TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky

habilitační práce

Petr VOLF

září 2000

TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií

RANDOM POINT PROCESSES AND MODELS OF HAZARD FUNCTIONS,

with application to statistical reliability analysis and quality control

Petr VOLF

habilitační práce

září 2000

RANDOM POINT PROCESSES AND MODELS OF HAZARD FUNCTIONS,

with application to statistical reliability analysis and quality control

Petr VOLF

Summary:

The main objective of the present thesis is the presentation of contemporary statistical methodology for systematic modelling and analysis of processes of events and of risk of these events, with application to the models of hazard rates (intensities) of failures and other events influencing the reliability of a system. The thesis describes the ideas and techniques developed in the corresponding fields of mathematical statistics and theory of probability, namely of the event-history analysis, survival analysis, random point processes (e.g. counting processes). The main features of the approach are explained, relevant theoretical and methodological results are presented, with particular attention to the contributions of the author. The models and methods are illustrated with the help of examples from the area of reliability analysis and control, and also of other areas. Namely, two last chapters deal with the cases and models of survival of technical devices. A selection of author's relevant publications is attached.

The methods described in the thesis are collected from many sources (cf. references), however, some of the methods and the most of the practical procedures used here has been developed, or at least adapted, by the author and his collaborators. In order to apply the methods numerically, a rather large set of algorithms had to be prepared and programmed, mostly in MATLAB. A number of references are given to other related author's papers discussing, solving, illustrating (and bringing a deeper mathematical theory to) some particular aspects of the whole theme. The examples accompanying the thesis are both artificial (though motivated by real cases), their aim is to illustrate the use of the models and techniques presented in the thesis, and the real-data cases, as in Part II.

Key words: data analysis, mathematical statistics, quality control, reliability, random point process, hazard rate, intensity of failures.

Introduction – mathematical statistics, probability, and models of reality

In R. A. Fisher's terminology (e.g. 1970), the science of statistics is essentially taken as a branch of applied mathematics. Really, the most important aim of statistical methodology is to produce and offer the tools for qualified data analysis. However, contemporary mathematical statistics is much more than "mere methodology". The development of theory (of proper use and behaviour of methods, of new ideas how to deal with data analysis problems) is closely connected with analytic mathematics. Naturally, the most close connection exists with the theory of probability, because the probabilistic models are prevailingly used for the description of the processes generating observed data. On the other hand, the approaches (and their theory) developed originally in classical statistical analysis are now extensively employed in rather general areas of "information processing", data mining, system identification, classification or pattern recognition problems, as well as the results of theory of probability are used in general systems modelling, simulation and control.

Although the progress in most areas of human activity may be regarded as a continuous process, we have to admit that in the history there were always some remarkable points at which a development progressed by a jump, by some new solution (or at least an idea of solution) of an important problem.

One of the most remarkable examples of fast transfer of new developments in mathematical probability theory to applied statistical methodology is the use of counting processes in the *event-history analysis*. By this we understand the study of a collection of objects or individuals, each moving among a finite (usually small) number of states. An "event-history analyst" is particularly interested in times of events (transitions) and in risks of these events. Compared to other branches of statistics, this area is characterized by the dynamic temporal aspect reflecting the same aspect of reality. That is why the main characteristic describing the process of events and quantifying the chance of occurrence of an event "just now" is the intensity, modelled with the help of a hazard function.

Originally, the statistical survival analysis has been developed and applied in biostatistics, medical research, as well as in the field of reliability analysis (cf. Andersen et al. 1993, Arjas 1989). The standard tools were connected either with discrete-time life tables or with continuous-time Poisson processes of events. Today, the event history analysis finds applications in all areas where the sequences of events are observed and examined, including social sciences, economics, demography. The counting processes model (generalizing the Poisson process) can significantly enrich both the description of real situations and the methodology of analysis, because the counting processes offer the connection of models of individual fates of examined objects with the model of behaviour of aggregated variables, of a system, and also the dependence of actual risk of an event on influential

factors and on the history of the system. In recent years, the flexibility of the models and their natural description of real-case processes has also been recognized in the area of insurance and financial mathematics. These fields have much common with the reliability and survival analysis (the sequence of events, risk of certain events, e.g. of large claims or of the ruining event), and naturally generalizes to models of risk and reliability of economic, financial, social systems, on both micro and macro levels. While in the past decades the complexity of structure of dependencies (on one side) and the lack of tools and techniques did not allow to go much further than to mere definitions of different versions of continuous time processes (see for instance Snyder, 1975), these methodological, theoretical and computational tools are now available or at least are developing.

Thus, the large progress has been achieved in the use of models with random effects (frailty models, e.g. Arjas and Andreev, 2000), and in the development of models using the counting processes as the basis for new classes of random processes (compound or cumulative processes of increments at random times, processes of extremes and records, see for instance Embrechts et al (1997), Grandell (1997), papers of T. Scheike (e.g. 1994), and also a series of papers of Volf (1996 - 2000). The books of Hoyland and Rausand (1994), Aven and Jensen (1999) represent the development in the field of reliability analysis. The random shocks models and models of gradual degradation of systems are at present another target of reliability studies, again with the use of random point processes and random sums (e.g. Kahle, 2000).

The objective of this study consists in the collection and presentation of the statistical methodology for the analysis and systematic modelling (via the hazard-based models dealing prevailingly with the continuous time) of the processes of events and of duration (i.e. the times to events). We shall concentrate to all important phases and steps of the model construction. In many instances it is not possible to transfer the methodology uncritically from one field of application to the other. It could result in statistical models having only weak motivation in actual problems and assuming such forms of data that are too ideal in comparison with the reality. Therefore, the proper (mostly step-wise) model construction is the task of the highest importance. The information (given by the data) has to be analyzed from different points of view, the parts of the model estimated, and the fit of the model checked, in a repeated (iterative) way. Finally, the main aim of the analysis is to prepare the "synthesis", which enables us to predict reliably the behaviour of the system under various conditions, and to suggest a proper *control strategy*.

Hence, the study is devoted mainly to the methodology of statistical inference for counting processes models with *regression* (i. e. models considering the factors affecting the behaviour of the process) and to examples showing the practical value of this methodology.

CONTENTS

PART I
COUNTING PROCESSES MODELS AND STATISTICAL EVENT-HISTORY ANALYSIS

1 Motivation from the field of reliability analysis

The reliability, as an important characteristics of the quality, means the ability of a device (a component, a system) to perform a required function for a stated period of time. Contemporary reliability (and quality) analysis has a very wide range of application areas. The main objective is not only to provide information, but also to construct models as a basis for decision, for control. Such a model (mathematical, mostly functional) has to be able to work with uncertainty (caused e.g. by the lack of knowledge, approximations). This is with advantage described with the help of probabilistic models containing random variables and processes.

The reliability itself can be quantified in different ways. We shall, for instance, consider the variable 'time to failure' (more generally – time to event, time *of* event), modelled as a random variable. Its probability distribution is mostly characterized by the hazard function $h(t) = -d\ ln(1 - F(t))/dt = \frac{dF(t)}{dt}/(1 - F(t))$, where $F(t)$ is the distribution function. The sense is that $h(t) \cdot dt$ quantifies the conditional probability of failure in a short interval $[t, t + dt)$ provided the object has survived up to t.

There are some basic examples of hazard functions. The most simple case is the constant hazard function corresponding to the exponential distribution of time-to-event. The Weibull distribution, gamma distribution are other examples of probability distributions popular (and useful) for modelling the distribution of the times to events (e.g. to failures) in reliability testing or in survival analysis. The present study will deal with a more general, nonparametrized or semiparametric, models of hazard rates. Inevitably we shall be confronted with the dilemma between the ability of the model to describe and to explain the reality and between the practical possibility to work with such a model, to identify it and to evaluate it. In other words, the choice is (and the trade-off should be made) between the completeness (flexibility) and the tractability of the model.

Let us recall some typical shapes of general hazard functions. In the field of reliability testing (and also in survival analysis of biological objects) the hazard rate of failure of a device is often "tube" shaped (cf. Figure 1a). Its shape is connected with the risk of failure during various periods of lifetime. The hazard rate describing the risk of death during our lifetime has also a very similar form (cf. Gavin et al, 1993). Let us imagine that the fates of two distinct groups of people, living in different conditions (geographical, environmental, ...), are compared. As a rule, the hazard rates are of the same shape, but for one group the rate can be higher than for the second group. Approximately $h_j(t) \sim c_j\, h_0(t)$, $j = 1, 2$, where $h_0(t)$ is some basic hazard rate and c_j are values expressing the difference of various groups or variability of objects, the

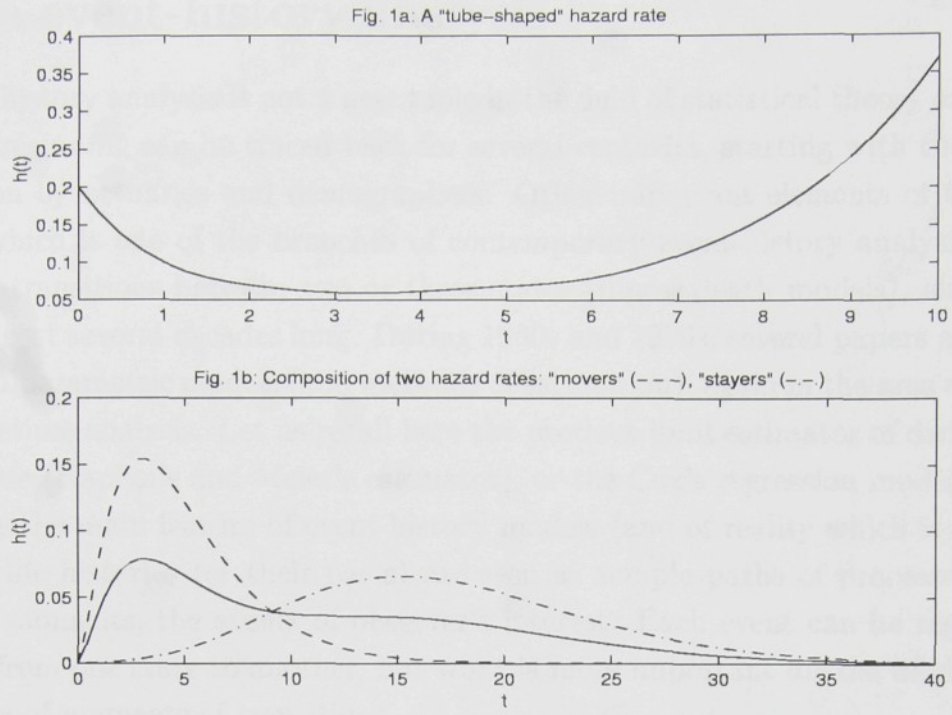


Figure 1: Examples of hazard functions

heterogeneity of the examined sample. Thus, the heterogeneity can be attached to the model of hazard in such a multiplicative way. From another point of view, c can be regarded as a random variable, *the random factor*, possessing a probability distribution throughout the population. It is also called the *frailty* variable, see Andersen et al (1993), Arjas and Andreev (2000).

Figure 1b displays another example of hazard rate, now from the field of labour force dynamics studies. Such a hazard rate can be a composition of two hazard rates specific to quite different individuals. For instance, among the people searching a new job there are two prevailing groups. The first one are people with ability, qualification and will to change their state (so called "movers"), the second group, on the contrary, are the "stayers". That is why their group-specific hazard rates of success of the search are quite different one from the other.

The approach based on hazard rates is used already in Barlow and Proschan (1967). The authors examined connections between hazard rates of components and of a system composed from these components, they modelled changes of hazard rates caused by failures of some components or by their exchanges. This monograph is today considered to be a classical book on the mathematical reliability analysis. Last two decades brought a very fast growth of both theory and methodology in the fields of random point processes and of event-history analysis. Naturally, the statistical methodology employed in the contemporary reliability analysis and quality control reflects this recent development. The present thesis will cover at least a part of this development. First, we shall recall the main ideas of the statistical event-history models.

2 The event-history analysis

The event-history analysis is not a new topic in the field of statistical theory and methodology. Its beginning can be traced back for several centuries, starting with the life-tables examination by actuaries and demographers. Other important elements of the survival analysis (which is one of the branches of contemporary event-history analysis), like the models for transitions between two or three states (illness-death models), also have the history at least several decades long. During 1960s and 1970s, several papers appeared on non- or semiparametric models (and methods of estimation), again in the area of biostatistics and lifetime analysis. Let us recall here the product-limit estimator of distribution of survival time (Kaplan's and Meier's estimator), or the Cox's regression model.

What is the main feature of event-history models (and of reality which is modelled)? Individual life histories (or their parts) are seen as sample paths of processes bringing, in random moments, the events of observer's interest. Each event can be regarded as a transition from one state to another, but what is more important for the model, it is the randomness of moments of transitions.

The main characteristic of the model of such a process of events is its rate or intensity at which the events of a certain kind occur. In biostatistics, the goal is usually in observing the time until a single nonrepeatable event (e. g. the death). In contrast, in the field of social or demographic studies, several kinds of events may be followed (e. g. transition among several labor force states, several important events in the life), some of them repeatable (e. g. change of job). Event-history data gives the type of event along with the time at which is happened, so that each observed event has also its "mark" indicating what (and to whom) occurred.

Naturally, from the point of reliability, we are mostly interested in events like the failures (of components of a system), but also in certain qualitative aspects of examined process, e.g. a crossing of tolerance limits, or an indication of fatigue of a material, etc. From this point of view, the main objective of quality control is to reduce the risk (the intensity) of such events, either immediatelly or in a certain time horizon. The methodology presented here is based on probabilistic models and statistical data analysis and offers the tools solving the following problems:

1. The analysis of relationships and dependencies, and their modelling.

2. The prediction of the risk of events, based on the model and on known (or partially known, random) inputs.

3. Repeated testing the model adequacy, detection of changes, improvement of the model.

4. The optimal decisions (control), based on solutions of points 1 - 3, in the sense of the actions controlling the factors (inputs) we are able to influence.

3 Counting process – definition and examples

The Poisson process is the most often used random point process, the counting process is its direct generalization. The n-variate counting process $N(t) = \{N_1(t), N_2(t), \ldots, N_n(t)\}$ has n components $N_i(t)$, each of them counting the number of a (registered, observed) specified event. So that with the help of $N(t)$ we can follow n types of events, or we can follow the occurrence of a certain type of event, for n objects. The time t runs *continuously*, as a rule from 0 to some finite T at which the data are collected. The time t could be the calendar time, or it could be a kind of relative time running individually for each object of the study. It is assumed that $N_i(0) = 0$ at the beginning and that N_i has jump of size $+1$ at the moment when the followed event is observed. Further, it is assumed that there are no two events at the same moment. This assumption is not fully realistic, because in praxis the data often contain a number of ties, of equal values. We shall return to this problem later. Another special feature of the event-history data is that one is rarely able to observe complete history. Therefore, the data are not complete, some event may not be observed, the part of the data is *censored*.

If only one object is followed (or a set of independent objects), arbitrary shift of the time is allowed, so that the observation of the object begins at 'its' $\tau = 0$. However, when a set of objects is examined, the histories of the objects may depend on each other, or on common covariates, in such a case the time shifts are not admissible.

The moments of events are random. This randomness is modelled with the help of *hazard functions*. In the setting of counting process $N_i(t)$, let $h_i(t)$, the hazard function, be simply a nonnegative, bounded, smooth (e. g. continuous) function defined for $t \in [0, T]$.

Let us now define, for each $N_i(t)$, the (random) indicator process $I_i(t)$ in such a way that $I_i(t) = 1$ if $N_i(t)$ is observed at moment t, $I_i(t) = 0$ otherwise – i. e. after the event which means the end of observation of i-th object, or during the period in which the fate of i-th object is censored. In other words, $N_i(t)$ is exposed to the risk of the count only if $I_i(t) = 1$. The function $\lambda_i(t) = I_i(t)\, h_i(t)$ is called the *intensity*, it is the variable (in general, it is the random process) which governs the behaviour of $N_i(t)$. In such a way, the notion of intensity is generalized, in comparison with the definition given in Section 1. Finally, define the *cumulative intensity* as $L_i(t) = \int_0^t \lambda_i(s)\, ds$. The process $M_i(t) = N_i(t) - L_i(t)$ is the martingale, so that it has several favourable properties, e. g. its mean is zero, its increments from t to $t + s$ are uncorrelated with its past up to t. Processes $L_i(t)$ are continuous and nondecreasing, $N_i(t)$ are taken as continuous from the right side, meanwhile $I_i(t)$ are continuous from the left. An example of trajectories of these processes is on Figure 2.

From the mathematical point of view, the behaviour of $N_i(t)$ is connected with the intensity $\lambda_i(t)$ in the following well known manner:

$$\Pr\{N_i(t+\Delta) - N_i(t) = 1 \,|\, \sigma(t)\} = \lambda_i(t) \cdot \Delta + o(\Delta), \tag{1}$$

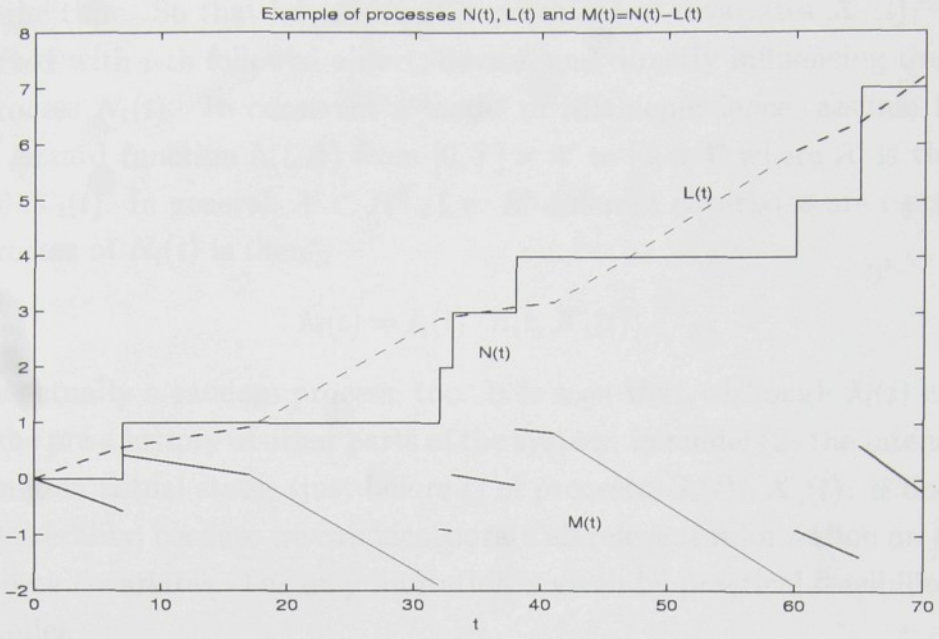


Figure 2: Example of trajectories of processes N(t), L(t) and M(t)=N(t)-L(t)

i. e. the probability, that $N_i(t)$ "counts" during a very short interval $[t,\ t+\Delta]$, conditioned by the history of the system before t, equals approximately $\lambda_i(t)$ multiplicated by the length of this interval. Formally, we can write $\Pr\{dN_i(t) = 1 \,|\, \sigma(t)\} = \lambda_i(t) \cdot dt$. Here $\sigma(t)$ (a σ-algebra of events) contains the relevant events which have occurred before t and which can influence the future. In probability theory, such a sequence of σ-algebras is called the *filtration*. From this scheme it is seen that, in general, the intensity (and therefore the behaviour) of a particular counting process $N_i(t)$ can depend on the situation of other components of $N(t)$, in such a dynamical way. This dependence can be, with advantage, expressed with the use of the *regression model*.

In the standard regression setting, the regression means the functional relationship between the explained (output, dependent) variable and a set of covariates (input, "independent" variables), as a rule contaminated by a random uncertainties. The choice of such regression models is very rich and contains models linear, nonlinear, nonparametric, onedimensional, multidimensional, and a large set of corresponding techniques of estimation, including the robust or simulation techniques. One example of nonlinear robust regression for lifetime data is in Volf and Picek (2000). However, in the present thesis we shall consider the regression in the framework of models for hazard functions.

4 Regression models

Let us now consider a hazard function depending on a set of explaining variables, predictors, covariates. The purpose of regression models is to explain (at least a part of) the heterogeneity observed in the data. The values of predictors can be fixed, deterministic, or they can be again given by observed random variables or processes, i.e. they can also

depend on the time. So that let us consider processes of covariates $\boldsymbol{X}_i(t)$, $i = 1, \ldots, n$, each connected with i-th followed object, device, and directly influencing the intensity of counting process $N_i(t)$. To construct a model of this dependence, assume that there is a (smooth) hazard function $h(t, \boldsymbol{x})$ from $[0, T] \times \mathcal{X}$ to $[0, \infty)$, where $\mathcal{X}$ is the domain of values of all $\boldsymbol{X}_i(t)$. In general, $\mathcal{X} \subset \boldsymbol{R}^K$, i. e. K different covariates are considered. The intensity process of $N_i(t)$ is then

$$\lambda_i(t) = I_i(t) \cdot h(t, \boldsymbol{X}_i(t)), \tag{2}$$

so that it is actually a random process, too. It is seen that, although $\lambda_i(t)$ can generally depend on the pre-t history of other parts of the system, in model (2) the intensity depends explicitly only on actual states (just before t) of processes $I_i(t)$, $\boldsymbol{X}_i(t)$. It does not mean any loss of generality, because we can incorporate all relevant information on the past into properly chosen covariates. The only limitation is given by practical feasibility of analysis of such a model.

In order to make the setting more tractable (from the points of mathematical theory), it is assumed that, similarly as processes $I_i(t)$, the processes $\boldsymbol{X}_i(t)$ are bounded and have their trajectories continuous from the left side. Further it is assumed that $I_i(t)$ is observable throughout the whole $[0, T]$ and that $\boldsymbol{X}_i(t)$ is observed at least when $I_i(t) = 1$.

Let us again answer the question why we are so eager to have a good model.

- First, if the good, reliable model is available (at least approximately, e. g. obtained by a procedure of estimation), we can say which covariate is relevant, crucial for the event of our interest, and in which sense. Of course, the combined influence of the factors has to be examined carefully, including the factors not considered by the present model, the heterogeneity (random, frailty) factors.

  Briefly, the good model is crucial for good analysis of (and for understanding) the system behaviour.

- Second, if we are not satisfied with the performance of the system, we can suggest a control action. From the knowledge of the model, the probability of results of this action can be evaluated. In other words, the models enables us to predict the behaviour of the system (the hazard of events of our interest) under certain combination of covariates.

- Third, two or more systems (machines, companies, groups of people, geographical areas) can be compared, various hypotheses can be tested, e.g. on best reliability of one over the other. Of course, the test of the primary interest is the test whether the chosen model is appropriate, whether it fits to the data (whether it really is good), whether it fits to the new data, too (the change detection), to each subgroup of the data (the stability test or the outlier detection).

The model for hazard function specifies the form of $h(t, \boldsymbol{x})$, i. e. the form of dependence on t and on (components of) $\boldsymbol{x}$. The parametric model means that this dependence is

expressed via a finite number of parameters. In Section 1 the most frequent types of parametric distributions have been mentioned, e. g., the exponential, Weibull, lognormal, gamma etc. For instance, the hazard rate of Weibull distribution is $h_W(t) = \alpha\beta t^{\beta-1}$, where α, β are parameters. If α (or β or both) depends on a covariate $\boldsymbol{x}$, the model expresses the regression of hazard. Again, functions $\alpha(\boldsymbol{x})$ or $\beta(\boldsymbol{x})$ can be specified with the help of other parameters, or they can remain in the general form. The model with both nonparametrized and parametrized parts is called a *semiparametric model.*

The use of parametric models is much more popular in common statistical praxis (and it is also easier). However, the analyst should be aware of the danger of improper choice of the model. Therefore it should be preferred to perform at least the preliminary analysis (for instance a graphical one) on a rather general, nonparametrized level. Even the nonparametrized models offer a rich choice of structure of dependence. For instance, we have seen in Section 1 how the "heterogeneity-specific" part is joined to the model in the multiplicative way. Hence, the same idea of separation of common hazard rate from the influence of covariates leads to the multiplicative model, called also the proportional hazard model,

$$h(t, \boldsymbol{x}) = h_0(t) \cdot \exp(b(\boldsymbol{x})).$$

Function $h_0(t)$ is a *baseline hazard function*, $b(\boldsymbol{x})$ is a regression - or *response function.* Both functions can be further specified, for instance the additive response function is frequently considered, i. e. $b(\boldsymbol{x}) = \sum_{j=1}^{K} b_j(x_j)$. Such a model is then totally multiplicative. As a rule, the estimates of nonparametrized functions are obtained in a graphical form. If the graph has any familiar shape (e. g. if it is close to linear, parabolic, exponential function), the next step consists in reparametrization (i. e. specification) of the model and in a new estimation step (now in a parametrized setting). It is seen that the component functions b_k are given up to an additive shift. If we wish to remove this ambiguity, we have to normalize the functions (e. g. we may demand $\sum_{i=1}^{n} b_k(X_{ki}(T)) = 0,\ k = 1, \ldots, k$ or "stabilize" the values of component functions in another way).

If the response function is parametrized, we obtain the semiparametric Cox's model. Its most popular form assumes a linear response function, namely $h(t, \boldsymbol{x}) = h_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x})$. In Murphy and Sen (1991) the variant with time-dependent "parameter" $\boldsymbol{\beta} = \boldsymbol{\beta}(t)$ is employed, see also Gamerman and West (1987), or models with varying coeficients, Hastie and Tibshirani (1993). Such a model then expresses the changes of the effect of a covariate during the period of the study. An alternative regression model for counting process was introduced by Aalen (1980). He suggested an additive form of hazard function

$$h(t, \boldsymbol{x}) = \beta_0(t) + \sum_{j=1}^{K} \beta_j(t)\, x_j,$$

see also Andersen et al (1993), Volf (1996). We have already seen the natural origin of the multiplicative form of regression model where the components of model intensify the influence of each other. The Aalen's model corresponds to another (more independent)

co-influence of covariates. Of course, the functions should be such that $h(t, \boldsymbol{x}) \geq 0$ on $[0, T] \times \mathcal{X}$.

There is no sharp boundary between nonparametric and parametric specifications of (a part of) a model. The nonparametric model is actually comparable with a parametric model with very high number of parameters. One class of models uses the response functions constructed from a basis of functional units, e.g. the regression splines, radial basis functions, goniometric functions, polynomials etc. Actually, the wavelet functions or feedforward neural networks are other examples of such models. The problem with such models is that, first, the parameters are 'nonlinear' and, second, that we do not know the optimal number of units, in order to obtain a good model and, simultaneously, not to overfit. Modern computational tools are able to solve such complicated optimization problems. One of tools is the Markov chain Monte Carlo (MCMC) method. For an overview see Volf and Linka (1997).

Example 1.1. The theme of the first part of the thesis will be illustrated by an example accompanying and explaining all steps of analysis and modelling. As the examples solved in parts two and three deal with the reliability of technical devices or materials, here we shall deal with analysis of "reliability" of a man as an operator of a controlled system. Namely, let us consider the following situation:

In the process of training of future operators - persons supervising and controlling a complex system (e. g. an automated control system of a plant), the candidates are tested with the help of simulated signals on their control desk, like false alarms etc. The adequacy of their response is observed. Let us image such a test during which the operator obtains randomly (with given intensity) certain alarm signals and two types of events are registered - either the delay of reaction, or the wrong reaction, the mistake. The experiment is long enough (5 hours, say), so that the fatigue increases and influences the response. If n persons are followed, n counting processes are observed in the interval [0, $T = 300$ minutes], with possible 0, 1, or more events of type 1 (delay) or 2 (mistake). We assume that no observation is censored before T, all indicators $I_i(t) = 1$, $i = 1, 2, \ldots, n$, in $t \in [0, T]$ (and $I_i(t) = 0$ for $t > T$).

The risk of the event (mistake, delay) depends on a number of factors. We shall select and follow only several of them. The first one is the age of the operator. Let us denote it X_1. It has a continuous domain of values between 20 and 60 years, say, and remains constant during the experiment. Further, let covariate $X_2 = 1, 2, 3, 4$ denote four different regimes (or conditions for the operator's work) e. g. comfortable, noisy, with higher temperature, with nonpractically designed control desk, respectively. Finally, let $X_3 = 0$ or 1 denote the non experienced or experienced operator. So that covariates X_2, X_3 are categorical, with values specific for each operator, but constant over time.

The example will continue in the next section where the sample of corresponding data will be described and analyzed.

4.1 The problem of competing risks

The case of competing risks arises when the end of the observation of an object, i.e. the switch of corresponding $I(t)$ from 1 to 0, can be caused by one from a set of possible events, which, moreover, may be mutually dependent. One event then actually censores the other. Such a situation can arise also in our Example: If there is no response at given time period, the event is denoted as "delay" and the correctness of delayed response is not checked (though the event "wrong response" can be presented, too).

The most frequent case of competing risks is caused by the presence of dependent (informative) censoring. The problem of competing risks has extensively been discussed in literature (cf. Arjas, 1989). Generally, the task of identification of marginal models of mutually dependent risks has not been solved sufficiently. To successful solution, a model of mutual dependence of examined events should be available, which actually leads to the case of bivariate (multivariate) lifetime data. Such cases are also a frequent theme of statistical research (for instance, the case of fully parametric model has been solved in Han and Hausman, 1990). The situation is worse when such a global model is not available. However, the definition of counting process contains one strong (but natural) condition – each marginal hazard at moment t is fully determined by the pre-t history. So that the dynamics of the process is given by a sequence of conditional hazards (actually, it is a continuous stream of conditional distributions of the nearest future – cf. (1) in Section 3). These conditional hazards (of competing events) are identifiable, moreover they are much more convenient for modelling the real expectations (i. e. the behaviour of real processes) than the models of marginal hazards. In general, it is possible to say that the Bayes approach is the way how the most of conditional information obtained from the data can be exploited and utilized for the conclusions on the joint and marginal distributions. Modern computational methods combine the Bayes scheme with the intensive simulation procedures, namely with Markov chain Monte Carlo algorithms (see also Volf and Linka, 1997). With their aid, optimal models are 'generated' from the data and conditional submodels. Naturally, neither these methods can overcome the lack of information and problems of nonuniqueness of solutions.

5 On techniques of estimation

The methodology of estimation of (the parts of) hazard rate models is collected elsewhere. Mainly in monographs devoted to statistical survival analysis (in a general sense), e. g. in Fleming and Harrington (1991) or in Andersen et al. (1993). The basic methods are well known, therefore we do not intend to discuss here extensively the most popular estimators. Let us only mention the famous Kaplan and Meyer product limit estimator of distribution function of the time to failure, and the Nelson–Aalen estimator of the cumulative hazard function. The same concerns to estimation in the framework of the most familiar regres-

sion models, i.e. of Cox's semiparametric model, and also of Aalen's additive hazard model. The situation is worse in the case of a more general nonparametric setting. We have already seen several "specimen" from the rich collection of hazard regression models. The aim of the analysis, at this stage, is to estimate consistently the unknown parts of the model (parameters, response functions, baseline hazard function etc.). The consistency means that with increasing amount of information, e. g. with increasing number of observed objects, the estimate converges to the "true" model (of course, in an ideal world of mathematical models). Let us again emphasize that the specification of the model is a sequential procedure which should start on a rather general (e. g. nonparametrized) level. There are only a few leading ideas on which the methods of statistical estimation are based, namely the least squares method and its modifications (including the robust variants), the maximum likelihood principle (generalized for nonparametric estimation), and also the Bayesian approach. In the sequel we intend to present (and to demonstrate on examples) some less common applications of the maximum likelihood estimation techniques, especially in the nonparametric setting.

