Petr Volf; Tomáš Kouřim

A model and application of binary random sequence with probabilities depending on history

Kybernetika, Vol. 60 (2024), No. 1, 110–124

Persistent URL: http://dml.cz/dmlcz/152349

KYBERNETIKA — VOLUME *60* (2024), NUMBER 1, PAGES 110–124

A MODEL AND APPLICATION OF BINARY RANDOM SEQUENCE WITH PROBABILITIES DEPENDING ON HISTORY

Petr Volf and Tomáš Kouřim

This paper presents a model of binary random sequence with probabilities depending on previous sequence values as well as on a set of covariates. Both these dependencies are expressed via the logistic regression model, such a choice enables an easy and reliable model parameters estimation. Further, a model with time-depending parameters is considered and method of solution proposed. The main objective is then the application dealing with both artificial and real data cases, illustrating the method of model evaluation and its use.

Keywords: recurrent events, discrete time process, binary sequence, varying probabilities, logistic regression, time-dependent parameters

Classification: 62J12, 62N02, 60G50

1. INTRODUCTION

This contribution presents a model of binary random sequence with probabilities depending on the sequence history. An inspiration may be found in modeling of certain sports matches development. For instance, let us consider a tennis match. The sequence of its games can be viewed as a random walk with steps ± 1. Thus, a principal characteristics is here the probability $P_t = P(X_t = 1)$, where the random variable X_t denotes the result ± 1 of tth game. As a rule, this probability depends on a number of factors, both constant and changing ones, as are for instance the players ranking, their actual form, circumstances of the match (e. g. the field surface), etc. Hence, a quite natural model formulation leads us to the logistic regression, meaning that $logit\, P_t = \alpha_t + \boldsymbol{\beta}'\boldsymbol{z}(t)$, where $\boldsymbol{z}(t)$ are actual values of covariates and α_t stands for a "baseline" term.

In Kouřim [5], analyzing a large set of tennis grand-slams data, it was revealed, however, that the probability P_t depended also on previous results X_{t-1} (just memory length 1 was considered there). This dependence on the process past values was expressed via a mechanism modeling the baseline component self-development.

A variant, namely a sequence of values 1 or 0, can model a series of recurrent events (e. g. failures, repairs in a reliability study). Here, the value 1 denotes an event occurrence, 0 then means no event in time interval t. Thus, such a sequence corresponds to

DOI: 10.14736/kyb-2024-1-0110

the discrete time recurrent events counting process model, where both event occurrence and absence change future event probability.

The case studied here can be viewed from (at least) two different points of view. The first one is the framework of longitudinal data analysis. The second one consists in a discrete time variant of survival (or event history) analysis, we have preferred the second way. One of typical features of longitudinal data analysis is a presence of individual (or random) effects. This approach, originally derived for linear regression models framework, is now well developed also for binary longitudinal data (see for instance Fitzmaurice et al. [3]). On the other hand, the same is possible in the survival analysis and counting processes setting, through the frailty models (for instance Kalbfleisch and Prentice [4]), in econometric applications also called "unobserved heterogeneity" (for instance Winkelmann [12]). However, to use this effectively, one needs to have sufficiently rich data. As we were afraid that it was not the case of the data of our application, that is why this concept of frailty was not used. Though, formally, there is no problem to integrate it to logistic model. Instead, the main objective of our modeling is to describe a case where the probability of event occurrence depends on previous probabilities and events, and to show a reasonable application of this model. From this point of view, our model is a discrete time version of "self-exciting" point processes in the sense of Hawkes [2].

An initial analysis of such models performance was provided in Kouřim and Volf [6]. The properties of models were studied, their limit (large sample) properties were derived theoretically, while their behavior in small time horizon was examined graphically and with the aid of simulations. Then, in the next paper of Volf and Kouřim [9], the logistic form of the model was introduced, though without regression yet. Under this formulation the task of parameter estimation can be solved rather comfortably, namely by the standard maximum likelihood estimation (MLE) approach, yielding simultaneously the asymptotic confidence intervals of parameters. Finally, the model studied in the present contribution enriches the setting considered there, adding the regression part of the model. Even in a rather general case, when the model parameters are allowed to be time-dependent, their reliable estimation is performed easily in the generalized linear model framework.

Let us also mention here several recent papers and monographs dealing with the models and analysis of discrete random time series. The term "self-excited" discrete valued process is used quite frequently today, however in a slightly different sense, see for instance Möller [7] dealing with discrete valued ARMA processes and with their regime switching caused by the process development (so called SETAR processes). The paper of Davis and Liu [1] contains a rather broad definition of a discrete-time process dynamics. Formally, our definition is covered as well. The monograph of Ch. Weiss [11] offers a thorough overview of models for discrete valued time series, focusing also on discrete count data and categorical processes. Models are accompanied by a number of real examples. The problem of process prediction and the test of model fit is discussed as well.

The rest of this paper is organized as follows: Next section recalls the model formulation. Further, the method of the ML estimation in the framework of the logistic form of the generalized linear models (GLM) is described and broadened to the case of

time-dependent model parameters. Methods of both parametric and non-parametric estimation of these functional parameters will be proposed and their performance checked. Model properties and methods of its parameters estimation will first be illustrated with the aid of a randomly generated example. Finally, a real data case consisting of observation of recurrent events, namely repeated attacks of a disease, will be presented and solved. Hence, the main goal of this contribution is to recall the model introduced in previous papers, extend it to regression case (which is rather straightforward) and to show its applicability by analyzing a rather interesting real data set.

