis for all calculations were – bought specifically for this purpose – individual data from the BBGD households' budget research carried out by the CSO in 2008.

2. Logistic regression model (logit model)

Logit model is used for studying the relationship between the binary variable Y – which assumes only two values symbolically marked as 1, 0 – and variables X_1, X_2, ..., X_m which can be both quantitative and qualitative variables.

[1] In accordance with this scale the first adult person in a household is attributed with value of 1, every next with 0.7, and every child under 14 with the value of 0.5.

[2] Such poverty line is set by the Polish CSO for the purpose of the domestic analyses of relative poverty sphere. The relative poverty line employed by EUROSTAT for the purpose of international comparison is established as a percentage (usually 60%) of the equivalent income median, for the calculation of which the modified OECD type 0.5/0.3 scale is used.

What we want to find is the relationship between the probability of Y assuming the value 1 and the value of explained variables X_j.

Let $p = P(Y = 1)$, $\left(\dfrac{p}{p-1}\right)$ denote odds that Y assumes the value 1 and let x_j be the value of variable X_j than the *logit model* is formulated as:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = a_0 + \sum_{j=1}^{m} a_j x_j = X^T \cdot A \tag{1}$$

where A stands for the parameter vector of the A=[a_0, a_1, ..., a_m] model, and X^T for the explanatory variables vector. Using, for example, the maximum likelihood method (ML) one is able to estimate the vector of parameters A, and then calculate probability p according to the formula

$$P(Y=1) = p = \frac{e^{X^T \cdot A}}{1+e^{X^T \cdot A}} \tag{2}$$

Directional parameter a_j has the following interpretation: if the value of x_j increases by 1 unit, the odds that Y = 1 increases e^{a_j} times.

3. The determinants of poverty

This part is devoted to the analysis of the dependence between the households' risk of poverty and various features characterizing those households. Dependent variable Y is defined as follows:

$$Y = \begin{cases} 1 \text{ (when the household is poor)} \\ 0 \text{ (when the household is not poor).} \end{cases}$$

By means of available data and classifications employed in households' budget research (BBGD) in 2008, qualitative characteristics have been taken into consideration as explanatory variables and have been ascribed with categories as follows:

- variable TS determines the social and economic type of household, where:
  TS1 stands for households of workers,
  TS 2 for households of farmers,

TS3 for households of people with private enterprises,
TS4 for households of pensioners,
TS5 for households whose members do not work and maintain themselves due to social benefit;

- variable M that determines the location of household, where:
  M1 stands for cities with population of more than 100 thousand people,
  M2 stands for towns with population of fewer than 100 thousand people,
  M3 stands for villages;
- variable R that determines the region in which the household is located, where:
  R1 stands for the Central Region including łódzkie and mazowieckie voivodeships,
  R2 stands for the South Region including małopolskie and śląskie voivodeships,
  R3 stands for the East Region including lubelskie, podkarpackie, świętokrzyskie, and podlaskie voivodeships,
  R4 stands for the North-West Region including wielkopolskie, zachodniopomorskie, and lubuskie voivodeships,
  R5 stands for the South-West Region including dolnośląskie and opolskie voivodeships,
  R6 stands for the North Region including kujawsko-pomorskie, warmińsko-mazurskie, and pomorskie voivodeships;
- variable L that assumes value 1 if the household possesses savings (including deposit accounts in banks and other institutions, life insurances) and otherwise assumes value 0.

Moreover, two quantitative characteristics have been taken into account:

X_1 – the size of the household measured by the number of people in the household, where X_1={1, 2, 3, 4, 5, 6, 7+] and 7+ stands for a household of seven people or more, and

X_2 – number of children under 14 in the household, X_2 = [0, 1, 2, 3, 4, 5+], 5+ stands for a household in which there are at least five children under 14.

The structures of households in regard to the aforementioned characteristics as well as to the fact of household being considered poor or not are presented in Table 1.