An overview of standard techniques of estimation of hazard-based models (including the Cox's and Aalen's ones) is summarized for instance in Volf (1992), however, the main body of that paper is devoted to the *local scoring* procedure – i.e. to the method based on locally employed maximum likelihood principle. Let us explain the method briefly. For instance, in the setting of counting processes observed in $[0, T]$, the conditional likelihood, given the intensities $\lambda_i(t)$, is a generalization of Poisson likelihood function, namely

$$\mathcal{L}^c = \prod_{i=1}^{n} \left[\prod_{t<T} \lambda_i(t)^{dN_i(t)} \exp\left\{ - \int_0^T \lambda_i(t)\, dt \right\} \right].$$

Here $\lambda_i(t) = I_i(t) \cdot h(t, X_i(t))$, $dN_i(t) = 1$ just at the moment of count of $N_i(t)$, $dN_i(t) = 0$ otherwise. Let us consider a one-dimensional covariate $X(t)$ and a multiplicative hazard function $h(t, x) = h_0(t) \cdot \exp b(x)$. The maximum likelihood estimates (MLE) of functions $h_0(t)$, $b(x)$ are functions maximizing $\mathcal{L}^c$. In praxis we look for maximizers of $\log \mathcal{L}^c$, namely of

$$\log \mathcal{L}^c = \sum_{i=1}^{n} \left[\int_0^T \{\log h_0(t) + b(X_i(t))\}\, dN_i(t) - \int_0^T h_0(t) \exp b(X_i(t))\, I_i(t)\, dt \right].$$

In the case of multiplicative model, the Cox's partial likelihood function is available for the estimation of $b(x)$ independently on $h_0(t)$. The partial likelihood can be derived directly from $\mathcal{L}^c$ as a 'profile' likelihood for $b(x)$, with the aid of the MLE principle. The logarithm of partial likelihood is

$$\ell^P = \sum_{i=1}^{n} \int_0^T \log \left[\frac{\exp b(X_i(t))}{\sum_{j=1}^{n} \exp b(X_j(t))\, I_j(t))} \right] dN_i(t),$$

and an "ideal" estimate of function $b(x)$ should maximize it. Then, the cumulative baseline hazard function $H_0(t) = \int_0^t h_0(s) ds$ is estimated by the Breslow–Crowley estimator (also

derived from maximization of $\mathcal{L}^c$):

$$\hat{H}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{j=1}^n \exp \hat{b}(X_j(s))\, I_j(s))},$$

where $\hat{b}$ is estimated function b. Then, baseline hazard function can be estimated by a standard kernel smoothing of steps of $\hat{H}_0(t)$, namely:

$$\hat{h_0}(t) = \frac{1}{d}\int_{-\infty}^{\infty} W(\frac{t-s}{d}) d\hat{H}_0(s),$$

where $W(z)$ is a kernel function.

How to cope with the task of estimation of a nonparametrized function $b(x)$ from the log-likelihood (here, from ℓ^P)? The idea of local scoring (Hastie and Tibshirani 1986, 1990) is based on sequential local maximization. If we wish to estimate the value of $b(x)$ at point $x = z$, say, we take function $b(x)$ as a constant b_z in a chosen window (neighbourhood) O_z around z. Then b_z is treated as a parameter and we solve the equation $d\ell^P/db_z = 0$, using only data with $X_i(t) \in O_z,\ t \in [0, T]$. Then the window is moved to another point, etc. Thus, we walk through the whole domain of x, repeatedly, until convergence. The method can be enriched by the use of weighting kernel in each window. If covariate $\boldsymbol{x} \in R^K$, a purely multiplicative model considers additive function $\mathbf{b(x)} = \sum_{j=1}^K b_j(x_j)$. Then the algorithm of local scoring iterations contains also a loop dealing with one component after another (see also Volf 1992, 1993a, Part II here). It is seen that functions b_j are determined up to additive constants, therefore it is necessary to normalize them for instance by some 'initial' condition like $b_j(0) = 0$ and similarly.

It is also evident that the method has much common with the standard kernel smoothing of curves. However, the standard kernel smoothing is based on the weighted averaging in each window, meanwhile the local scoring employs the MLE technique. A variant of the local scoring, the technique of local maximum of likelihood computed with the help of an algorithm proposed in Volf (1992, 1993a), is applied to the solution of the lifetime estimation problem in Part II.

As it has already been said, another way to estimation of nonparametrized function $b(x)$ may consist in approximation for $b(x)$ by a parametrized function. Every smooth function can be well approached by a linear combination from some basis of functions. For instance, the splines are the popular choice. Stone (1986, 1994) used the approximation of regression function by splines in the framework of exponential family of distributions. He proved the consistency of this approximation provided the parameters of splines were estimated by (global) maximum likelihood method. Thus, the reparametrization may be considered as an alternative way to solution. From this point of view, the simplest approximation, namely the regressogram (histogram), is a trivial spline, with order 0.

A widespread discussion runs about the advantages and capabilities of both approaches – splines and kernel-like smoothing, see already a discussion to paper of Hastie and Tibshirani (1986). We here do not intend to contribute to arguments of any side, our opinion

is that every well-working method is valuable. Although some data-analysts (when joking) claim that one data may be analyzed only once and only by one method – in order to avoid contradictions and problems with interpretation of results.

Example 1.2. Let us now describe the process of data analysis and model specification, and let us show the methods and results of estimation of hazard functions and other model's characteristics, in the case described by our Example 1.1. For illustrative purposes, the data were generated artificially.

The structure of the data was the following: Number of tested persons $n = 52$, together 71 events were observed (namely 35 delays, 36 mistakes, 12 persons passed the test without any delay or mistake). The design of covariates was balanced, in the sense of more or less uniform distribution of their values and the mutual independence of covariates. We started the solution by the evaluation of the dependence of the hazard rate on time and by other rather simple explorative steps. Figure 3 shows the distribution of times of events, Figure 4 then the estimate of cumulative hazard function and of hazard function - intensity of events, in $t \in [0, 300]$.

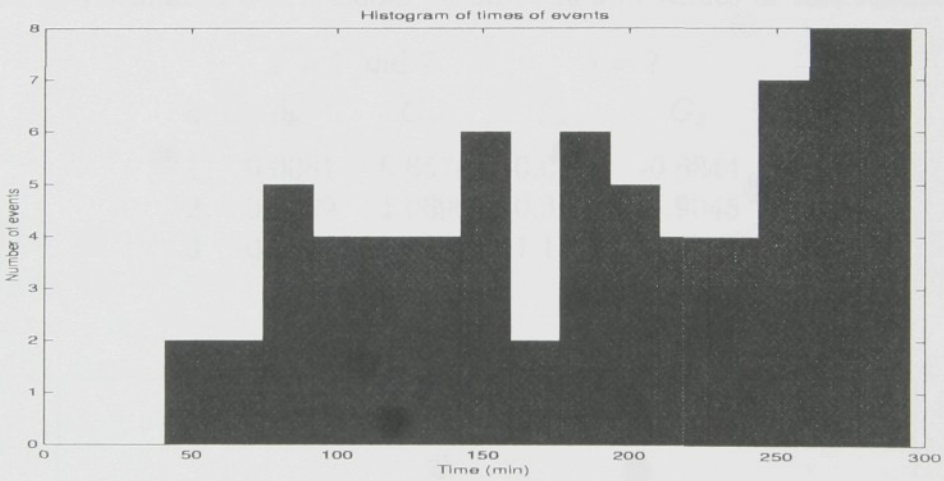


Figure 3: Distribution of times of events

Then, the dependence on covariates was analyzed, first in the framework of standard Cox's regression model i. e.

$$h(t, \boldsymbol{x}) = h_0(t) \exp \sum_{j=1}^{3} \beta_j x_j.$$

Moreover, we consider two types of events, therefore the same type of model can be proposed and evaluated for each type of event separately. The results are collected in Table 1, first for both types of events together (we denoted the indicator of event type by $\delta = 1$ or 2). Then the parameters were estimated solely for the event $\delta = 2$ (the event "mistake").

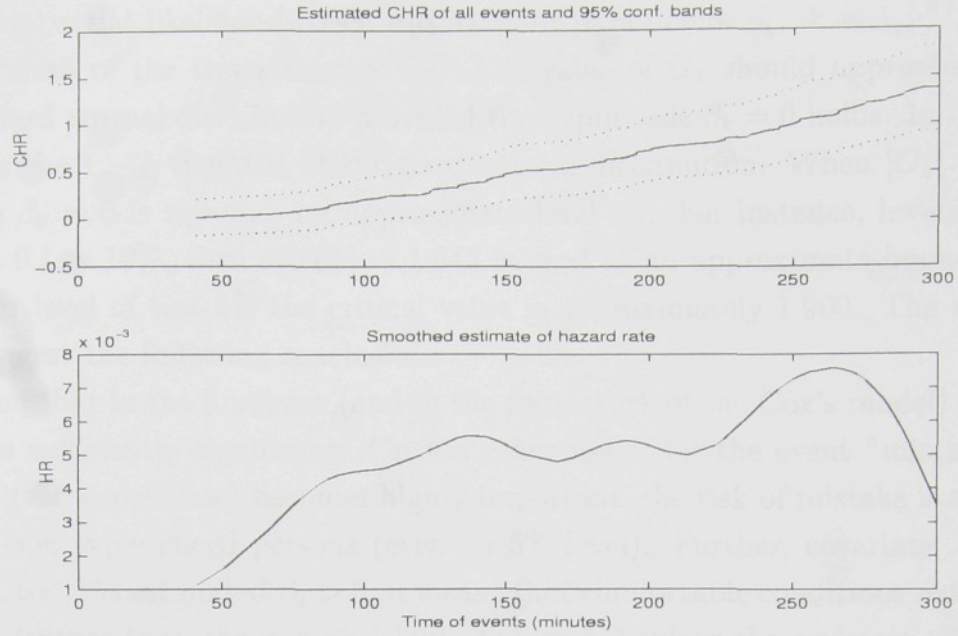


Figure 4: Estimated cumulative hazard function and hazard function

Table 1. Estimated Cox's model parameters and values of test statistics

	$\delta = 1$ and 2		$\delta = 2$	
k	β_k	G_k	β_k	G_k
1	0.0081	0.6578	-0.0116	-0.6844
2	0.1309	1.0690	0.3538	1.9045
3	-0.3325	-1.2031	-1.1473	-2.4751

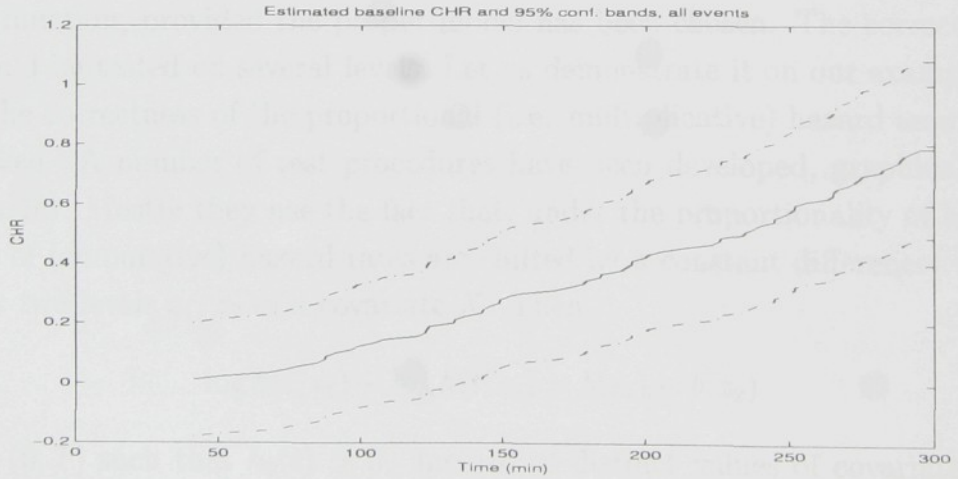


Figure 5: Estimate of cumulative baseline hazard function, in Cox's model

In the framework of Cox's model, the estimates of Cox's parameters β are asymptotically normal (cf. Andersen et al, 1993). That is why we are able to compute the test statistics having approximately Gaussian or chi-squared distribution. Let us again return to the example (and assume that the Cox's model is the right one for our data). Table 1

displays the partial likelihood-based estimates of parameters β_k, $k = 1, \ldots, 3$, together with the values of the test statistics G_k. The value of G_k should approximately come from standard normal distribution provided the hypothesis $\beta_k = 0$ holds. In other words, let $q(\alpha)$ be the $1 - \alpha$ quantile of standard normal distribution. When $|G_k| > q(\alpha)$, the hypothesis $\beta_k = 0$ is rejected, on approximate level 2α. For instance, let us choose the level $2\alpha = 0.1 = 10\%$, then $q(0.05) \doteq 1.645$ is used as an approximate bound of critical interval, for level of test 5% the critical value is approximately 1.960. The results from Table 1 suggest the following conclusions:

It is seen that in the first case (and in the framework of the Cox's model) no separate covariate is sufficiently significant. On the other hand, for the event "mistake" the covariate X_3 (the experience) becomes highly important, the risk of mistake is significantly higher for non-experienced persons (even on 5% level). Further, covariate X_2 becomes significant, too. As estimated $\beta_2 > 0$, it means that comfortable conditions reduce the risk of mistake (rather than the risk of delay). Figure 5 displays the estimate of cumulative baseline hazard function, for the first case (both event categories together).

After the first stage of estimation, in the next section we shall analyze the results and examine whether the model with time-dependent parameters is more adequate, i.e. whether the importance and influence of certain covariate changes with elapsed time.

6 Methods of statistical tests

Once the model is selected, its evaluation yields the information on actual dependencies. For instance, the general nonparametric approaches, e.g. the local likelihood (or the local scoring, or the moving window) method of estimation, are able to reveal the shape of the regression function, provided the proper model has been chosen. The correctness of the model should be tested on several levels. Let us demonstrate it on our example.

First, the correctness of the proportional (i.e. multiplicative) hazard assumption has to be checked. A number of test procedures have been developed, graphical as well as numerical ones. Mostly they use the fact that, under the proportionality of hazards, the logarithms of (cumulative) hazard rates are shifted by a constant difference. Namely, let us consider two levels z_1, z_2 of a covariate X. Then

$$\log h(t, z_1) - \log h(t, z_2) = b(z_1) - b(z_2)$$

for all $t \in (0, T]$ such that $h_0(t) > 0$. Instead of distinct values of covariate the strata around certain values may be considered. Only for the sake of simplicity, let us check the assumption about proportional hazard dependence on covariates graphically. The cumulative hazard functions have been estimated using the Nelson–Aalen estimator (separately for two groups of data).

Figure 6 compares logarithms of cumulative hazard functions (of wrong or delayed responses), for two different age groups, for value $X_2 = 1$ versus $X_2 > 1$, and also for

$X_3 = 0$ versus $X_3 = 1$. The graphs support more or less the hypothesis of proportional hazards with respect to the covariate X_1, the age. On the other hand, the proportionality of hazards dependence on other two covariates may be doubted, the plots indicate that the proportions of hazards differ in different time domains. That is why we decided to consider an enriched model allowing the dependence of parameters on time, the results are collected in the next paragraph.

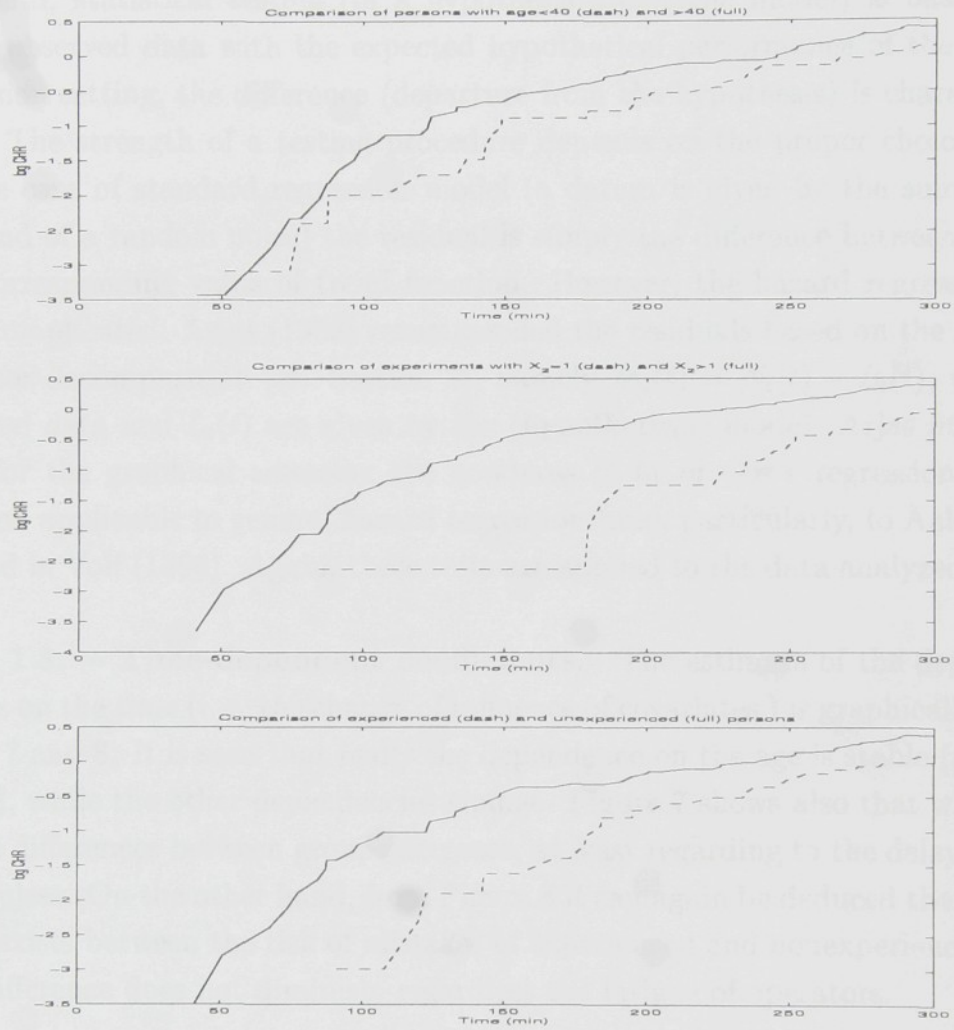


Figure 6: Comparison of logarithms of cumulative hazards

Second, the significance of regression can be doubted. The hypothesis is tested that the dependence of response on corresponding covariates x_ℓ is negligible. The test is performed with the help of asymptotic normality of estimators of Cox's model parameters (i.e. we test the hypothesis that they are zero), conclusions of such a test were already discussed above, in connection with the values in Table 1.

With some license, the conclusions of such a test may be acceptable even when the Cox's model is far from reality. However, how the (non)linearity of regression function should be checked? It is the third question to be answered by a test. One possibility is suggested in Stone (1986). Let us consider a polynomial form of the regression function,

estimate the parameters – coefficients of the polynomial. Then let us test the hypothesis that the coefficients of order higher than one are zero.

Other way of testing can be based on the linear approximation analysis of nonparametrically estimated regression function, which was used also in Volf (1992, 1993a, b). However, the results depend strongly on the "smoothing policy" during the local likelihood iterations.

In general, statistical testing (of a hypothesis about the model) is based on comparison of observed data with the expected hypothetical performance of the system. In the regression setting, the difference (departure from the hypothesis) is characterized by *residuals*. The strength of a testing procedure depends on the proper choice of residuals. In the case of standard regression model (a datum is given by the sum of a trend function and of a random noise) the residual is simply the difference between the datum and the corresponding value of trend function. However, the hazard regression models are more complicated. Arjas (1988) recommended the residuals based on the martingale-compensator decomposition (cf. Section 3), namely $M_i(t) = N_i(t) - L_i(t)$, where $N_i(t)$ are observed data and $L_i(t)$ are given by the (hypothetical) model. Arjas utilized these residuals for the graphical assessing the goodness of fit of Cox's regression model. A modification applicable to general hazard regression (and, particularly, to Aalen's model) is described in Volf (1996). Again, these tests are applied to the data analyzed in Part II.

Example 1.3. – Time-dependent coefficients. The estimate of the dependence of parameters on the time (i. e. the change of influence of covariates) is graphically displayed in Figures 7 and 8. It is seen that really the dependence on the age is stable (and not too important), while the other dependencies change. Figure 7 shows also that with growing fatigue the differences between groups decrease, at least regarding to the delayed reaction to the impulses. On the other hand, from Figure 8 it can again be deduced that the bigger difference exists between the risk of mistakes of experienced and nonexperienced persons, and this difference does not diminish, regardless the fatigue of operators.

7 Analysis of discrete-time (or grouped) data

We have seen that the hazard models of the counting processes support the characterization of individual histories, that the counting process of a system (of a group of objects) is obtained by summation of "individual" counting processes (i. e. of processes registering specified events of specified objects).

The examples have also shown that not too rough discretization of the reference time does not matter. On the contrary, sometimes the analyst discretizes the time domain artificially in order to construct a grid for histogram-like estimators of time-dependent components of model.

Another situation is encountered when not only the period of study is split into a

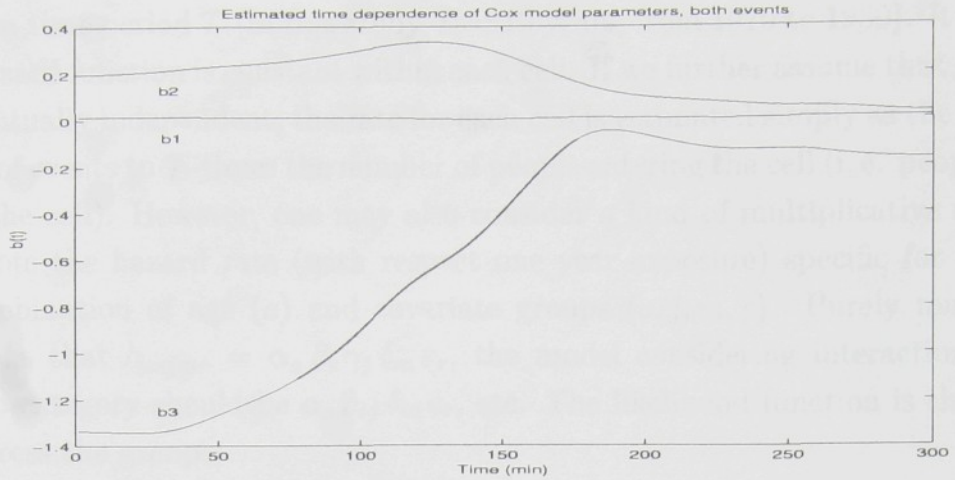


Figure 7: Estimated time-dependence of Cox's model parameters

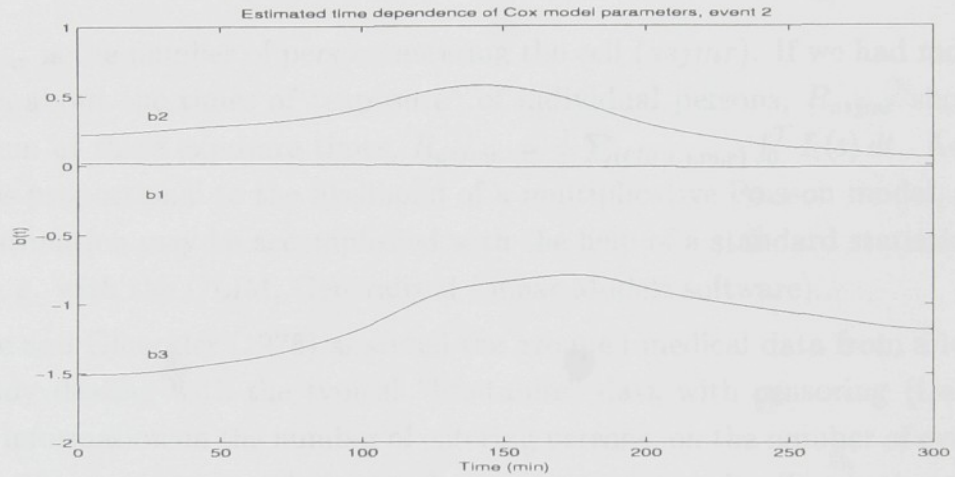


Figure 8: Estimated time-dependence of parameters of Cox's model for the risk of event "mistake"

sequence of intervals but, moreover, the individual data are summarized (grouped) in these intervals. We then have to deal with the *grouped data*, the structure typical for life tables. The values of covariates have to be grouped as well, so that the data are a result of a multidimensional grouping of events. As the covariates are either categorical or discretized, a rather general discrete model may then be constructed with the help of finite number of parameters. Let us now show these aspects with the aid of an example, from the field of sociological studies.

Example 2 (Andersen et al., 1993, example VI.1.3.). The example deals with the rate of suicides among nonmanual workers in Denmark. The persons had been grouped to the age groups (w.r. to age at the beginning of the study), several categorical covariates, e. g. sex, job category, marital status, regional characteristics. Thus, a set of groups (cells) has been obtained by a cross-classification with respect to all these categories. For each cell, the data contain the number of exposed people and the number of events (the suicides)

during given time period $\mathcal{T}$ (in this study $\mathcal{T} = 10$ years, from 1970 to 1980). It is assumed that the hazard function is constant within each cell. If we further assume that the rates in cells are mutually independent, the rate for each cell is estimated simply as the proportion of number of events to $\mathcal{T}$-times the number of people entering the cell (i. e. people exposed to risk in the cell). However, one may also consider a kind of multiplicative model. Let h_{asjmr} denote the hazard rate (with respect one year exposure) specific for a cell with certain combination of age (a) and covariate groups (s, j, m, r). Purely multiplicative model means that $h_{asjmr} = \alpha_a \beta_s \gamma_j \delta_m \varepsilon_r$, the model considering interactions between sex and job category should be $\alpha_a \emptyset_{sj} \delta_m \varepsilon_r$, etc. The likelihood function is the following product across the groups,

$$\prod_{\{a,s,j,m,r\}} \left\{ (h_{asjmr})^{N_{asjmr}} \exp(-h_{asjmr} \mathcal{T} R_{asjmr}) \right\}$$

where R_{asjmr} is the number of persons entering the cell $(asjmr)$. If we had more detailed information about the times of "exposure" of individual persons, R_{asjmr} should be the averaged sum of these exposure times, $R_{asjmr} = \frac{1}{\mathcal{T}} \sum_{i \in (a,s,j,m,r)} \int_0^{\mathcal{T}} I_i(t)\, dt$. Actually, the likelihood is proportional to the likelihood of a multiplicative Poisson model, so that the practical estimation may be accomplished with the help of a standard statistical analysis software (e. g. with the GLIM, Generalized Linear Models software).

Prentice and Gloeckler (1978) analysed the grouped medical data from a longitudinal clinical study dealing with the typical "life-tables" data with censoring (i. e. the data containing information on the number of entering persons, on the number of events and on the number of censorings at each time and covariate-specific cell). The authors considered the possibilities how to fit the Cox's regression model to the grouped data. It is due to the discrete character of the data that the Cox's model is equivalent to the multiparameter multiplicative model from the preceding Example 2. Again, the precision of estimation is limited by the noncomplete information about actual periods of exposure of the persons under study (i. e. about exact moments of events and censoring). A method of statistical tests for grouped survival data has been proposed in Volf (1982).

On the other hand, we should not forget that the Cox's model assumes the existence of *continuous* baseline hazard function $h_0(t)$. By the techniques described in Example 2 only the "crude" estimation h_0^* of h_0 is obtained (in Example 2, α_a is actually such an approximation to $h_0(a)$). In fact, this crude estimate is a histogram (with more or less dense nodes). So that the next step can consist in the smoothing of the estimate, mostly with the help of the standard kernel smoothing procedures. The kernel smoothing of hazard rates from the step-wise crude estimates (or from step-wise estimates of cumulative hazard rates) ranks among the common techniques in the field of event-history analysis. The properties of estimators are described elsewhere (e. g. in Andersen et al., 1993, cf. also Volf, 1992), in the continuous time setting the consistency of estimation is guaranteed. In the field of mathematical demography, the smoothing procedure is called *graduation*

(cf. Gavin et al., 1993). The same procedure may be applied to the smoothing of the crude (group-specific) estimates of covariate effects provided the natural scale of covariate is continuous and is discretized artifically by grouping.

Notice that in the case of Example 2 the model can be reformulated in such a way that the age is just one of the covariates, meanwhile the reference time is missing. The role of reference time could be played by the calendar time from $t = 0$ ($\sim$ 1970) to $t = \mathcal{T}$ ($\sim$ 1980), but the analysts assumed that the risk of suicide did not change during the followed time period.

In such a case as described by Example 2, when we know only the initial numbers of exposed objects (here, persons) and the dynamics of the process of events is not taken into account, it could be more convenient to abandon the hazard rate models and to describe directly the probability of occurrence of the event (in given time period), for instance with the help of logistic regression model (for an example of nonparametric logistic regression model, see Volf, 1993a).

When speaking about the secondary smoothing of estimated dependence of hazard function on a discrete-valued covariate, it should be emphasized that the local scoring procedure enables us to accomplish the smoothing automatically. It depends only on the choice of the width of window. If the window is narrow, the local scoring solve standard parametric task of maximum likelihood estimation (estimating one parameter after another, or one subset of parameters after another subset, not all parameters jointly). If the windows are wider than the distance between categories (for instance, if covariate "age" is grouped to two years classes and the corresponding window is chosen greater than four), the smoothing is performed.

The methods of analysis of grouped data are actually outside the scope of the present study. However, in many cases, the procedures described here can easily be modified and employed for such a situation. The limitations of the counting processes approach to the statistical life-history analysis are widely discussed by Andersen et al. (1993). They believe that the most nonfavourable cases are caused by such a lack of information that no other techniques can reach significantly better results.

"Detailed life history data may be given a thorough analysis using the methods based on counting processes. However, if less precise information is available, then alternative techniques are necessary. Some such techniques can also be based on counting process idea, but not in the simple form..." Andersen and Borgan (1985)

8 Control of the process of events' intensity

If the intensity process depends on covariates, the natural way how to control this process is the control the covariates, i.e. at least these which can be influenced by our decisions and actions. The sense of a control consists in the reduction of the number of nonfavourable events (i.e. failures), as well as in the optimization of the cost of events. If the intensity of

failures is known deterministic function (under different control conditions) the problem can be treated as the standard problem of optimization of exchanges and supplies, with Poisson process of events. However, in the more general concept considered here, the intensity process is a random process, influenced by random inputs - covariates, and by the random indicator process, too. Hence, for optimal decisions, the development of covariate processes has to be estimated, predicted (Aven, 2000). Then, we can again deal with the intensity, conditionally on given inputs, as with the determined function, and therefore solve the problem of *prediction* of the process of events, in the sense of the mean number of expected events, or in the sense of 90% (for instance) prediction intervals or bands. The prediction interval at a given t says that with 90% probability the number of events up to t will be in a given interval (e.g. lower than a given value - it is a one-sided interval). On the other hand, the meaning of a prediction band is that the whole process of events will be with 90% probability in a given band (or below a given curve forming a one-sided band).

We shall show an example of a procedure of prediction of failures and of their costs in the framework of the real data case treated in Part II. The objective of experiments with operators, as described in example 1.1. - 1.3., was also to find the optimal conditions for their work and, consequently, to guarantee such conditions in the real situation. Let us therefore return to this example once more:

Example 1.4. We want to compare the performance of operators under different conditions. In order to do such a comparison, we should first be able to predict the properties of corresponding counting process. So that let us for instance consider the counting process with the cumulative hazard rate (intensity) $H(t, \boldsymbol{x}) = H_0(t) \cdot \exp(b_1 x_1 + b_2 + b_3)$, i. e. for the "optimal" case such that $X_1 = X_2 = 1$. There are in fact two ways how to solve the problem of prediction. One solution is the simulation of a sample of trajectories of the process. The second approach will compute CHR for each person (i. e. for his value X_1) and then average them. For instance, the expected (mean) number of events in an interval $[0, S]$, with ideal values of covariates (i.e. $X_2 = 1$ and $X_3 = 1$) is

$$\frac{1}{n}\sum_{i=1}^{n} H(S, x_i) = H_0(S)\exp(b_2 + b_3) \cdot \frac{1}{n}\sum_{i=1}^{n}\exp(b_1 x_{1i}). \tag{3}$$

Let us present here some results obtained from the first (simulation) approach. Figure 9 shows 200 randomly generated realizations of counting processes. The cumulative hazard rate was the "optimal" $H(t, \boldsymbol{x})$ given above, where $H_0(t)$ and parameters b_j were estimated (in Example 1.2.). Values of X_1 - the age - were generated uniformly from interval [20,60]. Full lines connect the averages and empirical 90% quantiles, respectively. The "quantile" line thus shows approximate 90% one-sided prediction intervals at each t (the line is secondary smoothed).