2. MODEL FORMULATION

Let us consider a sequence of binary random variables X_t, $t = 1, 2, 3, \ldots$ attaining values $1, 0$ (or $1, -1$). It is assumed that the probabilities $P_t = P(X_t = 1)$ follow a logistic regression model, i. e. that their logits have the following form:

$$logit\, P_t = a(t, z(t)) = \alpha_t + \boldsymbol{\beta}'\boldsymbol{z}(t), \;\; t = 1, 2, \ldots \tag{1}$$

where α_t is a baseline part, $\boldsymbol{\beta}$ are regression parameters, and $\boldsymbol{z}(t)$ are covariates (observed, possibly K-variate, and allowed to be time-dependent). Further, the baseline part develops along the scheme proposed and studied already in Volf, Kouřim (2023). Namely, starting from an initial α_1:
1. In the case of steps $X_t = 1$ or 0:

$$\alpha_{t+1} = \alpha_t + c_1 X_t + c_2(1 - X_t) = \alpha_t + c_2 + X_t(c_1 - c_2)$$

$$= \alpha_t + c_2 + X_t d = \alpha_1 + tc_2 + d\sum_{s=1}^{t} X_s. \tag{2}$$

2. For the walk with steps $X_t = 1$ or -1:

$$\alpha_{t+1} = \alpha_t + c_1(1 + X_t)/2 + c_2(1 - X_t)/2 = \alpha_t + (c_1 + c_2)/2 + X_t(c_1 - c_2)/2$$

$$= \alpha_t + d_1 + X_t d_2 = \alpha_1 + d_1 t + d_2 \sum_{s=1}^{t} X_s. \tag{3}$$

Notice that the variables α_t (and therefore also probabilities P_t) form also a random sequence and are dependent on the whole history of $X_s, s < t$. Simultaneously, a two-variate sequence (P_t, X_t) is Markov, provided the values of covariates are given.

Parameters c_j, $j = 1, 2$ as well as α_1 can attain all real values (though values far from zero are not expected in real cases), hence it is quite natural to test whether they are significantly different from zero, or whether they are positive (negative), whether $c_1 = c_2$, etc. Notice also that $c_1 < 0$ reduces the probability of success $P_{t+1} = P(X_{t+1} = 1)$ after $X_t = 1$, while the value of c_2 shows the reaction of probabilities to the opposite result (0 or -1). It is seen that the model can be re-parametrized, in case 1 using parameters c_2 and $d = c_1 - c_2$, in case 2 with $d_1 = (c_1 + c_2)/2$, $d_2 = (c_1 - c_2)/2$.

In the sequel, we shall deal with the first case only, i. e. with the sequences having values 1 or 0. We shall also, at least for now, assume that parameters $\boldsymbol{\beta}, c_1, c_2$ are constant, this assumption will be relaxed later.

2.1. Maximum likelihood estimation

The likelihood function for one process running through times $t = 1, 2, \ldots, T$ equals

$$\mathcal{L} = \prod_{t=1}^{T} P_t^{X_t} \cdot (1 - P_t)^{(1-X_t)} = \prod_{t=1}^{T} \exp[a(t, \boldsymbol{z}(t))X_t] \cdot \frac{1}{\exp(a(t, \boldsymbol{z}(t))) + 1}.$$

As a rule, N processes are observed, let us denote them $X_{t,i}, t = 1, \ldots, T_i$, $i = 1, \ldots N$. The log-likelihood function in logistic model then equals

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \{X_{t,i} a_{t,i} - \ln(\exp(a_{t,i}) + 1)\},$$

where $a_{t,i}$ are the logits of probabilities $P_{t,i} = P(X_{t,i} = 1)$ in the ith process and time t. It follows from (1) that they are

$$a_{t,i} = \alpha_{t,i} + \boldsymbol{\beta}' \cdot \boldsymbol{z}_i(t) = \alpha_1 + c_2 \cdot (t-1) + d \cdot Y_{t-1,i} + \boldsymbol{\beta}' \cdot \boldsymbol{z}_i(t), \tag{4}$$

where we denoted $Y_{t,i} = \sum_{s=1}^{t} X_{s,i}$ ($Y_{0,i} = 0$).

From the form of (4) it is seen that the task of estimation of the whole "parameter" $\Theta = (\alpha_1, c_2, d, \boldsymbol{\beta}')$ can also be solved in the framework of logistic regression model. The design matrix entering the logistic maximum likelihood estimation procedure has $N_T = \sum_{i=1}^{N} T_i$ rows and $K + 3$ columns (K is the dimension of covariates $\boldsymbol{z}$), at each row (t, i) containing

$$(1, t-1, Y_{t-1,i}, z_{1,i}(t), \ldots, z_{K,i}(t)).$$

Hence, the row corresponding to object i and time t depends also on past members of ith sequence up to time $t-1$. However, at t the value of "covariate" $Y_{t-1,i}$ is already known. Therefore, for each fixed i the relevant likelihood part is again obtained as a product of components corresponding to different t-s, as they represent conditional distribution of $X_{t,i}$ given the past. It means that the likelihood can be written in a standard way, as a product of components corresponding to different t and i. Notice also that the large-sample properties are connected here with $N \to \infty$, while lengths of sequences are assumed to be finite, bounded uniformly. In fact, such a conditioning is common also in the continuous time survival analysis, where the intensity at t may depend on the process history up to t^-.