Table 1. The structures of households in regard to various socioeconomic characteristics and whether they belong to poverty sphere

Classes of households with regard to:	Percentage of households [%]	Percentage of households [%]	
		poor ($Y = 1$)	not poor ($Y = 0$)
Household type TS:	**100.00**	**14.32**	**85.68**
TS1	49.96	13.8	86.2
TS2	5.36	22.68	77.32
TS3	6.63	7.79	92.21
TS4	34.35	13.13	86.87
TS5	3.70	31.88	68.11
Location M:	**100.00**		
M1	29.06	6.71	93.29
M2	28.81	11.89	88.11
M3	42.13	21.23	78.77
Region R:	**100.00**		
R1	21.53	10.06	89.94
R2	20.09	13.44	86.56
R3	17.78	19.39	80.61
R4	15.49	13.66	86.33
R5	10.68	12.43	87.57
R6	14.43	17.72	82.28
Possessed savings L:	**100.00**		
L = 1	20.46	7.29	92.71
L = 0	79.54	16.12	83.87
Number of households studied in BBGD research	**37 358**	**5 348**	**32 010**

Source: own calculations based on the BBDG data.

The data presented in Table 1 shows that among the households studied in BBGD research

- nearly 50% were households of workers,
- the majority (58%) of households were located in cities,
- the largest percentage (21.53) was the group of households from the Central Region,
- only 7.3% of households had savings and among those, the majority were bank deposit accounts (65%),

- more than 14% of the households studied in the research belonged to the relative poverty sphere.

The grounds for marking a household as poor was the relative poverty line set at the level of 50% of the average equivalent expenditure of households. This line, determined by means of the original OECD type 0.7/0.5 scale and on the basis of the data from the 2008 BBGD research, amounted to 575.2 PLN. Households whose real expenditure calculated for an equivalent unit was lower than the established poverty line were labeled poor – that is belonging to the sphere of relative poverty.

The data presented in Table 1 served as a basis for calculation of the values of the test statistic χ^2 – which for different characteristics (amounting to 849.26, 506.02, and 216.74, respectively) were much higher than the critical values responding to various levels of significance. These values advocated rejecting the hypothesis about labeling a household as poor being independent from other characteristics shown in Table 1.

The analysis of relative poverty risk in households has also been done by means of logistic regression model, in which the probability of a household being labeled as poor is dependent on the type of the household (variable TS), location (including both the kind of location M and region R), possessed savings (variable L = 1) as well as the size of the household (variable X1) and the number of children under 14 (variable X2).

Reference households consisted of one person worker households with no children under 14, located in the Central Region, in cities of over 100 thousand people, with no savings. The results of the estimation of the *logit model* are presented in Table 2. All parameter estimates are statistically significant, which means that the variables taken into consideration in this model have a significant influence on the probability of a household being labeled as poor.

Table 2. The results of the estimation of logistic regression model for the probability of a household being labeled as poor

Explanatory variables	Estimate of parameter a_i	Standard error	p	Odds ratio
Constant	–3.549	0.066	0.000	0.029
TS2	–0.169	0.063	0.007	0.844
TS3	–0.682	0.081	0.000	0.506
TS4	0.528	0.040	0.000	1.695
TS5	1.559	0.069	0.000	**4.752**
M1	–0.552	0.053	0.000	0.576
M3	0.427	0.052	0.000	1.532
R2	0.232	0.056	0.000	1.261
R3	0.363	0.064	0.007	1.438
R4	0.211	0.055	0.000	1.235
R5	0.174	0.013	0.000	1.189
R6	0.502	0.021	0.012	1.653
L	–0.817	0.049	0.000	0.442
X1	0.40	0.050	0.000	1.492
X2	0.053	0.039	0.000	1.055
Fit measures	χ^2	Total loss		
	3764.5	13 458.882		$p = 0.0000$

Source: own calculations on the basis of BBGD data.

When it comes to a group of reference households determined in this way, the positive values of parameter estimates indicate that households that are characterized by a higher probability of being labeled as poor in comparison to the reference households are households of types TS4 and TS5, located in villages, in any region but the Central Region. The probability increases along with the increasing number of people in a household as well as with the increasing number of children under 14.

Should one want to analyze the odds ratio presented in Table 2, it can be stated that

– if households are of the same type and they are located in the same class of area and in the same region, the chance of a household being considered poor increases 1.5 times per every additional person and the increase in the number of children results in the chance increased by 5.5 pp,

- if households are of the same size, with the same number of children (under 14) and are located in the same region and area of the same class, TS5 type of households are at greatest risk (the odds ratio amounts to 4.75),
- if households differ in the class of the localization only, the chance of reaching poor status is almost 1.5 times higher for village households than for those located in towns with population of fewer than 100 thousand people.

The negative values of parameter estimates in regard to other variables indicate that the decrease in the chance of reaching the poor status is caused – among others – by the fact that the household is a household of people with private enterprises or farmers, that it has savings, and that it is located in a city of more than 100 thousand people. This is depicted by the probabilities of reaching poor status for different groups of households calculated on the basis of an estimate logit model. The negative values of parameter estimates in regard to variables TS2, TS3, M1 and L are reflected in the lowest probabilities of reaching poor status by the households characterized by these variables.