The mean number of events in $[0, 300]$, estimated from our analyzed data, is 71/52=1.365, while the expected (mean) number of events in the case of optimal covariates X_2 and X_3

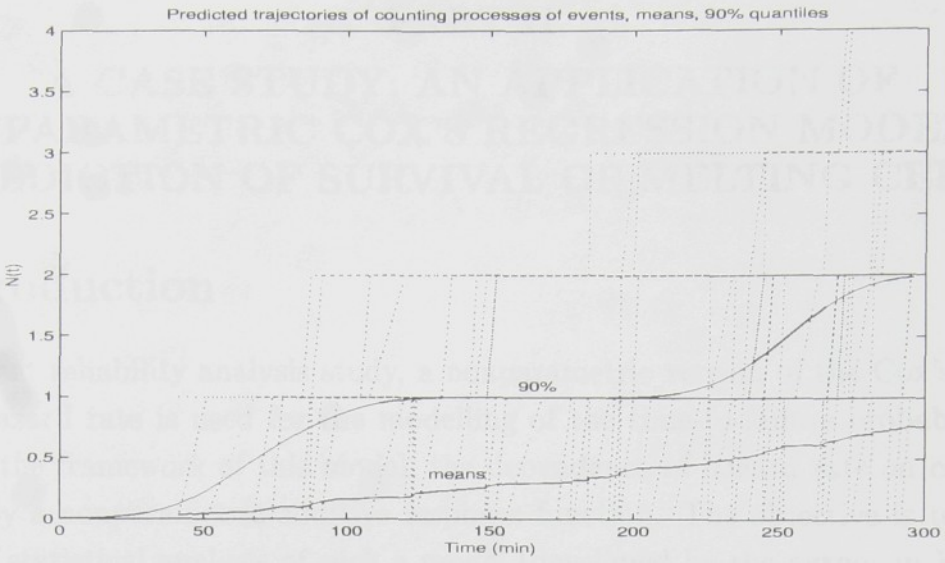


Figure 9: Simulated counting processes, averages and 90% empirical quantiles of numbers of events

is 0.926 (it is seen also from Figure 9.). Moreover, let us now imagine that after 2.5 hour the operator is changed. Then the mean number of events in first 150 minutes, computed from the data of our sample, is 0.481, while the mean number of events for persons with $X_2 = X_3 = 1$ is again lower, namely 0.334. So that the selection of operators and their timely exchange will reduce the risk of their wrong (or delayed) reaction considerably.

9 Conclusion of Part I

The main purpose of the present part of the study consisted in reviewing the methodology of statistical event-history analysis (including the author's results) and in suggestions of modifications convenient for the use of this methodology in praxis. In order to make the techniques applicable, the computational programs were prepared. They follow the algorithms described in the study and are adapted to particular data (e. g. to the data of Example 1, and to the cases analyzed in the following Parts II and III). As the standard procedures are now available in professional software systems (e. g. S-plus), including for instance the analysis of Cox's model with time-dependent covariates, our effort here has been concentrated onto their modifications, and, mainly in the next part, to the nonparametric techniques, namely to the maximum local likelihood (local scoring) approach, and to the graphical goodness-of-fit testing methods.

We said that (up to now) the counting processes models were used mainly in the area of biostatistics, medicine research, and, newly, also in reliability. However, it is clear that these models and analyzes are applicable in many other fields. Their list is rather long. We have already mentioned here the financial and economic management, ecology, security and defense, demography. Actually, it concerns to all areas where one deals with the forecasting of events, with the reliability and risk analysis in the random environment.

PART II

A CASE STUDY: AN APPLICATION OF NONPARAMETRIC COX'S REGRESSION MODEL TO PREDICTION OF SURVIVAL OF MELTING CELLS

1 Introduction

In the present reliability analysis study, a nonparametric version of the Cox's regression model of hazard rate is used for the modelling of the time-to-failure probability distribution. In the framework of this model, the dependence of hazard rate on covariates is described by a nonparametric additive response function. The objective is to recall the methods of statistical analysis of such a model (developed by the author in Volf, 1993a, 1996) and to adapt them to the real–case study of the Reynolds Metals Company data. The method of estimation of the response function uses a kernel-like approach which maximizes, in an iterative way, the local partial likelihood. Then, the goodness-of-fit is tested by the procedure based on the properties of generalized residuals.

At present, there exist a well developed methodology of nonparametric estimation in generalized regression models (i.e. in the exponential family of models, including the Cox one). Actually, the method used in the present study is a variant of the local scoring procedure proposed by Hastie and Tibshirani (1990) for generalized additive models. Other modifications are proposed for instance in Gentleman and Crowley (1991), O'Sullivan (1993), Fan and Gijbels (1996). These techniques were derived from the local smoothing (i.e. kernel, local polynomials etc.) procedures used successfully in the framework of standard (normal) regression. Nevertheless, the main problems remain, namely the problems of quality (e.g of consistency) of estimates and of convergence of algorithms at all. We shall return to the problem of optimality of nonparametric procedures in these "likelihood' regression models in the concluding remarks.

The organization of the study is following: In subsections of Part 2 the model is formulated and the method of moving window (maximizing iteratively weighted local likelihood) is described. Then, a considerable part of the study is devoted to the analysis of a real case data known as the Reynolds Metals Company data (Part 3).

The method is applied to the evaluation of losses caused by the reduction of lifetimes of damaged electrolytic cells, this reduction being caused by an uncontrolled shutdown of cells during a strike in an aluminium smelter. The lifetime distribution of the cells is modelled, the model characteristics are estimated, and then the goodness-of-fit of the model is tested. The case has been statistically examined by several analysts, cf. contributions of Kalbfleisch and Struthers, and of Thomas, both published in CS, 1982. Nevertheless, the case is so interesting from the point of statistical data analysis, that it can be used as a bench-mark example for comparision of analytic techniques. Thus, today's interest in these data is due to recent development of methodology of nonparametric identification

of general regression models. Thus, in the paper of Arjas and Liu (1995) the Bayesian approach is used to estimation of nonparametrized hazard rate, with the help of Gibb's sampling procedures. The novelty of our solution is that we consider a more detailed and more flexible (nonparametric) model then the preceding authors, we use the moving window method, and, finally, we also check the fit of the model by the statistical tests. We shall use the goodness-of-fit tests for the hazard-regression models based on the properties of generalized residuals. The graphical version has originally been proposed by Arjas (1988), numerical test for the parametric Cox's model has been developed by Marzec and Marzec (1993). Then, Volf (1996) derived the test for the Aalen additive model and also proposed the general graphical procedure of testing the fit of a nonparametric Cox's model. Such a version will be used in the present case.

2 Counting process and Cox's regression model

The notion of counting process as well as the Cox's regression model have been introduced already in Part I of the thesis. Nevertheless, let us here recall them briefly. A multivariate counting process $N(t) = N_1(t), \ldots, N_n(t)$ is a set of right-continuous random step functions on $[0, \mathcal{T}]$, with $N_i(0) = 0$ and with steps $+1$ at random moments. It is assumed that no two components step simultaneously. In the case analyzed here, $N_i(t)$ will jump from 0 to 1 at the moment of (observed) failure of the i-th device. The probability (the hazard) of the failure is modelled via the hazard function.

The Cox's model is a well known model for description of the dependence of the hazard function on a covariate $\mathbf{x}$. It specifies the form of the hazard function to $h(t, \mathbf{x}) = h_0(t) \exp(f(\mathbf{x}))$, where $h_0(t)$ is the baseline hazard function. The most frequently used standard semiparametric Cox's model has parametrized $f(\mathbf{x})$, mostly $f(\mathbf{x}) = \beta'\mathbf{x}$. We shall deal with the case when $f(\mathbf{x})$ is a nonparametrized smooth function. Its estimation can be based on a multidimensional moving window (or kernel) procedure (such a procedure being a direct generalization of a one-dimensional case). However, a multidimensional covariate causes the data sparse and the multidimensional kernel approach loses its effectivity. Therefore, we shall consider a less general model of additive influence of covariates, namely the model with $f(\mathbf{x}) = \sum_{j=1}^{K} f_j(x_j)$. The analyst has to identify suitable functions $f_1, \ldots, f_K$ and also function $h_0(t)$, or its cumulative version $H_0(t) = \int_0^t a(s)\,\mathrm{d}s$. Evident ambiguity (with respect to additional constants in f_j's) can be overcome by a proper normalization of these functions.

Simultaneously, we shall admit the time-dependent covariates $\mathbf{X}_i(t)$, $i = 1, \ldots, n$ (in general, they can also be random). Then, the behaviour of component $N_i(t)$ is governed directly by the (possibly random) intensity

$$\lambda_i(t) = h(t, \mathbf{X}_i(t)) \cdot I_i(t) = h_0(t) \cdot \exp f\,(\mathbf{X}_i(t)) \cdot I_i(t),$$

$i = 1, \ldots, n$, $t \in [0, \mathcal{T}]$, where $I_i(t)$ is an indicator process, $I_i(t) = 1$ if the i-th object is

in the risk set at moment t, $I_i(t) = 0$ otherwise. In other words, $I_i(t) = 0$ either after the failure of i-th object or after the i-th object is censored. The inference is based on the Cox's partial likelihood (cf. Andersen et al.,1993). Let us again recall its logarithm:

$$\ell_n = \sum_{i=1}^{n} \int_0^{\mathcal{T}} \ln \frac{\exp f(\mathbf{X}_i(t))}{\sum_{j=1}^{n} \exp f(\mathbf{X}_j(t)) I_j(t)} \mathrm{d}\, N_i(t).$$

Notice that $dN_i(t) = 1$ at points of events, $dN_i(t) = 0$ otherwise, so that the integral transforms to a sum. $A(t) = \int_0^t a(s)\,\mathrm{d}s$, When an estimate of function f is available, we can use the following generalized maximum likelihood estimator of the cumulative baseline hazard function

$$\hat{H}_0(t) = \int_0^t \frac{\mathrm{d}\bar{N}(s)}{\sum_j \exp \hat{f}(\mathbf{X}_j(s)) I_j(s)}, \tag{1}$$

where $\bar{N}(t) = \sum_{i=1}^{n} N_i(t)$. $\hat{H}_0(t)$ is then a nondecreasing stepwise function, with steps at points of observed events (i.e. points of counts of $N(t)$). From the increments of $\hat{H}_0(t)$ an estimation of function $h_0(t)$ can be obtained, with the aid of a smoothing procedure.

Let us now describe the idea of the moving window estimation of function $f(\mathbf{x})$, following Volf (1993a).

2.1 Moving window estimation procedure

Let us first consider a one-dimensional covariates $X_i(t)$ with values in some finite interval $\mathcal{X} \subset R$, for each i and $t \in [0, \mathcal{T}]$. If we wish to estimate the value of $f(x)$ at a point $x = z$, we take f as a constant f_z in some chosen neighborhood (window) around z, say, in $\mathcal{O}(z)$. Then f_z is treated as a parameter, we have to solve the equation $\partial \ell_n \,/\, \partial f_z = 0$, where

$$\frac{\partial \ell_n}{\partial f_z} = \sum_i \int_0^{\mathcal{T}} \left\{ \mathbf{1}[X_i(t) \in \mathcal{O}_z] - \frac{\exp f_z \cdot \sum_j I_j(t) \cdot \mathbf{1}[X_j(t) \in \mathcal{O}_z]}{\sum_j \exp\{f(X_j(t))\} \cdot I_j(t)} \right\} dN_i(t). \tag{2}$$

Hastie and Tibshirani (1990) recommend to compute also the second derivatives and to use the Newton–Raphson procedure. Then, the step from r-th to $(r+1)$-st iteration is given by the following expression:

$$f_z^{(r+1)} = f_z^{(r)} - \frac{\partial \ell}{\partial f_z} \Big/ \frac{\partial^2 \ell}{\partial f_z^2},$$

where the derivatives are evaluated at $f^{(r)}(x)$. Such a method is called the local scoring procedure. We derive the scheme of iteration directly from (2). Iteration starts from a chosen initial function, e.g. $f^{(0)}(x) \equiv 0$. Then we update it sequentially, in the following way: At each point z, we set

$$f^{(r+1)}(z) = \\ = -\ln\left[\sum_i \int_0^{\mathcal{T}} \frac{\sum_j \mathbf{1}\left[X_j(t) \in \mathcal{O}(z)\right] I_j(t)}{\sum_j \exp\{f^{(r)}(X_j(t))\} I_j(t)} \mathrm{d}N_i(t) \Big/ \sum_i \int_0^{\mathcal{T}} \mathbf{1}\left[X_i(t) \in \mathcal{O}(z)\right] \mathrm{d}N_i(t) \right]. \tag{3}$$

It is seen that we need not register all trajectories of $X_j(t)$, but only their values $X_j(T_i)$ at T_i - the moments of observed counts of $N_i(t)$.

Let us now consider the case of multidimensional covariate processes $\mathbf{X}(t) = (X_1(t), \dots X_K(t))$ and suppose the additive form of function f, namely $f(\mathbf{x}) = \sum_{j=1}^{K} f_j(x_j)$. In order to derive the procedure of iteration, let us imagine that when estimating, say, function $f_\ell(x_\ell)$, we have already estimated all necessary values of functions $f_m(x_m)$, $m = 1, 2, \dots, K$, during the preceding steps. Quite analogically to the one-dimensional case, from the equation $\partial \ell_n / \partial f_\ell(z) = 0$ we can obtain the following scheme of the moving window estimation procedure:

$$f_\ell^{(r+1)}(z) = -\ln \left[\sum_i \int_0^\mathcal{T} \frac{R_\ell(z, f^{(r)}, t)}{S_0(f^{(r)}, t)} \, \mathrm{d}N_i(t) \Big/ \sum_i \int_0^\mathcal{T} \mathbf{1}\left[X_{\ell i}(t) \in \mathcal{O}_\ell(z)\right] \mathrm{d}N_i(t) \right],$$

where now

$$R_\ell(z, f, t) = \sum_{j=1}^{n} \mathbf{1}\left[X_{\ell j}(t) \in \mathcal{O}_\ell(z)\right] \cdot \exp\left\{ \sum_{k=1}^{K} \mathbf{1}[k \neq \ell] \cdot f_k(X_{kj}(t)) \right\} \cdot I_j(t), \qquad (4)$$

$S_0(f, t) = \sum_{j=1}^{n} \exp\{f(\mathbf{X}_j(t))\} \cdot I_j(t)$, and $\mathcal{O}_\ell(z)$ is a chosen window around z in the domain of ℓ-th covariate.

The loop iterates through $\ell = 1, \dots, K$ and updates the values of $f_\ell^{(r+1)}(z)$ at every chosen z (we need at least the values at all observed $X_{\ell j}(T_i)$ provided $I_j(T_i) = 1$, $i, j = 1, \dots, n$). Then we proceed to the $(r+2)$-nd step of 'outer' iteration. The procedure may start from $f_1 = \cdots = f_K \equiv 0$ or from another convenient initial guess.

In order to control the progress of iteration (and convergence), a following method can be employed: After each r-th loop of iterations, optimal (in the least squares sense) lines are led through just estimated values of component functions f_k. We pursue the development and convergence of parameters of these lines. On this basis we decide either to continue or to stop the estimation process.

It is seen that the method can easily be enlarged to the kernel or to the nearest neighbours estimation procedures. The choice of the window-widths is a matter of experience (there exists a well developed theory of optimal window-width selection, for the standard kernel regression cases). We prefer the use of simple kernels (e.g. a triangular one) in combination with the nearest neighbours approach, i.e. we prescribe a minimal number of data points in the window.

2.2 Remarks on convergence of the procedure

As regards the consistency of estimates (i.e. of the large sample convergence to the 'true' hazard function), in the case of standard Cox's model the existence of consistent sequence of estimates is guaranteed theoretically (see Andersen et al, 1993). On the other hand, even in this standard semiparametric case the actual solution is obtained by an iteration procedure depending strongly on the actual data and on starting points. Iteration then

can lead to a local optimum or to an oscillation without any convergence at all. One possibility how to overcome such a case is to start the procedure several times, under different initial and control conditions. These problems are even more actual in the case considered here, where the optimization in a local window is combined with repeated shifts of this window and with sequential local updating of component functions f_k). The problem of convergence and of consistency of estimates is discussed elsewhere, for instance in Hastie and Tibshirani (1990), see also Volf (1993a). However, neither for the exponential family of models, nor for the Cox's model, a definite result is available. Nevertheless, the experience with the method presented here is encouraging, it has been tested successfully by a number of simulated as well as of real-data examples.

3 A case study

In 1967, a strike at a Quebec aluminium smelter resulted in the uncontrolled shutdown of electrolytic cells. The company claimed that the shutdown caused the shorter operating lives of cells operating at the time. The case led to a legal action and initialized a need of a deep statistical analysis of the data, in order to confirm expected higher failure rate after the intervention (shutdown) and to estimate statistically the losses caused by this (eventual) higher rate. The more details about the case, as well as the complete data were published in "Case Studies in Data Analysis" (CS, 1982), a section of Canadian Journal of Statistics, V. 10 (1982). The table of data contains three parts. In the first subtable, there are data on 395 cells of standard types, of which 297 experienced the shutdown. There are 20 types of standard cells (denoted A1 – A20). The second subtable refers to 104 cells of experimental design, their types are labelled as B, $C, \ldots,$ K. The survival of cells was measured in days, the highest observed time to failure was 2 541 days. The installation times differed from 2 287 to only 3 days before the shutdown (no cell installed after the strike is considered). The survival times are known, noncensored.

From all these 499 cells, 349 were in circuit at the moment of intervention. The cells of the third group, 73 experimental cells labelled from L do O, were installed and had failed before the intervention.

From description in CS, 1982 it is not clear how long the interruption lasted. It is therefore assumed that the aging of cells had been "stopped" during this period, that the failure rate of a nonfunctioning cell is negligible.

It is supposed that each type of cell may have specific properties influencing its survival time. There arises the question whether the third group of cells (no experiencing the intervention) is worth to be taken into account, not bringing any information about the influence of the shutdown. However, as soon as the model with a common baseline hazard is considered, even these data contribute to the estimation of the baseline characteristic.

3.1 Choice of the model

In the first part of their study, Kalbfleisch and Struthers (1982) estimated and compared the age-specific hazard rates before and after intervention. The fact of experience of intervention was treated as a $\{0,1\}$ covariate in Cox's model, two-sample test showed substantial increase of aggregated hazard rate after the intervention. Other covariates have been considered, namely the age of the cell at the moment of intervention and the time from intervention (provided the cell experienced it at all). Thus, the changes of the hazard rate in the course of individual time have been examined attentively, while the types of cells have not been considered.

Thomas (1982) used also the standard Cox's model, considering the following covariates: date of installation, experience of intervention (0 or 1), subtype of cell (the types of cells were aggregated to 6 subclasses). The pairwise interactions of these three covariates were considered, too. Thomas remarked that the statistical tests revealed a lack of fit of the model. It could be caused by a nonoptimal choice of covariates or by a nonoptimal structure of the chosen model. It seems to be more appropriate to consider two hazard rates, one for non-damaged cells, and the second as a function of the time after intervention. These two hazard rates can be connected in a multiplicative way, so that one is regarded as the baseline hazard, the second as a function describing the covariate effect of the time after intervention.

Following this idea, we shall work with the following model of the hazard rate of failure of i-th cell, using the age of the cell as a reference time t:

$$\lambda_i(t) = h_0(t) \cdot \exp\left\{b(t-U_i)\cdot 1[t > U_i] + c(x_i)\right\}, \quad t \in [0, T_i], \quad \lambda_i(t) = 0 \text{ otherwise}, \quad (5)$$

where T_i is the survival time of i-th cell. U_i is now the age of i-th cell at the moment of intervention and x_i is the type of cell i. The values from 1 to 34 are assigned to types A1,...,A20, B, C,...,N,O. While the covariate x is categorized, the time is regarded as a continuous variable, functions h_0, b are assumed to be continuous and bounded.

The validity of the model was tested. Let us postpone the problem of testing to the concluding part of analysis. Now let us apply the moving window technique to the estimation of nonparametric regression functions b and c.

3.2 The procedure of estimation

Denote $f(t,u,x) = b(t-u)1[t > u] + c(x)$. As each cell encountered exactly one failure, at moment T_i, Cox's partial likelihood is now

$$\ell = \sum_{i=1}^{n} \ln \frac{\exp(f(T_i, U_i, x_i))}{S(i)},$$

where $S(i) = \sum_{j=1}^{n} \exp(f(T_i, U_j, x_j)) I_j(T_i)$. The computations start from b, $c \equiv 0$. Let us imagine that we wish to estimate value of function b at a fixed point s (> 0). Therefore

we regard (for the moment) function $b(\cdot)$ as a constant b_s in a chosen window $\mathcal{O}(s)$ around s. Quite analogically to (2), from equation $\partial\ell \,/\, \partial b_s = 0$ we obtain

$$\sum_{i=1}^{n} \left\{ 1[T_i > U_i] \cdot 1[(T_i - U_i) \in \mathcal{O}(s)] - \frac{R(s,i)}{S(i)} \cdot \exp b_s \right\} = 0, \tag{6}$$

where $R(s,i) = \sum_{j=1}^{n} 1[(T_i - U_j) \in O(s)] \cdot 1[T_i > U_j] \cdot \exp c(x_j) \cdot I_j(T_i)$. An iteration step, derived directly from (6), is then:

$$b_s^{(r+1)} = -\ln \left\{ \sum_i \frac{R(s,i)}{S(i)} \Big/ \sum 1[T_i > U_i]\, 1[(T_i - U_i) \in \mathcal{O}(s)] \right\},$$

where naturally the right side uses the estimates of functions b, c obtained from the preceding steps. The task of local estimation of function c is solved in the same way. As the variable x is categorized to $M = 34$ classes, the task is equivalent to solution in the framework of Cox's model with M parameters $c_1, \ldots, c_M$ and 0,1-valued covariates. We have

$$\frac{\partial \ell}{\partial c_m} = \sum_{i=1}^{n} \left\{ 1[x_i = m] - \frac{S(m,i)}{S(i)} \exp c_m \right\}, \tag{7}$$

$$\text{where } S(m,i) = \sum_{j=1}^{n} 1[x_j = m] \cdot \exp\left\{ b(T_i - U_j) \cdot 1[T_i > U_j] \right\} \cdot I_j(T_i),$$

and the iteration step is

$$c_m^{(r+1)} = -\ln \left\{ \sum_i \frac{S(m,i)}{S(i)} \Big/ \sum_i 1[x_i = m] \right\}.$$

Results: The progress of iteration was controlled and its convergence observed from the changes of estimated parameters c_m. Originally, function $b(s)$ was estimated at equidistant points, at each 10 (days). The full domain of s was from 0 to 1 837, so that we obtained 183 values. Then the estimate was secondary smoothed, i.e. the values were averaged (in a weighted way) in a moving window. The graphs of estimated functions b and c are displayed in Figures 1a, 1b.

After we decided to stop the iterations (when changes of values of c_m were less than 0.1%), we computed the estimate of cumulative baseline hazard function $H_0(t)$ in accordance with (1). From it, by a kernel smoothing of its increments, we obtained a graph of estimate of $h_0(t)$ in Figure 1c.

In the framework of Cox's regression model, the procedure of estimation yields also the asymptotic confidence regions for true values of parameters and for true cumulative baseline hazard function. It follows from the asymptotic normality of estimates (cf. Andersen et al. 1993). In our case, the application of these asymptotic results is rather limited, because we use a nonparametrized function b(s) in model (5). This could be overcome with the help of a rougher, histogram-like approximation to b(s), or by bootstrapped construction of confidence regions. However, even if we have confidence bounds

for 'true' characteristics of the model, the transformation of these intervals (in order to obtain confidence intervals for the mean number of lost days, say) is a rather hazy and ambiguous problem. That is why the confidence intervals were not constructed.

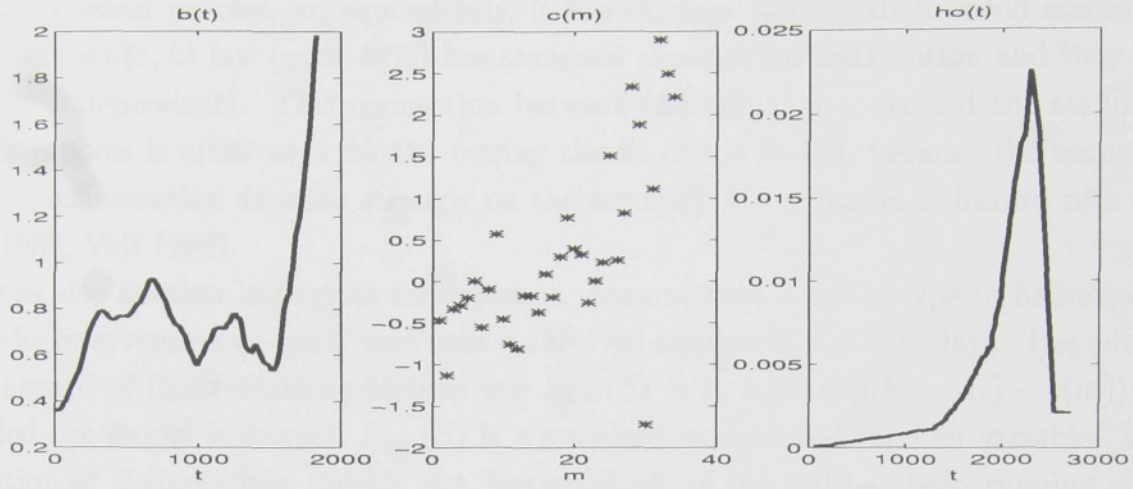


Figure 1: Estimates of functions $b(s)$, $c(m)$ and baseline hazard rate $h_0(t)$

3.3 Evaluation of the cost of intervention

Method 1. Let $T(m,U)$ be a random variable – the remaining time to failure of the cell type m, which has survived up to age U and which is supposed not to be involved in the intervention. The hazard rate of distribution of $T(m,U)$ is $h_0(s+U)\cdot\exp c_m$, $s\geq 0$, its estimate is available. The data contains n_m cells of type m which passed the intervention at ages U_i, their remaining survival times $S(m,i)=T_i-U_i$, $i=1,\ldots,n_m$, have been observed. Then the total loss (in days) of cells of type m is given by random variable $D_m=\sum_{i=1}^{n_m}(T(m,U_i)-S(m,i))$. The mean number of lost days is directly $E\,D_m=\sum_i(E\,T(m,U_i)-S(m,i))$.

Theoretically, the mean remaining lifetime is $E\,T(m,U)=-\int_0^\infty s\,d\,P_{U,m}(s)$, where

$$d\,P_{U,m}(s)=d\,P_m(s+U)/\,P_m(U)\quad\text{and}\quad P_m(t)=P_0(t)^{\exp c(m)},$$

$P_m(t)$ is the survival function (for nondamaged cells of type m) and $P_0(t)=\exp(-H_0(t))$ is the baseline survival function. The estimate of the mean is given by the following sum

$$\hat{E}\,T(m,U)=-\sum_{j=1}^{n}(T_{(j)}-U)\,1[T_{(j)}>U]\cdot\Delta\,\hat{P}_m(T_{(j)})\,\big/\hat{P}_m(U),$$

where $T_{(j)}$ are ordered all survival times, $T_{(1)}\leq T_{(2)}\leq\ldots\leq T_{(n)}$, and where

$$\Delta\,\hat{P}_m(T_{(j)})=\hat{P}_m(T_{(j)})-\hat{P}_m(T_{(j-1)}),\quad \hat{P}_m(T_{(0)})=1,\quad \hat{P}_m(t)=\exp\left(-\hat{H}_0(t)\cdot\exp\hat{c}(m)\right).$$

Method 2. It is worth to remind that $\sum_{i=1}^{k} T(m, U_i)$ $(k \leq n_m)$ can be regarded as a waiting time to k-th failure of a nonhomogeneous Poisson process. If $L(t)$ is a common cumulative hazard rate of independent identically distributed random variables T_i characterizing the times to failures, then $\sum_{i=1}^{k} L(T_i)$ is the waiting time to the k-th event of standard Poisson process, or, equivalently, it is a random variable distributed according to the gamma $(1, k)$ law (each $L(T_i)$ has standard exponential distribution and they are mutually independent). This connection between the actual process and the standard Poisson process is often used for the testing the fit of the model, because the accuracy of this transformation depends strongly on the accuracy of the model of hazard rate (cf. Arjas 1988, Volf 1996).

It has also another important consequence. Assume that a cell of type m had experienced the intervention at age U and then it survived another $S = T - U$ days. The cumulative hazard of this remaining lifetime was $L_{1,m}(S) = \int_U^T h_0(s) \exp(b(s - U) + c(m))\, ds$. Provided the model is correct, $L_{1,m}(S)$ is a standard exponential random variable. The realization of S shows how quickly the 'hazard clock' of the cell has been running after U.

If the intervention did not occurred, the cumulative hazard on the interval (U, t) would be $L_{2,m}(t - U) = \int_U^t h_0(s) \exp c(m)\, ds = (H_0(t) - H_0(U)) \cdot \exp c(m)$. So that $R = L_{2,m}^{-1}(L_{1,m}(S))$ is now the remaining survival time after U of the same cell but in an "another world" in which the cell did not pass the intervention. Again, provided the model is correct. Natural estimates are:

$$\hat{L}_{1,m}(S) = \sum_{i=1}^{n} 1[U < T_i \leq U + S] \exp(\hat{b}(T_i - U) + \hat{c}(m)) \Delta \hat{H}_0(T_i),$$

$$\hat{L}_{2,m}(s) = \exp \hat{c}(m) \cdot \sum 1[U < T_i \leq U + s] \, \Delta \hat{H}_0(T_i) \quad \text{and}$$

$$\hat{L}_{2,m}^{-1}(z) = \inf \left\{ s : \hat{L}_{2,m}(s) \geq z \right\},$$

eventually the inverse function can be computed with the help of interpolation. So that the estimate of $R - S$ yields another estimate of loses caused by the intervention. Notice that the evaluation of R from equation $L_{2m}(R) = L_{1m}(S)$ does not require the knowledge (estimate) of function c (while the former method of evaluation of expected remaining lifetimes did not use the function b).

The results of estimation of numbers of lost days are summarized in Table 1. Both methods yield approximately the same results. The estimates of cumulative baseline hazard function A, of function c and secondary smoothed estimate of function b have been used. The first two columns contain the results of method 1, namely the expected remaining lifetimes summarized for type m, $\sum_i \mathsf{E}\, T(m; U)$, and the differences from really achieved remaining survival times. The last two columns display the values of estimated remaining lifetimes, R_i, again summarized for types of cells, and summarized differences $R_i - S_i$. Both R_i and S_i are computed following the second method.

Cell Type	Method 1: Remain.Lifetimes	Differences	Method 2: Remain.Lifetimes	Differences
1	703.021	323.021	502.1910	122.1910
2	6186.977	1176.977	7214.6456	2204.6456
3	8018.959	2720.959	8403.7884	3105.7884
4	8691.149	2169.149	9625.9951	3103.9951
5	9457.718	3022.718	10077.0737	3642.0737
6	11735.956	3895.956	11530.0891	3690.0891
7	20500.358	4856.358	20920.3909	5276.3909
8	12074.879	4390.879	11402.1867	3718.1867
9	3501.335	876.335	3795.4695	1170.4695
10	3244.579	1368.579	2862.9982	986.9982
11	10865.187	2232.187	11583.7112	2950.7112
12	13970.679	2770.679	14346.7773	3146.7773
13	12199.764	3366.764	11973.1333	3140.1333
14	7221.806	1820.806	7461.6275	2060.6275
15	12802.812	3357.812	12653.0895	3208.0895
16	15184.571	4319.571	14294.2969	3429.2969
17	29215.036	7062.036	29166.3350	7013.3350
18	16568.215	3726.215	16690.5910	3948.5910
19	3193.496	528.496	3621.4483	966.4483
20	15309.948	3232.948	15793.7712	3716.7712
21	198.770	−835.230	1034.0000	0.0000
22	3649.869	2142.869	2233.7557	726.7557
23	1507.566	391.566	1613.0126	497.0126
24	250.948	250.948	0.0000	0.0000
25	890.102	264.102	909.0553	283.0553
26	5237.939	2283.939	4338.3263	1384.3263
27	2712.441	690.441	2914.5540	892.5540
28	1869.895	413.895	2100.9605	644.9605
29	1534.350	176.350	1949.4194	591.4194
30	3442.019	1273.019	2294.1161	125.1161
T O T A L:	241840.344	64270.344	243316.8094	65746.8094

Table 1.