From the logistic form of the model it further follows that both the first and second derivatives of L are tractable and the MLE as well as the asymptotic variance matrix of estimates can be computed with the aid of a convenient numerical procedure (e. g. the Newton–Raphson algorithm). Moreover, these algorithms are included standardly in data-analysis software packages, mostly as a part of methods for generalized linear models. Numerical examples presented here will utilize the Matlab function *glmfit.m*.

2.2. Case of time-dependent parameters

In many instances the impact of process history to its future steps could change during observation period and therefore the time-dependent parameters $c_1 = c_1(t)$, $c_2 = c_2(t)$

should be considered. Then $d = c_1 - c_2 = d(t)$ as well. Furthermore, the impact of covariates can change as well, it means that regression parameters $\boldsymbol{\beta} = \boldsymbol{\beta}(t)$ are allowed to be time-dependent, too. This phenomenon is observed rather frequently in lifetime, social or demographic studies. It opens a question of a flexible estimation. The problem is solved quite similarly as in other regression model cases: Either the parameters-functions are approximated by certain functional types (polynomial, combination of basic functions, regression splines) or constructed by a smoothing method, similar to moving window or kernel regression approach. The method described in Murphy and Sen [8] is of such a type and concerns the Cox regression model. All these approaches can be adapted easily to the logistic model form, just the number of parameters will be larger than in a constant-parameters model.

While inserting time-dependent parameters into a standard logistic regression model is rather straightforward (this model will be utilized, too, for comparison of results of final real-data example), in the case of our model given by (1) and (2), expression (2) has to be re-formulated, in the following way:

$$\alpha_{t+1} = \alpha_t + c_1(t) \cdot X_t + c_2(t) \cdot (1 - X_t)$$

$$= \alpha_t + c_2(t) + d(t) \cdot X_t = \alpha_1 + \sum_{s=1}^{t} c_2(s) + \sum_{s=1}^{t} d(s) \cdot X_s. \tag{5}$$

Here α_1 is still an initial value of intercept, while both other parameters c_1, c_2, as well as regression parameter(s) $\boldsymbol{\beta}$, are allowed to be time-dependent. Then the log-likelihood has the form (compare it with (4)):

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \{X_{t,i} a_{t,i} - \ln(\exp(a_{t,i}) + 1)\},$$

where now

$$a_{t,i} = \alpha_1 + \sum_{s=1}^{t-1} c_2(s) + \sum_{s=1}^{t-1} d(s) \cdot X_{s,i} + \boldsymbol{\beta}(t)' \boldsymbol{z}_i(t), \tag{6}$$

with $a_{1,i} = \alpha_1 + \boldsymbol{\beta}(1)' \boldsymbol{z}_i(1)$.

In order to specify the form of design matrix in the present case, let us as an example consider just linear parametric functions $c_2(s), d(s)$: $c_2(s) = \gamma_0 + \gamma_1 \cdot s$, $d(s) = \delta_0 + \delta_1 \cdot s$; let also $\boldsymbol{\beta}(s) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \cdot s$. Then

$$a_{t,i} = \alpha_1 + \gamma_0(t-1) + \gamma_1 \sum_{s=1}^{t-1} s + \delta_0 \sum_{s=1}^{t-1} X_{s,i} + \delta_1 \sum_{s=1}^{t-1} s X_{s,i} + (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 t)' \boldsymbol{z}_i(t).$$

Corresponding design matrix has therefore again $N_T = \sum_{i=1}^{N} T_i$ rows, however now $2K + 5$ columns, each row (t, i) containing

$$\Big(1, t-1, \sum_{s=1}^{t-1} s, \sum_{s=1}^{t-1} X_{s,i}, \sum_{s=1}^{t-1} s X_{s,i}, \boldsymbol{z}_i'(t), t\boldsymbol{z}_i'(t)\Big),$$

with $\boldsymbol{z}_i'(t) = (z_{i,1}(t), \ldots, z_{i,K}(t))$.

As regards a non-parametric moving window method, it can consist of following steps: The estimation procedure starts from an initial estimate of parameter α_1 (obtained e. g. from constant model or polynomial model described above). Then other parameters c_2, d, $\boldsymbol{\beta}$ are estimated repeatedly, like constant parameters, however with data weighted by a Gauss kernel centered sequentially at M time points S_m, $m = 1, 2, .., M$ selected inside $[1, T = \max T_i]$. In such a way, M preliminary rough estimates of parameters, $c_2(S_m)$, $d(S_m)$, $\boldsymbol{\beta}(S_m)$ are obtained. After that, these rough estimates are smoothed secondary, again with a Gauss kernel, to obtain smooth curves of $c_2(t)$, $d(t)$ and $\boldsymbol{\beta}(t)$ given at all $t = 1, 2, \ldots T$. Initial estimate of α_1 can be varied in order to obtain maximal possible value of the log-likelihood. The procedure result depends on the choice of "window width" parameter, i. e. the standard deviation of Gauss density used as the kernel function. The implementation to Matlab is rather easy, as the function *glmfit.m* is able to work with different weights assigned to each data-point.