Calculations done (using parameters a_i and formula 2) for households of different demographic composition and different status of possessed savings indicate that the highest probability of a household being labeled as poor characterizes households that are maintained by means of social benefit, located in villages, in the East Region. On the other hand, the lowest probability of reaching poor status is attributed to households of people with private enterprises, located in cities of more than 100 thousand people in the Central Region.

4. Conclusions

The main aim of this paper was to evaluate the significance of influence of chosen socioeconomic characteristics attributed to households in Poland on the probability of a household being labeled as belonging to the sphere of relative poverty. Logistic regression model has been used for that purpose. As a consequence, results have been obtained that made it possible to put forward the following conclusions:

- all variables taken into consideration in the analyses of relative poverty had a significant influence on the probability of a household being labeled as poor,
- among the characteristics that increase the risk of reaching the poor status one should include the size of the household and the number of children under 14, while among the characteristics that reduce the risk – possessed savings,
- in 2008 the households at greatest risk were those maintained by means of social benefit, located in villages, in the East or North Region, with no savings,
- the smallest risk of a household being labeled as poor was limited to one person households with private enterprises, with savings, located in cities of more than 100 thousand people, in the Central Region.

References

Agresti A. *Categorical Data Analysis*, Wiley, New York 1990.

Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Monografie i Opracowania 490, SGH, Warszawa 2002.

Panek T. (Ed.), *Statystyka społeczna*, PWE, Warszawa 2007.

Rusnak Z., *Statystyczna analiza dobrobytu ekonomicznego gospodarstw domowych*, Prace Naukowe Akademii Ekonomicznej nr 1182, AE, Wrocław 2007.

ON DETECTING A GRADUAL CHANGE IN AN OPEN-END SETTING

Hella Timmermann (University of Cologne)

1. Introduction

A lot of research in change point analysis focuses on the detection of an abrupt change, whereas in the case of gradual changes fewer results are known. If one expects a gradual, i.e. slowly increasing, change it seems reasonable to choose a weighted test statistic, putting the heaviest weight on the last observation, where the “size” of the change is the largest. This idea was carried out by Hušková and Steinebach [2] in an iid-setting and later by Steinebach in [4] for a general stochastic process satisfying some (weak) invariance principle. These two papers are concerned with a posteriori procedures for the testing of changes, that is, one has a complete data set to be analyzed after the observation has terminated. Chu et al. in [1], on the other hand, initiated a discussion of the monitoring problem to detect structural breaks in linear models, i.e., for the sequential detection of abrupt changes in the (unknown) model parameters. Taking up this idea, Steinebach and Timmermann in [5] approached the problem of detecting a gradual change sequentially, constructing a so called closed-end-test, i.e. a test with a fixed sample size (tending to infinity). The aim of the present work is to drop this assumption and construct a test for an open-end setting.

2. Setting of the problem

Assume we sequentially observe a stochastic process with a possible change in the drift parameter, i.e.

$$Z(t) = \begin{cases} bY(t) + at & 0 \le t \le T^*, \\ bY(T^*) + b^*Y^*(t - T^*) + at + \Delta_{m,\delta,\gamma}(t - T^*) & T^* < t < \infty, \end{cases}$$

where a, b, b^*, T^* are unknown parameters, a, b, b^* being constant and the "change-point" T^* being larger than some training period m; where m is known. The "change" $\Delta_{m,\delta,\gamma}(t)$ needs to be either non-negative or non-positive function also depending on m and some further constants $\delta \neq 0$ and $\gamma > 0$; with $\left|\Delta_{m,\delta,\gamma}(t)\right|$ being monotonically increasing, e.g. $\Delta_{m,\delta,}\ (t) = \delta \sum_{i=1}^{\lfloor t \rfloor}(i/m)$. We are interested in testing

H_0: $T^* = \infty$ *versus* $H_1^+: T^* < \infty,\ \Delta_{m,\delta,\gamma}(t) > 0\ \ \forall\, t > 0$ "one-sided change"

or

H_0: $T^* = \infty$ *versus* $H_1\ : T^* < \infty,\ \Delta_{m,\delta,\gamma}(t) > 0\ \ \forall\, t \neq 0$ "two-sided change".