Our results are not far from the estimate (in much rougher, semiparametric model) of Kalbfleisch and Struthers (1982)). Their estimate of lost days was 82 653.5.

3.4 Testing the goodness–of–fit

Let us now consider the problem how to asses the fit of the model of hazard rate. The main idea of the time-scale transformation (i.e. transformation to standard Poisson process) has already been mentioned in the preceding part. For the more precise formulation of

the testing method, see for instance Arjas (1988), Marzec and Marzec (1993) – the case of Cox's model with the estimator plugged in the test statistics, or, in a rather general setting, Volf (1996). The sample of examined objects is divided into two or more strata and in each stratum S the counting process of ordered observed failure times is examined. Let these times be $0 \leq T_{1,S} \leq T_{2,S} \leq \ldots$. Their transformation to the times $\mathcal{T}_{k,S}$ of events of a standard Poisson process is given (provided the model holds) by

$$\mathcal{T}_{k,S} = \sum_{i \in S} \int_0^{T_{k,S}} \lambda_i(t)\, dt.$$

Their estimates are obtained from the estimates of components of the model (5), i.e. $\hat{\mathcal{T}}_{k,S} = \sum_{i \in S} \sum_j 1[T_j \leq \min(T_{k,S}, T_i)] \cdot \exp\left\{\hat{b}(T_j - U_i)\, 1[T_j > U_i] + \hat{c}(x_i)\right\} \cdot \Delta \hat{H}_0(T_j)$. These $\hat{\mathcal{T}}_{k,S}$ are plotted graphically, the "ideal" value of $\mathcal{T}_{k,S}$ should be k. If the plot of $\hat{\mathcal{T}}_{k,S}$ differs from k significantly, it is an indication that the model does not correspond to the data. As the test uses estimated response functions, and, in general, there does not exist a relevant theory of large sample properties of local likelihood estimates, the test is used as an explorative tool, in its graphical version. Such a graphical testing procedure was used in order to check the fit of (estimated) model (5). The data were stratified with respect to the types of cells. We did the same kind of test for different combinations of strata – even the least favourable results of tests did not contradict to the model. For instance, Figure 2a shows the plots for cell types 1–4 (121 cells, high survival, mostly without intervention) and cell types 21–30 (104 cells, lower survival, mostly with intervention). Then, in Figure 2b there are the plots for cells of types 11–20 (121 cells with rather high survival, in spite of the intervention) and cells 31–34 (73 experimental cells with low survival, without intervention). It is possible to say that the model fits well for all types of cells.

Figure 2c displays the plots for cells which have passed the intervention and for cells which have not. The picture shows that for our data the actual hazard of damaged cells was lower than it was assumed by the model (and was higher for nondamaged cells).

The reason was rather simple and natural. There was a high positive correlation between the survival of a cell and the event that this cell passed the intervention. In other words, the cells which had higher survival (caused only by a random fluctuations) were more likely to survive to (and to pass) the moment of intervention. We tested this connection with the help of 2×2 contingency table (Table 2), with resulting chi-squared test highly significant (i.e. rejecting the hypothesis of independence).

	$T \leq 1\,300$	$T > 1\,300$	Totals
Int.	114	235	349
No int.	184	39	223
Totals	298	274	572

TABLE 2. Resulting value of test statistics $\chi^2_{(1)}$ is 135.46.

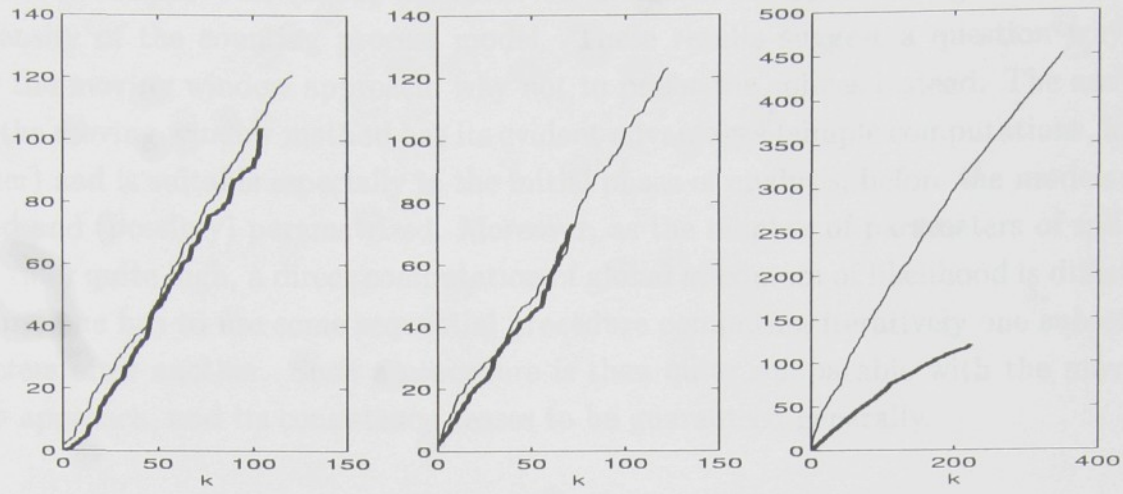


Figure 2: Graphical goodness-of-fit tests. Plots of $\hat{T}_{kS}$ for: a) thick - cell types $21-30$, thin - types $1-4$; b) thin - types $11-20$, thick - types $31-34$; c) thick - cells without intervention, thin - cells which passed intervention

4 Conclusion of Part II

We have described a method of nonparametric estimation of Cox's regression model, a method based on a moving window concept. The procedure is iterative, its idea consists in a sequential steps toward maximization of local partial likelihood. The performance of the method has been demonstrated on solution of a complex real-data case.

Nonparametric estimation has its advantages (consisting mainly in its generality), but its weak sides are well known, too. One of them consists in a rather slow global consistency of nonparametric estimates, while an estimate of parametrized model is able to reflect the main features of model very quickly. The nonparametric estimate depends more strongly on the data and their local nonregularities. For instance, the estimate of baseline hazard rate $a(t)$ in our example decreases at the right end of the time interval. It is caused only by the lack of data (of nondamaged cells) surviving for so long.

Although the moving window procedure is a very flexible and easily computable method, its consistency in general likelihood-based models is not guaranteed. On the other side, in the cases of such complex models,one can use, as an alternative, the reparametrization approach (e. g. approximation of response function by regression splines). Then the problem is transferred to the solution of a standard (but multiparametric) task of maximization of global (relevant) likelihood.

Stone, in a series of papers (e.g. Stone, 1994) has examined the family of exponential regression models and their approximation with polynomial splines. The parameters of splines are estimated via the global maximum likelihood method. Stone has shown theoretical consistency of such an approach. Kooperberg et al. (1993) have proved such a consistency of spline approximation for the case of nonparametric Cox model with

constant covariates. Volf (1994) then derived a similar result for the general case of the intensity of the counting process model. These results suggest a question why to employ the moving window approach, why not to prefer the splines instead. The answer is that the moving window method has its evident advantages (simple computations, local character) and is suitable especially in the initial phase of analysis, before the models are specified and (possibly) parametrized. Moreover, as the number of parameters of splines is (as a rule) quite high, a direct computation of global maximum of likelihood is difficult. Therefore, one has to use some sequential procedure computing iteratively one subset of parameters after another. Such a procedure is then quite comparable with the moving window approach, and its consistency ceases to be guaranteed generally.

PART III
A COUNTING PROCESS MODEL OF RELIABILITY OF A PARALLEL LOAD-SHARING SYSTEM

1 Introduction

In the present part we model the reliability (survival) of a system composed from m parallel identical components. The reliability is understand as a resistance of the system against a load (strength, stress) causing its failure, the reliability of the system is derived from the reliability of the components.

We assume that the system is tested by a load increasing from 0 up to the level breaking the system (i.e. all its components) – or up to a given maximal load $S_{\max}$ when the experiment is terminated. Let the testing experiment be relatively fast, so that the time of duration of the stress does not influence the survival. We use more or less the standard survival analysis approach and the counting processes model, however, instead the time, the load per one component is the variable of interest. Simultaneously, we consider a rather simple scheme of re-distribution of the load among the components, namely the Daniels load-sharing model, see Crowder et al (1991). We assume that the breaking strengths of individual components are the independent and identically distributed random variables, and that at each moment the load applied to the system is divided equally among the (unbroken) components. The same model has been used, for instance, in Belyaev and Rydén (1997). The global load affecting the system is observed. However, as the break of a component leads to an immediate re-distribution of the load to the other components (so that to the abrupt increase of the load per each component), the consequence can be the immediate break of several of remaining components. Therefore, in such a case of multiple breaks we observe directly the strength causing the break of the first component only. Moreover, we often are not able to register the order in which the components broke, but in the case of identical components the information on the order is not important. The statistical analysis will use only the values of directly observed (i.e. noncensored) data. The censoring of the other breaks will be expressed by a properly defined observability indicator process, as it is common in the setting of the counting process model. Thus, we actually omit a part of information. On the other hand, the estimation and the proofs of the large-sample properties are then much more straightforward.

We have to admit that we deal with a rather simple description of a parallel system, that a more accurate model should consider for instance the (irreversible) consequences of shocks caused by the abrupt increase of the load per one component. Some cases, for instance the case of a wire or a textile yarn composed from a set of strands, are even more complicated due the elasticity of the material, or due the mutual dependence of twisted

strands. Therefore, the present model is just a step to the more profound investigation of reliability of systems.

The theme of reliability of a system composed from parallelly organized units has already been studied by a number of authors. In most instances the time to break under a constant stress has been analyzed. The random process approach was used for instance in Daniels (1989), who examined the behaviour of maxima of certain Gaussian processes and with their aid he modelled the breaking strengths of a bundle of fibres. As we have already said, the starting point of our analysis is the model of the counting process characterized by a nonparametric hazard function. In the same setting, Belyaev and Rydén (1997) proved the uniform consistency and the local asymptotic normality of the Nelson–Aalen estimate of the cumulative hazard function (C.H.F.) characterizing the probability distribution of the breaking strengths of components.

Our objectives are mostly methodological, in the sense that we want to propose a set of methods for modelling, computing, testing and simulation of reliability of a parallel system. Quite naturally, certain theoretical problems have to be solved, too. In Section 2 the counting process model of the breaking strengths of the components in a load-sharing system is recalled. In the beginning of Section 3, still following Belyaev and Rydén (1997), we present the Nelson–Aalen estimator of C.H.F. of the breaking strength of one component. The main results are concentrated in sections 3 and 4. After a rather trivial statement of Lemma 1, we offer a modified proof of uniform consistency in Theorem 1. Then the main Theorem 2 proposes the weak convergence of residual process to a Wiener process, on a whole interval. In Section 4 this global asymptotic normality is utilized for the formulation of a goodness-of-fit test. The proofs use the relevant results and theory available in a number of papers and monographs dealing with the counting processes models (e. g. Andersen and Borgan (1985), Fleming and Harrington (1991), Andersen et al (1993)).

The problem how to derive the probability distribution of the breaking strength of the system, if the distribution of breaking strengths of its components is known, is discussed in Section 5. We recall both the computation approach proposed already in Suh et al (1970) and the simulation method, and we compare them. In concluding Example 3 we consider also a more general case of two types of units with proportional hazard rates, and we analyze such a situation as a simple case of the Cox's hazard regression model (Andersen and Gill, 1982).

2 The counting process of breaks of components

Let us first consider one component and the random variable U – its breaking strength. We assume that U has a continuous distribution on $[0, \infty)$ with a distribution function $F(u)$, density $f(u)$, hazard function $h(u) = \frac{f(u)}{1-F(u)}$ defined on $u \in [0, S]$ such that $F(S) < 1$. By

$H(u) = \int_0^u h(v)\,\mathrm{d}v$ we denote the cumulative hazard function. The 'fate' of a component during the increase of the load affecting it, u, is described by two random processes, by the counting process $N^1(u)$ and the indicator $I^1(u)$. $I^1(u) = 1$ if the load u affecting the component is observed, otherwise $I^1(u) = 0$. Namely, $I^1(u) = 0$ if the component is already broken or if the experiment is terminated. We assume that the trajectories of $I^1(u)$ are left-continuous. As regards $N^1(u)$, $N^1(0) = 0$ and $N^1(u)$ jumps to 1 at the load level u_b causing the observed break of the component (i.e. provided $I^1(u_b) = 1$). Trajectories of $N^1(u)$ are taken as right-continuous. The above description is actually the standard scheme of survival analysis, where the increasing load per component stands instead of time. As we assume a continuous distribution of U, we also consider a continuous scale of u. The difference in comparison with the standard survival analysis scheme consists in that we allow for abrupt jumps-up of the actual load affecting the component – in these intervals we set $I^1(u) = 0$, too.

2.1 The model of parallel system

Let us now consider the system composed from m components, let the breaking strengths of components be described by i.i.d. random variables U_j, $j = 1, \ldots, m$, with distribution given by $f(u)$, $F(u)$, $h(u)$, $H(u)$, respectively. The following example illustrates the structure of observed data. Let us imagine that the breaks of components occurred for K 'global' loads affecting the system, $0 < s_1 < s_2 < \ldots < s_K < S_{\max}$, that on levels s_j the numbers k_j of components broke, with $\sum k_j = m$. Therefore, just before the first break the load per each component was $u_1 = s_1/m$, while just before the moment of the second break it was $u_2 = s_2/(m - k_1)$ (naturally affecting only $m - k_1$ remaining components) and, finally, immediately before the moment of the last break the load per each of last k_K components was $u_K = s_K/k_K$. As regards the observed breaks (i.e. the breaks caused by **known** loads per component), we actually observed only K of them, caused exactly by loads u_j. Other breaks (if $k_j > 1$) were caused by unknown (unobserved) loads per component from intervals $(u_1, \overline{u}_1 = s_1/(m - k_1 + 1))$, $(u_2, \overline{u}_2 = s_2/(m - k_1 - k_2 + 1))$, ... $(u_K, \overline{u}_K = s_K)$, respectively for $k_1 - 1, k_2 - 1, \ldots, k_K - 1$ components. Our first aim is to analyze the distribution of U_j on $[0, S]$. We assume that the maximal load per system $S_{\max}$ is sufficiently large (e.g. $S_{\max} > S \cdot m$) in order not to terminate experiments too early.

Remark: Taking into account the assumption that the probability distribution of U_j is continuous, then (theoretically) there cannot occur two breaks at the same load per component level u. In other words, components break one after another, not simultaneously (though sometimes we are not able to distinguish their order). The intervals of breaking strengths can be specified even more precisely than $(u_k, \overline{u}_k)$ above. However, as it has been said, our solution will not use the information about interval-censored strengths ex-

plicitly, but through a properly defined observability indicator process only. That is why we do not discuss the details of interval censoring here.

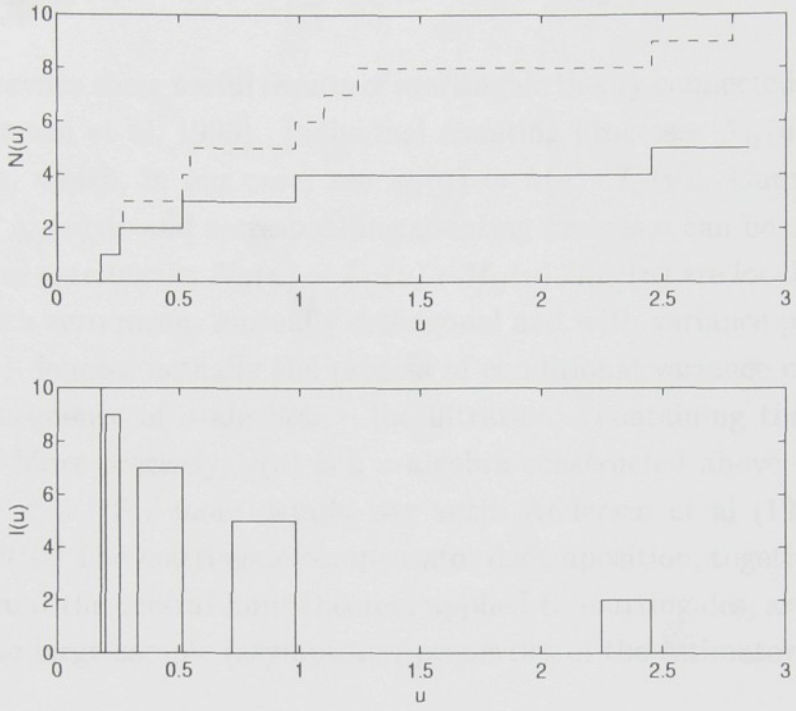


Figure 1: Counting process $N(u)$ and indicator process $I(u)$.

Thus, the data used for analysis consist of $I(u) = \sum_{j=1}^{m} I_j^1(u)$, $N(u) = \sum_{j=1}^{m} N_j^1(u)$, i.e. the sums of indicators and observed counting processes of the components. By u we again denote the load per (unbroken) component. For the better explanation of the structure of observed data, let us display one example graphically. A system composed from $m = 10$ components has been simulated. Distribution of U_j followed the exponential distribution with the mean one. $K = 5$ successive breaks has been observed. Figure 1 shows the indicator process $I(u)$, the counting process $N(u)$ of observed breaks (full line), and also the underlying counting process of actual breaks of components (dashed line), which, in the real data cases, is not observed.

3 Estimator of C.H.F., asymptotic properties

In the present part the estimator of the cumulative hazard function $H(u)$ (of distribution of the breaking strength of one component) is recalled and the uniform consistency and asymptotic normality of this estimator on interval $[0, S]$ are proven. Let us consider that n identical and independent systems are tested. Denote by U_{ij} the random variables – breaking strengths, by $N_{ij}(u)$, $I_{ij}(u)$ related individual counting and indicator processes

for the j-th component of the i-th system ($j = 1, 2, \ldots, m,\ i = 1, 2, \ldots, n$). Further, denote

$$N_i(u) = \sum_{j=1}^{m} N_{ij}(u), \quad I_i(u) = \sum_{j=1}^{m} I_{ij}(u), \quad N(u) = \sum_{i=1}^{n} N_i(u), \quad I(u) = \sum_{i=1}^{n} I_i(u).$$

Let us first revoke some useful results of martingale theory connected with the counting processes (Andersen et al, 1993). Individual counting processes $N_{ij}(u)$ are governed by their intensities, which, in our case, are $\lambda_{ij}(u) = h(u) \cdot I_{ij}(u)$. Cumulative intensities are $L_{ij}(u) = \int_0^u \lambda_{ij}(v)\, \mathrm{d}v$ and corresponding counting processes can be decomposed to the compensator and martingale, $N_{ij}(u) = L_{ij}(u) + M_{ij}(u)$. $M_{ij}(u)$ are local square-integrable martingales, with zero mean, mutually orthogonal and with variance process $\langle M_{ij} \rangle(u) = L_{ij}(u)$. Here $\langle\ \rangle$ denotes actually the process of conditional variance conditioned by the nondecreasing sequence of σ-algebras – the filtration – containing the *observed* history of the process. More precisely, $\sigma(u)$ is a σ-algebra constructed above the trajectories of $N_{ij}(v), I_{ij}(v), v < u$. For more details, see again Andersen et al (1993), Fleming and Harrington (19910. The martingale-compensator decomposition, together with the law of large numbers and the central limit theorem applied to martingales, are the basis for the derivation of the large-sample (asymptotic) properties of the estimator.

3.1 Nelson–Aalen estimator of C.H.F.

The most common estimator of the cumulative hazard function is the Nelson–Aalen one

$$\hat{H}_n(u) = \int_0^u \frac{1[I(v) > 0]}{I(v)} \mathrm{d}N(v),$$

where we set $0/0 = 0$. It is seen that the ability of the estimator to approximate well the 'true' $H(u)$ depends on the indicator process, i. e. on the observability of the counting processes for all values of strength u in the interval of interest $[0, S]$.

Let us assume that the number of tested systems, n, tends to infinity. Then it is also desirable that at each point $u \in [0, S]$ the number of observed unbroken components is of order n. Lemma 1 shows that such a property follows from the initial simple assumptions of identical, independent and continuous distribution of U_{ij} together with the assumption that $F(S) < 1$.

Lemma 1. There exists, with probability 1, a limit

$$\lim_{n \to \infty} \frac{I(u)}{n} = r(u)$$

for $u \in [0, S]$. Moreover, this limit is uniform w.r. to $u \in [0, S]$ and $r(u) \geq \varepsilon$ for some $\varepsilon > 0$.

Proof:

1. For each fixed u, $I_i(u)$ are *i.i.d.* random variables, with values from $\{0, 1, 2..., m\}$. Hence, the law of large numbers yields the almost sure convergence

$$\frac{1}{n}\sum_{i=1}^{n} I_i(u) \to r(u) = E\{I_1(u)\}.$$

2. From assumptions that random variables U_{ij} are *i.i.d.* and that $F(S) < 1$ it follows that for each $u \leq S$ there is a positive probability $[1 - F(u)]^m$ that all components of the system survive u. Therefore, $r(u) \geq m \cdot [1 - F(u)]^m \geq m \cdot [1 - F(S)]^m > 0$. The last expression can be used as ε in the Lemma 1.

3. It remains to prove the uniformity of convergence. Random functions $I_i(u)$ are mutually independent, with the same distribution. They have maximally $2m$ finite jumps (m down and up). We shall use the results collected in Hoffmann-Jorgensen (1994, Vol II, Parts 9.13 to 9.17). We can imagine that each trajectory of $I_i(u)$ is given by a bounded, piecewise-constant nonrandom function $b(\mathbf{v}, \mathbf{k}; u)$, where $\mathbf{v}, \mathbf{k} = (v_1, v_2, ..., v_m, k_1, k_2, ..., k_m)$ are the realizations of random vector $\mathbf{V}$,$\mathbf{K}$= $(V_1, K_1, V_2, K_2, ..., V_m, K_m)$, K_j are levels of $I_i(u)$ and V_j are its points of jumps down. Again, for different i-s these vectors are *i.i.d.*, $\mathbf{K}$ is bounded, $\mathbf{V}$ has continuous distribution. Hence, the proof of Lemma 1 follows from the boundedness in the mean and from a.s. piecewise continuity of functions $b(\cdot; u)$ with respect to the distribution of $\mathbf{V}$,$\mathbf{K}$ (cf. again Hoffmann-Jorgensen, 1994, Theorem 9.17 on uniform convergence).

From the uniform convergence and from the boundedness of jumps of $I(u)$ it also follows that the limit function $r(u)$ is continuous on $[0, S]$.

Remark 1. The statement of Lemma 1 implies that for every $\delta > 0$, for sufficiently large $n > n_\delta$, $Pr\{I(u) \geq \delta$ on the whole $[0, S]\} = 1$. Hence, with probability one it also holds that $1[I(u) = 0] = 0$ on $[0, S]$, and also $\sqrt{n}\int_0^S 1[I(u) = 0]\mathrm{d}H(u) = 0$. Such a property corresponds to one of conditions required in Andersen et al (1993, Th. IV.1.1 and IV.1.2) for the consistency and asymptotic normality of the Nelson-Aalen estimator.

3.2 Asymptotic properties

1. Consistency: The uniform consistency of $\hat{H}_n$ has already been proved in Belyaev and Rydén (1997). We shall prove the same result with the aid of Lemma 1. Let us examine residuals $\hat{H}_n(u) - H(u)$ for $u \in [0, S]$. Denote $M_i = \sum_{j=1}^{m} M_{ij}$, $M = \sum_{i=1}^{n} M_i$. From the martingale-compensator decomposition we obtain

$$\hat{H}_n(u) - H(u) = \int_0^n \frac{\sum\sum \mathrm{d}N_{ij}(v)}{I(v)} 1[I(v) > 0] - H(u) =$$

$$= \int_0^u \frac{\sum\sum \mathrm{d}M_{ij}(v)}{I(v)} 1[I(v)>0] + \int_0^u \frac{\sum\sum \mathrm{d}L_{ij}(v)}{I(v)} 1[I(v)>0] - H(u) =$$

$$= \int_0^u \frac{\mathrm{d}M(v)}{I(v)} 1[I(v)>0] + \int_0^u \frac{\sum\sum h(v)\, I_{ij}(v)\, \mathrm{d}v}{I(v)} 1[I(v)>0] - H(u) =$$

$$= \frac{1}{n}\int_0^u \frac{\mathrm{d}M(v)}{I(v)/n} 1[I(v)>0] - \int_0^u 1[I(v)=0]\mathrm{d}H(v). \qquad (1)$$

Theorem 1. $\hat{H}_n(u)$ is an *a.s.*-consistent estimate of $H(u)$ on $[0,S]$. Moreover, this consistency is uniform w.r. to $u \in [0,S]$, i.e. $sup_{u\in[0,S]}|\hat{H}_n(u) - H(u)| \to 0$ *a.s.*

Proof: Processes $\frac{1}{n}\int_0^u \mathrm{d}M(v) = \frac{1}{n}\sum_{i=1}^n M_i(u)$ have zero mean, M_i are mutually independent. Moreover, as $M_i(u) = N_i(u) - \int_0^u h(s)I_i(s)\mathrm{d}s$, they are uniformly bounded on $[0,S]$. Therefore, at fixed u, $\frac{1}{n}\int_0^u \mathrm{d}M(s) \to 0$ *a.s.* (it follows from the law of large numbers).

Uniform convergence $sup_{u\in[0,S]}\frac{1}{n}\int_0^u \mathrm{d}M(s) \to 0$ *a.s.* can be proved similarly as in the preceding Lemma 1. We can represent $N_i(u)$ by a piecewise constant function $c(\mathbf{v};u)$ which has maximally m steps $+1$ at points $v_1, v_2, ..., v_m$, and $\int_0^u h(s)I_i(s)\mathrm{d}s$ can be represented by a continuous and bounded function $\int_0^u h(s)b(\mathbf{v},\mathbf{k};s)\mathrm{d}s$, where b is a function defined in the proof of Lemma 1, **v,k** are *i.i.d.* realizations of **V,K** (also the same as in the proof of Lemma 1). Then, Theorem 9.17 of Hoffmann-Jorgensen (1994) can again be applied to the proof of the *a.s.* uniform convergence $\frac{1}{n}\int_0^u \mathrm{d}M(s) \to 0$, on $[0,S]$.

From this and further from Lemma 1 and Remark 1 the statement of Theorem 1 follows immediately.

2. Asymptotic distribution: Let us now analyze the behaviour of the process $\sqrt{n}(\hat{H}_n(u) - H(u))$ on $[0,S]$, for $n \to \infty$. Similarly as in (1), we obtain

$$\sqrt{n}(\hat{H}_n(u) - H(u)) = \sqrt{n}\int_0^u \frac{1[I(v)>0]}{I(v)}\mathrm{d}M(v) - \sqrt{n}\int_0^u 1[I(v)=0]\mathrm{d}H(v). \qquad (2)$$

Taking into consideration the uniform convergence of $I(v)/n$ given in Lemma 1, the statement of Remark 1, and the boundedness of jumps of $\mathrm{d}M(v)$ (jumps are less or equal to m), we immediately obtain the following theorem specifying the asymptotic distribution of residual process.

Theorem 2. Random process $\sqrt{n}(\hat{H}_n(u) - H(u))$ converges weakly on $[0,S]$ to a Gauss random process with independent increments, zero mean and with variance function

$$w(u) = \int_0^u \frac{\mathrm{d}H(v)}{r(v)}.$$

In other words, the process is asymptotically distributed as $W(w(u))$, where $W(.)$ is a Wiener process.

Proof: The proof follows directly either from the central limit theorem for martingales (e.g. Andersen et al, 1993, part II.5.) or from Theorem 3.2. of Andersen and Borgan (1985). It remains to show the convergence of the variance process and to compute the exact form of its limit. The convergence follows from our Lemma 1 and from the boundedness of both $H(u)$ and $r(u)$ on $[0, S]$, namely

$$\begin{aligned} \text{as var} \;&\sim\; \text{var}\{\sqrt{n}\int_0^u \frac{\mathrm{d}M(v)}{I(v)}\} = n\,E\int_0^u \frac{\langle \mathrm{d}M\rangle(v)}{I^2(v)} \;= \\ &=\; n\,E\int_0^u \frac{\mathrm{d}L(v)}{I^2(v)} = E\int_0^u \frac{h(v)\,\mathrm{d}v}{I(v)/n} \to \int_0^u \frac{\mathrm{d}H(v)}{r(v)}. \end{aligned}$$

4 Goodness-of-fit test

Let the hypothetical model be given by the cumulative hazard function $H^0(u)$. We want to decide whether the data correspond to it. The data are represented by the observed trajectories of $N_i(u), I_i(u)$ $i = 1, \ldots, n$. The tests are quite naturally based on the comparison of $\hat{H}_n(u)$ with expected $H^0(u)$.

Graphical test: Let us order all observed strengths breaking the components into one nondecreasing sequence u_k, $k = 1, \ldots, K$. For the graphical comparison, we plot the values

$$L(u_k) = \int_0^{u_k} dH^0(v) I(v)$$

against $N(u_k) = k$ on the abscissa. If the model holds the residual process $L(u) - N(u)$ is a martingale. Then it is expected that the curve $L(u_k)$ will be close to the line $y(k) = k$. An opposite case (e.g. expanding distance of both curves) indicates that the model $H^0(u)$ does not correspond to the data. Approximate critical bounds for such a comparison can be derived e.g. from the following numerical procedure.

Numerical test: Numerical test is based on asymptotic distribution. From the result of Theorem 2 it follows that the process

$$D_n(u) = \sqrt{n}(\hat{H}_n(u) - H^0(u))\,/(1 + w(u))$$

is (if the model holds) asymptotically distributed as a Brownian bridge process $\mathcal{B}(\tau(u))$, where $\tau(u) = w(u)/(1 + w(u))$, $u \in [0, S]$. Hence, a test of Kolmogorov–Smirnov type can be used. From the theory of Brownian bridge it follows, for instance, that if $d \geq 0$,

$$P\left(\max_u D_n(u) \geq d\right) = P\left(\min_u D_n(u) \leq -d\right) \approx \exp(-2d^2)$$

approximately. So that the value $\exp(-2d^2)$, where d is the observed $\max_k |D_n(u_k)|$, is an approximate p-value for the test of hypothesis of the goodness-of-fit against a proper

one-sided alternative. Simultaneously it holds that the critical value of the two-sided test, on the level α, namely the value $d(\alpha)$ fulfilling

$$P\{sup_u|D_n(u)| > d(\alpha)\} = \alpha,$$

can be approximated by $d(\alpha) = \sqrt{\ln(\frac{2}{\alpha})\frac{1}{2}}$. It follows that the approximate $(1-\alpha)$ confidence region for 'true' $H(u)$ is the band $\hat{H}(u) \pm d(\alpha)(1+w(u))/\sqrt{n}$. A more precise critical values can be obtained from the relevant results on the Brownian bridge process and on its probability of crossing a given level. An example of the test is provided in Section 6, Example 2.

5 Distribution of the breaking strength of system

Let us now assume that we know the characteristics of breaking strengths distribution of individual components (e.g. the distribution function $F(u)$) and our aim is to compute the reliability for the whole system composed from m such components. Though such a problem has already been considered elsewhere, for instance in Suh et al (1970), we think that it is useful to recall this approach in order to complete the set of methods presented in the paper. More precisely, let the probability that the system will not survive the (global) load s be given by the distribution function $F_R(s) = P(R < s)$, where R is the random variable describing the breaking strength of the system. If we denote by $U_{(1)} < U_{(2)} < \ ... \ < U_{(m)}$ the order statistics created from the random strengths breaking individual components of the system, $U_1, U_2, \ ... \ U_m$, then evidently

$$F_R(s) = P(R < s) = P\{\cap_{k=1}^{m}[U_{(k)} < \frac{s}{m-k+1}]\},$$

which can be computed from the joint distribution of order statistics $U_{(1)}, ..., U_{(m)}$. Though such a distribution is well-known (see e.g. Rao, 1965, Ch. 3.6.), the computation of the joint distribution function is not easy. In our case a sequential computation yields that

$$F_R(s) = m!\ A_m(s),$$

where $A_0(s) = 1$ and

$$A_k(s) = \sum_{j=1}^{k} \frac{(-1)^{j-1}}{j!} A_{k-j}(s) F^j(\frac{s}{k}). \tag{3}$$

Another, simple and universal approach to the evaluation of distribution of random variable R consists in the simulation. The following example illustrates and compares both methods.