3. ARTIFICIAL NUMERICAL EXAMPLE

The objective of examples with randomly generated data is, firstly, to study the behavior of modeled processes, and, secondly, to examine how well the MLE performs. We shall present here just one such example, with constant parameters, as, in the next part, the main goal will be the analysis of quite interesting real data case.

The data of our example were generated from the model with initial $a_1 = 1$, constant parameters $c_1 = -0.6$, $c_2 = 0.3$ as well as constant (just 1-dimensional) regression parameter $\beta = 0.5$. Covariates were generated as a $N \times T$ matrix of random numbers between 0 and 2, hence in fact they were changing with time. Here N is the number of sequences, T is the length of each.

Two cases were compared, in the first one just $N = 20$ walks, each being of length $T = 20$, then the second experiment contained a larger set, $N = 100$, of longer sequences, $T = 100$. The ML estimates, together with their standard errors based on approximate normality of the MLE, are displayed in Table 1.

N, T	a_1	c_2	d	β	$c_1 = c_2 + d$
20, 20	0.8084	0.2409	-0.7666	0.6714	-0.5257
st. dev.:	0.3104	0.0578	0.0801	0.1039	0.1116
100, 100	0.9975	0.2978	-0.8940	0.4831	-0.5962
st. dev.:	0.0796	0.0096	0.0282	0.0391	0.0186

Tab. 1. Estimated parameters and asymptotic standard deviations.

It is seen that even for a case with small number of observations the estimates are quite reasonable, simultaneously all values are significant statistically. And, as expected, the results of the larger data case are more precise.

Figure 1 then shows the development of a_t and P_t, namely their averages and variances from generated 100 walks. It is seen that both stabilize rather quickly, which is the consequence of negative c_1 and positive c_2 reducing P_{t+1} after event $X_t = 1$ and increasing it after $X_t = 0$.

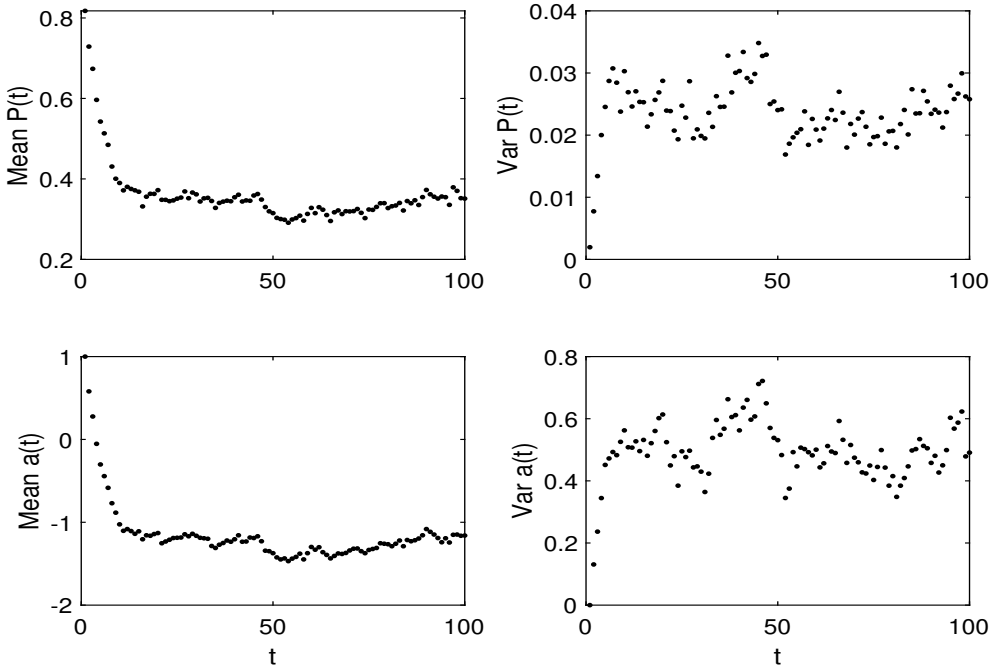

Fig. 1. Sample means and variances of a_t and P_t.

4. REAL DATA EXAMPLE

The data were analyzed originally already in 1980 and concern to patients with superficial bladder tumors. They are in detail described in Kalbfleisch and Prentice [4], Ch. 9. The data set contains records of $N = 86$ patients. Just some of them had repeated tumors occurrence, the observation started after the first tumors were removed. The time was measured in months, maximal length of record was 64 months, while maximal time of repeated tumor occurrence was 53. One patient with very short time of observation (just several days, without recurrence) was not included to present analysis, hence records of 85 persons was examined.

The patients were randomly separated to 2 groups, 38 were treated by the thiotepa, they had together 45 recurrences, while 47 patients were assigned to placebo, in this group 87 recurrence times were observed, i. e. 132 recurrences together, in average 1.553 relatively to one person (1.1842 in the first group, 1.8511 in the second). Number of observed repetitions varied from 0 (19 in the placebo group, 20 in the other group) to

maximal number 9 (one case). Just in 8 cases there were 5 or more recurrence times, while the 4th repetition occurred 14 times. Therefore, probably, another statistical analysis in Wei et al. [10] took maximally 4 repetitions into account. Except the covariate $Z_1 = 0$ for placebo and $Z_1 = 1$ for the treatment, there were other 2 discrete covariates, Z_2 and Z_3 corresponding, respectively, to the number of initial tumors and the size of the largest initial tumor.