For our asymptotic analysis below we assume throughout that the following (weak) invariance principles (with rate) hold: There exist (standard) Wiener processes $\{W_m(t)\}_{t \ge 0}$ and $\{W_m^*(t)\}_{t \ge 0}$ such that, for some $\kappa < 1/2$, as $m \to \infty$,

$$\sup_{0<t<\infty} \frac{|Y(t) - W_m(t)|}{t^\kappa} = \mathcal{O}_P(1)$$

$$\sup_{0<t<\infty} \frac{|Y^*(t) - W_m^*(t)|}{t^\kappa} = \mathcal{O}_P(1)$$

Example. Assume we have $X_i = \varepsilon_i + a + \delta((i - T^*)/m)_+^\gamma$, where $\{\varepsilon_i\}$ are iid with $E(\varepsilon_i) = 0$, $0 < \mathrm{Var}(\varepsilon_i) = b^2 < 1$ and $\mathrm{E}(\varepsilon_i^{1/\kappa}) < 1$ for some κ < 1 2: Further be $\delta \neq 0$, > 0 and x_+ the positive part of x: Then the process $Z(t) = \sum_{i=1}^{\lfloor t \rfloor} X_t$ fulfills the assumptions above.

Like in [1] our test statistic will be based on weighted sums of the increments $Z_i = Z(i) - Z(i - 1)$: The idea is to decide one by one, i.e. with each newly observed Z_i, whether or not the initial structure is still valid (null hypothesis) or a change has occurred (alternative). In order to take the gradual structure of a possible change into account, we put the heaviest weight on the last observation. Thereby, for the sake of generality, we make use of a weight-function $g(x)$ satisfying the following regularity conditions:

Let g: $[0, \infty) \to [0, \infty)$ be increasing and differentiable. Further, g needs to be chosen in such a way that for the function

$$G(t) = \int_0^t g^2(x)dx \tag{1}$$

it holds that $G([1, \infty)) = [G(1), \infty)$. These assumptions are for instance fulfilled if $g(x) = x_+^\lambda$, where x_+ denotes the positive part of x and $\lambda > 0$: We stop our monitoring procedure, if the detectors

$$T_k = \frac{\sum_{i=1}^k g\left(\frac{i}{m}\right)(Z_i - a)}{b\sqrt{m}}, \quad k \geq m,$$

are large in the sense that T_k divided by some boundary function h_c exceeds 1, so our stopping times are

$$\tau_m^+ = \inf\left\{k \geq m \middle| \frac{T_k}{h_c(G(k/m))} \geq 1\right\} \quad \text{(one-sided alternative)}$$

or

$$\tau_m = \inf\left\{k \geq m \middle| \frac{|T_k|}{h_c(G(k/m))} \geq 1\right\} \quad \text{(two-sided alternative)}$$

with inf $\emptyset := \infty$, $G(t)$ as in (1) and $h_c(t)$ specified in Theorem 3.1 and Remark 3.1 below. Note that these stopping times are constructed for known ("in-control") parameters a and b. In Theorem 3.2, we will replace these usually unknown parameters by suitable estimates.

3. Results

The following theorem shows, how the boundary function $h_c(t)$ needs to be chosen such that the test attains a prescribed level α asymptotically.

Theorem 3.1. Let $\{W(t)\}_{t\geq 0}$ be a Wiener process. With the notation and assumptions of the previous section, as well as, for $m \to \infty$,

$$\sup_{m\leq s<\infty} \frac{s^{\kappa} g(s/m)}{\sqrt{m} h_c(G(s/m))} \to 0,$$

$$\sup_{m\leq s<\infty} \sup_{0\leq \xi<1} \frac{g'((s-\xi)/m)\cdot 1/m \cdot \sqrt{s \log\log(s/m)/m}}{m h_c(G(s/m))} \to 0,$$

$$\sup_{m\leq s<\infty} \left| 1 - \frac{h_c(G(s/m))}{h_c(G(\lfloor s\rfloor/m))} \right| \to 0,$$

it holds under the null hypothesis that

$$\lim_{m\to\infty} P(\tau_m^+ < \infty) = P\left(\sup_{G(1)\leq t<\infty} \frac{W(t)}{h_c(t)} \geq 1 \right), \tag{2}$$

$$\lim_{m\to\infty} P\big(\tau_m < \infty\big) = P\left(\sup_{G(1)\leq t<\infty} \frac{|W(t)|}{h_c(t)} \geq 1 \right). \tag{3}$$