6 Examples

Example 1

Let us consider a system composed from $m = 10$ components and assume that the breaking strength of each component (i. e. random variable U_{ij}) has the standard exponential distribution (i. e. with $EU_{ij} = 1$). We simulated the breaks of $n = 200$ such systems. The results observed for one of them are already in Figure 1. Naturally, the global load under which the system broke was observed, too. We thus obtained a sample of $n = 200$ independent realizations r_i of random variable R – the breaking strength of the system. The empirical distribution function $\hat{F}_R(s) = \frac{1}{n}\sum 1[r_i < s]$ constructed from this sample is displayed in Figure 2a and compared with $F_R(s)$ computed from (3). Other empirical characteristics can be easily derived, too. For instance, the estimate of the cumulative hazard function can be obtained either as $\hat{H}_R(s) = -\ln(1 - \hat{F}_R(s))$ (see Figure 2b) or directly from the ordered sample: Let (i) be the order of r_i in $r_1, r_2, \ldots, r_n$, then the standard Nelson-Aalen estimator is $\tilde{H}_R(s) = \sum_{i=1}^{n} \frac{1[r_i<s]}{n-(i)+1}$.

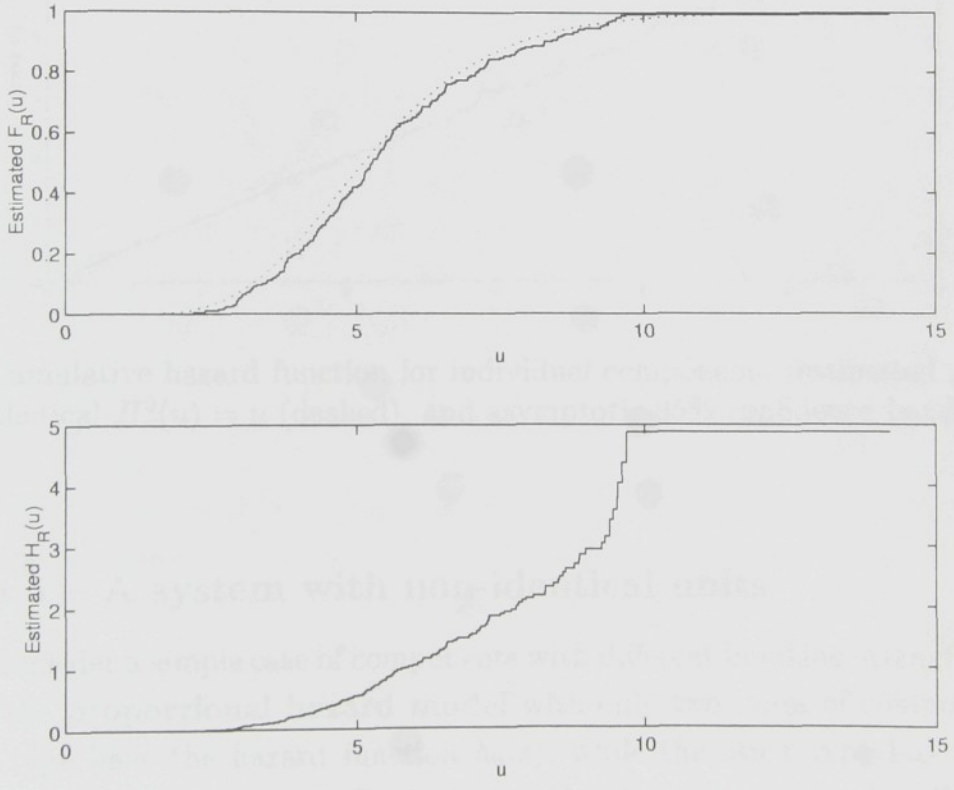


Figure 2: Estimated $\hat{F}_R(s)$ and $\hat{H}_R(s)$ of distribution of random variable R, compared with dotted $F_R(s)$ computed from (3).

Example 2

In this example of the numerical goodness-of fit test we use the data simulated in Example 1. We intend to test the hypothesis $\mathbf{H}_0$ that the data really correspond to the

standard exponential distribution, on interval $u \in [0,5]$. Therefore, we should compute $\hat{H}_n(u)$, estimate $w(u)$ by

$$\hat{w}_n(u) = \int_0^u \frac{n \cdot \mathrm{d}\hat{H}_n(v)}{I(v)},$$

and find the maximum of $|D_n(u)|$, assuming that the hypothetical C.H.F. of standard exponential distribution is $H^0(u) = u$.

The maximal and minimal observed values of $D_n(u)$ on $[0,5]$ were $d^+ = 0.0158$, $d^- = -0.0711$. We then computed approximate critical value for the test level $\alpha = 5\%$, $d(\alpha) = 1.3581$. As it was considerably greater than $d = max(d^+, -d^-) = 0.0711$, the hypothesis $\mathbf{H_0}$ was not rejected (on approximately 5% level of test significance). Estimated $\hat{H}_n(u)$ together with hypothetical $H^0(u) = u$ and approximate 95% confidence bands are displayed in Figure 3.

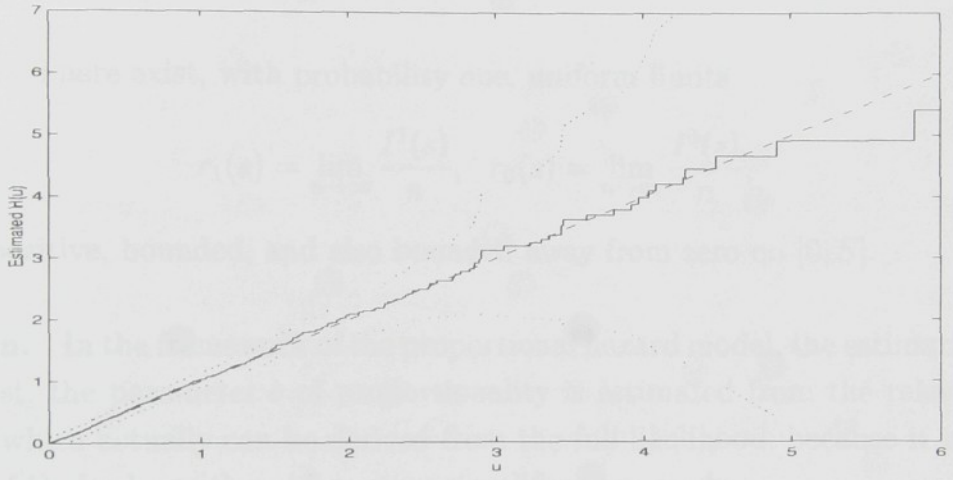


Figure 3: Cumulative hazard function for individual components: estimated $\hat{H}_n(u)$ (full line), hypothetical $H^0(u) = u$ (dashed), and asymptotic 95% confidence bands for $H(u)$ (dotted).

Example 3 – A system with non-identical units

Let us now consider a simple case of components with different breaking strengths, namely the case of the **proportional hazard model** with only two types of components. Let a standard type have the hazard function $h_0(s)$, while the other type has the hazard function $h_1(s) = c \cdot h_0(s)$, $c \geq 0$. Equivalently, the situation can be described via the hazard functions of components $h(s) = h_0(s)exp(bX)$, where b =lnc and X is a random variable with $p = P(X = 1), 1-p = P(X = 0)$. In the follow up we shall assume that the configuration of x's is known, i.e. we are able to match a certain value x to each observed broken or censored component. Then the case can be regarded as a simple version of Cox's regression model (with only two levels of regressor). The objective of statistical data analysis is to estimate parameter b and function $h_0(s)$, respectively its cumulative

version $H_0(u) = \int_0^u h_0(s)\mathrm{d}s$, on [0,S]. It is well known (cf. Andersen and Gill, 1982) that such an estimation problem is solvable consistently, moreover with estimates possessing the property of asymptotic normality.

Assumptions. We assume that $H_0(S) < \infty$ and that $1 > p > 0$. These assumptions actually suffice for the validity of conditions (given in Andersen and Gill, 1982) ensuring the desirable large sample properties of estimates.

Let us denote

$$I^1(s) = \sum_i \sum_j I_{ij}(s)\mathbf{1}[X_{ij} = 1], \quad I^0(s) = \sum_i \sum_j I_{ij}(s)\mathbf{1}[X_{ij} = 0],$$

let $N^1(s)$ and $N^0(s)$ be defined in a similar way. Then a variant of Lemma 1 (with a quite analogical proof) holds:

Lemma 2. There exist, with probability one, uniform limits

$$r_1(s) = \lim_{n\to\infty} \frac{I^1(s)}{n}, \quad r_0(s) = \lim_{n\to\infty} \frac{I^0(s)}{n},$$

which are positive, bounded, and also bounded away from zero on $[0, S]$.

Estimation. In the framework of the proportional hazard model, the estimation has two stages. First, the parameter b of proportionality is estimated from the relevant partial likelihood (which actually can be derived from the full likelihood, because it is a 'profile' likelihood of b). Its logarithm, after some simplification, reads

$$lnL_p(b) = \int_0^S b\mathrm{d}N^1(s) - \int_0^S ln\{I^0(s) + exp(b)I^1(s)\}\mathrm{d}N(s). \tag{4}$$

Optimal $\hat{b}$ (the maximizer of (4)) is obtained from the solution of equation $\mathrm{d}lnL_p/\mathrm{d}b = 0$, via the Newton-Raphson algorithm (or via another iterative procedure). In such a simple case considered here the solution is unique and, as the second derivative of (4) is negative, the maximum of $lnL_p(b)$ can be well reached practically from arbitrary starting value of the iteration procedure. In practical examples, the Newton-Raphson algorithm converged as a rule in less than 10 steps. The next stage consists in the estimation of the cumulative baseline hazard function, by the Newton-Aalen type estimator (in a regression context called the Breslow-Crowley one):

$$\hat{H}_0(u) = \int_0^u \frac{1[I^0(s) + I^1(s) > 0]}{I^0(s) + exp(\hat{b})I^1(s)}\mathrm{d}N(s). \tag{5}$$

As we have already said, the large sample properties (consistency and asymptotic normality) follow from the results derived for the more general case of Cox's model.

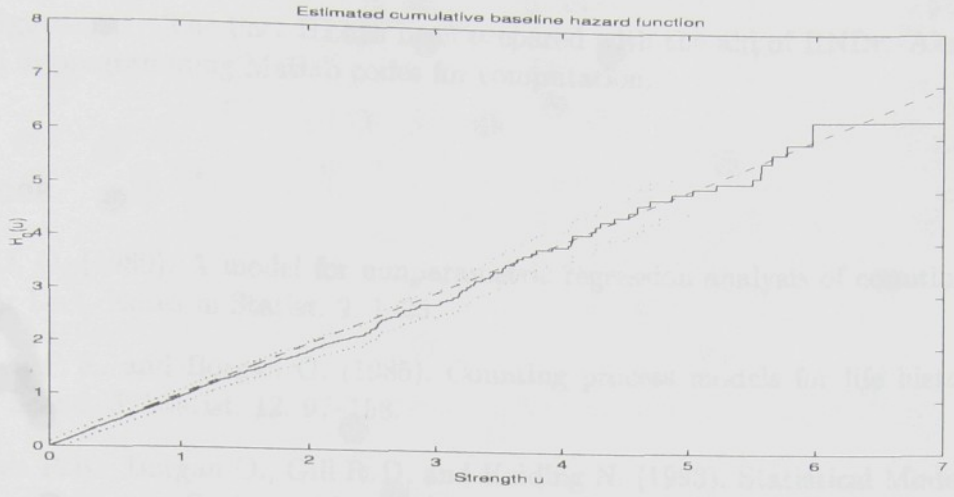


Figure 4: Final estimate of cumulative baseline hazard function, with asymptotic confidence bands (dotted) and $H_0(u) = u$ (dashed).

Numerical example: We generated a sample of 200 'systems', each composed from 10 components with the breaking strengths given by exponential distributions: 5 standard components had the mean 1, 5 stronger components had the mean 2. So that $H_0(u) = u$ and the proportionality parameter $c = exp(b) = 0.5$. After a fast and short iteration we obtained the estimate of $b = -0.7072$, i.e. of $c = 0.4930$, with approximate 95% confidence interval, based on the asymptotic normality of estimate of b, (0.4358, 0.5577). Estimated cumulative baseline hazard function in Figure 4 shows a linear trend with the slope close to one. Approximate 95% confidence bands for $H_0(u)$ computed in accordance with the results of Andersen and Gill (1982) are displayed by dotted lines.

7 Conclusion of Part III

We presented a set of procedures for the probabilistic modelling and statistical analysis of the breaking strengths in a system of parallel components. The data were treated as the lifetime data, with the increasing load per one component as the leading variable. We studied the asymptotic behaviour of the estimator of the cumulative hazard function of breaking strength distribution. In particular, asymptotic normality of the estimator on the whole interval was proved and on this basis the goodness-of-fit test was proposed. Such a test can for instance be useful for the assessing the agreement of observed breaking strength data with the expected resistance of the system (e.g. with the resistance guaranteed by the producer).

In the concluding example we also showed how the approach used in the present work could be generalized to a system of units with nonequal reliability. Another generalization can consist in models considering simultaneously the load, the time and/or the cumulated load, in the framework of the hazard regression model.

Acknowledgement: The Part III has been prepared with the aid of RNDr. Aleš Linka who collaborated in programming Matlab codes for computation.

References

[1] Aalen O. O. (1980). A model for nonparametric regression analysis of counting processes. Springer Lect. Notes in Statist. 2, 1–25.

[2] Andersen P. K. and Borgan O. (1985). Counting process models for life history data: A review. Scand. J. Statist. 12, 97–158.

[3] Andersen P. K., Borgan O., Gill R. D. and Keiding N. (1993). Statistical Models Based on Counting Processes. Springer, New York.

[4] Andersen P. K. and Gill R. D. (1982). Cox's regression model for counting processes: A large sample study. Annals of Statistics 10, 1100–1120.

[5] Arjas E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. J. Amer. Statist. Assoc. 83, 204–212.

[6] Arjas E. (1989). Survival models and martingale dynamics. Scand. J. Statist. 16, 177–225.

[7] Arjas E. and Andreev. A. (2000). Predictive inference, causal reasoning, and model assessment in nonparametric Bayesian analysis: a case study.Lifetime Data Analysis 6.

[8] Arjas E. and Liu L. (1995): Assessing the losses caused by an industrial intervention – a hiearchical Bayesian Approach. J. Royal Statist. Society, ser. C, 44, 357–368.

[9] Aven T. (2000). Reliability analysis as a tool for expressing and commonicating uncertainty. In Recent Advances in Reliability (eds. N. Limnios and M. Nikulin). Birkhauser, Boston, 23–38.

[10] Aven T. and Jensen U. (1999). Stochastic Models in Reliability. Springer, New York.

[11] Barlow R.E. and Proschan F. (1967). Mathematical Theory of Reliability. Wiley, New York.

[12] Belyaev Yu. K. and Rydén P. (1997). Non-parametric estimators of the distribution of tensile strengths for wires. Research Report 1997-15, University of Umea.

[13] Crowder M. J., Kimber A. C., Smith R. L. and Sweeting T. L. (1991). Statistical Analysis of Reliability Data. Chapman and Hall, London.

[14] Daniels H. E. (1991). The maximum of a Gaussian process whose mean path has a maximum, with an application to the strength of bundles of fibres. Adv. Appl. Probab. 1991, 315–333.

[15] Embrechts P., Klüppelberg C. and Mikosch T. (1997). Modelling Extremal Events. Springer.

[16] Fan J. and Gijbels I. (1996): Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

[17] Fleming T.R. and Harrington D.P. (1991): Counting Processes and Survival Analysis. Wiley, New York.

[18] Fisher R. A. (1970). Statistical Methods for Research Workers, 14th ed., Hafner, New York.

[19] Gamerman D. and West M. (1987). An application of dynamic survival models in unemployment studies. The Statistician 36, 269–274.

[20] Gavin J., Haberman S. and Verrall R. (1993). Moving weighted average graduation using kernel estimation. Insurance: Mathematics & Economics 12, 113–126.

[21] Gentleman R. and Crowley J. (1991): Local full likelihood estimation for the proportional hazard model. Biometrics 47, 1283–1296.

[22] Grandell J. (1997). Mixed Poisson Processes. Chapman and Hall.

[23] Han A. and Hausman J. A. (1990). Flexible parametric estimation of duration and competing risk models. J. of Applied Econometrics 5, 1–28.

[24] Hastie T. and Tibshirani R. (1986). Generalized additive models; with discussion. Statist. Science 1, 297–318.

[25] Hastie T. and Tibshirani R. (1990). Generalized Additive Models. Wiley, New York.

[26] Hastie T. and Tibshirani R. (1993). Varying-coefficient models. J. R. Statist. Soc. B 55, 757–796.

[27] Hoffmann-Jorgensen J. (1994). Probability with a View Towards Statistics, Vol. II. Chapman and Hall, London.

[28] Hoyland A. and Rausand M. (1994). System Reliability Theory. Wiley, New York.

[29] Kahle W. and Wendt H. (2000). Statistical analysis of damage processes. In Recent Advances in Reliability (eds. N. Limnios and M. Nikulin). Birkhauser, Boston, 199–212.

[30] Kalbfleisch J. D. and Struthers C. A. (1982): An analysis of the Reynolds Metals Company data. In Case Studies in Data Analysis, Canad. J. Statist. 10, 237–259.

[31] Kooperberg C., Stone C. J. and Truong Y. K. (1993). The L_2 rate of convergence for hazard regression. Techn. Report Calif. Univ. No 390.

[32] Lawless J. F. (1995). The analysis of recurrent events for multiple subjects. Applied Statistics 44, 487–498.

[33] Marzec L. and Marzec P. (1993): Goodness-of-fit inference based on stratification in Cox's regression model. Scand. J. Statist. 20, 227–238.

[34] Murphy S. A. and Sen P. K. (1991). Time-dependent coefficients in a Cox-type regression model. Stoch. Proc. and Their Applications 39, 153–180.

[35] O'Sullivan F. (1993): Nonparametric estimation in the Cox model. Annals Statist. 21, 124–145.

[36] Prentice R. L. and Gloeckler L. A. (1978). Regression analysis of grouped data with application to breast cancer data. Biometrics 34, 57–67.

[37] Rao C. R. (1965). Linear Statistical Inference and Its Application. Wiley, New York.

[38] Scheike T.H. (1994). Parametric regression for lokgitudinal data with counting process measurement times. Scand. J. Statist. 21, 245–263.

[39] Singpurvala N. D. (1995). Survival in dynamic environments. Statist. Science 10, 86–103.

[40] Snyder D. L. (1975). Random Point Processes. Wiley, New York.

[41] Stone C.J. (1986). The dimensionality reduction principle for generalized additive models. Annals Statist. 14, 590–606.

[42] Stone C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. Annals of Statist. 22, 118–194.

[43] Suh M. V., Bhattacharyya B. B. and Grandage A. (1970). On the distribution and moments of the strength of a bundle of filaments. J. Appl. Probab. 7, 712–720.

[44] Thomas D. C. (1982): Case analysis using Cox's model. In Case Studies in Data Analysis. Canad. J. Statist. 10, 237–259.

_ _ _ _ _

[45] Volf P. (1982). Estimates of distribution function and tests of fit from censored and grouped data. In Proceedings of 9-th Prague Conference on Info Theory, Academia Praha, 281–286.

[46] Volf P. (1992). Regression models in survival analysis. In Proceedings of Robust'92. JČMF Prague, 220–235 (in Czech).

[47] Volf P. (1993a). Moving window estimation procedures for additive regression function. Kybernetika 29, 389–400.

[48] Volf P. (1993b). Proportional intensity model for regression in event-history analysis. Informatica 4, 111–125.

[49] Volf P. (1994). Dimension reducing approach in statistical regression analysis. In Proceedings of IEEE – CMP'94 Workshop. UTIA AS CR, Prague, 71–76.

[50] Volf P. (1996). Analysis of generalized residuals in hazard regression models. Kybernetika 32, 501–510.

[51] Volf P. (2000). On cumulative process model and its statistical analysis. Kybernetika 36, 165–176.

[52] Volf P. and Linka A. (1997). On MCMC Methods in Bayesian Regression Analysis and Model Selection. Lecture Notes on MCMC, Summer School of IASC, Univ. of Padova. 33 pp.

[53] Volf P. and Linka A. (1998). Two applications of counting processes. Research Report ÚTIA AV ČR, No 1935.

[54] Volf P. and Linka A. (2000). On reliability of system composed from parallel units. Submitted to Intern. J. on Reliability, Quality and Safety Engineering.

[55] Volf P., Smid J., Markham B. and Seiferth J. (1997). Modelling and change detection from satellite instrumentation callibration coefficients time series: An example based on Landsat-5 thematic mapper data. In Earth Observing Systems. The Spie Paper No 3117, San Diego, 197–205.

[56] Volf P. and Picek J. (2000). Nonlinear quantile regression survival analysis. In Proceedings of Mathem. Methods in Reliability, Univ. of Bordeaux 2000, Vol. 2, 864–867 .

ATTACHMENT

Attachment contains 3 selected publications of Petr Volf:

1. Volf P.: Moving window estimation procedures for additive regression function. Kybernetika 29, 1993, 389–400.
2. Volf P.: Analysis of generalized residuals in hazard regression models. Kybernetika 32, 1996, 501–510.
3. Volf P.: On cumulative process model and its statistical analysis. Kybernetika 36, 2000, 165–176.

In the first paper, the method of local scoring is described. It is used for the estimation of (nonparametrized) regression functions in general regression models. A modification of the method is proposed, its use in the case of logistic regression model and nonparametric Cox's regression model is presented.

The second paper proposes a generalization of the method of statistical test (of goodness-of-fit) in the case of models of hazard functions. Namely, the cases of Cox's model and of additive Aalen's model are discussed. Both graphical and numerical versions (the latter one based on limit properties) of tests are presented.

The third paper deals with the process of random sums driven by the counting process, i.e. the process with increments at random moments. Again, the methods of model specification, evaluation and testing are derived. An example of analysis of financial time series is given. The most contemporary author's research in this area is focused on the models of degradation of reliability of technical systems (e.g. caused by a sequence of shocks and similar events).

KYBERNETIKA — VOLUME 29 (1993), NUMBER 4, PAGES 389 – 400

MOVING WINDOW ESTIMATION PROCEDURES FOR ADDITIVE REGRESSION FUNCTION[1]

PETR VOLF

The general additive regression function $b(\mathbf{x}) = \sum b_j(x_j)$ is considered and subjected to nonparametric estimation. The method of estimation is inspired by the regressogram approximations to the components of regression function. The procedure using the moving window is then derived, it naturally generalizes to a kernel approach. The method can be applied to the likelihood-based models, in which the value of regression function is a parameter of likelihood of a response variable Y. Suggested moving window algorithm is a variant of Hastie and Tibshirani's [3] local scoring procedure. In order to discuss the quality of obtained results, the method is compared with the approximation by regression splines, treated successfully by Stone [6]. An example illustrates the solution for the logistic regression, the proportional hazard regression model is also examined.

1. INTRODUCTION

The methods for nonparametric estimation and smoothing of curves are in the centre of attention of the data analysts for a long time. The modern equipment enables the statistician to examine the data attentively and to do deep preliminary analysis. Hence the nonparametric estimation of the covariate effect is at least a part of preliminary examination.

Let us consider a pair (X, Y) of real-valued random variables. In a regression model, X is called a covariate, meanwhile Y is a response variable. Let the general regression function be some smooth function $b(x)$, describing the dependence of a response variable Y on a covariate X. Likelihood-based regression model means that the value $b(x)$ is a parameter of likelihood for Y given $X = x$. Examples of this are the normal regression model, in which the regression function stands for $\mathsf{E}(Y|X = x)$, or the logistic regression model. We shall also mention the proportional hazard regression model for survival data.

If the observation is represented by a random sample (X_i, Y_i) of extent n, often the logarithm of likelihood can be expressed as

$$\ell_n = \sum_{i=1}^{n} \ell_1(Y_i, b(X_i)), \tag{1}$$

[1]This work was supported by Czech Academy of Sciences grant No. 27 557.

where ℓ_1 is a loglikelihood for one realization of Y, conditioned by a value of X. How to cope with the task of estimation of function b from the log-likelihood? One way may consist in approximation for $b(x)$, by a parametrized function. Every smooth function can be well approached by a linear combination from some basis of functions. For instance, the splines are the popular choice. Sleeper and Harrington [5] illustrate successfully the flexibility of regression splines in the analysis of the form of hazard ratio. Stone [6] used the approximation of regression function by splines in the framework of exponential family of distributions. He proved consistency of this approximation provided the parameters of splines were estimated by (global) maximum likelihood method. Thus, the reparametrization may be considered as an alternative way to solution. From this point of view, the regressogram is a trivial spline, with the order 0.

A widespreaded discussion runs about advantages and capabilities of both approaches – splines and kernel-like smoothing, cf. also discussion to paper of Hastie and Tibshirani [3]. The author does not intend to contribute to arguments of any side, his opinion is that every well-working method is valuable. Although some data-analysts (when joking) claim that one data may be analysed only once and only by one method – in order to avoid contradictions and problems with interpretation of results.

Our approach to estimation of regression function starts from a regressogram approximation. Then it proceeds to the moving window concept, considering simultaneously the additive regression function $b(\mathbf{x}) = \sum_j b_j(x_j)$ in the case of multi-dimensional covariate. It is necessary to stress at once that the additive model can include various transformations of covariates, their interactions (e. g. $x_1 \cdot x_2$), or, say, two-dimensional covariate, so that its idea seems to be sufficiently wide and flexible.

The general features of the method are described in the second part. Part 3 deals with the case of multi-valued logistic regression model. Part 4 considers a rather general case of a counting process with intensities fulfilling the proportional hazard model. The properties of solution are discussed in Part 5. Finally, an example with artificial data is solved numerically and discussed briefly.

Although the moving window procedure is a very flexible and easily computable method, its consistency is not guaranteed by any theoretical result. Only for the case of the normal regression model, in a more general concept of the Alternating Conditional Expectations (ACE) algorithm, Breiman and Friedman [2] show that the solution obtained by the moving window smoothing is the best additive approximation to $\mathsf{E}(Y|\mathbf{X})$. It means also that if $\mathsf{E}(Y|\mathbf{X} = \mathbf{x})$ is an additive function, it is consistently estimable by the moving window approach.

For a more general family of models, the results of Stone [6], mentioned above, can be used in order to support our conviction about the quality of the moving window smoothing. We discuss the connection between the moving window concept and the approximation by regression splines–polynomials on fixed windows.

Volf in [8] deals exclusively with the proportional hazard regression models and solves several simulated examples, in order to show a good performance of the method. In the example illustrating the paper of Sleeper and Harrington [5], the result of smoothing by the splines is compared graphically with the result obtained

by the local scoring.

2. LIKELIHOOD-BASED ESTIMATION PROCEDURE

Let us first consider the one-dimensional covariate X, with values in some finite interval $\mathcal{X} \subset R$. The construction of a regressogram means that the domain $\mathcal{X}$ is divided into M disjoint intervals I_m (their choice depends on the analyst), the function $b(x)$ is approximated as $\sum_{m=1}^{M} \beta_m \cdot \mathbf{1}[x \in I_m]$. Now, after inserting into the loglikelihood, the parameters β_m are estimated in ordinary way, which searches for solution of the equations $\partial \ell_n / \partial \beta_m = 0$, $m = 1, \dots M$. If the loglikelihood is of the form (1), then its first and second derivatives are

$$\frac{\partial \ell_n}{\partial \beta_m} = \sum_{i=1}^{n} \mathbf{1}[X_i \in I_m] \cdot \ell_1'(Y_i, \beta_m) \tag{2}$$

$$\frac{\partial^2 \ell_n}{\partial \beta_m \, \partial \beta_k} = \sum_{i=1}^{n} \mathbf{1}[X_i \in I_m] \cdot \ell_1''(Y_i, \beta_m) \quad \text{for } m = k, \quad = 0 \text{ otherwise.}$$

The step from estimation of the regressogram to the moving window estimation is quite straightforward. If we wish to estimate the value of $b(x)$ at a point $x = z$, we take b as a constant b_z in some chosen neighborhood (window) around z, say, in $\mathcal{O}(z)$. Then b_z is treated as a parameter, we have to solve the equation $\partial \ell_n / \partial b_z = 0$.

If the loglikelihood has the form (1), then

$$\frac{\partial \ell_n}{\partial b_z} = \sum_{i} \mathbf{1}\left[X_i \in \mathcal{O}_z\right] \cdot \ell_1'(Y_i, b_z). \tag{3}$$

Now, (3) contains only the derivatives of a "local" loglikelihood. It is the basis for the idea of the local scoring (or local likelihood) algorithm. However, when the form (1) does not hold, the derivatives do not contain the local results only. It is clearly visible in Example 2 which deals with the proportional hazard regression model.

Example 1. Logistic regression with two-valued response.
Let $\mathrm{P}(Y = 0 \,|\, x) = 1/(1 + \exp b(x))$, $\mathrm{P}(Y = 1 \,|\, x) = 1 - \mathrm{P}(Y = 0 \,|\, x)$. Then

$$\ell_n = \sum_{i=1}^{n} \left\{ b(X_i) \cdot \mathbf{1}[Y_i = 1] - \ln(1 + \exp b(X_i)) \right\}.$$

Example 2. Proportional hazard model for survival times and for i.i.d. sample $\{Y_i, \delta_i, X_i, i = 1, \dots, n\}$, where Y_i is an observed value and δ_i is the indicator of censoring. It means that $\delta_i = 1$ when Y_i is observed survival time, $\delta_i = 0$ if Y_i is less than survival time, the ith observation is censored at time moment Y_i. The inference for the hazard proportion $b(x)$ is based on the logarithm of Cox's partial

likelihood (cf. Andersen and Gill [1]), namely on

$$\ell_n = \sum_{i=1}^{n} \delta_i \ln \left\{ \frac{\exp b(X_i)}{\sum_{j=1}^{n} \exp b(X_j) \cdot I_j(i)} \right\},$$
$$\text{where } I_j(i) = 1 \text{ if } Y_j \geq Y_i, \quad I_j(i) = 0 \text{ otherwise.}$$

This partial likelihood has not the form (1). Nevertheless, let us compute its first derivatives with respect to value b_z in a neighbourhood $\mathcal{O}(z)$ of a point $z \in \mathcal{X}$:

$$\frac{\partial \ell_n}{\partial b_z} = \sum_i \delta_i \left\{ \mathbf{1}[X_i \in \mathcal{O}_z] - \frac{\exp b_z \cdot \sum_j I_j(i) \cdot \mathbf{1}[X_j \in \mathcal{O}_z]}{\sum_j \exp b(X_j) \cdot I_j(i)} \right\}. \tag{4}$$

The numerical iteration is the most frequently used way how to solve the likelihood equations. As a rule, the procedures need the second derivative of the loglikelihood, which in the case (1) yields

$$\frac{\partial^2 \ell_n}{\partial b_z^2} = \sum_i \mathbf{1}\{x_i \in \mathcal{O}_z\} \cdot \ell_1''(Y_i, b_z).$$

When the Newton–Raphson procedure is applied, the step from sth to $(s+1)$-st iteration is given by the following expression:

$$b_z^{(s+1)} = b_z^{(s)} - \frac{\partial \ell}{\partial b_z} \Big/ \frac{\partial^2 \ell}{\partial b_z^2}, \tag{5}$$

where the derivatives are evaluated at $b^{(s)}(x)$.

Hastie and Tibshirani [3] recommend to incorporate a smoothing directly into every step (5), they suggest the modification

$$b_z^{(s+1)} = \text{ smooth } \left[b_z^{(s)} - \frac{\partial \ell}{\partial b_z} \Big/ \text{ smooth } \left(\frac{\partial^2 \ell}{d b_z^2} \right) \right].$$

The notion of smoothing can have a very wide meaning, from weighted mean to, say, local parametrized regression.

Both examples mentioned above allow also another iteration procedure, which differs from (5) and which does not use the second derivatives. Moreover, after smoothing the results at each point, we shall "secondarily" smooth the final result. The procedure will be described in the following parts of the paper.