We shall first repeat briefly results presented in Kalbfleisch and Prentice [4], considering continuous time setting. Their basic model was the Cox one, with indicated three covariates constant in time. Let us recall here that the Cox model specifies the form of hazard rate of new event occurrence at time t for object i as

$$h_i(t) = h_0(t) \cdot \exp(\boldsymbol{\beta}' \boldsymbol{z}_i(t)),$$

where $h_0(t)$ is a common, baseline hazard rate, $\boldsymbol{\beta}$ are regression parameters and $\boldsymbol{z}_i(t)$ are values of covariates corresponding to object i at time t^-.

Further, the model was enhanced by terms expressing possible changes of hazard rate after each disease attack: The hazard rate was multiplied by the term $\exp(\gamma_j)$ for subjects experiencing jth recurrence (i. e. the repetition of attack), at time $t_{i,j}$. Again, just consequences of first four recurrences were modeled, leading in fact to four new covariates – indicators $v_{i,j}(t) = 1[t > t_{i,j}]$ for the ith subject. Thus, the hazard rate of a new tumor occurrence for subject i at time t was

$$h_i(t) = h_0(t) \cdot \exp(\boldsymbol{\beta}' \boldsymbol{z}_i) \cdot \exp(\boldsymbol{\gamma}' \boldsymbol{v}_i(t)), \tag{7}$$

where $\boldsymbol{z}_i = (z_{i,1}, z_{i,2}, z_{i,3})'$, $\boldsymbol{v}_i(t) = (v_{i,1}(t), v_{i,2}(t), v_{i,3}(t), v_{i,4}(t))'$ were values of covariates described above, and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ were corresponding Cox model regression parameters, i. e. 3 and 4 dimensional vectors. In fact, parameters $\boldsymbol{\gamma}_j$ of model (7) correspond to changes of additional parameters estimated in Kalbfleisch and Prentice [4]. We think that this incremental form has a better interpretation, reflecting the impact of each tumor repetition separately.

Table 2 shows the results. The first two lines are results of our re-analysis of model (7), lines 3 and 4 are taken directly from Table 9.3 of Kalbfleisch and Prentice [4]. The question of main interest was whether the impact of the medication, i. e. the first covariate, was provable, in other words, whether the parameter α_1 was statistically significant. And it is seen from Table 2 that there is not a definite answer, at leas nor in these models framework.

Covariate Z_2, the number of initial tumors, was proved to increase significantly the risk of recurrences, while Z_3, the size of tumors, not. Further, the risk of further incidences appear to be increasing after the first or second recurrence, the influence of next repetitions does not seem to be significant (it is necessary to take also into account a rather small number of cases with more than 3 recurrences).

In fact, first two models have slightly different interpretation. While in the simpler model parameter β_1 characterizes a common impact of the therapy to all repeated events, in the full model this is combined with possible consequences of already occurred tumor repetition. The result of estimation in a simpler model containing just involved 3 covariates Z_1, Z_2, Z_3 (rows 3–4 of Table 2) supports the significance of Z_1. Rows 5–6 of Table 2 contain again results of own analysis taking into account just the first tumor

	β_1	β_2	β_3	γ_1	γ_2	γ_3	γ_4
estimate	-0.2882	0.1590	-0.0043	0.5129	1.1397	-0.3057	0.0925
st. dev.	0.1958	0.0490	0.0676	0.2580	0.3054	0.3558	0.3819
estimate	-0.524	0.201	-0.040				
st. dev.	0.187	0.044	0.065				
estimate	-0.5176	0.2360	0.0679				
st. dev.	0.3158	0.0761	0.1012				

Tab. 2. Estimated parameters and asymptotic standard deviations: full model (rows 1,2), model with just $Z_1 - Z_3$ (rows 3,4), model for the first recurrence (rows 5,6).

occurrence. It is seen that, again, from this point of view, the impact of Z_1 is not significant. This result actually corresponds to one of results of Wei et al. [10], they studied separately the occurrence of $j-$th repetitions, for $j = 1, 2, 3, 4$.

Thus, the main question to be answered is the influence of covariate Z_1. A conjecture is that the impact of the medication is delayed, in the Cox model setting it means that the parameter β_1 may be time-dependent. In fact, we have not seen such an analysis up to now. As the main theme of the present study is the presentation of a binary sequence model, in fact a discrete-time counting process model, we shall deal with this question using first a standard logistic regression model, then the model proposed here in Section 2.

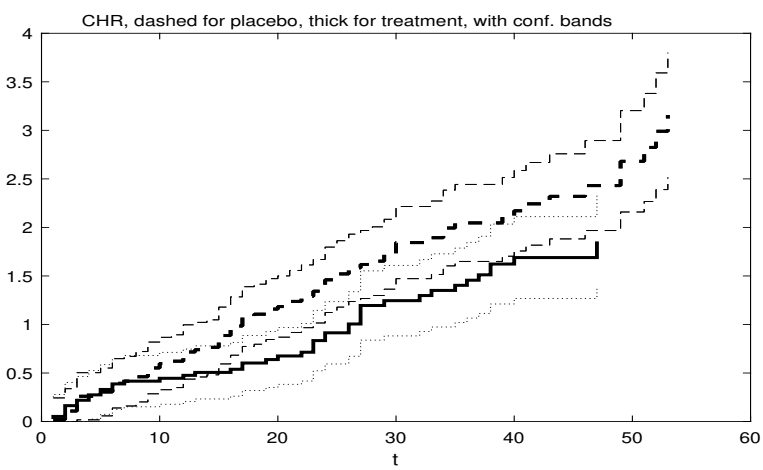

Fig. 2. Graphs of cumulative hazard functions estimated separately in groups with placebo ($Z_1 = 0$), (dashed curve, with dashed 95% confidence bands), and with treatment ($Z_1 = 1$) (thick, with dotted 95% confidence bands).