Remark 3.1. Possible boundary functions $h_c(t)$ are, for example, $h_c(t) = c\,t$ (for (2)) and $h_c(t) = \sqrt{\boldsymbol{t(c^2 + \log(t))}}$ (for (3)), because according to [3], Example 1 and 3, it holds that

$$P\left(\sup_{t_0\leq t<\infty} \frac{W(t)}{c\,t} \geq 1 \right) = 2\big(1 - \Phi(c\sqrt{t_0})\big),$$

$$P\left(\sup_{t_0\le t<\infty}\frac{|W(t)|}{\sqrt{t(c^2\,+\,\log(t))}}\ge 1\right)=2\left(1-\Phi\left(\sqrt{c^2+\log(t_0)}\right)+\varphi(a)\sqrt{\frac{c^2+\log(t_0)}{t_0}}\right),$$

where t_0, $c > 0$, or $t_0 > e^{-c^2}$ respectively, and Φ and φ are the distribution function and the density function of the standard normal distribution.

Now we replace the usually unknown parameters a and b by certain estimates. We estimate a by the empirical mean of the observations made up to the present time point, i.e.

$$\hat{a}_k=\frac{1}{k}\sum_{i=1}^{k}\big(Z(i)-Z(i-1)\big)=\frac{Z(k)-Z(0)}{k}.$$

As to b we need an estimate $\hat{b}_k$ which satisfies under the null hypothesis, as $m \to \infty$,

$$\sup_{k\ge m}\left|1-\frac{b}{\hat{b}_k}\right|=o_P(1) \qquad (4)$$

For Theorem 3.4 below, we also need the same rate to hold under the alternative, which can be achieved by only taking observations into account, obtained during the training period, i.e. we use an estimate $\hat{b}_k = \hat{b}_m$ for all $k \ge m$. A possible choice for $\hat{b}_m$ is the empirical variance of the increments of $Z(t)$ taken over larger intervals, say of length $h = h_m$, where $m^{\kappa}\sqrt{\log(m/h)/h} \to 0$ (see (3.23) in [5].

Plugging in the above estimates we get the following detectors:

$$\hat{T}_k=\frac{\sum_{i=1}^{k}g(i/m)(Z_i-\hat{a}_k)}{\hat{b}_k\sqrt{m}},\qquad k\ge m.$$

Incorporating $\hat{a}_k$ in the limiting behavior of the test statistic results in using a slightly modified version of our boundary function, namely we need to replace $G(t)$ by

$$\tilde{G}(t) = \int_0^t g^2(x)dx - \frac{1}{t}\left(\int_0^t g(x)dx\right)^2. \qquad (5)$$

Thus we now consider the modified stopping times:

$$\hat{\tau}_m^+ = \inf\left\{k \geq m \,\middle|\, \frac{\hat{T}_k}{h_c(\hat{G}(k/m))} \geq 1\right\},$$

$$\hat{\tau}_m = \inf\left\{k \geq m \,\middle|\, \frac{|\hat{T}_k|}{h_c(\hat{G}(k/m))} \geq 1\right\}.$$

for which we obtain the following result corresponding to Theorem 3.1.

Theorem 3.2. Assume the assumptions of Theorem 3.1 hold with $\tilde{\boldsymbol{G}}$ instead of G and let $\tilde{\boldsymbol{G}}$ ([1, ∞)) = [$\tilde{\boldsymbol{G}}$(1), ∞). Moreover we assume to have estimates $\hat{\boldsymbol{b}}_{\boldsymbol{k}}$, which fulfill the rate of (4). Then we have under H_0

$$\lim_{m\to\infty} P(\,\hat{\tau}_m^+ < \infty) = P\left(\sup_{\tilde{G}(1)\leq t<\infty} \frac{W(t)}{h_c(t)} < 1\right),$$

$$\lim_{m\to\infty} P(\,\hat{\tau}_m < \infty) = P\left(\sup_{\tilde{G}(1)\leq t<\infty} \frac{|W(t)|}{h_c(t)} < 1\right).$$

Next, we state two results under the alternative, namely the consistency of the testing procedure and the asymptotic distribution of the stopping time under H1: The two theorems below are only stated for unknown parameters, yet hold for known parameters under even slightly milder conditions.