Let us now consider the K-dimensional covariate $\mathbf{X}$, with values in some bounded interval $\mathcal{X} \subset R_K$. When the dimension of $\mathbf{X}$ increases, the data are sparse and the method using the K-dimensional windows becomes ineffective. Then the additive "hypothesis" is available. The general additive regression model means that the regression function is

$$b(\mathbf{x}) = \sum_{k=1}^{K} b_k(x_k).$$

The component functions b_k should be nonparametrically estimated. The technique is essentially the same as for the one-dimensional case, but the inner loops has

to be incorporated to the procedure. This loop computes (at each sth step of the "outer" loop) successively all $b_k^{(s)}$, $k = 1, \ldots, K$, at all chosen points z_k. At least the values at all realized points x_{ki} are needed for further computation. Here k denotes the component, i denotes the case, $i = 1, \ldots, n$.

Let z be a point from the domain of X_1, say. The derivation of loglikelihood (1) with respect to $b_1(z)$ now yields

$$\frac{\partial \ell_n}{\partial b_1(z)} = \sum_{i=1}^{n} \mathbf{1}\left[X_{1i} \in \mathcal{O}_z\right] \cdot \ell_1'\left(Y_i,\, b_1(z) + \sum_{k=2}^{K} b_k(X_{ki})\right).$$

It is seen that the actual estimates (i.e. estimates obtained from the last preceding step) of other component functions b_k, $k = 2, 3, \ldots, K$, have to be available.

3. LOGISTIC REGRESSION MODEL

Let Y be a random variable with $M + 1$ possible values from $\{0, 1 \ldots, M\}$. The logistic model describes the dependence of probability distribution of Y on a (K-dimensional) covariate $\mathbf{X}$. The model assumes that

$$\mathrm{P}(Y = 0 \,|\, \mathbf{X}) = 1/S(\mathbf{X}), \qquad \mathrm{P}(Y = y \,|\, \mathbf{X}) = \exp(C(y, \mathbf{X}))/S(\mathbf{X}),$$

$$\text{with} \quad S(\mathbf{X}) = 1 + \sum_{m=1}^{M} \exp C(m, \mathbf{X}),$$

when $y = 1, 2, \ldots, M$. Moreover, the additive version of the model considers additive functions $C(m, \mathbf{x}) = \sum_{k=1}^{K} C(m, k, x_k)$. The form of the log-likelihood has been sketched in Example 1, now it enlarges to

$$\ell_n = \sum_{i=1}^{n} \left\{ \sum_{m=1}^{M} \mathbf{1}[Y_i = m] \cdot \sum_{k=1}^{K} C(m, k, X_{ki}) - \ln S(\mathbf{X}_i) \right\}. \tag{6}$$

Our task consists in successive estimation of all functions $C(m, k, x)$ as a functions of $x = x_k$. Let us imagine that we have already got some estimates of the regression functions from the sth step of the outer loop. In order to proceed with $(s + 1)$-st step of estimation, we need to know the estimates of $C(m, k, x)$ at all realized points x_{ki} $i = 1, \ldots, n$, $k = 1, \ldots, K$. Let z be a point in the domain of X_j, $\mathcal{O}_z$ be its neighborhood (an interval around z). The actualized $(s + 1)$-st estimation of $f_m = C(m, j, z)$ is obtained from the solution of (local) likelihood equation

$$\frac{\partial \ell_n}{\partial f_m} = \sum_{i=1}^{n} \mathbf{1}\left[X_{ji} \in \mathcal{O}_z\right] \left\{ \mathbf{1}[Y_i = m] - \frac{\exp(f_m) \cdot \exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} \right\} = 0, \tag{7}$$

where $C_j(m, \mathbf{x}_i) = \sum_{k=1}^{K} \mathbf{1}[k \neq j]\, C(m, k, x_{ki})$. We can estimate the value of f_m simultaneously for all $m = 1, \ldots, M$. The equations (7), in which j and z are fixed, can be solved separately for each m, or it can be solved as an M-dimensional equation. When computing the example described in Part 6, we used separate evaluation

for one value of m after another. The procedure then proceeds to another point z in the domain of X_j. When the values of $C(m, j, x_{ji})$ in all realized points x_{ji} and for each m are estimated, $i = 1, \ldots, n$, $m = 1, \ldots, M$, the algorithm starts to compute the estimates of $C(m, j+1, x_{j+1,i})$. All these computations are a part of the inner loop. It iterates through all $j = 1, \ldots, K$. Only then the algorithm may proceed to a further $(s+2)$-nd step of the outer loop, which again runs for $j = 1$ to K. The iterations are repeated until the convergence of all estimated functions $C(m, k, \cdot)$. How can be the convergence of functions checked and recognized? After every step, for every component $C(m, j, \cdot)$ we can construct the optimal least squares line through the points $C(m, j, x_{ji})$, $i = 1, \ldots, n$. The changes of the parameters of the line can serve as a criterion of iteration progress and as an indicator of convergence.

The usual way how to solve (7) consists in an iteration with the help of the second derivative of loglikelihood, for instance it may follow the scheme (5). In our example with the logistic model,

$$\frac{\partial^2 \ell_n}{\partial f_m^2} = \sum_{i=1}^{n} \frac{\exp(f_m) \cdot \exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} \left\{ \frac{\exp(f_m) \cdot \exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} - 1 \right\} \cdot \mathbf{1}[X_{ji} \in \mathcal{O}_z],$$

if again $f_m = C(m, j, z)$, j and z are fixed. But the form of equation (7) suggests also another procedure of iterative estimation. If (7) is solved directly for f_m, it yields

$$f_m = -\ln \left\{ \sum_{i,z} \frac{\exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} \Big/ \sum_{i,z} \mathbf{1}[Y_i = m] \right\}, \tag{8}$$

where the sums are through $\{i = 1, \ldots, n : X_{ji} \in \mathcal{O}_z\}$. The "inner" iteration again proceeds through all $m = 1, \ldots, M$, then through all $z = x_{ji}$ (realized points), and it renovates successively the estimates of component functions for $j = 1, \ldots, K$.

4. PROPORTIONAL HAZARD REGRESSION MODEL

The model is a popular choice for the description of covariate effect in life events history analysis. Especially, the Cox model is an often used representative of the model. It is able to analyse the censored data, its semiparametric form can be identified easily. However, the Cox model restricts the log hazard ratio to be linear in the covariates. A proportional hazard model considering a more general hazard function has an intensity $\lambda(t|x) = a(t) \cdot \exp(b(\mathbf{x}))$, where $b(\mathbf{x})$ is an unspecified smooth function. The estimation of proper function b can be based on K-dimensional kernel procedure (Volf [7]).

However, a more-dimensional covariate causes the data sparse and the global kernel approach loses its effectivity. Therefore, let us return to the model of the additive influence of covariates to the log hazard. Now the log hazard ratio has K components, $b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j)$. The analyst has to identify suitable functions $b_1, \ldots, b_K$ and also the underlying common hazard function $a(t)$, or better, its cumulative version $A(t) = \int_0^t a(s)\,\mathrm{d}s$. Evident ambiguity (with respect to additional constant in b_j's) can be overcome by proper normalization of the functions. Volf [8]

describes a method of estimation for a particular but frequent case, when each object has constant values of covariates. The method consists in alternating sequential computing of functions b_j and A. The procedure has been tested successfully, by simulated examples as well as by real data.

In the sequel, we shall consider a more general design, based on the counting processes. We have simultaneously to enlarge the model and to allow the time-dependent (random) processes of covariates $\mathbf{X}_i(t)$, $i = 1, \ldots, n$. In fact, such a system ceased to have the proportional hazards, although, for fixed $\mathbf{X} = \mathbf{x}$, the proportional hazard model holds. The counting process $N(t) = N_1(t), \ldots, N_n(t)$ is a set of right-continuous random step functions on $[0, \mathcal{T}]$, with steps $+1$. It is assumed that no two components step simultaneously. In this model, the components need not to be i.i.d., the recurrent jumps are allowed. $N_i(t)$ simply counts the events of ith kind or of ith object in the life history.

The model is fully described by the (random) hazard rates for counting processes $N_i(t)$, namely $\lambda_i(t) = a(t) \cdot \exp b\,(X_i(t)) \cdot I_i(t)$, $i = 1, \ldots, n$, $t \in [0, \mathcal{T}]$, where $I_i(t)$ is an indicator of risk set. It means that $I_i(t) = 1$ if the ith object is in the risk set at moment t, $I_i(t) = 0$ otherwise. The inference is based on Cox's partial likelihood. Its logarithm is

$$\ell_n = \sum_{i=1}^{n} \int_0^{\mathcal{T}} \ln \frac{e^{b(X_i(t))}}{\sum_{j=1}^{n} e^{b(X_j(t))} I_j(t)} \mathrm{d}\, N_i(t).$$

By the way, if we define again the underlying baseline cumulative hazard function $A(t) = \int_0^t a(s)\,\mathrm{d}s$, there exists its generalized maximum likelihood estimator $\hat{A}(t) = \int_0^t \frac{\mathrm{d}\bar{N}(s)}{\sum_j \exp b(X_j(s)) I_j(s)}$, where $\bar{N} = \sum_{i=1}^{n} N_i$. In the frame of Cox's model, this estimator is strongly consistent and asymptotically normal. However, here the analogy with the survival time model ends.

Let us now return to the idea of the kernel (moving window, or m-nearest neighbor) estimate for function $b(x)$. Inspired by (4) of Example 2, dealing with the one-dimensional case, we may suggest the iteration scheme $b^{(s+1)}(z) = h\left(b^{(s)}, z\right)$, where

$$h(b, z) =$$
$$= -\ln\left[\sum_i \int_0^{\mathcal{T}} \frac{\sum_j \mathbf{1}\left[X_j(t) \in \mathcal{O}(z)\right] I_j(t)}{\sum_j \exp b(X_j(t)) I_j(t)} \mathrm{d}N_i(t) \Big/ \sum_i \int_0^{\mathcal{T}} \mathbf{1}\left[X_i(t) \in \mathcal{O}(z)\right] \mathrm{d}N_i(t)\right]. \quad (9)$$

Here $\mathcal{O}(z)$ denotes our moving window-neighborhood of point $z \in \mathcal{X}$.

Do not forget that $N_i(t) - s$ are the step-wise functions, with steps $+1$ at the moments of "counts". In a survival time model it corresponds to moments $(Y_i, \delta_i = 1)$. It is seen, that we need not register all trajectories of $X_j(t)$, but only their values $X_j(S_i)$, where S_i are the moments of counts of $N_i(t)$, and we are able to register them only if $I_j(S_i) = 1$.

Let us now consider the situation with multidimensional covariate processes $\mathbf{X}(t) = (X_1(t), \ldots X_K(t))$ and suppose the additive form of function b, $b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j)$. Then the inner loop has to be incorporated to our iteration scheme. It

computes successively all components $b_1^{(s)}$ to $b_K^{(s)}$, then we proceed to the $(s+1)$-st step of "outer" iteration.

In order to obtain a generalization for iteration (9) with an additive regression function, let us imagine that when estimating, say, function $b_\ell(x_\ell)$, we have already estimated all functions $b_m(x_m)$, $m = 1, 2, \ldots, K$, during the preceding step of the outer loop.

Quite analogically to the one-dimensional case, from the equation $\partial \ell_n / \partial b_\ell(z) = 0$ we can suggest the following scheme for the moving window estimation procedure:

$$b_\ell^{(s+1)}(z) = -\ln \left[\sum_i \int_0^T \frac{R_\ell(z, b^{(s)}, t)}{S_0(b^{(s)}, t)} \, \mathrm{d}N_i(t) \Big/ \sum_i \int_0^T \mathbf{1}\left[X_{\ell i} \in \mathcal{O}_\ell(z)\right] \mathrm{d}N_i(t) \right],$$

where now

$$R_\ell(z, b, t) = \sum_{j=1}^n \mathbf{1}\left[X_{\ell j}(t) \in \mathcal{O}_\ell(z)\right] \cdot \exp\left\{ \sum_{k=1}^K \mathbf{1}[k \neq \ell] \cdot b_k(X_{kj}(t)) \right\} \cdot I_j(t),$$

$S_0(b, t) = \sum_{j=1}^n \exp\{b(\mathbf{X}_j(t))\} \cdot I_j(t)$ and $\mathcal{O}_\ell(z)$ is a chosen window around z in the domain of ℓth covariate.

The inner loop now iterates through $\ell = 1, \ldots, K$ and gives the values of $b_\ell^{(s+1)}(z)$ at every chosen z (we need at least the values at all observed $x_{\ell j}(T_i)$ provided $I_j(T_i) = 1$, where T_i are the moments of counts, $i, j = 1, \ldots, n$). The first inner loop may start from $b_1 = \cdots = b_K \equiv 0$ or from another convenient initial guess.

5. REMARKS ON CONSISTENCY

Stone [6] has examined the family of exponential-type regression models. Their loglikelihood has the form (1) with

$$\ell_1(Y, \mathbf{X}) = c(\theta(\mathbf{X})) \cdot Y + d(\theta(\mathbf{X})), \tag{10}$$

where c, d, e are known functions, θ is a regression function of our interest. The functions c, d are required to be twice continuously differentiable, with $c' > 0$. Stone has proved that under mild conditions a unique (as to the additive shift) additive function $b(\mathbf{x}) = \sum_{j=1}^K b_j(x_j)$ exists, closests to $\theta(\mathbf{x})$ in the sense of the Kullback–Leibler distance. Leaving this aspect of the problem apart, we assume that the regression function has already the additive form. From this point of view, the second result of Stone [6] is important. Stone has considered the polynomial splines (of chosen order) approximating each component b_j. Thus, the model is reparametrized by a finite number of parameters, they are then estimated by means of the standard (global) maximum likelihood method. It suffices to assume that:

1. The distribution of $\mathbf{X}$ is absolutely continuous on $\mathcal{X}$, with its density bounded away from zero and infinity.

2. Function b is Lipschitz continuous on $\mathcal{X}$.

3. The knots of the splines are chosen equidistantly and their number is proportional to n^γ, where γ is chosen properly from (0,1).

4. There are positive constants r and R such that

$$\mathsf{E}(\exp(s\,Y)|\mathbf{X}=\mathbf{x}) \leq R \quad \text{for} \quad |s| \leq r \quad \text{and} \quad \mathbf{x} \in \mathcal{X}.$$

Then, when n increases to infinity, this approximation by splines yields the consistent estimates of functions b_j. Stone gives also the order of convergence.

By the way, even the $M+1$ valued logistic regression model can be regarded as an M-dimensional representant of the exponential family. Let us recall its loglikelihood (6). It has the form (1), with

$$\ell_1(Y,\mathbf{X}) = \sum_{m=1}^{M} Y_m^* b^m(\mathbf{X}) - \ln\left\{1 + \sum_{m=1}^{M} \exp b^m(\mathbf{X})\right\},$$

where $Y_m^* = \mathbf{1}[Y=m]$.

Sometimes the likelihood is more complicated, e. g. in the case of the proportional hazard regression model. New results of Kooperberg et al. [4] prove consistency even for the spline approximation of hazard regression.

Let us now try to transfer the results of Stone to the moving window estimation. Let us recall again the regressogram approximation, in the case of loglikelihood (1) and one-dimensional covariate X. Its (global) maximum likelihood solution (2) leads in fact to the local likelihood iterations, because the matrix of the second derivatives is diagonal. Thus, the only difference between (3) and (2) consists in the use of the moving window instead of the fixed one. That is why the consistency property of Stone (which applies also to the regressogram – a "trivial" spline of order 0) holds also for the moving window solution. The sufficient conditions are the same as above, instead of increasing number of knots the decreasing width of window has to be considered, proportional to $n^{\gamma-1}$. Again, the result is well known in the case of the kernel estimation of regression function $\mathsf{E}(Y|X=x)$ in the normal regression model.

Unfortunately, the same statement does not hold when the additive regression function of multi-dimensional covariate is considered. Even in the framework of the exponential family of models, Hastie and Tibshirani [3] express a mere "conjecture" that their result of local scoring does not differ significantly from the approximation by splines.

However, the number of the splines-generated parameters is high, the direct computational task of global maximum likelihood would be too large. Therefore one should search for some sequential procedure, computing iteratively one subset of parameters after another. Such a procedure is again comparable with the local likelihood approach. But the optimality of such a procedure is not guaranteed.

6. NUMERICAL EXAMPLE WITH LOGISTIC MODEL

This part describes the data and the solution of an artificial example. Nevertheless, the case may represent a real situation. Let us imagine a company (say a kind of an

academic Institute), budget of which was affected by the economic problems in the country (probably the country from Central Europe). Therefore it was necessary to reduce the staff of the Institute. Simultaneously, some people are leaving the Institute voluntarily, they are searching for better paid jobs in the slowly developing private sector.

The sample has been collected during the critical period of the last two years. That is why the values of all covariates may be considered as constant in the time. The response variable characterizes the kind of leaving (or not leaving) the job during followed period. The data have the following structure:

$$\{\delta_i,\ X_{1i},\ X_{2i},\ X_{3i},\ X_{4i},\ i = 1, \ldots, n = 185\}.$$

Here n is the number of employeed, the response variable $\delta = 1$ when the employee was fired (42 cases), $\delta = 2$ when the individual left his job voluntarily (20 cases, retired employees are included in this group), $\delta = 0$ for remaining employees. The covariables have the following meaning: X_2 is the length of the previous employment in the Institute, up to the moment of event (in case of $\delta > 0$) or to the moment when the data have been collected ($\delta = 0$). It is measured in years. Its values are from 0 to 14. X_3 characterizes the category of the job: 1 – scientist (40 cases), 2 - specialist (98), 3 – administration (21), 4 – technical staff (16), 5 – unqualified assisting employees (10). $X_4 = 1$ for men (107), $= 2$ for women (78). X_1 is the age of the individual, again in years, at the moment of leaving the job or of collecting the data. Its range is from 20 to 60 years.

We wish to reveal and describe the influence of covariates on the probability of individuals to remain in the Institute, to be fired or to leave, respectively. The appropriate mathematical model is the model for the response variable (δ) and for its regression on the covariables $X_1, \ldots, X_4$.

For this example, the analysis of the dependence of the response on the covariates will be accomplished in the frame of the logistic model, by the iteration procedure (8). As the fourth covariate acquires two values only, its influence can be described fully by a linear function $b_4(x_4) = \alpha + \beta \cdot x_4$. This assumption can be incorporated into the computing procedure.

The results of estimation are summarized in Table 1 and Figure 1. After 9 iterations the convergence has been achieved. Table 1 displays the parameters of optimal lines (and correlation and variance analysis) led through estimated points of functions $C(m, j, x)$. This linear analysis has been done before a final (secondary) smoothing. Thus, the least squares procedure has been weighted by the number of tied values of a covariate. It concerned especially the third covariate. The weighting with respect to the variance of results in a window has not been considered. Figure 1 then displays smoothed estimates of functions $C(m, j, x)$ for first three covariates, i.e. $j = 1 \sim$ age, $j = 2 \sim$ duration of employment in the Institute, $j = 3 \sim$ category of employee. Two distinct events were considered, for $m = \delta = \{1, 2\}$ for two distinct reasons of departure.

The example has been constructed in order to demonstrate the usefulness of additive regression models and in order to check the procedure of solution. No test of significance of regression has been applied in order to support the conclusions.

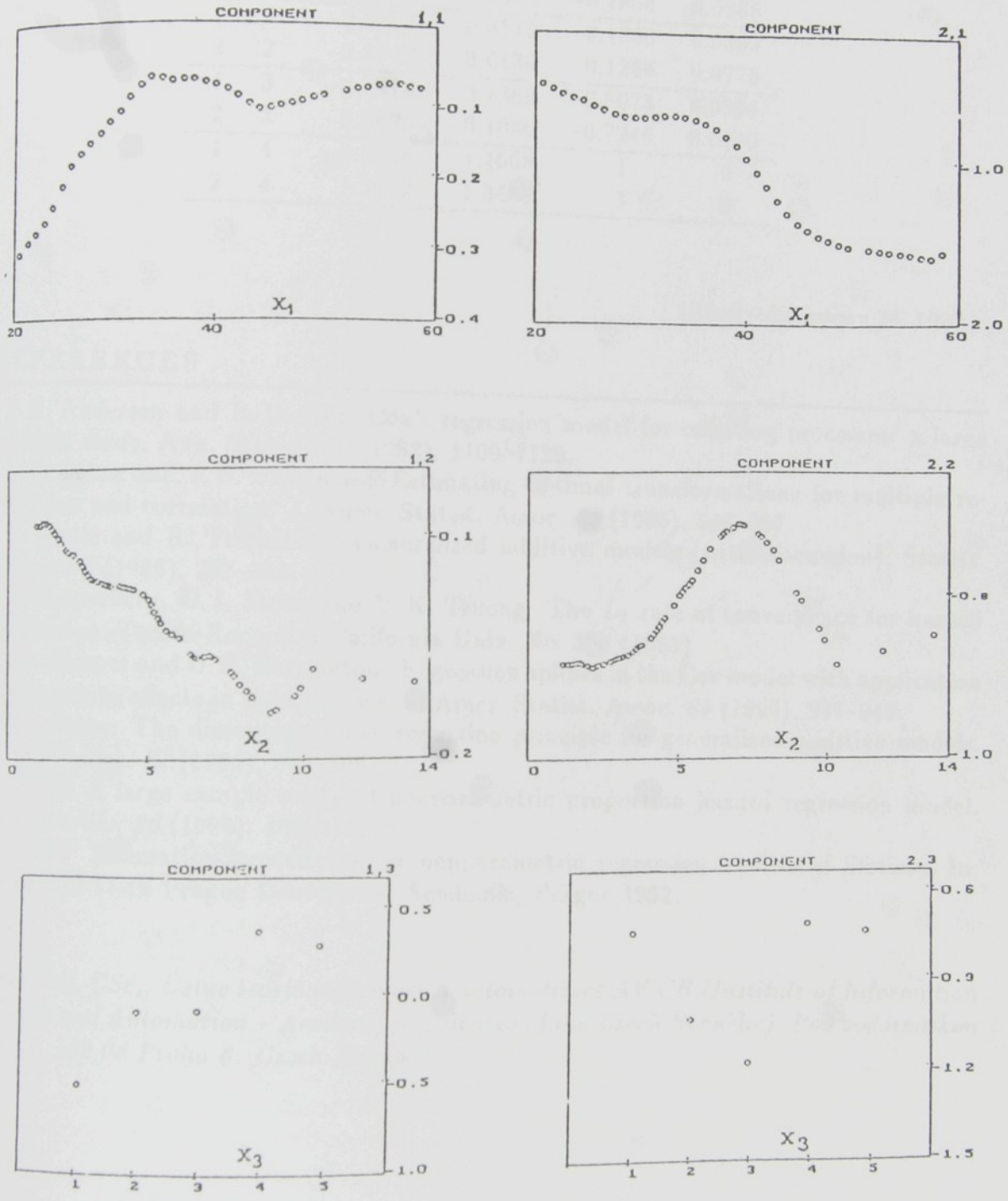


Figure 1.

Such a test might be accomplished in a traditional way, for a parametric logistic model, i.e. for a linear approximation to functions $C(m, j, x)$. The values of intercepts and slopes from Table 1 can be considered as preliminary estimates of the parameters of this parametric model.

Table 1.

m	j	intercept	slope	correl	var
1	1	−0.3434	0.0058	0.2805	0.0381
2	1	0.4507	−0.0386	−0.7858	0.0885
1	2	−0.0753	−0.0119	−0.1600	0.0383
2	2	−0.8912	0.0136	0.1286	0.0776
1	3	−0.7376	0.2369	0.8075	0.0394
2	3	−0.6800	−0.1006	−0.7246	0.0480
1	4	−0.5424	1.1668	1	0
2	4	1.3862	1.3485	1	0

(Received January 28, 1993.)

REFERENCES

[1] P. K. Andersen and R. D. Gill: Cox's regression model for counting processes: a large sample study. Ann. Statist. *10* (1982), 1100–1120.

[2] L. Breiman and J. H. Friedman: Estimating optimal transformations for multiple regression and correlation. J. Amer. Statist. Assoc. *80* (1985), 580–597.

[3] T. Hastie and R. Tibshirani: Generalized additive models (with discussion). Statist. Science *1* (1986), 297–318.

[4] C. Kooperberg, C. J. Stone and Y. K. Truong: The L_2 rate of convergence for hazard regression. Techn. Report of California Univ. No. 390 (1993).

[5] L. A. Sleeper and D. P. Harrington: Regression splines in the Cox model with application to covariate effects in liver disease. J. Amer. Statist. Assoc. *85* (1990), 941–949.

[6] C. J. Stone: The dimensionability reduction principle for generalized additive models. Ann. Statist. *14* (1986), 590–606.

[7] P. Volf: A large sample study of nonparametric proportion hazard regression model. Kybernetika *26* (1990), 404–415.

[8] P. Volf: Estimation procedures for nonparametric regression models of lifetime. In: Trans. of 11-th Prague Conference, Academia, Prague 1992.

Petr Volf, CSc., Ústav teorie informace a automatizace AV ČR (Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic), Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.

KYBERNETIKA — VOLUME 32 (1996), NUMBER 5, PAGES 501 – 510

ANALYSIS OF GENERALIZED RESIDUALS IN HAZARD REGRESSION MODELS

PETR VOLF

In the present paper, we consider a counting process and a model of its intensity. We introduce the generalized residuals measuring the deviation of observed times to counts from the expected times given by the model. These residuals are then used for assessing the goodness-of-fit of hazard regression models. The method is inspired by Arjas' [4] graphical procedure (dealing with Cox's model) and generalized to a quite general hazard regression case. The large sample properties of the test statistics are derived, they are then specified for the case of Aalen's regression model. The diagnostic ability of the method is illustrated by an example with simulated data.

1. INTRODUCTION

The purpose of the statistical event-history analysis consists in the examination of streams of events modelled with the help of the counting processes. The random behaviour of a counting process is as a rule characterized by a hazard function. Hazard regression models describe the case when the hazard function depends on values of covariates. The inferences are sought about the form of this dependence. The most popular representative of hazard regression models is the proportional hazard model of Cox. The present contribution is prevailingly devoted to the methods of diagnostics for quite general hazard regression models. We develop both graphical and numerical procedures of goodness-of-fit testing. Then, for the case of Aalen's model, we derive an asymptotic distribution of the test statistics with plugged-in estimate of the hazard function.

Our approach is inspired by the explorative and diagnostic methods presented for instance in Arjas [4]. The approach is based on the martingale-compensator decomposition of the counting process and on properly defined generalized residuals. Arjas dealt with graphical methods for assessing the fit of Cox's model. Later on, the large sample properties of Arjas' statistics were examined by Marzec and Marzec [9]. The main objective of the present paper is to generalize these results concerning both graphical and numerical procedures.

The paper is organized as follows: In Section 2 the notion of counting process is recalled, the hazard regression model is introduced and the process of generalized

residuals is defined. Part 3 is devoted to the graphical goodness-of-fit testing. We present a method applicable to quite general models. The large sample properties of the test statistics in the general case are derived in Part 4. Then, we consider also the situation when (a part of) the model is estimated. The case with 'plugged–in' estimator is solved for the Aalen's hazard regression model. A simple example with simulated data illustrates the usefulness of suggested approach.

2. MODEL AND RESIDUALS

Let $\boldsymbol{N}(t) = (N_1(t), \ldots, N_n(t))'$ be a multivariate counting process followed in the time period $[0, \mathcal{T}]$. It is supposed that $N_i(0) = 0$ and that $N_i(t)$ counts $+1$ when the ith individual encounters the (observed) event of interest. Further, it is assumed that there is maximally one count at one moment. The behaviour of $N_i(t)$ is governed by an intensity process $\lambda_i(t) = I_i(t) \cdot \lambda(t, \boldsymbol{X}_i(t))$, where $\lambda(t, \boldsymbol{x})$ is a bounded, non-negative, continuous hazard function, $\boldsymbol{X}_i(t)$ is a vector covariate process and $I_i(t)$ is a $\{0, 1\}$ valued process indicating whether $N_i(t)$ is at risk of count at moment t. In other words, $I_i(t) = 1$ when the ith individual is observed, $I_i(t) = 0$ otherwise.

The most popular hazard regression model is the Cox's proportional one, with $\lambda(t, \boldsymbol{x}) = \lambda_0(t) \exp(b(\boldsymbol{x}))$, where $\lambda_0(t)$ is a baseline hazard function. The most common semi–parametric version considers a linear specification $\boldsymbol{\beta}'\boldsymbol{x}$ of the function $b(\boldsymbol{x})$.

The intensity of the counting process has the interpretation $\lambda_i(t)\,\mathrm{d}t = P\,(\mathrm{d}\,N_i(t) = 1 \,|\, \sigma(t^-))$, where $\sigma(t)$ is a right continuous nondecreasing (w.r. to $t \in [0, \mathcal{T}]$) sequence of σ-algebras defined on the sample space of histories of $\boldsymbol{N}(t)$. More precisely, $\sigma(t)$ contains all (relevant) events which occurred up to time t. Processes $\boldsymbol{X}_i(t)$ and $I_i(t)$ are assumed to be left continuous and predictable w.r.to the sequence $\sigma(t)$, the trajectories of $N_i(t)$ are continuous from the right side.

Let us now recall the martingale-compensator decomposition of counting process, so basic for the theory of asymptotic normality and consistency of estimates. Define the cumulative intensities by $L_i(t) = \int_0^t \lambda_i(s)\,\mathrm{d}s$. The fact that $N_i(t)$ has an intensity process $\lambda_i(t)$ implies that $M_i(t) = N_i(t) - L_i(t)$ is a local square integrable martingale on $[0, \mathcal{T}]$, adapted to σ-algebras $\sigma(t)$. It holds that $E\,M_i(t) = 0$, the variance process $\langle M_i \rangle(t) = L_i(t)$, moreover, $M_i(t)$ is orthogonal to $M_j(t)$ $(i \neq j)$. Further details can be found for instance in Andersen and Gill [2], in Arjas [5], in Andersen et al [3].

Individual counting processes are connected through their common history stored in $\sigma(t)$. In the framework considered here, this connection is given by dependence of processes $I_i(t)$, $\boldsymbol{X}_i(t)$ on the past (up to t^-) of the system. Therefore, processes $N_i(t)$ are mutually conditionally independent provided the realizations of $\boldsymbol{X}_i(t)$ and $I_i(t)$ are known (of course, $\boldsymbol{X}_i(t)$ is needed only when $I_i(t) = 1$).

For the moment, let us imagine the case that each $N_i(t)$ has maximally one count (and that $I_i(t) = 0$ after the moment of the count). For each $i = 1, \ldots, n$ let us denote $S_i = \sup\{t \in [0, \mathcal{T}],\, I_i(t) = 1\}$ and define indicators $\delta_i = 1$ if S_i is the moment of count, $\delta_i = 0$ otherwise. Let us consider random variables T_i – waiting times to the counts of $N_i(t)$. The distribution of T_i is given by cumulative intensity $L_i(t) = \int_0^t \lambda_i(s)\,\mathrm{d}s$, the values of T_i's are observed with the right-sided censoring,

i.e. S_i, δ_i, $i = 1, \ldots, n$, are observed instead. When the conditional independence is taken into account, the following holds (cf. also Arjas, [5]):

Proposition 1. The couples of random variables $(L_i(S_i), \delta_i)$, $i = 1, \ldots, n$, are mutually (conditionally, for intensities $\lambda_i(t)$ given) independent, they are the results of the right-sided censoring of unit-exponentially distributed random variables $L_i(T_i)$.

Similarly, if the events are recurrent, $L_i(t)$ represents a transformation of the time scale. From each counting process $N_i(t)$ (having counts at moments T_{ij}^*, $j = 1, \ldots, N_i(T)$) a standard Poisson stream P_i $(L_i(t))$ is obtained, with counts at moments $L_i(T_{ij}^*)$. Again, these Poisson processes are mutually conditionally independent. It is seen that the testing the behaviour of the counting process can be based on comparision of $N_i(t)$, representing the data, with $L_i(t)$, representing the model.

Definition. The variable $L_i(t) - N_i(t)$, at a given t, is called the (generalized) residual. The process $L_i(t) - N_i(t)$, $t \in [0, T]$ is called the residual process.