4.1. Logistic regression with time-dependent parameters

Figure 2 displays Nelson – Aalen (NA) estimates of plain cumulative hazard functions (CHR), separately in both groups (therapy or placebo), with 95% confidence bands around them derived from the Kolmogorov-Smirnov statistics. It could be assumed that the values of other covariates were distributed randomly across subjects selected for therapy or placebo due to randomization.

Let us recall briefly the NA estimator form: Let $R(t)$ be a number of objects still at risk (of event occurrence) just before time t, $\Delta_i(t) = 1$ if object i experiences the event at time t, $\Delta_i(t) = 0$ otherwise. Then the CHR estimate equals

$$H(t) = \sum_{i=1}^{N} \sum_{s \le t} \frac{\Delta_i(s)}{R(s)}.$$

From Figure 2 it is seen that estimated CHR for placebo group (dashed curve) lies above the estimated CHR of the treatment group. However, evidently, confidence bands overlap. It is also seen that from the beginning both curves almost coincide, which could be interpreted again in the sense that the impact of Z_1 to first (early) repetitions is not strong, however is increasing later, reducing further recurrences. Simultaneously, this can also lead to doubts about suitability of proportional odds assumption of constant parameters logistic regression. There are essentially two ways how to overcome the non-proportionality of odds. Either another type of model could be taken into account, or, still in the framework of logistic model, at least the parameter for Z_1 should vary in time. We shall adopt the second approach, starting from a basic logistic model. Namely, let $P(t, \boldsymbol{z}) = P(X(t) = 1 | \boldsymbol{Z} = \boldsymbol{z})$ fulfil

$$P(t, z) = \frac{\exp(\alpha + \boldsymbol{\beta}' \boldsymbol{z})}{\exp(\alpha + \boldsymbol{\beta}' \boldsymbol{z}) + 1}, \tag{8}$$

where, after an initial analysis in constant parameters model providing a starting point, both the baseline component α as well as regression parameters $\boldsymbol{\beta}$ will be allowed to vary in time.

In the sequel, two indicators of model quality will be used. First, the p-values of tests of hypotheses whether a certain parameter is significant (i. e. significantly different from zero). These tests are based on asymptotic normal distribution of estimates in the MLE setting, and are used as a primary tool for models reduction. Then, the Akaike Information Criterion (AIC) will be used for an additional model ordering, as the models contain different numbers of parameters or are of different types. Namely, AIC$=-2L + 2K$, where K is the number of model parameters, L is the value of log-likelihood.

Constant parameters. In the standard logistic regression model taking into account all three covariates, obtained parameters estimates as well as corresponding p-values are displayed in first two rows of Table 3. It is seen that β_3 is not significant, while β_1, β_2 are, therefore the covariate Z_3 has been omitted. The result is in further two rows of the table. The last column contains achieved values of the log-likelihood and the AIC.

Further, to have a comparison with results displayed in Table 2, a model extension considering an incremental impact of recurrent events was added, similarly like in (7). Thus, additional covariates were $v_{i,j}(t) = 1[t > t_{i,j}]$ for the jth recurrence observed on ith subject. Just first two disease repetitions, i. e. $j = 1, 2$, were taken into account. The results are also a part of Table 3. It was revealed that just the first recurrence impact was significant. It is seen that the model has a slightly better fit (in the sense of the log-likelihood and AIC) than the standard model, however still worse than models with time-dependent parameters (see bellow), and also than the constant parameters model in part 4.2.

	α	β_1	β_2	β_3	γ_1	γ_2	L / AIC
estimate	-3.2915	-0.4545	0.1831	-0.0482			-458.302
p-value	< 0.0001	0.0267	0.0003	0.5023			924.604
estimate	-3.4059	-0.4547	0.1908				-458.535
p-value	< 0.0001	0.0269	0.0001				923.069
estimate	-3.5176	-0.3523	0.1558	-0.0483	0.3717	0.2051	-455.177
p-value	< 0.0001	0.0939	0.0037	0.5461	0.1351	0.4322	922.355
estimate	-3.6170	-0.3615	0.1631		0.4755		-455.764
p-value	< 0.0001	0.0848	0.0016		0.0193		919.528

Tab. 3. Estimated parameters and p-values of tests of their significance: standard logistic model (rows 1,2), reduced model with Z_1, Z_2 (rows 3,4), extended model reflecting the impact of first two recurrences, full in rows 5,6, and optimized, rows 7,8.

Cubic polynomials for parameters. In this model, even the intercept component may depend on time. Therefore, "parameters" α and β_1 were modeled as cubic polynomials, while the use of polynomial for β_2 did not improve the fit significantly. The dependence on Z_3 was omitted again. Thus, the full model had 9 parameters, obtained AIC=915.121 was rather high, while $L = -448.561$ was achieved. Moreover, some components in this model were not significant statistically. After model reduction, till all p-values were smaller than 0.05, an optimal model had

$\alpha(t) = -2.9978 - 0.0008 \cdot t^2$, with p-values < 0.0001, 0.0022, respectively,
$\beta_1 = -0.1976 \cdot t + 0.0118 \cdot t^2 - 0.0002 \cdot t^3$, with p-values 0.0031, 0.0106, 0.0293,
and $\beta_2 = 0.1789$, with p-value 0.0002.