Theorem 3.3. Let the assumptions of Theorem 3.2 hold. Further, assume there are an integer-valued, increasing function $N = N_m > T^*$ of m and estimates $\hat{\boldsymbol{b}}_{\boldsymbol{k}}$, such that, as $m \to \infty$,

$$\frac{g(N/m)\sqrt{N}}{\hat{b}_N\sqrt{m}\,h_c(\tilde{G}(N/m))} = \mathcal{O}_P(1),$$

$$\frac{\sum_{i=\lceil T^*\rceil}^{N}\left(g(i/m) - \sum_{j=1}^{N} g(j/m)/N\right)\left|\Delta_{m,\gamma}(i-T^*) - \Delta_{m,\gamma}(i-1-T^*)\right|}{\hat{b}_N\sqrt{m}\,h_c(\tilde{G}(N/m))} \xrightarrow{P} \infty$$

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 9 (15)

Then we have under H_1

$$\lim_{m\to\infty} P(\hat{\tau}_m^+ < N) = \lim_{m\to\infty} P(\hat{\tau}_m < N) = 1$$

which immediately implies that the test is consistent.

Remark 3.2. If $g = x^\lambda$ where $0 < \lambda \leq$ and $h_c(t)$ is chosen as in Remark 3.1 a possible choice of N is $N = \rho T^*$ where $1 < \rho < (\lambda + 1)^{1/\lambda}$.

Finally, we show that under stronger assumptions, i.e. essentially the so called "early-change-scenario" (see (7)) and a more precise knowledge about the kind of change (see (6)), we obtain the asymptotic distribution of the one-sided stopping times τ_m^+ and $\hat{\tau}_m^+$ under the alternative.

Theorem 3.4. Assume that the assumptions on $g(t)$ and $h_c(t)$ of Theorem 3.2 hold true. Further, let the boundary function h be continious at 1 and

$$\sup_{1\leq\xi\leq m^{-1/(2+2\gamma)}} |g'(\xi)| = \mathcal{O}_P(1),$$

$$\Delta_{m,\gamma}(i - T^*) - \Delta_{m,\gamma}(i - 1 - T^*) = \delta\left(\frac{i - T^*}{m}\right)^\gamma \quad \text{for some } \delta \in \mathbb{R}_+. \tag{6}$$

In the situation of an "early-change-scenario", i.e.

$$\frac{T^* - m}{m} \to \infty \quad \text{as } m \to \infty, \tag{7}$$

we obtain for all $x \geq -h_c(\tilde{G}(1)) = \tilde{G}(1)$

$$P\left(\frac{(\hat{\tau}_m^+ - T^*)_+^{1+\gamma}\left(g(1) - \int_0^1 g(x)dx\right)\delta}{m^{0.5+\gamma}(1+\gamma)\, b\, h_c(\tilde{G}(1))} - \frac{h_c(\tilde{G}(1))}{\tilde{G}(1)} \leq x\right) \to \Phi(x),$$

where $x_+ = \max\{x, 0\}$.

Theorem 3.4 yields the following confidence interval for the change point T^*.

Corollary 3.1. On setting

$$\chi = \left(\left(\frac{h(\tilde{G}(1))}{\sqrt{\tilde{G}(1)}} + \Phi^{-1}\left(2 - \alpha - \Phi\left(\frac{h(\tilde{G}(1))}{\sqrt{\tilde{G}(1)}} \right) \right) \right) \frac{(1+\gamma)b\, h(\tilde{G}(1)) m^{0.5+\gamma}}{\left(g(1) - \int_0^1 g(x)dx \right) \delta} \right)^{1/(1+\gamma)}$$

we obtain

$$\lim_{m\to\infty} P(\hat{\tau}_m^+ - \chi \le T^* < \hat{\tau}_m^+) = 1 - \alpha.$$

References

[1] Chu C.S.J., Stinchcombe M., White H., *Monitoring structural change*, “Econometrica” 1996, Vol. 64, pp. 1045–1065.

[2] Hušková M., Steinebach J., *Limit theorems for a class of tests of gradual changes*, “Journal of Statistical Planning and Inference” 2000, Vol. 89, pp. 57–77.

[3] Robbins H., Siegmund D., *Boundary crossing probabilities for the Wiener process and sample sums*, “Annals of Mathematical Statistics” 1970, Vol. 41 (5), pp. 1410–1429.

[4] Steinebach J., *Some remarks on the testing of smooth changes in the linear drift of a stochastic process*, Theory of Probabability and Mathematical Statistics 2000, Vol. 61, pp. 173–185.

[5] Steinebach J., Timmermann H., *Sequential testing of gradual changes in the drift of a stochastic process*, “Journal of Statistical Planning and Inference” 2011, Vol. 141, pp. 2682–2699.