Such a definition of residuals enable us to perform the analysis with the help of martingale-compensator decomposition. Sometimes, when we wish to compare the occurrence of events in various subclasses of individuals, it is convenient to define an aggregated counting process $\overline{N_S}(t) = \sum_{i \in S} N_i(t)$ – the sum through a chosen stratum $S \subset \{1, \ldots, n\}$. Its intensity process is given by $\overline{\lambda_S}(t) = \sum_{i \in S} \lambda_i(t)$ and the time scale transformation $\overline{L_S}(t) = \int_0^t \overline{\lambda_S}(s)\, ds$ yields again a standard Poisson stream of events.

3. GRAPHICAL GOODNESS–OF–FIT TESTING

In this section, we shall suggest a graphical procedure for testing the fit of general hazard-based regression model. Let $0 \leq T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)} \leq T$ denote the ordered times of counts of the whole observed system (actually, equalities are ruled out theoretically). Set $\overline{N}(t) = \sum_{i=1}^n N_i(t)$, $\overline{M}(t) = \sum_{i=1}^n M_i(t)$, $\overline{L}(t) = \sum_{i=1}^n L_i(t)$. We propose the graphical test based on (explorative) analysis of differences $\overline{L}(T_{(k)}) - \overline{N}(T_{(k)}) = -\overline{M}(T_{(k)})$. Notice that $\overline{N}(T_{(k)}) = k$, therefore we propose to plot $\overline{L}(T_{(k)})$ against k. In fact, we obtain nothing other than the chart of the time–transformed cumulative hazard function, which should be the cumulative hazard function of unit-exponential distribution, provided the model fits.

Suppose now that we wish to compare the behaviour of distinct groups of individuals. In this case, we split the n-sample into (two or more) strata and draw the plot separately for each of them. Of course, we plot the sums of functions $\overline{L_S}(t) = \sum_{i \in S} L_i(t)$, evaluating them at points $T_{(S,k)}$ – the kth moment of count in stratum S – and comparing them again with $\overline{N_S}(T_{(S,k)}) = k$.

Our method of testing generalizes the idea presented in Arjas [4]. In the framework of the Cox's model, Arjas used the variables $\overline{H_S}(T_{(k)})$, $\overline{H_S}(t) = \sum_{i \in S} H_i(t)$,

where $H_i(t) = \int_0^t p_i(s)\,\mathrm{d}\overline{N}(s)$ and $p_i(t) = I_i(t)\exp b(\boldsymbol{X}_i(t)) \,/\, \sum_{j=1}^n I_j(t)\exp b(\boldsymbol{X}_j(t))$. These variables are convenient because the knowledge (or estimation) of baseline hazard function is not required. The difference of $\overline{L_S}(t)$ from $\overline{H_S}(t)$ is again a martingale, so that the performance of both tests is comparable. In the case of the Cox's model, when in $\overline{L_S}(t)$ the baseline hazard function is replaced by its consistent estimate $\mathrm{d}\,\widehat{L}_0(t) = \mathrm{d}\overline{N}(t) \,/\, \sum_j I_j(t)\exp b(\boldsymbol{X}_j(t))$, both statistics coincide. Notice that $\overline{H}(t) = \sum_{i=1}^n H_i(t) = \overline{N}(t)$ directly, so that the statistics of this type cannot be used for testing without a stratification.

Let us now consider a following frequently encountered case: In the framework of a certain model, we wish to decide whether a specific form of submodel holds. For instance, let us assume that the general Cox's model fits and that we wish to assess whether the regression function $b(\boldsymbol{x})$ can be expressed as $\boldsymbol{\beta}'\boldsymbol{x}$ for some $\boldsymbol{\beta}$. Andersen and Gill [2] solved the task of maximal partial likelihood estimation of $\boldsymbol{\beta}$. The estimate $\widehat{\boldsymbol{\beta}}$ is shown to be consistent. Therefore, we replace $b(\boldsymbol{x})$ by $\widehat{\boldsymbol{\beta}}'\boldsymbol{x}$ in test statistics. The fit for stratified subsamples can be tested with the help of Arjas' statistics $\widehat{H_S}(t)$ (i.e. $\overline{H_S}(t)$ with $b(\boldsymbol{x}) = \widehat{\boldsymbol{\beta}}'\boldsymbol{x}$). Similarly, in a general case, an unknown part of a tested model (i.e. of $\overline{L_S}(t)$) may be replaced by its consistent estimator (provided such an estimator is available). Now the martingale decomposition (as well as Proposition 1, applied to the estimated model) holds only approximately, i.e. asymptotically. Nevertheless, the graphical procedure is still a useful indicator of validity of our hypothesis.

4. NUMERICAL TESTS AND LARGE SAMPLE PROPERTIES

Proposition 1 transforms the data into a sample of censored unit-exponential variables, provided $\lambda(t, \boldsymbol{x})$ is the "true" hazard rate. Therefore the test of fit of $\lambda(t, \boldsymbol{x})$ can be accomplished with the help of standard goodness-of-fit procedures adapted to censored data. For instance the modified Kolmogorov–Smirnov procedure can be used to assess the unit-exponentionality of $\{L_i(T_i)\}$. If the independent unit-exponentional waiting times are ranked into series, the waiting time to the kth sequential event is distributed according to the gamma $(1, k)$ law. The same holds for time $\overline{L_S}(T_{(S,k)})$ to kth event in an aggregated system $\{i \in S\}$ of parallelly running unit-exponential times. Simultaneously, $\overline{L_S}(T_{(S,k)}) - \overline{L_S}(T_{(S,k-1)})$ are distributed unit – exponentially and independently of $\sigma(T_{(S,k-1)})$.

All these properties should hold if the model $\lambda(t, \boldsymbol{x})$ is chosen properly. While the graphical testing methods are based directly on the properties of "exact" distributions, the numerical tests use as a rule the asymptotic laws, consequences of the central limit theorem. Their advantage is that they offer a quantified information about the magnitude of deviation from model. However, as soon as a part of the model is unknown and estimated, the asymptotics becomes rather hazy. The cases of plugged-in parameter estimator are discussed for instance in Khmaladze ([8] – the case of standard Kolmogorov–Smirnov statistics) or in Hjort ([7] – the case of estimated parameters in parametrized hazard function and in Cox's model). Both authors (and some others, cited in Hjort as well as in Arjas [4]) show that it is pos-

sible to construct asymptotic test of Kolmogorov–Smirnov type as well as χ^2-type test. However, the test statistics then need more computation than mere replacing the unknown parameter by an estimator.

The asymptotic behaviour of Arjas' residual process $\widehat{H_S}(t) - \overline{N_S}(t)$ applied to the Cox's model has been examined by Marzec and Marzec [9]. Two types of conditions have been specified, both following Andersen and Gill's [2] conditions of asymptotic stability and regularity. These conditions guarantee strong consistency of the estimator $\widehat{\boldsymbol{\beta}}$ and validity of the central limit theorem for martingales $M_i(t)$. The stronger set of conditions of Marzec and Marzec (claiming the uniqueness of limits of $\frac{1}{|S|}\sum_{i\in S} I_i(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}_i(t))$ independently of the strata S, provided $|S| \to \infty$, where $|S| = \sum_{i\in S} 1$) ensures the weak convergence of $n^{-\frac{1}{2}}(\widehat{H_S}(t) - \overline{N_S}(t))$ to a Gaussian process with independent increments. Therefore, the Kolmogorov–Smirnov asymptotic confidence regions can be constructed.

As has been pointed out, the stratification is a very helpful tool in the model diagnostics. In particular, the Arjas' statistics is derived to deal with stratified data. Nevertheless, in the sequel we shall leave the idea of stratification. It is not difficult to adapt all results to the stratified case. We shall briefly consider the general case of the hazard model, then we shall derive some asymptotic results for the regression model of Aalen [1].

Let us first repeat the basic assumptions about the boundness of intensity and covariate processes. These conditions, although slightly too strong, enable us to omit any additional condition of Lindeberg type (cf. again Andersen and Gill [2]).

Assumption 1. The covariate processes $\boldsymbol{X}_i(t)$ have their values in a bounded interval $\mathcal{X} \subset R^K$, for $t \in [0, \mathcal{T}]$.

Assumption 2. The hazard function $\lambda(t, \boldsymbol{x})$ is uniformly bounded on $[0, \mathcal{T}] \times \mathcal{X}$.

Further, let us formulate a general variant of a stability condition. Let P-lim denote the limit in probability:

Assumption 3. There exists a deterministic function
$\mathcal{W}(s) = P\text{-}\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I_i(s)\,\lambda(s, \boldsymbol{X}_i(s))$. The limit is uniform in $[0, \mathcal{T}]$.

Proposition 2. Under Assumptions 1–3 the process $n^{-\frac{1}{2}}\overline{M}(t)$ converges weakly on $[0, \mathcal{T}]$ to the Gaussian random process with independent increments and with variance function $\int_0^t \mathcal{W}(s)\,\mathrm{d}s$.

The proof follows immediately from the central limit theorem for martingales – cf. also Andersen and Gill [2], Marzec and Marzec [9].

Let us again return to the case of an unknown hazard function estimated by $\widehat{\lambda}$. Now, the analysis is based on the following 'empirical' residual process:

$$\int_0^t \sum_{i=1}^{n} \widehat{\lambda}_i(s)\,\mathrm{d}s - \overline{N}(t) = -\overline{M}(t) + \int_0^t \sum_{i=1}^{n} (\widehat{\lambda}_i(s) - \lambda_i(s))\,\mathrm{d}s. \tag{1}$$

The behaviour of such a process depends, obviously, on properties of the estimator of the hazard rate. In the following, we shall specify the asymptotic behaviour of this process for the case of Aalen's hazard regression model.

Aalen's regression model. Let the hazard function be $\lambda(t, \boldsymbol{x}) = \boldsymbol{\beta}'(t)\,\boldsymbol{x}$, both $\boldsymbol{\beta}$ and $\boldsymbol{x}$ being K-dimensional vectors. It follows that the individual intensity process is $\lambda_i(t) = I_i(t)\,\boldsymbol{\beta}'(t)\,\boldsymbol{X}_i(t)$, $i = 1, \ldots, n$. Let us assume that $\beta_1, \ldots, \beta_K$ are nonnegative, left-continuous functions, bounded on $[0, \mathcal{T}]$ – cf. Assumption 2. Further, assume that $X_{1i}(t) \equiv 1$, (so that $\beta_1(t)$ is an intercept function) and that $X_{2i}(t), \ldots, X_{Ki}(t)$ are the actual nonnegative covariate processes. The method of estimation of cumulative functions $B_k(t) = \int_0^t \beta_k(s)\,\mathrm{d}s$ is described (and consistency and asymptotic normality are shown) for instance in Andersen et al [3], part VII: Denote by $\boldsymbol{Z}(t)$ the $(K \times n)$ matrix having ith column $\boldsymbol{Z}_{\cdot i}(t) = \boldsymbol{X}_i(t) \cdot I_i(t)$. Then the simplest approach considers the estimator $\widehat{\boldsymbol{B}}(t) = \int_0^t \overline{\boldsymbol{Z}}(s)\,\mathrm{d}\boldsymbol{N}(s)$, where $\overline{\boldsymbol{Z}}(s) = (\boldsymbol{Z}(s)\,\boldsymbol{Z}'(s))^{-1}\boldsymbol{Z}(s)$, assuming that the inverse matrices exist for $s \in [0, \mathcal{T}]$. Notice, that $\overline{\boldsymbol{Z}}(s)$ is a generalized inverse matrix to $\boldsymbol{Z}'(s)$. Andersen et al [3] and others argue that this estimator, which is based on the simple least squares principle, does not take into account possible unequal variances of individual martingales $M_i(t)$. Therefore the weighted variant using $\overline{\boldsymbol{Z}} = (\boldsymbol{Z}\,\boldsymbol{W}\,\boldsymbol{Z}')^{-1}\,\boldsymbol{Z}\,\boldsymbol{W}$ should be preferred, with $\boldsymbol{W} = \mathrm{diag}[w_i(t)]$ a diagonal matrix of weights. The weights $w_i = 1/\lambda_i(t)$ are optimal theoretically ($w_i = 0$ should be set if $\lambda_i(t) = 0$). This choice can be achieved approximately with the help of a sequential procedure.

It holds that $\widehat{\boldsymbol{B}}(t) - \boldsymbol{B}(t) = \int_0^t \overline{\boldsymbol{Z}}(s)\,\mathrm{d}\boldsymbol{M}(s)$ and that under proper conditions $n^{\frac{1}{2}}(\widehat{\boldsymbol{B}}(t) - \boldsymbol{B}(t))$ is asymptotically distributed as a Gaussian process with independent increments. Its covariance function is then $P\text{-}\lim_{n\to\infty} n \int_0^t \overline{\boldsymbol{Z}}(s)\,\boldsymbol{D}(s, \boldsymbol{\beta})\,\overline{\boldsymbol{Z}}'(s)\,\mathrm{d}s$, provided such a limit exists and is regular. Here $\boldsymbol{D}(s, \boldsymbol{\beta})$ is the $(n \times n)$ diagonal matrix with components $\lambda_i(s) = \boldsymbol{\beta}'(s)\,\boldsymbol{X}_i(s)\,I_i(s)$. It is seen that if optimal weights $w_i = 1/\lambda_i$ are used, then $\overline{\boldsymbol{Z}}\,\boldsymbol{D}\,\overline{\boldsymbol{Z}}' = (\boldsymbol{Z}\,\boldsymbol{W}\,\boldsymbol{Z})^{-1}$.

The test statistics is now derived from $\overline{L}(t) = \sum_{i=1}^n \int_0^t I_i(s)\,\boldsymbol{X}_i'(s)\,\boldsymbol{\beta}(s)\,\mathrm{d}s$. When $\mathrm{d}\widehat{\boldsymbol{B}}(s)$ is inserted instead of $\boldsymbol{\beta}(s)\,\mathrm{d}s$, we obtain

$$\widehat{L}(t) = \sum_{i=1}^{n} \int_0^t I_i(s)\,\boldsymbol{X}_i'(s)\,\overline{\boldsymbol{Z}}(s)\,\mathrm{d}\boldsymbol{N}(s) = \sum_i \int_0^t \boldsymbol{Z}'_{\cdot i}(s)\,\overline{\boldsymbol{Z}}(s)\,\mathrm{d}\boldsymbol{N}(s),$$

where $\mathrm{d}\boldsymbol{N}(s) = (\mathrm{d}N_1(s), \ldots, \mathrm{d}N_n(s))'$. The difference from the observed number of counts up to t is

$$\widehat{L}(t) - \overline{N}(t) \;=\; \widehat{L}(t) - \overline{L}(t) - \overline{M}(t) = \sum_i \int_0^t \boldsymbol{Z}'_{\cdot i}(s)\left\{\mathrm{d}\widehat{\boldsymbol{B}}(s) - \mathrm{d}\boldsymbol{B}(s)\right\} - \overline{M}(t)$$

$$= \sum_i \int_0^t \left\{\boldsymbol{Z}'_{\cdot i}(s)\,\overline{\boldsymbol{Z}}(s)\,\mathrm{d}\boldsymbol{M}(s) - \mathrm{d}M_i(s)\right\} = \int_0^t \boldsymbol{i}'\left\{\boldsymbol{Z}'(s)\,\overline{\boldsymbol{Z}}(s) - \boldsymbol{I}\right\}\mathrm{d}\boldsymbol{M}(s),$$

where $\boldsymbol{i}' = (1, \ldots, 1)$ is the vector of dimension n, $\boldsymbol{I}$ denotes the identity matrix. Denote $\boldsymbol{u}'(s) = \boldsymbol{i}'(\boldsymbol{Z}'(s)\,\overline{\boldsymbol{Z}}(s) - \boldsymbol{I})$.

Assumption 4. Components $u_i(s)$ are bounded, uniformly with respect to $s \in [0, T]$ and to $i = 1, 2, \ldots$

Assumption 5. There exists a deterministic function

$$\mathcal{V}(s) = P\text{-}\lim_{n\to\infty} \frac{1}{n} \boldsymbol{u}'(s)\, \boldsymbol{D}(s, \boldsymbol{\beta})\, \boldsymbol{u}(s).$$

The limit is uniform w.r.to $s \in [0, T]$.

The assumptions claim actually the uniform regularity of matrices $\boldsymbol{Z}(s)\, \boldsymbol{W}(s)\, \boldsymbol{Z}'(s)$. They also assume existence of two limits, namely of

$$\frac{1}{n} \boldsymbol{i}'\, \boldsymbol{Z}'(s)\, \overline{\boldsymbol{Z}}(s)\, \boldsymbol{D}(s, \boldsymbol{\beta})\, \overline{\boldsymbol{Z}}'(s)\, \boldsymbol{Z}(s)\, \boldsymbol{i}, \quad \text{and of} \quad \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i'(s)\, \boldsymbol{\beta}(s)\, I_i(s).$$

Notice again that $\overline{\boldsymbol{Z}}\, \boldsymbol{D}\, \overline{\boldsymbol{Z}}' = (\boldsymbol{Z}\, \boldsymbol{W}\, \boldsymbol{Z}')^{-1}$ provided the optimal matrix $\boldsymbol{W}$ is used.

Proposition 3. Suppose the Assumptions 4, 5 hold. Then the process $n^{-\frac{1}{2}}(\widehat{L}(t) - \overline{N}(t))$ converges weakly to a Gaussian process with independent increments and with variance function $C(t) = \int_0^t \mathcal{V}(s)\, \mathrm{d}s$.

The proof follows again from the boundness of all processes and functions and from the fact that $\boldsymbol{D}(s, \boldsymbol{\beta})\, \mathrm{d}s$ is the conditional covariance matrix of $\mathrm{d}\boldsymbol{M}(s)$.

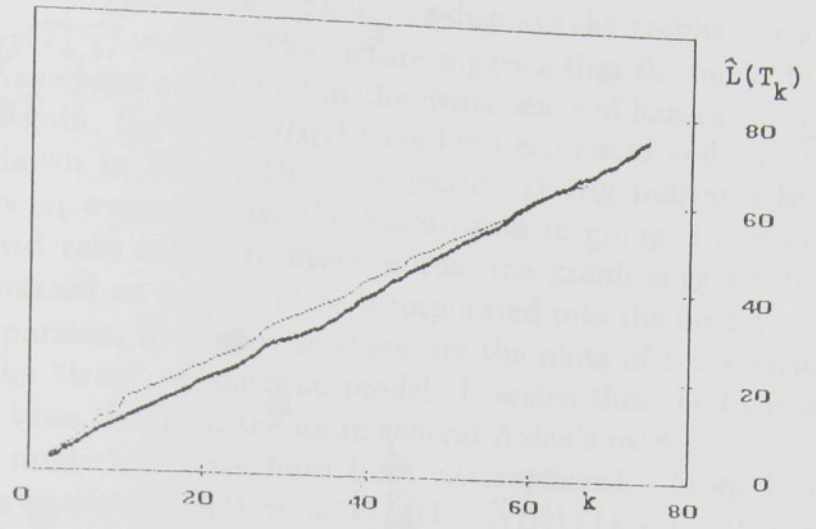
5. EXAMPLE

The method described in the present paper has been employed to analyse both simulated and real data. The application to regression diagnostics in the framework of the semi-parametric Cox's model is described in Arjas [4] as well as in Marzec and Marzec [9]. A more general multiplicative models are tested in the real data study of Volf [11].

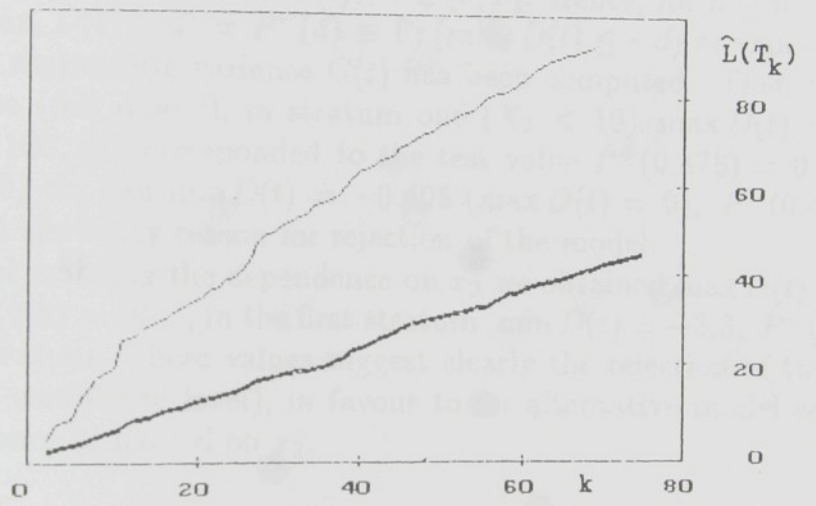
For the lack of space, let us present here one simple artificial example only. We simulated a sample ($n = 150$) of independent survival times fulfilling the Aalen's model with hazard function $\lambda(t, \boldsymbol{x}) = \beta_1(t)\, x_1 + \beta_2(t)\, x_2 + \beta_3(t)$. The values of covariates were generated uniformly, X_{1i} from (0,10), X_{2i} from (0,20), they were independent of time. The sample was not censored. We put $\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = 0.7$. Thus, the survival time had actually a very simple distribution, namely an exponential one with a constant hazard rate.

First, the "full" model has been estimated, by the method described in Part 4. The solution has been searched for in the set of general Aalen's models. We obtained the (nonparametric) estimates of cumulative functions $\widehat{B}_1(t)$, $\widehat{B}_2(t)$, $\widehat{B}_3(t)$. The first two of them were approximately linear (with slopes $\widehat{\beta}_1 \doteq 0.84$, $\widehat{\beta}_2 \doteq 1.08$), but $\widehat{B}_3(t)$ was far from a linear function. It could be caused by that the resulting general model corresponded to our data better than the original exponential one.

a)



b)



c)

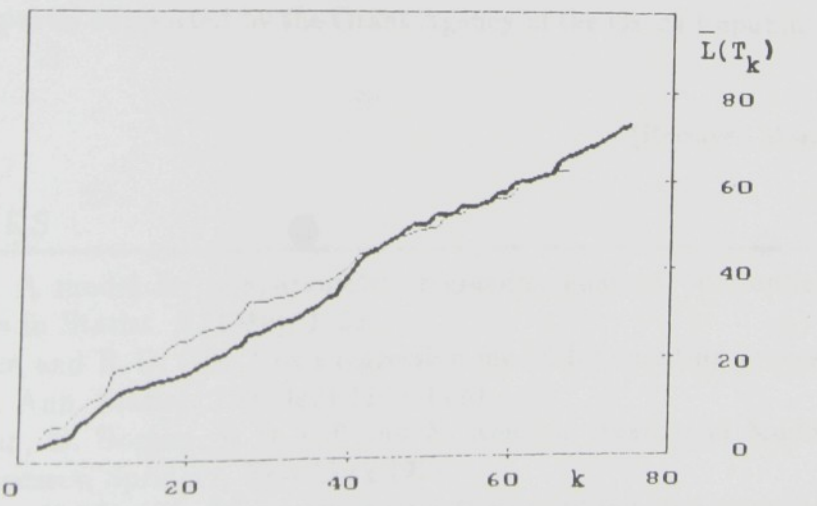


Fig. 1. Cumulative hazards: a) of the full model; b) of reduced model; c) of the 'true' model. Fine line – items with $X_2 < 10$, thick line – items with $X_2 \geq 10$.

For the purpose of diagnostics the sample has been divided into two strata, $S = \{i,\, X_{2i} < 10\}$, $\overline{S} = \{i,\, X_{2i} \geq 10\}$. Figure 1a displays the graphs of testing variables $\widehat{L}_S(T_{k,S})$ and $\widehat{L}_{\overline{S}}(T_{k,\overline{S}})$ vers. k. The picture suggests that the model fits well. Then, the same data have been analysed, but the dependence of hazard function on x_2 has been omitted. Again, $B_1(t)$ and $B_3(t)$ have been estimated and $\widehat{L}_S$, $\widehat{L}_{\overline{S}}$ computed. The result is shown in Figure 1b. The picture clearly indicates that the model considering only x_1 overestimates the hazard rate in group S and underestimates the actual hazard rate in $\overline{S}$. In other words, the graph suggests that a positive dependence of hazard on x_2 should be incorporated into the model.

For the comparison, in Figure 1c there are the plots of the statistics $\overline{L}(t)$ constructed from the "true" exponential model. It seems that the fit is slightly worse (but still good) than the fit of the more general Aalen's model.

Finally, the numerical tests have been accomplished. From Propositon 3 it follows that the statistics $D(t) = n^{-\frac{1}{2}}(\widehat{L}(t) - \overline{N}(t))\,/\,(1 + C(t))$ is asymptotically distributed (provided the Aalen's model holds) as a Brownian bridge process $\mathcal{B}(K(t))$, where $K(t) = C(t)/(1 + C(t))$, $t \in [0, T]$. Hence, for $d \geq 0$, it holds that $P^+(d) \equiv \Pr(\max_t D(t) \geq d) = P^-(d) \equiv \Pr(\min_t D(t) \leq -d) \approx \exp(-2d^2)$. First, the estimate of asymptotic variance $C(t)$ has been computed. Then we obtained, for the first case (full model), in stratum one ($X_2 < 10$) $\max D(t) \doteq 0.478$ and $\min D(t) \doteq -0.196$. It corresponded to the test value $P^+(0.478) \doteq 0.63$. In stratum 2 ($X_2 \geq 10$) we had $\min D(t) \doteq -0.405$ ($\max D(t) = 0$), $P^-(0.405) \doteq 0.72$. The test did not show any reason for rejection of the model.

For the model omitting the dependence on x_2 we obtained $\max D(t) \doteq 2.5$ corresponding to $P^+(2.5) \sim 10^{-4}$, in the first stratum, $\min D(t) \doteq -3.3$, $P^-(3.3) \sim 10^{-5}$ in the second stratum. These values suggest clearly the rejection of the model (on each reasonable confidence level), in favour to an alternative model considering a positive dependence of hazard on x_2.

ACKNOWLEDGEMENT

The research was partly supported by the Grant Agency of the Czech Republic under Grant No. 201/94/0322.

(Received March 21, 1995.)

REFERENCES

[1] O. O. Aalen: A model for nonparametric regression analysis of counting processes. Lecture Notes in Statist. *2* (1980), 1–25.

[2] P. K. Andersen and R. D. Gill: Cox's regression model for counting processes: A large sample study. Ann. Statist. *10* (1982) 1100–1120.

[3] P. K. Andersen, O. Borgan, R. D. Gill and N. Keiding: Statistical Models Based on Counting Processes. Springer, New York 1992.

[4] E. Arjas: A graphical method for assessing goodness of fit in Cox's proportional hazard model. J. Amer. Statist. Assoc. *83* (1988), 204–212.

[5] E. Arjas: Survival models and martingale dynamics. Scand. J. Statist. *16* (1989), 177–225.

[6] W. E. Barlow and R. L. Prentice: Residuals for relative risk regression. Biometrika *75* (1988), 65–74.
[7] N. L. Hjort: Goodness of fit tests in models for life history data based on cumulative hazard rates. Ann. Statist. *18* (1990), 1221–1258.
[8] E. V. Khmaladze: Martingale approach to the goodness of fit tests. Theory Probab. Appl. *26* (1981), 246–265.
[9] L. Marzec and P. Marzec: Goodness of fit inference based on stratification in Cox's regression model. Scand. J. Statist. *20* (1993), 227–238.
[10] T. M. Therneau and P. M. Grambsch: Martingale–based residuals for survival models. Biometrika *77* (1990), 147–160.
[11] P. Volf: Nonparametric Estimation of Hazard Regression, with Application to a Case Study. Research Report No. 1776, Institute of Information Theory and Automation, Prague 1993.

Petr Volf, CSc., Ústav teorie informace a automatizace AV ČR (Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic), Pod vodárenskou věží 4, 18208 Praha 8. Czech Republic.

KYBERNETIKA — VOLUME 36 (2000), NUMBER 2, PAGES 165 – 176

ON CUMULATIVE PROCESS MODEL AND ITS STATISTICAL ANALYSIS

PETR VOLF

The notion of the counting process is recalled and the idea of the 'cumulative' process is presented. While the counting process describes the sequence of events, by the cumulative process we understand a stochastic process which cumulates random increments at random moments. It is described by an intensity of the random (counting) process of these moments and by a distribution of increments. We derive the martingale – compensator decomposition of the process and then we study the estimator of the cumulative rate of the process. We prove the uniform consistency of the estimator and the asymptotic normality of the process of residuals. On this basis, the goodness-of-fit test and the test of homogeneity are proposed. We also give an example of application to analysis of financial transactions.

1. INTRODUCTION

A counting process is a stochastic point process registering random events and counting their number. The trajectory of such a process starts at zero and has jumps $+1$ at random moments. The main characteristic is the intensity of the stream of events. A review of theory and applications of counting process models is given, for instance, in Andersen et al [2], or in Fleming and Harrington [5].

In the present paper, we consider a random process

$$C(t) = \int_0^t Y(s)\,\mathrm{d}N(s), \tag{1}$$

where $N(t)$ is a counting process and $Y(t)$ is a set of random variables. We assume that the time runs through $[0, T]$ and starting value is again $C(0) = 0$. From this point of view, we deal with a process having the random increments at disjoint random moments. The process $C(t)$ will be called the cumulative process. In the special case when the process of times, $N(t)$, is the Poisson one, the process (1) is known as the compound Poisson process (see for example Embrechts et al [4]).

The objective of the present paper is to describe the process (1) with the aid of characteristics of both its components, i.e. the hazard function of $N(t)$ and the distribution of $Y(t)$. The paper is organized as follows: In part 2 the process (1) is defined more accurately. Its martingale-compensator decomposition is presented

and the variance process of the martingale is computed. Then the estimator of the mean trajectory of the process (actually representing the cumulative rate) is constructed and its uniform consistency is proved. In this we generalize the results of Volf [7] achieved for the case of underlying Poisson process of events. The main result consists in the derivation of the weak convergence of the residual process to a Wiener one. Finally, based on this convergence, a test procedure is proposed both for assessing the goodness-of-fit of the model and for testing the homogeneity of two processes. In this we follow the method of analysis of generalized residuals for counting processes proposed in Arjas [3] and also in Volf [6].

2. THE MODEL OF CUMULATIVE PROCESS

In order to define the process (1), we consider a measurable, nonnegative and bounded function $h(t), t \geq 0$, the hazard function, and the indicator process $I(t)$ which equals 1 if $N(t)$ is in the risk of count, $I(t) = 0$ otherwise. Actually, $I(t)$ is an indicator of observability of counting process $N(t)$. Then, the behaviour of the counting process $N(t)$ in (1) is governed by a random (in general) intensity process $\lambda(t) = h(t)I(t)$.

Further, let us consider a right–continuous nondecreasing sequence of σ-algebras, $\mathcal{S}(t)$, where each $\mathcal{S}(t)$ is defined on the sample space of $\{N(s), I(s), Y(s),\ 0 \leq s \leq t\}$. We assume that process $N(t)$ is $\mathcal{S}(t)$-measurable. Following Andersen and Borgan [1], we denote by $\mathrm{d}N(t)$ the increments of $N(t)$ over the small time interval $[t, t+\mathrm{d}t)$. Then we can write that $\lambda(t)\,\mathrm{d}t = P(\mathrm{d}N(t) = 1 \mid \mathcal{S}(t^-))$. The trajectories of $N(t)$ are right–continuous, the trajectories of $I(t)$ and also the "histories" collected in $\mathcal{S}(t^-)$ are left–continuous.

As regards the variables $Y(t)$, we assume that

1. $Y(t)$ are distributed with (unknown) densities $f(y;t)$.
2. Their means $\mu(t)$, variances $\sigma^2(t)$, and also $E(|Y(t)|^3)$ exist and are measurable and bounded functions on $[0, T]$.
3. Each $Y(t)$ is independent of $\mathcal{S}(t^-)$, i.e. of the history of the process $C(s)$ up to t (on the other hand, $\mathrm{d}N(t)$ can depend on history of $Y(s), s < t$).

Remark. Point 3 is a rather strong condition which in some cases is not fulfilled. On the other hand, we can imagine a number of processes (especially in the area of natural sciences) for which such an independence of increments on the history is a quite realistic property.

The assumption on existence and boundedness of the 3-rd absolute moments is actually a condition of the Lyapunov version of the central limit theorem. Here, it will be utilized for the proof of the condition of Lindeberg required in the central limit theorem for martingales (Proposition 3.).