While $L = -449.037$ was slightly smaller than above, AIC = 910.074 indicated that this reduced polynomial model should be preferred. Estimated $\alpha(t)$ and $\beta_1(t)$ are displayed in Figure 3. This figure contains also the bands consisting of (connected) 95% point-wise confidence intervals for 'true' functions values at fixed t-s, computed under the assumption of asymptotic normality of estimates, with the aid of estimated asymptotic variance matrix of the ML estimates.

Non-parametric estimation. The next step consisted in non-parametric estimation of model parameters as functions via the smoothing approach described in Sub-

section 2.2. First, this approach was used in the framework of standard model (8). Again, impact of Z_3 was omitted, estimates - functions are displayed in Figure 4, left part. Log-likelihood was, naturally, higher, when compared to parametric models above, $L = -439.142$.

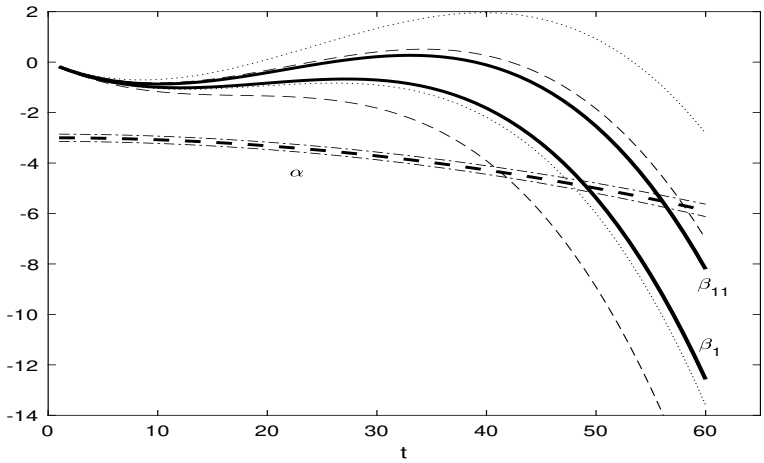

Fig. 3. Graphical comparison of cubic models components: α and β_1 are from standard logistic model (both with dashed bands connecting point-wise 95% confidence intervals). Function β_{11}, with dotted bands around it, is from 'our' model (5).

4.2. Results in the model with self-dependent probabilities

Model with constant parameters. The model described in (1), (2), and (4), compared to standard logistic model, had two more parameters, namely c_2 and d. First, all three covariates were considered, the results are displayed in Table 4, rows 1,2. The impact of Z_3 was again non-significant statistically, also the impact of Z_1 seemed to be rather weak.

Results of reduced model without covariate Z_3 are shown in rows 3, 4 of Table 4. It is seen that the results (both higher L and smaller AIC) suggest that the present model should be preferred to standard logistic model with constant parameters.

Interpretation of parameters values could be the following: $c_1 = c_2 + d$ is positive, it means that the event occurrence increases also its future probability, while negative c_2 decreases this probability after a time without new event.

Cubic polynomial for β_1. The next model contains constant parameters, too, except the parameter β_1 representing the time-dependent impact of medication on the event occurrence. Parameter α_1 is constant by definition, as it characterizes the initial probability P_1. Further, like in the standard model, use of cubic polynomial for β_2 did

	α	c_2	d	β_1	β_2	β_3	L / AIC
estimate	-2.8713	-0.0363	0.2142	-0.3677	0.1526	-0.0315	-450.073
p-value	< 0.00001	0.0001	0.0156	0.0791	0.0036	0.6636	912.146
estimate	-2.9455	-0.0365	0.2160	-0.3661	0.1575		-450.170
p-value	< 0.00001	0.0001	0.0146	0.0809	0.0021		910.239

Tab. 4. Estimated parameters and p-values of tests of their significance: full model (rows 1,2), reduced model with Z_1, Z_2 only (rows 3,4).

not improve the fit sufficiently, all its non-constant components were non-significant. Covariate Z_3 was not taken into account.

After removal all statistically non-significant components, the model consisted of following estimates (corresponding p-values are in brackets):

$\alpha_1 = -2.8081\,(< 0.0001),\ c_2 = -0.0448\,(0.0001),$
$d = 0.2237\,(0.0126),\ \beta_2 = 0.1510\,(0.0032),$
$\beta_1 = -0.1908 \cdot t + 0.0123 \cdot t^2 - 0.000187 \cdot t^3$, (p-values = 0.0051, 0.0103, 0.0233).

Log-likelihood was = -447.030, model had 7 parameters, hence the AIC = 908.059. Both values are slightly more favorable than the result of standard model with cubic components, despite that the standard model had time-dependent intercept parameter. Figure 3 compares graphically both estimates of $\beta_1(t)$. Notice that the time-development of dependence on Z_1 is quite comparable in both models and supports the conjecture on delayed effect of the medication.