Let us now recall the martingale-compensator decomposition of the counting process $N(t)$. Define first its cumulative intensity $L(t) = \int_0^t \lambda(s)\,\mathrm{d}s$. Then $N(t) = L(t) + M(t)$, where $M(t)$ is a martingale adapted to the σ-algebras $\mathcal{S}(t)$ (i.e. it

is $\mathcal{S}(t)$-measurable). It holds that $E\,M(t) = 0$ and variance processes $\langle N\rangle(t) = \langle M\rangle(t) = L(t)$, where the notation $\langle N\rangle(t)$ means $var(N(t)|\mathcal{S}(t^-))$. Similarly as the paths of $N(t)$, the paths of $M(t)$, and also of $C(t)$, are right–continuous (while the paths of $L(t)$ are continuous).

The first task is to derive a compensator of the process $C(t)$. We utilize the decomposition $N(t) = L(t) + M(t)$ of the counting process. Denote $Y^*(t) = Y(t) - \mu(t)$. Then we can decompose

$$C(t) = \int_0^t (Y^*(s) + \mu(s))\,\mathrm{d}N(s) = \int_0^t \mu(s)\,\mathrm{d}L(s) + \mathcal{M}(t),$$

where

$$\mathcal{M}(t) = \mathcal{M}_1(t) + \mathcal{M}_2(t) = \int_0^t Y^*(s)\,\mathrm{d}N(s) + \int_0^t \mu(s)\,\mathrm{d}M(s).$$

Proposition 1. The processes $\mathcal{M}(t)$, $\mathcal{M}_1(t)$, $\mathcal{M}_2(t)$ are martingales adapted to σ-algebras $\mathcal{S}(t)$ on $[0,T]$, the variance process of $\mathcal{M}(t)$ is

$$\langle\mathcal{M}\rangle(t) = \int_0^t (\sigma^2(s) + \mu^2(s))\,\mathrm{d}L(s).$$

Proof. Evidently, $E\,\mathcal{M}(t) = 0$. As regards the property defining the martingale, we have for $0 < s < t$ that

$$E(\mathcal{M}(t)\,|\,\mathcal{S}(s)) = \mathcal{M}(s) + E\left(\int_s^t \mathrm{d}\mathcal{M}(\tau)|\mathcal{S}(s)\right) = \mathcal{M}(s),$$

because $E\left(\int_s^t \mathrm{d}\mathcal{M}(\tau)|\mathcal{S}(s)\right) = 0$ holds for both parts of $\mathcal{M}(t)$: For $\mathcal{M}_1(t)$ it follows from the centering of $Y^*(t)$ and from the independence of $Y^*(t)$ on $\mathrm{d}N(t)$. Properties of $\mathcal{M}_2(t)$ follow directly from properties of $M(t)$. From the independence of $Y^*(t)$ on the past up to t it also follows that $Y^*(t)$ is orthogonal both to $\mathrm{d}M(t)$ and to $\mathrm{d}N(t)$, distribution of $\mathrm{d}N$ and $\mathrm{d}M$ being given by predictable process $\mathrm{d}L$. Therefore

$$\langle\mathcal{M}_1\rangle(t) = \int_0^t \sigma^2(s)\,\mathrm{d}L(s), \quad \langle\mathcal{M}_1, \mathcal{M}_2\rangle(t) = 0.$$

Further, from martingale properties of $M(t)$ we have that $\langle\mathcal{M}_2\rangle(t) = \int_0^t \mu^2(s)\,\mathrm{d}L(s)$. Then

$$\begin{aligned}\langle\mathcal{M}\rangle(t) &= E\,\{\mathcal{M}_1(t)^2 + \mathcal{M}_2(t)^2 + 2\mathcal{M}_1(t)\,\mathcal{M}_2(t)\,|\,\mathcal{S}(t^-)\}\\ &= \int_0^t \sigma^2(s)\,\mathrm{d}L(s) + \int_0^t \mu^2(s)\,\mathrm{d}L(s).\end{aligned}$$

□

Corollary. Process $\int_0^t \mu(s)\,\mathrm{d}L(s)$ is the compensator of process $C(t)$ on $[0,T]$.

Evidently, process $\int_0^t \mu(s)\,\mathrm{d}L(s)$ fulfils the conditions of compensator. Its substraction from $C(t)$ yields a martingale, process is predictable and its paths are uniformly continuous on $[0,T]$ (which is more than is needed).

3. LARGE SAMPLE PROPERTIES

In the follow-up, let us imagine that n realizations $C_i(t)$ of a cumulative process $C(t)$ are observed in interval $[0,T]$. More precisely, we observe moments of events T_{ij} of counting processes $N_i(t)$, corresponding indicators $I_i(t)$ and "jumps" $Y_i(T_{ij})$, $(i = 1,\ldots,n,\ j = 1,\ldots,n_i = N_i(T))$. Formally, observed trajectories are

$$C_i(t) = \int_0^t Y_i(s)\,\mathrm{d}N_i(s) = \sum_{j=1}^{n_i} Y_i(T_{ij}).$$

It is assumed that random variables $Y_i(t), i = 1,2,\ldots,n$ have the same distributions, with densities $f(y;t)$ and moments $\mu(t)$ and $\sigma^2(t)$. Further, we assume that $Y_i(t)$ are independent of the common history of the processes $N_i(s), Y_i(s), I_i(s), s < t, i = 1,2,\ldots,n$ stored now in σ-algebras $\mathcal{S}(t^-)$. Finally, we assume that $N_i(t)$ are characterized by the same hazard function $h(t)$. Corresponding intensities of $N_i(t)$ are then $\lambda_i(t) = h(t)I_i(t)$. As it is assumed that the hazard function is finite, the compensator is a continuous process. The consequence is also that there are not two events at one moment and, further, that for $i \neq j$ $\mathrm{d}\langle M_i, M_j\rangle(t) = 0$, $\mathrm{d}\langle \mathcal{M}_i, \mathcal{M}_j\rangle(t) = 0$ and even $\mathrm{cov}\{Y_i(t)\,\mathrm{d}\mathcal{M}_i(t), Y_j(t)\,\mathrm{d}\mathcal{M}_j(t)|\mathcal{S}(t^-)\} = 0$.

The likelihood process (which is actually the generalization of the likelihood of Poisson process) is

$$\mathcal{L} = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\{\lambda_i(T_{ij})f(Y_i(T_{ij});T_{ij})\}\cdot\exp\left\{-\int_0^T \lambda_i(t)\,\mathrm{d}t\right\}.$$

It is seen that the part containing the intensities and the part containing the distribution of Y's can be separated (and therefore both characteristics can be estimated independently). Denote by $\mathcal{L}(f) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} f(Y_i(T_{ij});T_{ij})$. Then we obtain the following log-likelihood:

$$\ln\mathcal{L} = \sum_{i=1}^{n}\left\{\int_0^T \ln\lambda_i(t)\,\mathrm{d}N_i(t) - \int_0^T \lambda_i(t)\,\mathrm{d}t\right\} + \ln(\mathcal{L}(f)). \tag{2}$$

In the case of parametrized function f, its parameters can be estimated from the maximum likelihood estimation procedure based on $\mathcal{L}(f)$ only. In a nonparametrized case, estimates of functions $\mu(t)$, $\sigma^2(t)$ can be obtained with the help of a smoothing (kernel) technique. Even the density $f(y;t)$ can be then estimated via the kernel method. In the follow-up, we shall try to characterize the process $C(t)$ jointly and to derive some asymptotic properties which depend on $h(t)$, $\mu(t)$ and $\sigma^2(t)$.

3.1. Estimates and their convergence

Let us recall here the well-known Nelson–Aalen estimator of the cumulative hazard function $H(t) = \int_0^t h(s)\,\mathrm{d}s$:

$$\widehat{H}_n(t) = \sum_{i=1}^{n}\int_0^t \frac{1[I(s) > 0]}{I(s)}\,\mathrm{d}N_i(s),$$

where $I(s) = \sum_{i=1}^{n} I_i(s)$ characterizes the risk set at moment s. Let us make the following assumption:

A1. There exists the limit $r(s) = \lim_{n\to\infty} \frac{I(s)}{n}$ in probability such that
a) the limit is uniform on $[0,T]$,
b) $1 \geq r(s) \geq \varepsilon$ on $[0,T]$, for some $\varepsilon > 0$.

Then it is proved elsewhere (for instance in Andersen and Borgan [1], Andersen et al [2]) that $\widehat{H}_n(t)$ is a consistent estimate of $H(t)$. Further, such an estimate is asymptotically normal on $[0,T]$, namely $\sqrt{n}(\widehat{H}_n(t) - H(t)) \approx \sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{\mathrm{d}M_i(s)}{I(s)}$ converges weakly to a Wiener process with variance function $\int_0^t \frac{\mathrm{d}H(s)}{r(s)}$, when $n \to \infty$. It follows from the central limit theorems for martingales (e. g. Andersen et al [2], chapter II). Moreover, it is due A1 and due the boundedness of all involved functions that $\widehat{H}_n(t)$ is a uniformly consistent estimator of $H(t)$ on $[0,T]$ (see also Winter, Földes and Rejtö [8], and their variant of Glivenko–Cantelli theorem).

Inspired by these results, we consider the average of observed processes

$$\overline{C}_n(t) = \sum_{i=1}^{n} \int_0^t \frac{1[I(s) > 0]}{I(s)} Y_i(s)\,\mathrm{d}N_i(s)$$

as an estimator of the function $K(t) = \int_0^t \mu(s)\,\mathrm{d}H(s)$. Actually, $K(t)$ represents the cumulative rate describing the risk and the mean size of jumps of $C(t)$.

From A1 and boundedness of $H(s)$ and $\mu(s)$ we easily see that $P\left\{\int_0^T 1[I(s) = 0]\,\mathrm{d}s = 0\right\} \to 1$ for $n \to \infty$, so that even $P\left\{\sqrt{n}\int_0^t 1[I(s) = 0]\,\mathrm{d}K(s) = 0\right\} \to 1$, uniformly w.r.t. $t \in [0,T]$.

Proposition 2. Under A1, $\overline{C}_n(t)$ is a uniformly consistent estimate of $K(t)$ on $[0,T]$, that is

$$\lim_{n\to\infty} \sup_{t\in[0,T]} |\overline{C}_n(t) - K(t)| = 0 \text{ in probability.}$$

Proof.

$$\overline{C}_n(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{I(s)} \{\mu(s)\,\mathrm{d}L_i(s) + \mathrm{d}\mathcal{M}_i(s)\}$$

$$= K(t) + \int_0^t \sum_{i=1}^{n} \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)} - \int_0^t 1[I(s) = 0]\,\mathrm{d}K(s). \tag{3}$$

Here $\mathcal{M}_i(t)$ are square integrable martingales, with the same distribution as $\mathcal{M}(t)$ defined in the preceding part. They are mutually orthogonal, $\langle \mathcal{M}_i, \mathcal{M}_j \rangle(t) = 0$ for $i \neq j$, their variances are uniformly bounded on $[0,T]$. From Lenglart's inequality (cf. Andersen and Borgan [1], or Andersen et al [2], part II.5.2) it follows that for each $\delta, \varepsilon > 0$ and for sufficiently large $n > n(\delta, \varepsilon)$

$$P\left(\sup_{t\in[0,T]} |\frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_i(t)| \geq \varepsilon\right) \leq \frac{\delta}{\varepsilon^2}.$$

This, together with the uniform convergence of $\frac{I(s)}{n}$ assumed in A1, leads to the convergence of $\sup_{t\in[0,T]} |\int_0^t \sum_{i=1}^n \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)}|$ in probability to zero. □

Proposition 3. Under A1 the process $\sqrt{n}(\overline{C}_n(t) - K(t))$ converges weakly on $[0,T]$ to a Wiener process (i. e. the continuous Gaussian process with zero mean and independent increments) which has the variance function $W(t) = \int_0^t \frac{1}{r(s)}(\mu^2(s) + \sigma^2(s))\,\mathrm{d}H(s)$.

Proof. From (3) we have that

$$\sqrt{n}(\overline{C}_n(t) - K(t)) = \sqrt{n}\int_0^t \sum_{i=1}^n \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)} - \sqrt{n}\int_0^t 1[I(s)=0]\,\mathrm{d}K(s).$$

The convergence follows from the central limit theorem for martingales. We use the version stated in Andersen et al [2], namely Theorem II.5.1 (Rebolledo's). The proof requires two convergences in probability to hold, namely that for all $t \in [0,T]$

(i) $\langle M^{(n)}\rangle(t) \to W(t)$,

(ii) $\langle M_\varepsilon^{(n)}\rangle(t) \to 0$ (Lindeberg condition).

(i) Here $M^{(n)}(t) = \sqrt{n}\int_0^t \sum_{i=1}^n \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)}$. From Proposition 1 and from A1 it follows that

$$\langle M^{(n)}\rangle(t) = \int_0^t \sum_{i=1}^n \frac{(\sigma^2(s)+\mu^2(s))\,\mathrm{d}L_i(s)}{I^2(s)/n} =$$

$$= \int_0^t \frac{(\sigma^2(s)+\mu^2(s))\,1[I(s)>0]}{I(s)/n}\,\mathrm{d}H(s) \to W(t)$$

in probability, (i) is proved.

(ii) By $M_\varepsilon^{(n)}(t) = \int_0^t \sum_{i=1}^n \frac{\sqrt{n}}{I(s)} Q_{i,\varepsilon}(s)\mathrm{d}\mathcal{M}_i(s)$, with $Q_{i,\varepsilon}(s) = 1[|\sqrt{n}\mathrm{d}\mathcal{M}_i(s)/I(s)| > \varepsilon]$, we mean the process (martingale) containing all jumps of $M^{(n)}(t)$ larger than chosen $\varepsilon > 0$. For "zero-one" random variables $Q_{i,\varepsilon}(s)$ we have from the Chebyshev inequality that, for each $k = 1, 2, \ldots$

$$E(Q_{i,\varepsilon}^k(s)|\mathcal{S}(s^-)) = P(Q_{i,\varepsilon}(s) = 1|\mathcal{S}(s^-)) =$$

$$= P\left\{|\frac{\sqrt{n}\,\mathrm{d}\mathcal{M}_i(s)}{I(s)}| > \varepsilon|\mathcal{S}(s^-)\right\} \le \left\{\frac{n}{\varepsilon^2}\frac{(\mu^2(s)+\sigma^2(s))}{I^2(s)}\right\}\mathrm{d}L_i(s), \qquad (4)$$

where the expression in brackets is of size $\mathcal{O}_P\left(\frac{1}{n}\right)$ uniformly in $s \in [0,T]$, as a consequence of A1 (by the notation $B_n \sim \mathcal{O}_P(a_n)$, for sequences B_n of random variables and a_n of numbers, we mean that the sequence B_n/a_n is asymptotically bounded in probability).

From the conditional orthogonality of increments $\mathrm{d}\langle \mathcal{M}_i, \mathcal{M}_j\rangle(s)$ for $i \neq j$, we obtain that

$$\langle M_\varepsilon^{(n)}\rangle(t) \leq \int_0^t \frac{n}{I^2(s)} \sum_{i=1}^n E\{Q_{i,\varepsilon}^2(s)(\mathrm{d}\mathcal{M}_i(s))^2 | \mathcal{S}(s^-)\}.$$

Further, the Hölder inequality yields that

$$E\{Q_{i,\varepsilon}^2(s)(\mathrm{d}\mathcal{M}_i(s))^2 | \mathcal{S}(s^-)\} \leq [E\{|\mathrm{d}\mathcal{M}_i(s)|^3 | S(s^-)\}]^{\frac{2}{3}} \cdot [E\{Q_{i,\varepsilon}^6(s) | S(s^-)\}]^{\frac{1}{3}}.$$

From our assumptions on the boundedness of (absolute) moments of variables $Y(s)$ up to the 3-rd moment it follows that $E\{|\mathrm{d}\mathcal{M}_i(s)|^3 | S(s^-)\} = E\{|Y_i(s)|^3\}\mathrm{d}L_i(s) + \mathcal{O}((\mathrm{d}s)^2)$, taking into account that $\mathrm{d}L_i(s) = h(s)I_i(s)\mathrm{d}s$. Finally, we obtain

$$\langle \mathcal{M}_\varepsilon^{(n)}\rangle(t) \leq \int_0^t \frac{n}{I^2(s)} \sum_{i=1}^n \{E|Y_i(s)|^3 \mathrm{d}L_i(s) + \mathcal{O}((\mathrm{d}s)^2)\}^{\frac{2}{3}} \{\mathcal{O}_P(\frac{1}{n})\mathrm{d}L_i(s)\}^{\frac{1}{3}}$$

$$\sim \int_0^t \frac{n}{I(s)} \{E|Y_1(s)|^3 h(s)\,\mathrm{d}s\}^{\frac{2}{3}} \mathcal{O}_P(n^{-\frac{1}{3}})(h(s)\,\mathrm{d}s)^{\frac{1}{3}} \sim \int_0^t \frac{n}{I(s)} \mathcal{O}_P(n^{-\frac{1}{3}}) h(s)\,\mathrm{d}s,$$

which is of size $\mathcal{O}_P(n^{-\frac{1}{3}})$ uniformly in $t \in [0, T]$. This proves the condition (ii). □

4. STATISTICAL TESTS

In the following part, the asymptotic normality of the residual process will be utilized for the construction of statistical tests, namely the goodness-of-fit test and the test of homogeneity of two samples of cumulative processes.

From two parts of martingale $\mathcal{M}(t)$ the first reflects the variability of Y's and the second equals $\mu(t)$-times the 'residual' martingale known from the counting processes scheme. Hence, the variance function of residuals $\sqrt{n}(\overline{C}(t) - K(t))$ contains also two parts, expressed by σ^2 and μ^2. They can significantly influence the power of test procedures. Therefore, for the purpose of tests, we recommend to normalize the residuals, i. e. to divide them by $\sqrt{\sigma^2 + \mu^2}$.

4.1. The goodness-of-fit test

Arjas [3] and later Volf [6] derived goodness-of-fit tests for the counting processes model, and generalized them for the case of hazard regression models (namely Arjas considered the Cox model, Volf a general case of hazard regression model and the Aalen model). From this point of view, the case considered here is much simpler, because the regression is not involved.

Let the model be given by functions $H(t)$, $\mu(t)$, $\sigma^2(t)$, we want to decide whether the data correspond to it. The data are represented by the observed trajectories $C_i(t)$ and indicators $I_i(t)$, $i = 1, \ldots, n$. The tests are quite naturally based on the comparison of $\overline{C}_n(t)$ with expected $K(t)$. The process of differences $\overline{C}_n(t) - K(t)$ is called the residual process.

Graphical test: Let us order all moments of events into one nondecreasing sequence T_k, $k = 1, \ldots, K$. For a graphical comparison, we plot $K(T_k) = \int_0^{T_k} \mu(s)\,\mathrm{d}H(s)$ and $\overline{C}(T_k)$ into one figure, against k on the abscissa. If the model fits the residual process is a martingale asymptotically tending to zero. Then it is expected that both plots will be close to each other. An opposite case (i.e. an increasing distance of both curves) indicates that the model $K(t)$ does not correspond to the data. Of course, a more precise test will need a specification of critical limits for the distance of compared curves. Such critical bounds can be derived from the large sample properties, for instance in the following way.

Numerical test: Numerical test is based on asymptotic distribution. Define the normalized residual process by

$$R_n(t) = \int_0^t \frac{\mathrm{d}(\overline{C}_n(s) - K(s))}{\sqrt{\mu^2(s) + \sigma^2(s)}}.$$

From Proposition 3 it follows that $\sqrt{n}R_n(t)$ is asymptotically distributed as a Wiener process with the variance function $V(t) = \int_0^t \mathrm{d}H(s)/r(s)$. Then the process

$$D_n(t) = \sqrt{n}R_n(t)\,/(1 + V(t))$$

is (if the model holds) asymptotically distributed as a Brownian bridge process $\mathcal{B}((V(t)/(1 + V(t))$, in $t \in [0, T]$. Hence, a test of Kolmogorov–Smirnov type can be used. From the theory of Brownian bridge it follows, for instance, that for any $d \geq 0$,

$$P\left(\max_t D_n(t) \geq d\right) = P\left(\min_t D_n(t) \leq -d\right) \approx \exp(-2d^2)$$

approximately. So that the value $\exp(-2d^2)$, where d is the observed $\max_k |D_n(T_k)|$, is an approximate P-value for the test of hypothesis of the goodness-of-fit against a proper one-sided alternative. Unknown limit function $r(s)$ needed for computation of $V(t)$ is consistently estimated by $I(s)/n$ from A1.

4.2. Test of homogeneity

Besides the goodness-of-fit tests, we can also consider the tests of homogeneity. They compare two (sets of) realizations of the process. Both graphical and numerical comparison can be based on slight modifications of the methods described above. On the other hand, the performance of a test of homogeneity is influenced by the fact that, as a rule, certain characteristics of the joint model have to be estimated. The properties of the test procedure then depend strongly on the properties of the estimator.

Let us consider two independent sets of cumulative processes, $C_i^{(k)}(t)$, $k = 1, 2$, $i = 1, 2, \ldots, m_k$, each representing a certain model characterized by $H^{(k)}(t), \mu^{(k)}(t)$, $\sigma^{(k)}(t)$. The test of homogeneity assesses the hypothesis $\boldsymbol{H}_0$ that $H^{(k)}(t)$, $\mu^{(k)}(t)$, $\sigma^{(k)}(t)$ are the same for $k = 1, 2$, on a given interval $[0, T]$. To confirm it, we analyse

the averaged processes

$$\overline{C}^{(k)}(t) = \int_0^t \sum_{i=1}^{m_k} \frac{Y_i^{(k)}(s)\,\mathrm{d}N_i^{(k)}(s)}{I^{(k)}(s)}\, 1[I^{(k)}(s) > 0],$$

and their difference

$$\sqrt{\frac{m_1 m_2}{m}} \left\{ \overline{C}^{(1)}(t) - \overline{C}^{(2)}(t) \right\}$$

$$\approx \sqrt{\frac{m_1 m_2}{m}} \left\{ K^{(1)}(t) - K^{(2)}(t) + \int_0^t \sum_{i=1}^{m_1} \frac{\mathrm{d}\mathcal{M}_i^{(1)}(s)}{I^{(1)}(s)} - \int_0^t \sum_{i=1}^{m_2} \frac{\mathrm{d}\mathcal{M}_i^{(2)}(s)}{I^{(2)}(s)} \right\},$$

where $m = m_1 + m_2$. Let us now assume that:

1. $\boldsymbol{H}_0$ holds, so that $K^{(1)}(t) = K^{(2)}(t)$, μ, σ and H are common for both processes.

2. m_1, m_2 tend to infinity in such a way that $\frac{m_1}{m} \to \alpha \in (0,1)$.

3. Assumption A1 is fulfilled for both sets of processes (possibly with different $r^{(k)}(s)$).

Then, from Proposition 3 it follows that

$$\sqrt{\frac{m_1 m_2}{m}} \int_0^t \frac{\mathrm{d}(\overline{C}^{(1)}(s) - \overline{C}^{(2)}(s))}{\sqrt{\mu^2(s) + \sigma^2(s)}} \tag{5}$$

tends weakly to the Wiener process with zero mean and variance function $V^*(t) = (1-\alpha)V^{(1)}(t) + \alpha V^{(2)}(t)$, where $V^{(k)}(t) = \int_0^t \mathrm{d}H(s)/r^{(k)}(s)$, $k = 1, 2$. In order to estimate (5), we need the estimates of joint characteristics of the processes. As regards $H(t)$, the Nelson–Aalen estimator is available, cf. part 3.1. The moments $\mu(t)$ and $\sigma^2(t)$ can be estimated e. g. with the help of the moving window (or kernel) approach. We can then compute (approximatelly) the test process

$$D_m(t) = \sqrt{\frac{m_1 m_2}{m}} \int_0^t \frac{\mathrm{d}(\overline{C}^{(1)}(s) - \overline{C}^{(2)}(s))}{\sqrt{\mu^2(s) + \sigma^2(s)}} / (1 + V^*(t)), \tag{6}$$

which again behaves asymptotically as the process of Brownian bridge. Therefore, the test of $\boldsymbol{H}_0$ is then performed in a quite similar way as the goodness-of-fit test, i. e. by evaluation of $d = \max |D_m(t)|$ on $[0, T]$ and taking $\exp(-2d^2)$ as an approximate P-value of the test against a one-sided alternative.

5. EXAMPLE OF THE TEST OF HOMOGENEITY

As an example, let us consider one-day processes of financial transactions performed via credit cards at two different gas stations, both for $m_1 = m_2 = 100$ days. We follow both the number of transactions (forming the counting process) and also the

cumulation of transferred amounts (from this we obtain the cumulative process), $t \in [0, 24]$ hours. Figure 1 contains the averaged processes $\overline{N}^{(k)}(t) = \frac{1}{m_k} \sum_{i=1}^{m_k} N_i^{(k)}(t)$ and estimated corresponding hazard functions $\widehat{h}^{(k)}(t)$, $k = 1, 2$. Estimates of hazard function were obtained from the estimated cumulative hazard functions by a smoothing (kernel) technique.

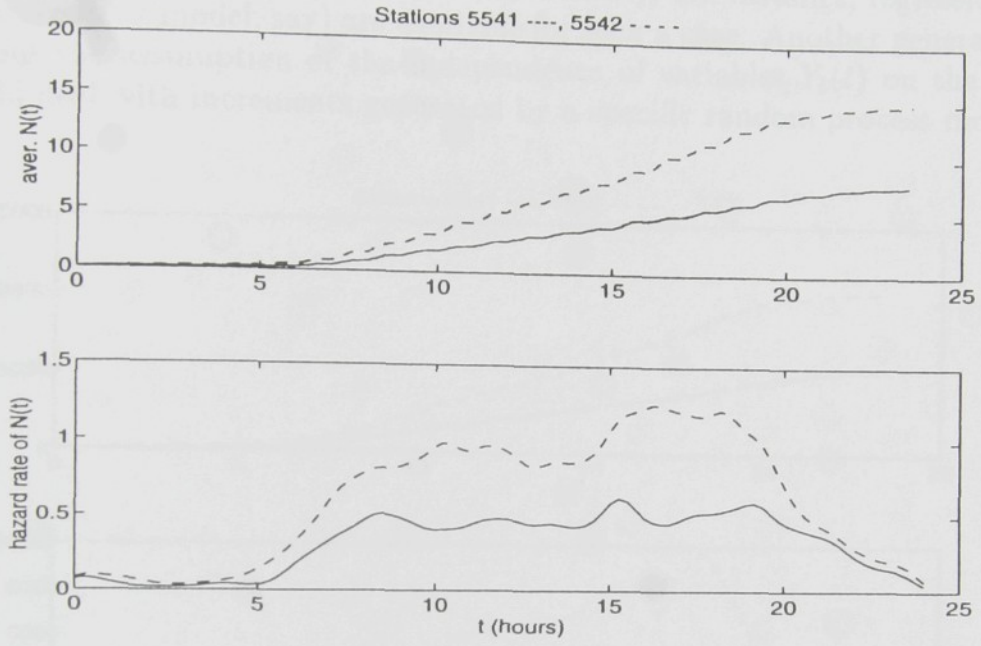


Fig. 1. Averaged counting processes and estimates of their hazard rates.

Figure 2 shows the averaged cumulative processes $\overline{C}^{(k)}(t) = \frac{1}{m_k} \sum_{i=1}^{m_k} C_i^{(k)}(t)$, and estimated and smoothed derivatives of functions $K^{(k)}(t)$.

From the graphical comparison we already see the difference between both sets of processes. By the numerical test of homogeneity computed in accordance with (6) we obtained that the minimum of $D_n(t)$ was -3.841, which was highly significant (P-value was $\sim 10^{-13}$). Functions $\mu(t)$ and $\sigma^2(t)$ were estimated with the aid of the moving window procedure, $H(t)$ by the Nelson–Aalen estimator.

6. CONCLUSION

The main advantage of the counting processes is their dynamics resulting from the conditioning the actual intensities by the history of the system. This area of statistical methods has a well developed theoretical background as well as the techniques of computational analysis.

The main purpose of the paper was to describe and analyze the random process (called here the cumulative process) consisting in the combination of the counting process with the process of random increments. Such models are suitable for description of many real-world technological, environmental, biological (and also financial) processes. We derived tools for modelling and statistical analysis of such situations,

namely we proposed the estimator of the rate of the cumulative process and proved its large sample properties. These properties were utilized in the proposal of procedures for the test of agreement of the data with the cumulative process model and for the tests of homogeneity of two cumulative processes.

As regards the generalization of the case studied in the present paper, the first one should consider a functional model for the hazard function describing also the influence of the history of $C_i(t)$ on the actual intensity. For instance, regression models (variants of Cox model, say) are available for such a case. Another generalization should omit the assumption of the independence of variables $Y_i(t)$ on the history and should deal with increments generated by a specific random process model.

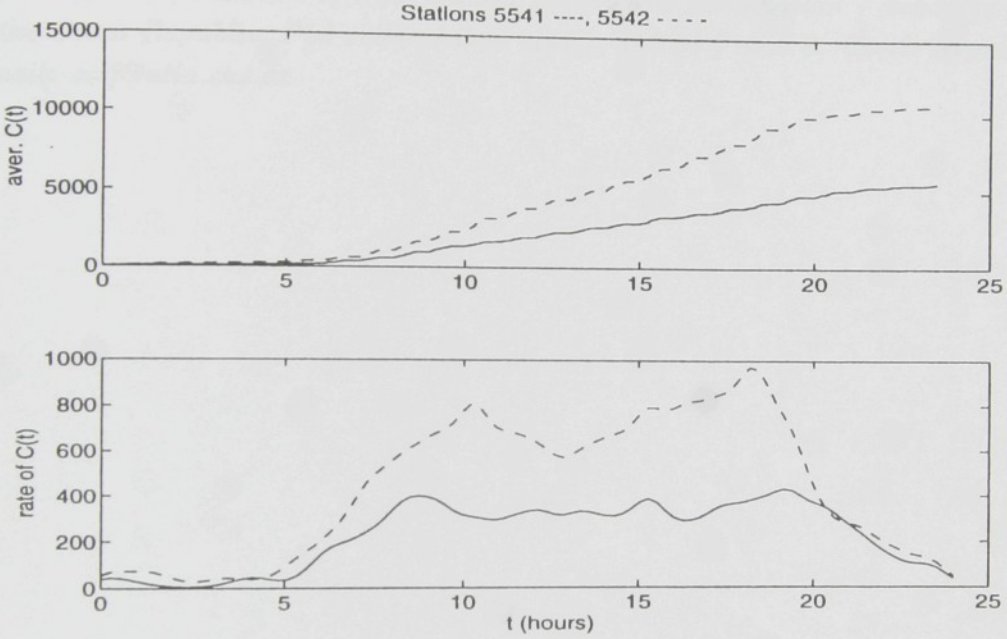


Fig. 2. Averaged cumulative processes and estimates of their rates.

ACKNOWLEDGEMENT

The present study was supported by the Grant Agency of the Czech Republic under grants 201/97/0354 and 402/98/0742.

(Received October 20, 1998.)

REFERENCES

[1] P. K. Andersen and O. Borgan: Counting process models for life history data: A review. Scand. J. Statist. *12* (1985), 97–158.

[2] P. K. Andersen, O. Borgan, R. D. Gill and N. Keiding: Statistical Models Based on Counting Processes. Springer, New York 1993.

[3] E. Arjas: A graphical method for assessing goodness of fit in Cox's proportional hazards model. J. Amer. Statist. Assoc. *83* (1988), 204–212.

[4] P. Embrechts, C. Klüppelberg and T. Mikosch: Modelling Extremal Events. Springer, Heidelberg 1997.
[5] T. R. Fleming and D. P. Harrington: Counting Processes and Survival Analysis. Wiley, New York 1991.
[6] P. Volf: Analysis of generalized residuals in hazard regression models. Kybernetika *32* (1996), 501–510.
[7] P. Volf: On counting process with random increments. In: Proceedings of Prague Stochastics'98, Union of Czech Math. Phys., Prague 1998, pp. 587–590.
[8] B. B. Winter, A. Földes and L. Rejtö: Glivenko–Cantelli theorems for the product limit estimate. Problems Control Inform. Theory *7* (1978), 213–225.

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