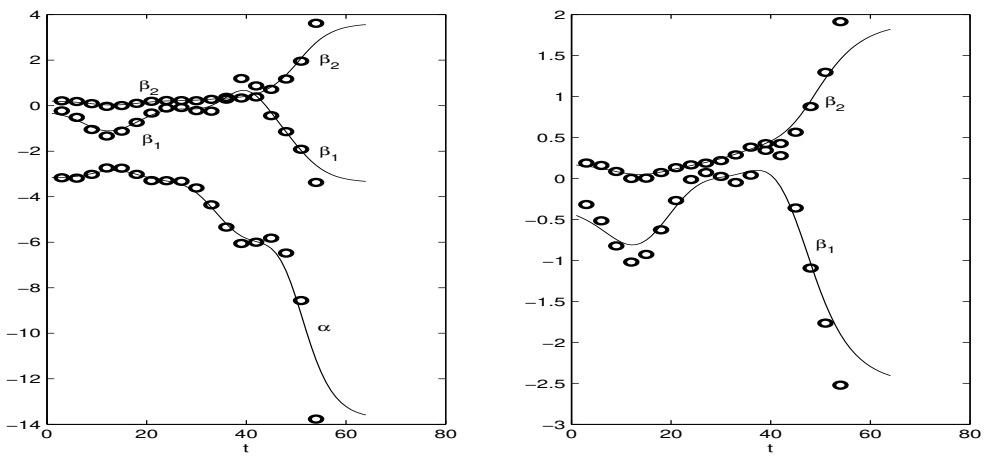

Fig. 4. Graphical comparison of models components estimated via kernel smoothing: Left – standard logistic model, right – our model. Circles come from repeated solution of weighted logistic regression, lines were obtained by a secondary smoothing. Notice different scale of vertical axis.

In the next attempt the model was enriched by cubic polynomials for parameters c_2, d. However, the model was then evidently over-parametrized, its reduction led us back to the model preferring constant c_2 and d, no further significant improvement was achieved.

Non-parametric estimation. The last analysis step consisted in a non-parametric estimation of selected model parameters as functions of time, again via the smoothing approach described in Subsection 2.2. The starting value α_1 was kept constant, while, initially, the parameters c_2 and d were allowed to vary in time. However, their variability was rather small, hence we report here just their mean values. Covariate Z_3 was again omitted.

Optimal values were: $\alpha_1 = -2.835$, means of c_2, $d = -0.055$, 0.096, respectively, functions β_1, β_2 are displayed in Figure 4, right part. These estimates led to likelihood value $L = -437.276$, i. e. again slightly better than the value achieved in the standard logistic model setting.

5. CONCLUSION

This work is a continuation of our recent results on random binary sequences with varying probabilities. Models proposed in the present paper offer an explicit description of impact of process history to actual count probabilities. A generalization may consist of considering a longer memory, we have explored just models with memory 1. On the other hand, the present model includes also an impact of covariates on probability logits. The model form makes it easy to utilize logistic regression approach and corresponding computation procedures. In this framework, certain observable events from the process history could be taken as covariates, too. The method has been successfully tested on a set of randomly generated data. Further, the model has also many possible uses in real life applications. Such a type of random sequence describes especially well processes where either a single or just a small number of events can significantly affect the process future development. Such applications can be found in reliability analysis, econometric studies, and very often in medical data analysis, as shown in the example solved in the last section.

(Received June 13, 2023)

REFERENCES

[1] R. A. Davis and H. Liu: Theory and inference for a class of observation-driven models with application to time series of counts. Statistica Sinica *26* (2016), 1673–1707. DOI:10.5705/ss.2014.145t

[2] A. G. Hawkes: Spectra of some self-exciting and mutually exciting point processes. Biometrika *58* (1971), 83–90. DOI:10.1093/biomet/58.1.83

[3] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware: Applied Longitudinal Analysis. Wiley, Hoboken 2004.

[4] J. D. Kalbfleisch and R. L. Prentice: The Statistical Analysis of Failure Time Data. Wiley, New York 2002.

[5] T. Kouřim: Random walks with memory applied to grand slam tennis matches modeling. In: Proc. MathSport International 2019 Conference (e-book). Propobos Publications 2019, pp. 220–227.

[6] T. Kouřim and P. Volf: Discrete random processes with memory: Models and applications. Appl. Math. *65* (2020), 271–286. DOI:10.21136/AM.2020.0335-19

[7] T. A. Möller: Self-exciting threshold models for time series of counts with a finite range. Stoch. Models *32* (2016), 77–98. DOI:10.1080/15326349.2015.1085319

[8] S. A. Murphy and P. K. Sen: Time-dependent coefficients in a Cox-type regression model. Stoch. Proc. Appl. *39* (1991), 153–180. DOI:10.1016/0304-4149(91)90039-f

[9] P. Volf and T. Kouřim: A model of discrete random walk with history-dependent transition probabilities. Commun. Statist. – Theory and Methods *52* (2023), 5173–5186. DOI:10.1080/03610926.2021.2004425

[10] L. T. Wei, D. Y. Lin, and L. Weissfeld: Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J. Amer. Statist. Assoc. *84* (1989), 1065–1073. DOI:10.1080/01621459.1989.10478873

[11] Ch. H. Weiss: An Introduction to Discrete Valued Time Series. Wiley, New York 2018.

[12] R. Winkelmann: Econometric Analysis of Count Data. Springer, Berlin 2008.

Petr Volf, The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: volf@utia.cas.cz

Tomáš Kouřim, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague. Czech Republic.
e-mail: kourim@outlook.com